IEEE COMMUNICATIONS MAGAZINE

February 2017, Vol. 55, No. 2

- People-Centric IoT
- IoT in 5G: Industrial Challenges and Business Opportunities
- Securing the Internet of Things in a Quantum World
- Resource Management in Massive Wireless IoT Systems
- UAV-Based IoT Platform: A Crowd Surveillance Use Case

IEEE ComSoc IEEE Communications Society
A Publication of the IEEE Communications Society
www.comsoc.org

IEEE ComSoc™
IEEE Communications Society

Join our Community!

IEEE

Networking • Conference Discounts • Technical Publications • Volunteer

Member Benefits and Discounts

Valuable discounts on IEEE ComSoc conferences
ComSoc members save on average $200 on ComSoc-sponsored conferences.

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

*2015 Journal Citation Reports (JCR)

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

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

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

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

Join today at www.comsoc.org

IEEE COMMUNICATIONS MAGAZINE

FEBRUARY 2017, vol. 55, no. 2

www.comsoc.org/commag

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

ANNUAL SUBSCRIPTION: $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.

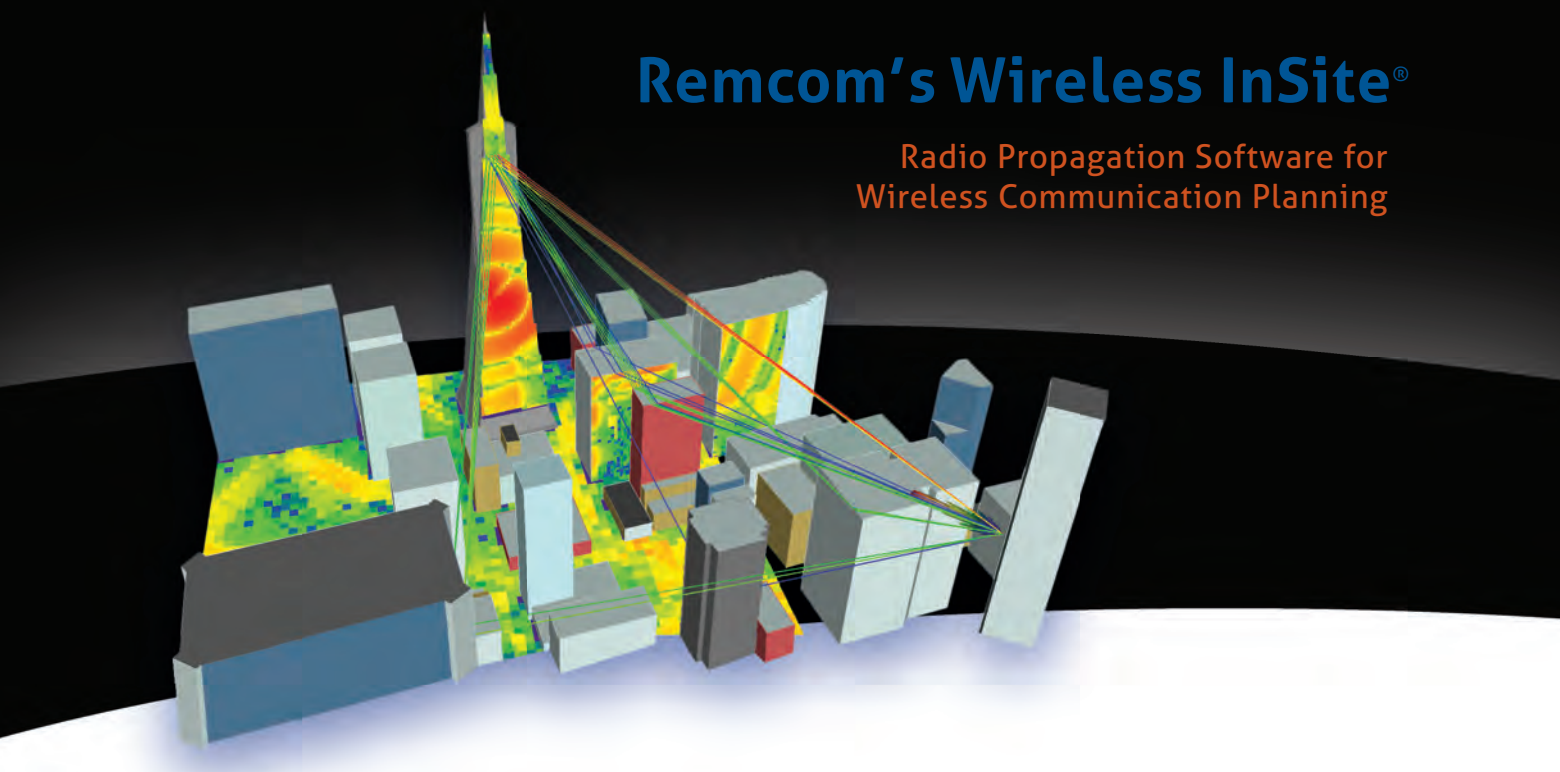
ADVERTISING: Advertising is accepted at the discretion of the publisher. Address correspondence to: Advertising Manager, *IEEE Communications Magazine*, 3 Park Avenue, 17th Floor, New York, NY 10016.

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

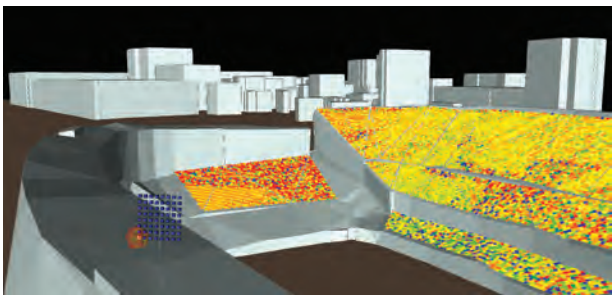
BPA WORLDWIDE

Remcom's Wireless InSite®

Radio Propagation Software for Wireless Communication Planning



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



MIMO capability predicts received power and complex channel matrix throughout Soldier Field stadium

Predictive Simulation for Telecommunications and Wireless Networks:

- 5G MIMO simulation
- Macrocell and small cell coverage
- Urban multipath and shadowing
- Indoor WiFi
- Wireless backhaul
- LTE and WiMAX throughput analysis
- Ad hoc networks and D2D communication

See all the latest enhancements at | **www.remcom.com/wireless-insite-features** ›››

REMCOM®

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

COMMUNICATIONS, COMSOC, AND BIG DATA: WHAT AND WHAT NEXT?

Last year in the President's Pages we covered three new technology areas in which ComSoc was involved: 5G, IoT, and Fog. This article describes Big Data, a relatively new area for ComSoc. Leading the IEEE Big Data Initiative is David Belanger, who is also assisting ComSoc by taking a critical role in guiding and supporting its transition from an Initiative to a viable program.

David Belanger is currently a Senior Research Fellow at Stevens Institute of Technology. He continues his work in Big Data Technology, Applications, and Governance, and is a leader in the Business Intelligence & Analysis Master's Degree program. He leads the IEEE Big Data Initiative (bigdata.ieee.org), which addresses activities ranging from a new data repository to standards and educational activities in support of IEEE in the area of Big Data. In addition, he is involved in several national and international activities supporting Big Data issues in policy and education; sits on the advisory boards of several companies, journals, and university programs; and is active in research and speaking on topics related to Big Data. He currently holds 31 patents.

Dr. Belanger retired as Chief Scientist of AT&T Labs, and Vice President of Information, Software, & Systems Research. He created the AT&T InfoLab, a very early (1995) participant in "Big Data" research and practice. Prior to that, he led the Software Engineering Research Department at Bell Labs. He holds a Ph.D. in Mathematics from Case Western Reserve University. Among his awards are the AT&T Science and Technology Medal for contributions in very large scale information mining technology; AT&T Fellow for "lifetime contributions in software, software tools, and information mining"; and the IEEE Communications Society Industrial Innovator Award.

Harvey Freeman

David Belanger

For those of us who have been involved in the evolution toward Big Data over the past two decades, one of the most eye opening changes has been the ability to do things that were previously too difficult, too costly, or just impossible. An example is the ability to discover rare events, e.g., fraud, from massive amounts of data. Prior to the ability to move and store terabytes of data, analysts were required to sample data, apply aggregate thresholding tests to the data, and/or spend large amounts on hardware and communications. Even though the amount of data has changed by several orders of magnitude over two decades, it is now possible to isolate specific instances by their own, specific behaviors. Combining this flow of data with advanced analytics, sometimes deep learning, has created opportunities in areas like medicine, video analysis, finance, telecom, etc., many of which are just emerging into production.

This revolution in the use of data is dependent on a variety of things, including: advances in the cost and power of the computers used, particularly in parallel and distributed computing; the availability of software that scales to manage the huge amounts of data; enormous improvements in the analytic techniques available; and improvements in the understanding of valuable uses for the data. At the front of this list needs to be the ability to move the data so that it is available where and when it is most useful, and the role that communications, and the communications industry, has played in the evolution of Big Data.

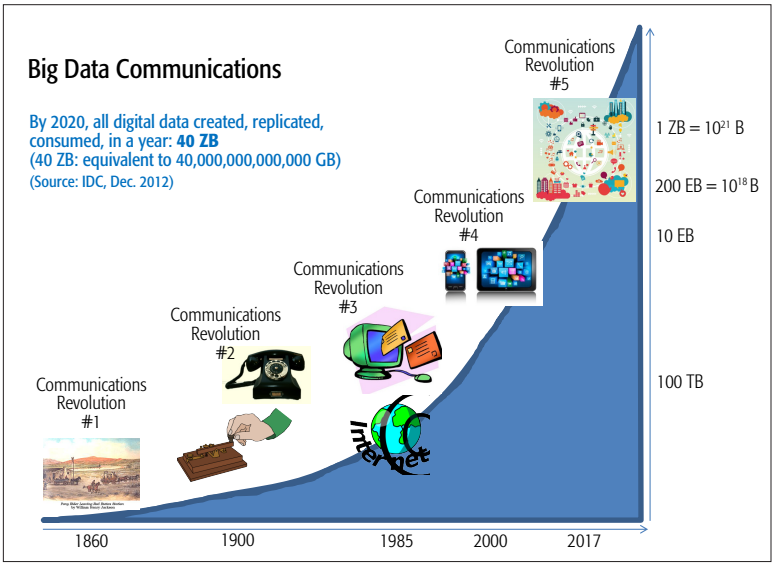
Over time, the role of communication has been central to the evolution of Big Data in at least three ways. The first, and probably most obvious, is the movement of the data itself. The amount of bandwidth, both for access and long haul, has been increasing exponentially for at least two decades, as has the ease of access to the bandwidth. Looking at the accompanying figure, one can clearly see a correlation between step function changes in the communications technologies available, e.g., Internet, IP, 3/4G cellular, WiFi, and the amount of data that is available to be analyzed. It is also not hard to argue that, as the access to these technologies has become easier, e.g., web, smartphones, and apps, the amount of data generated by and results consumed by the "crowd" has exploded.

But this is not the whole story. There are very significant changes in the sources of the data, and in the nature of the data from those sources. Transactional data has long been the core of data used for corporate analysis, e.g., credit card swipes, telephony call detail data, and operational data from industries ranging from finance to manufacturing to web purchases to hospitals. This type of data is still at the heart of much corporate data analysis. Increasingly though, it is augmented by data that is far more detailed, perhaps from search engines, social networks, or even clickstreams and cookies, and is often unstructured, e.g., Twitter tweets and other textual data. This data is detailed enough to allow analysts to consider the behavior of users, as well as their purchasing decisions. We are now at the beginning of yet another step function change in the underlying communications technology, and this will, in turn, drive a new generation of applications, a new and much larger set of sources for data and consumers of the results of analysis, and even more convenient and widespread access to the data and its results. These include: social networking, 5th generation wireless (5G), software defined networks (SDN)/network function virtualization (NFV), Fog, and IoT (Internet of Things)/CPS (cyber-physical systems) among other technologies. Some of these have already changed the way we communicate, and the way we do business. We are only at the tip of the iceberg.

A second way that communications has influenced the evolution of Big Data is the contributions made by communications corporations to the technologies and applications in Big Data. During the early days of the current Big Data, e.g., mid 1990s, the technologies and uses of Big Data were led by a few industries. Significant among them were telecom and finance. Joined by the various cyber industries a bit later and medicine more recently, they still are among the leaders in this space. The

applications that drive the communications industry include: network operations, service operations, customer marketing and service, fraud and security, location based applications, and many more. These applications have been driven by successively larger and more compelling sources of data. On top of this are ever larger, and incredibly rich, sources of data created by very different forms of communications such as social networking (e.g., Facebook, Twitter).

The third way that communications has contributed to Big Data is one that is often overlooked. The policy implications of the availability of such large, and very detailed, sets of data, along with the associated advances in analytics applied to the data, have been significant, critical, and often contentious. These include things like governance, policy, compliance, organization, privacy/security, and the societal impact of applications. What we can expect in these areas is going to determine where big data can actually be leveraged. Some big issues are just starting to evolve, and policy has not yet kept up. An easy place to think about this is in the area of location data. Accurate locations of individuals using GPS or other technologies are now held by a variety of organizations. This data has many critical applications for first responders and law enforcement organizations, as well as less critical but important and widely used applications in social interactions, marketing, manufacturing, and many more.



We are now on the cusp of yet another set of changes in the communications systems that will certainly increase the amount, and value, of big data by several orders of magnitude. In fact, it is likely to lead to what McKinsey and others are referring to as the digitization of our world. The communication technologies involved are led by 5G wireless, which is scheduled to be widely available in the US, and several other countries, within the next few years; and the associated Internet of Things, a potentially far larger, and very different, version of a "crowd" (a "crowd" of things as well as people), which will once again fundamentally change how we think about data, its availability, and its uses. We can expect that the communications infrastructure will support terabit/second communications, tens of billions of sensors operating off very low power sources, and latency in the single digit milliseconds. This will surely support new levels of predictive analysis. It also means that real-time, immersive, augmented reality and gaming applications across wide areas will appear. One could think of, for example, computer games that can be played involving physical as well as virtual motion, across continents, in real time. This could include the use of huge databases that support realism in terms of video, audio, as well as structured data.

As you might expect, the IEEE Communications Society (ComSoc) is very active in Big Data on a number of fronts. Several of our volunteers, including myself, are actively involved in the IEEE Big Data Initiative (bigdata.ieee.org) with activities including: support/proposal of new and existing publications, conferences, and workshops; new standards and educational proposals; an exciting new data repository (https://ieee-dataport.org/) now being trialed; outreach (e.g., Collabratec, Facebook, Twitter, podcasts); analytics contests at selected IEEE conferences; and many other activities. In addition, ComSoc has a Technical Committee on Big Data (TCBD (http://bdpan.committees.comsoc.org/)) actively involved in many Big Data issues related to communications.

Big Data is a prominent topic in nearly every major ComSoc conference, including Globecom, ICC, and NOMS, in the form of speakers, panelists, special sessions, and contests. In addition, ComSoc offers a large suite of webinars, webcasts, white papers, local and international conferences, and other educational programs available to all members on the topics described in this column. To go more deeply into these and other communication areas, I encourage you to visit www.comsoc.org and look at the large catalog of learning material in the areas discussed, and more.

Looking to the future, I expect ComSoc to be deeply involved in a variety of Big Data activities. For example, as the IEEE Big Data Initiative transitions from Initiative status, ComSoc will have a critical role as one of the primary IEEE Societies guiding and supporting that transition. ComSoc has been among the leading Societies in support of the activities of the initiative, and along with several other IEEE Societies who have been active in the Initiative, it will be imperative that, through volunteers, funding, and active support, the various activities already initiated be expanded and made permanent. For example, the data repository IEEE Dataport is now in trial. Success will depend on an active user community, both contributors and users of the data, and on involvement of volunteers to provide guidance in the product evolution, and where necessary, funding until it becomes self-supporting. The analytical contests currently being held at ComSoc and other IEEE Society conferences are aimed at learning how IEEE can more effectively use its valuable storehouse of data. This process will take a while, and we need to be leaders as it evolves through providing sources of data, expertise of volunteers, and continued conference support. In addition, Big Data will continue to be an essential component of conferences, standards, and, educational activities. ComSoc, because of the role that communications plays in Big Data and its future, must take a leadership role going forward. Some newer challenges, for example ensuring the quality and survivability of large, complex Big Data systems, are areas that ComSoc should take the lead in addressing.

So what can we expect in the combination of communications and big data? As described above, some trajectories are clear. Much faster, lower latency, lower-power communications leading to orders of magnitude more data of all forms. Also, the emergence of increasingly powerful techniques for management, analysis, and visualization of data, leading to applications that we probably can't even imagine today. With the increase in machine learning based applications embedded in large systems of systems, we can expect much more research in deploying and operating such systems. Finally, we can expect a dynamism that we haven't come close to yet, driven by more open data about anything and everything, more innovative consumer and business applications, and a world in which we can communicate with nearly everything.

A HISTORY OF RADIO WAVE PROPAGATION: FROM MARCONI TO MIMO

BY J. BACH ANDERSEN

ABSTRACT

Radio waves from a few kilohertz to millimeter-wave frequencies play a key role in modern wireless communications. The development over the last 120 years is traced with an emphasis on communication aspects and physical phenomena rather than theory. The early years were characterized by experiments with no theory and lack of knowledge of ionospheric propagation. High frequency (HF) propagation via the ionosphere at HF frequencies meant global communications for thousands of kilometers. Another natural medium is the atmosphere near the Earth's surface, the troposphere, leading sometimes to anomalous phenomena, but it is also important for satellite signals near the horizon.

Propagation over man-made structures like in an urban environment is covered by the simple Hata equations for first generation cellular systems. Higher generations must include delay information to accurately describe propagation, and the Hata-like equations may be extended into the millimeter frequency range. Indoor propagation may also be covered by a diffuse impulse response. Finally, the promise of increased spectral efficiency is given by multiple-input multiple-output (MIMO), if certain conditions of uncorrelated antenna signals are fulfilled.

THE EARLY YEARS: EXPERIMENTS AND THEORIES

The experimental work on electromagnetic waves started in 1888, when Heinrich Hertz verified Maxwell's theory that the waves propagated with the velocity of light. It is interesting to observe that Hertz had little feeling for the communication possibilities. When asked about the application of the waves in telegraphy, he answered:

*However, the vibrations of a transformer or telegraph are far too slow; take for example a thousand in a second, which is a high figure, then the wavelength in the ether would be 300 km and the focal length of the mirror must be the same magnitude. If you could construct a mirror as large as a continent, you might succeed with such experiments, but it is impractical to do anything with ordinary mirrors, as there would not be the least effect observable. (*Quoted from [1]).

Clearly, the notion of a carrier was not understood. Only seven years later, in 1895, a 20-year-old Italian, Guglielmo Marconi, kicked off the wireless world by establishing a 2 km link behind a hill [2]. It took place from his parents' home, Villa Griffone, near Bologna. The return path was acoustic, a gunshot from an assistant. After that, events moved quickly for Marconi, extending the range to tens of kilometers and eventually to the famous transatlantic transmission in1902. The application was wireless telegraphy, which of course was opposed vigorously by the established cable companies. Other opponents were the scientists who could prove that it was not possible to overcome the curvature of the earth over such large distances. The presence of the ionosphere was not known at that time, so many questioned the results in the beginning. About the same time as Marconi, the scientist Alexander Popov from Russia established a wireless link by improving the receiver apparatus, the so-called coherer [3]. It is not considered fruitful to discuss who was first; the inventions were "in the air." Marconi directed his efforts toward commercial success; Popov was more a scientist.

The transmitter was a multiple sparks generator coupled to a resonant circuit and an antenna, and the following conclusions were obtained by Marconi:

- The antenna should be a vertical wire, and the length of the wire determined the range: the longer the wire, the larger the range.
- Although the spark gap transmitter produced a wide range of frequencies, the propagated frequency band, while still wide, was determined by the circuit and the antenna.

It is clear that the engineer inventor was winning over the theoreticians, who were explaining the results years later. Another interesting controversy was over the apparently simple problem of propagation from a vertical dipole over a flat finitely conducting surface. The Norton surface wave result was published as late as 1936; see [4] for interesting details. The radiation pattern (the sky wave) is easily computed using the Fresnel reflection coefficients.

Despite the huge success of the Marconi system, there were some inherent serious problems related to the transmitter system. They were not apparent if you were the only one in the world, but the incoherent pulsed signal caused the following problems:

- The radiated energy was spread over a large range of frequencies, an inefficient spread spectrum.
- There was a lack of privacy.
- There was no means of selecting the wanted from the unwanted signal from two or more transmitters.

R. Fessenden (Canada) is credited for introducing coherent transmitters, audio broadcasts, and inventions of the heterodyne principle, so gradually the sparks died out and were eventually forbidden.

INTERNATIONAL SCIENTIFIC COOPERATION, URSI, INTERNATIONAL UNION OF RADIO SCIENCE

The first General Assembly of the Union was held in July 1922 in Brussels. At that time, only four National Committees had been formed officially: Belgium, France, the United Kingdom, and the United States. However, the following new Committees adhered to the Union during the same year: Australia, Spain, Italy, Japan, and the Netherlands. We find that, although only as observers, two scientists from Norway participated actively in the work of the Assembly.

The Agenda of that first Assembly had been drawn up by

Figure 1. Marconi's lab on the second floor of Villa Griffone. From the window one can see the Celestini Hill at a distance, the natural obstacle that obstructed the line-of-sight propagation [2].

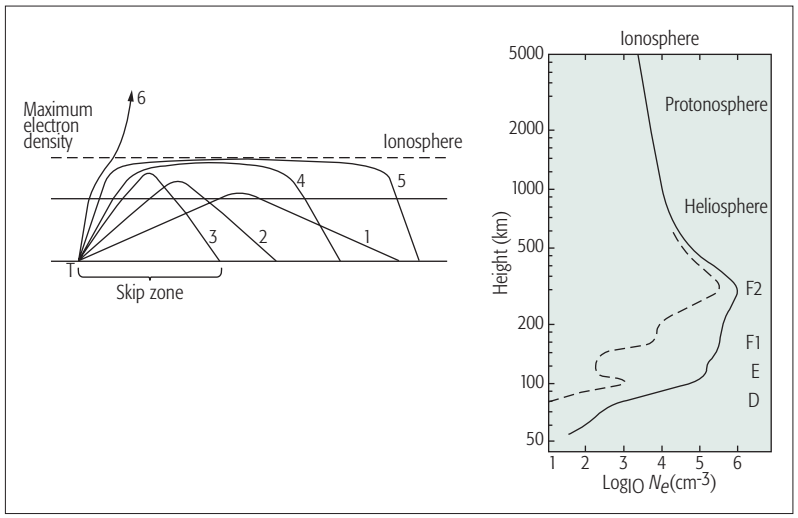


Figure 2. a) Ray paths in the ionosphere; b) height vs. electron density [5] (with permission from Dover Publications).

General Ferrie and Prof. Goldschmidt, to be elected later as President and Secretary General of the Union, respectively. Among the topics to be considered by the Commissions, General Ferrie cited:

- Measurements of the electromagnetic field and its variations
- Study of variations in radio direction finding measurements
- Study of statics and disturbances in general

It was considered that it would not be desirable for URSI to cover tubes since this might have implied a more industrial character, which had to be excluded.

The scientific Commissions formed in 1922 were as follows:

- Measurements Methods and Standardisation
- Radio Propagation, with two Sub-Commissions on the electromagnetic field and on radio direction finding, respectively (taken from www.ursi.org)

URSI is still very much active; the XXXIInd URSI General Assembly and Scientific Symposium (GASS) will be held in Montreal in 2017.

The Ionosphere

In 1918 G. N. Watson published a solution to the diffraction around the spherical Earth with the conclusion that the resulting fields in the shadow region were much smaller than experimentally observed ones, so a different explanation was necessary.

It was not until the 1920s that the presence of ionized regions in the upper atmosphere was scientifically established, where names like Appleton and Chapman were important. The free charges like electrons are created by ionizing radiation from the Sun at different heights, typically between 50 and 400 km above the Earth's surface, varying over the seasons, monthly and diurnally. Most phenomena may be explained by the simple expression

$$n^2 = \varepsilon = 1 - \frac{f_p^2}{f^2}$$

where n is the index of refraction, ε is relative permittivity, f_p^2 is proportional to the electron density, and f is the frequency. f_p is called the plasma frequency. The relative permeability is one, so the medium is non-magnetic. The equation indicates an unusual medium with propagation for $f > f_p$ ($0 < \varepsilon < 1$) where most normal media will have $\varepsilon > 1$, and exponential decay for $\varepsilon < 0$. A more exact equation would include effect of losses and influence of the Earth magnetic field. If the geo-magnetic field is important, the link is no longer reciprocal, a property that is important for time-division duplex systems.

Figure 2a shows different ray paths for a given frequency and various angles of departure, and Fig. 2b the electron density and height for the various layers. Due to the less-than-one index of refraction the paths are refracted away from the normal and eventually reflected except ray 6, which passes through due to the steep incidence, rays 1, 2, and 3 are normal long ranging rays, while 4 and 5 are high-ray paths. Note that there is a certain region near the transmitter, the skip zone, where there are no rays and covered only by the surface wave. The useful range may be several thousand kilometers, and the useful frequency range will be in the HF region from 3 to 30 MHz. The lower frequencies from 300 kHz to 3 MHz suffer from heavy absorption during the day, while the still lower frequencies offer too small bandwidth.

Marconi realized the benefit of higher frequencies in the 1920s and established good links over several thousand kilometers with moderate power and simpler antennas, with a system called the Short-Wave Beam System [6] with frequencies above 3 MHz. Today, ionospheric transmissions have limited applications due to the presence of satellites.

Meteor Burst Propagation

A great number of meteors constantly enters the Earth's atmosphere and thereby creates ionized trails that may be exploited for communications [7]. Of course, the intermittent nature of the link limits the applications. They operate with carrier frequencies from 30 to 100 MHz with rates between a few tens and a few hundred bits per second. Maximum path length is about 2000 km. Typical waiting times are between a few seconds and a few minutes. One use is communicating data from unmanned measurement stations.

The Troposphere

The troposphere is the lowest portion of the atmosphere, up to about 10 km. Unexpected anomalous propagation for short radio waves beyond the horizon by radars during World War II served to generate research in tropospheric propagation.

The humidity, temperature, and density of the lower atmosphere determine the index of refraction N and play an import-

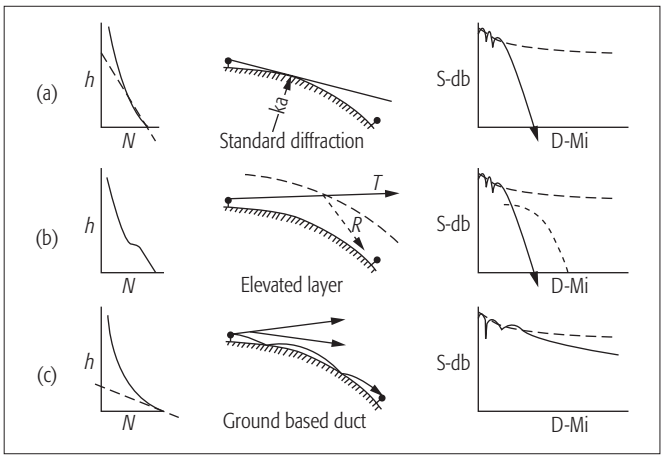


Figure 3. Tropospheric propagation mechanisms for various values of refraction index N. D is distance, h height [8].

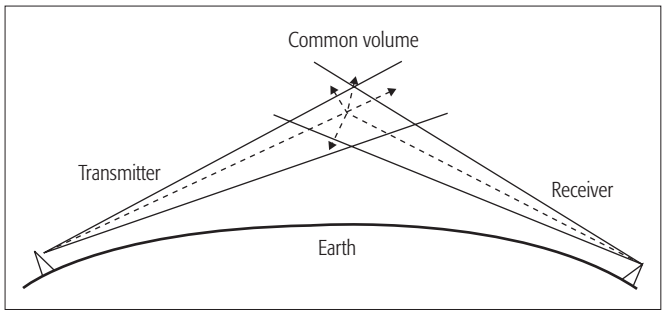


Figure 4. Troposcatter paths [9].

ant role in communications, as sketched in the diagrams of Fig. 3 [8].

In (A) the index of refraction decreases linearly with height, which is the standard situation, and it may be shown that this gives rise to a downward bending of a ray, so the radio horizon is further away than the geometrical horizon. Instead of bending the ray, one may introduce an effective larger Earth radius, typically 4/3 of the true radius for the standard atmosphere, and the rays may then be drawn as straight. The fading before the horizon is due to the interference between the direct and the ground reflected ray (third column).After the horizon the power drops rapidly. The refraction effects have an influence on near horizon satellites.

The second group (B) illustrates an elevated layer due to an abrupt change in the temperature and water vapor content. The third group(C) illustrates a waveguide effect or duct giving rise to propagation around the surface with relatively little attenuation. It is most likely to occur over coastal regions where warm dry air masses flow over a cooler sea[8]. The mechanisms are infrequent and unpredictable for long-range communication systems. There is, however, one mechanism that has led to a useful and reliable link: Troposcatter, first introduced or discovered in the 1950s.

As indicated in Fig. 4, the link is established by scattering from inhomogeneities in the lower atmosphere. A high gain transmit antenna is directed slightly above the horizon and illuminates the scatterers, and a small part of the forward scattered energy is picked up by the receiver high gain antenna. A few gigahertz are used, and data rates above 20 Mb/s may be achieved with latency of a few milliseconds, making the technique attractive for military systems. In addition, the probability of intercept is low [10].

The atmosphere in itself contains a number of different gases and water particles depending on the humidity. They contribute to the overall attenuation, as shown in Fig. 5, of relevance for propagation from satellites and for propagation along ground. The most severe peaks are the water resonance at 22 GHz and the dry air resonance at 60 GHz with attenuations of 0.2 dB/km and 15 dB/km, respectively [10].

In the satellite case rain and ice particles severely affect radio links above 10 GHz as shown in a comprehensive study from 1982 [11] by Cox and Arnold.

It is expected that fifth generation (5G) mobile communications will utilize millimeter waves, so the peaks should be avoided. On the other hand, taking advantage of the peaks will limit interference from neighboring cells in cellular systems.

PROPAGATION IN THE URBAN ENVIRONMENT

The first cellular networks introduced in the early 1980s were narrowband analog systems. Examples are Advanced Mobile Phone System (AMPS) in the USA and Nordic Mobile Telephone (NMT) in the Nordic countries, and the use was mainly for cars: a vehicular system.

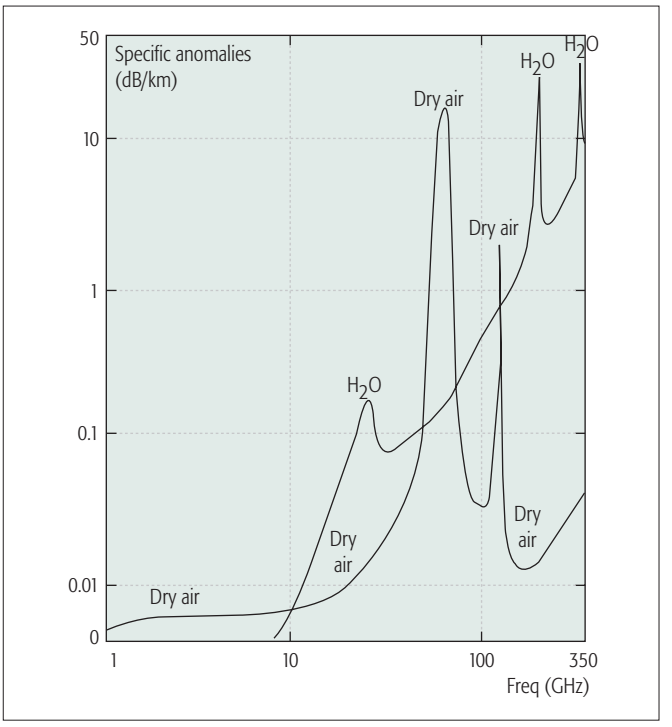


Figure 5. Atmospheric attenuation vs. frequency [10].

The most important propagation parameter for a wireless system is the power density, determining the coverage from a base station. In 1968 Okumura did an extensive set of measurements in Japan for different frequencies from150 to 1500 MHz, and as a function of antenna height and environment. In 1980 Hata [11] transformed the data into a set of simple equations like $L = \gamma d^{\alpha}$ where L is the mean path loss, the ratio between the transmitted and received power; α is the power law exponent equal to 2 for free space and equal to 4 for a flat lossy ground — for the urban and suburban environments, the value lies between 3 and 4. The factor γ depends on the environment, and d is distance. It is an interesting observation that α is independent of frequency, only dependent on the base station antenna height. Expressed in dB, the Hata formula reads

$$PL(d) = PL_0 + 10\ \alpha\ \log10(d/d_0) + S_\sigma$$

where S is a Gaussian zero mean variable with a standard deviation σ depending on the environment, explaining the slow fading, and d_0 a reference distance.

The simple power law has shown validity in many other situations, even indoors, and for urban structures quite different from those in Japan. It is worth noting that delay information is not included.

For the propagation specialists, there have been many other results from ray tracing, diffraction theory for over roofs, and integral equations for rural propagation, and others. They serve more to understand the physics, but they all lack the simplicity of the Hata model.

There is also a fine structure in the received power due to multipath, that is, waves are arriving from a number of different directions with different amplitude and phases, and they will sometimes enhance the signal, sometimes combine destructively to a small value; the phenomenon is usually called fast fading. The most severe case leads to Rayleigh fading, as illustrated in Fig. 6, for the case of f equal to 900 MHz and a vehicle speed of 120 km/h. Less severe fading is the case if there is a direct line of sight from the base antenna to the mobile antenna. The mean

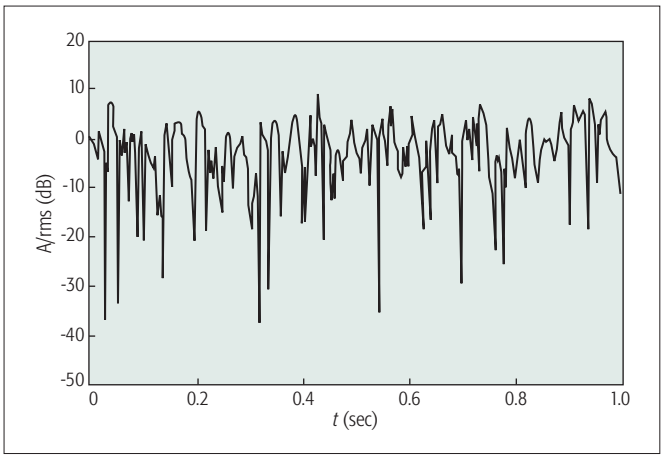


Figure 6. Rayleigh fading for f =900 MHz and for a vehicle speed of 120 km/h.

power is measured by averaging over a few wavelengths. The fast fading may be mitigated by using diversity, while the slow fading is unavoidable.

The faster the speed and the higher the frequency, the more rapid the fluctuations. We can also interpret the figure as a spatial distribution with x-axis distance instead of time, so if standing still one might be so unfortunate as to suffer a 30 dB loss. The remedy is to move a little.

DIGITAL TECHNIQUES

It was clear that international cooperation was necessary to define the next system. It started in 1982 with setting up a working group (Groupe Special Mobile, in short GSM), also known as Global System for Mobile Communication. The cooperation was successful, and spectrum was allocated in the 900 MHz band and later in 1993 extended to the 1800 MHz band. The first mobile 2G call was in 1991.

Compared to the narrowband fading discussed above, the fading is less severe, since individual frequency fades are only part of the total bandwidth; there will be other frequencies that do not fade. The fading is thus frequency selective.

Combined delay-Doppler scattering function was first measured in 1973 [13]. One early work on impulse propagation was done in 1975 by Cox [14]. The original narrow impulse is smeared out in time due to the scattering from buildings and other objects, and this will lead to intersymbol interference depending on the details of the system. It is customary to use the root mean square (rms) delay spread as a characteristic measure of the spread. In the case of Fig. 7 it is 1.2 μs.

Considerable work on propagation supporting the development of GSM under the European Union (EU) was done under the Cooperation in Science and Technology (COST) framework supported meetings and other collaborations.

The trend was to higher and higher carrier frequencies with the promise of higher bandwidth, and recently the millimeter-wave frequencies were studied in various environments [15] by Rappaport in 2013. Modeling the path loss is again done with a Hata-like model with an example shown in Fig. 8 based on formulas in [16]. Atmospheric attenuation is not included.

The path loss increases with the square of the frequency, so if we have an acceptable link budget at 5 GHz as an example, we need an additional power of 100 times or 20 dB for 50 GHz. This path loss is an average over both the slow fading due to the changing environment and fast fading as discussed above.

The experiments are based on static measurements and steerable antennas searching for the maximum power[15]. In

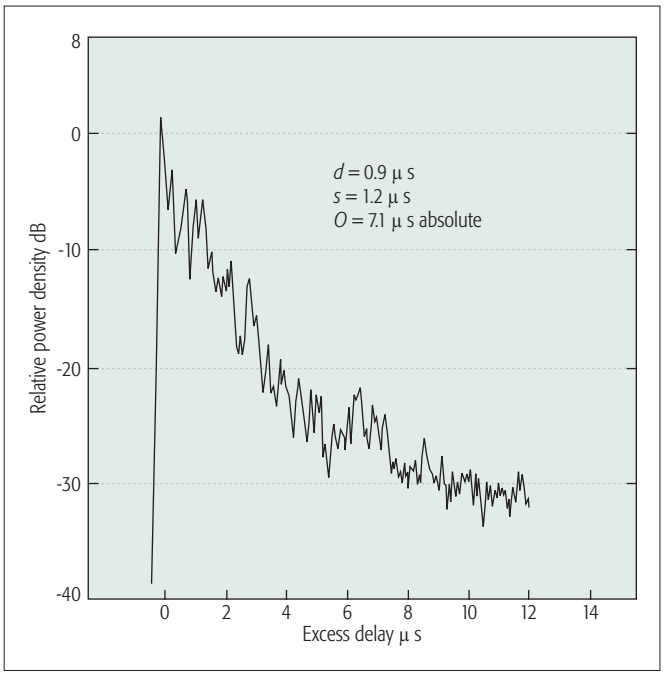


Figure 7. Average power delay profile measured in New York City [14] by Cox in 1975. The frequency is 910 MHz.

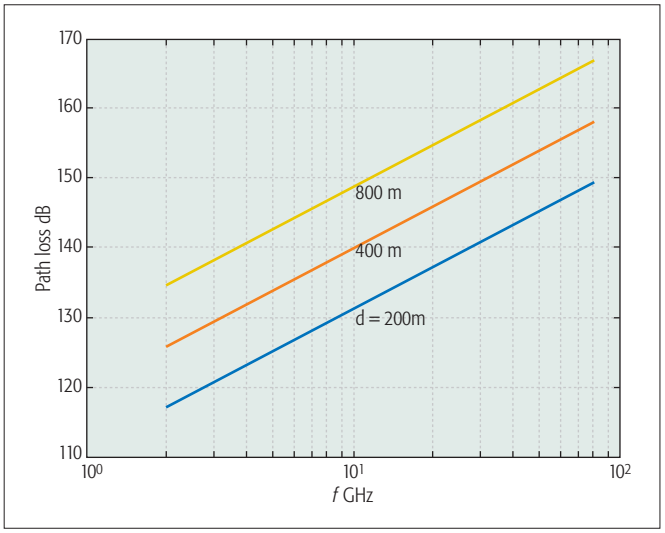


Figure 8. Path loss vs. frequency for a non-line-of-sight urban case, based on [16].The slow fading is not shown.

a true mobile case the antennas must deliver the missing 20 dB jointly for transmitter and receiver. This calls for an adaptive array with many elements with the trouble of having to find the beamforming factors, or more simply, just beam scanning. As if this was not enough, there is an additional problem: the Doppler spread increases linearly with frequency, so the beamforming must be updated every few millimeters.

PROPAGATION IN THE INDOOR ENVIRONMENT

The growing interest in the wireless industrial environment with machine or robot connections [17] or the office environment has led to new models of propagation. Saleh and Valenzuela [18] suggested in 1987 that the mean impulse response consists of clusters with each cluster having its own response with exponential decay.

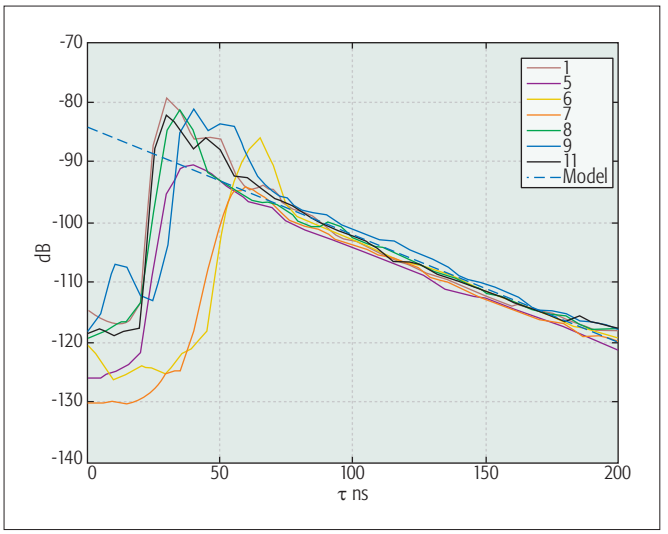


Figure 9. The average power delay profiles vs. delay in nanoseconds for different positions in an 11 m × 19 m furnished room. The dashed line is the theoretical reverberation model: frequency 5.8 GHz, bandwidth 100 MHz [20].

The Hata model was also used by Ghassemzadeh *et al.* [19] for different types of indoor environments, commercial buildings, and single-family homes.

An alternative theory by Andersen in 2007 [20] is to consider the room as a cavity with lossy walls where Fig. 9 shows the average power profile at different locations. It is noteworthy that for a given delay on the tail, the scattered diffuse power is independent of position.

It is also interesting that radio wave propagation in a room is similar to acoustical wave propagation for the same wavelengths.

MIMO

Winters [21] pointed out in 1987, before it was called MIMO (Multiple Input Multiple Output), that for two arrays with M elements each, up to M independent channels can be established in the same bandwidth.

MIMO is not really a topic only for propagation, but is intimately connected with antennas, so we need to digress slightly from our main issue and introduce some antenna topics. The problem may be illustrated as in Fig. 10, where two arrays are communicating via a number of scatterers. The arrays have M and N elements, and for convenience we assume the left array to be transmitting. In the example, $M = 5$ and $N = 2$. The path shown is just an example; all elements couple with all elements via many scatterers. The interesting case is when the angular spread is large seen from both arrays, also expressed as low correlation between the elements. This will be the typical case for outdoor and indoor situations when there is no direct line of sight. The result is a number of independent parallel channels equal to min(N,M) leading to an increase in spectral efficiency. If the angular spread is small seen from one side, there is only one channel [22].

An alternative situation is where one array is replaced by many users, and the base station consists of many, maybe hundreds, of elements, called massive MIMO [23].

CONCLUDING REMARKS

We have been looking back in time, including the most recent achievements. It is relevant to look forward and mention a few things we have not covered. The Internet of Things, IOT, requires widespread propagation in an environment. Wearable antennas need studies of propagation near the human body. The use of optical frequencies may be a solution in certain cases. It is certain that the history of wave propagation does not end here.

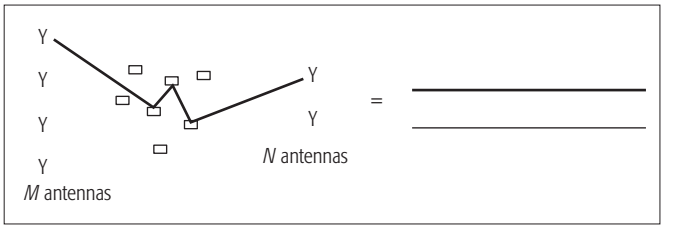


Figure 10. Two linear arrays of M and N elements in a scattering environment. For $N = 2$ there are two independent channels where the relative gains depend on the angular spreads of the scatterers seen from the arrays.

REFERENCES

[1] H. Sobol,"Microwave Communications — A Historical Perspective," *IEEE Trans. Microwave Theory and Techniques*, vol. 32, no. 9, Sept. 1984, pp 1170–81.
[2] G. Falciasecca, "Marconi's Early Experiments in Wireless Telegraphy, 1895," *IEEE Antennas & Propagation Mag.*, vol. 52, no. 6, Dec. 2010, pp 220–21.
[3] O. G. Vendik, "Contribution of Prof Alexander S Popov to the Development of Wireless Communications," *Euro. Microwave Conf.*, vol 2, 1995, 895–902.
[4] J. R.Wait,"The Ancient and Modern History of EM Ground-Wave Propagation," *IEEE Antennas and Propagation Mag.*, vol. 40, no. 5, Oct. 1998, pp. 7–24.
[5] K. Davies, *Ionospheric Radio Propagation*, Dover, 1966.
[6] W. J. Baker, *A History of the Marconi Company*, Methuen & Co Ltd, 1970.
[7] I. A. Glover, "Meteor Burst Propagation," *Electronics & Commun. Engineering J.*, Aug. 1991, pp. 185–92.
[8] J. Chisholm, "Progress of Tropospheric Propagation Research Related to Communications Beyond the Horizon," *IRE Trans. Commun. Systems*, vol. 4, no. 1, 1956, pp. 6–16.
[9] E. Dinc and O. B. Akan, "Coherence Time and Coherence Bandwidth of Troposcatter Links for Mobile Receivers," *IEEE Vehic. Tech. Mag.*, vol 18, June 2015, pp. 86–92.
[10] ITU-R Rec. P.676-3, "Attenuation by Atmospheric Gases," 1997.
[11] D. C. Cox and H. W. Arnold, "Results from the 19- and 28-GHz COMSTAR Satellite Propagation Experiments at Crawford Hill," *Proc. IEEE*,vol. 70, no. 5, May 1982.
[12] M. Hata, "Empirical Formula for Propagation Loss in Land Mobile Radio Services," *IEEE Trans. Vehic. Tech.*, vol. 29, no 3, Aug. 1980, pp. 317–25.
[13] D. C. Cox, "A Measured Delay-Doppler Scattering Function for Multipath Propagation at 910 MHz in an Urban Mobile Radio Environment," *Proc. IEEE*, Apr. 1973.
[14] D. C. Cox and R. P. Leck, "Correlation Bandwidth and Delay Spread Multipath Propagation Statistics for 910-MHz Urban Mobile Radio Channels," *IEEE Trans. Commun.*, vol. 23, 11, Nov. 1975.
[15] T .S. Rappaport *et al.*, "Millimeter Wave Mobile Communications for 5G Cellular: It Will Work!," *IEEE Access*, vol.1, 2013, pp. 335–49.
[16] S. Sun *et al.*, "Investigation of Prediction Accuracy, Sensitivity, and Parameter Stability of Large-Scale Propagation Path Loss Models for 5G Wireless Communications," *IEEE Trans. Vehic. Tech.*, vol. 65, no. 5, May 2016, pp. 2843–60.
[17] M. Cheffena, "Propagation Channel Characteristics of Industrial Wireless Sensor Networks," *IEEE Antennas & Propagation Mag.*, Feb. 2016, pp. 66–73.
[18] A. A. M. Saleh and R.A. Valenzuela, "A Statistical Model for Indoor Multipath Propagation," *IEEE JSAC*, vol. 5, 2, Feb. 1987, pp 128–37.
[19] S. S. Ghassemzadeh *et al.*, "An Empirical Indoor Path Loss Model for Ultra-Wideband Channels," *J. Commun. and Networks*, vol 5, no 4, Dec. 2003.
[20] J. Bach Andersen *et al.*, "Room Electromagnetics," *IEEE Antennas & Propagation Mag.*, vol. 49, no. 2, Apr. 2007.
[21] J. H Winters, "On the Capacity of Radio Communication Systems with Diversity in a Rayleigh Fading Environment," *IEEE JSAC*, vol. 3, no. 5, June 1987, pp. 871–78.
[22] J. Bach Andersen, "Antenna Arrays in Mobile Communications: Gain, Diversity, and Channel Capacity," *IEEE Antennas & Propagation Mag.*, vol. 42, no 2, Apr. 2000, pp. 12–16.
[23]E. G. Larsson *et al.*, "Massive MIMO for Next Generation Wireless Systems," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 186–95.

BIOGRAPHY

JØRGEN BACH ANDERSEN (M'68–SM'78–F'92–LF'02) (jba@es.aau.dk) received the M.Sc. and Dr.Techn. degrees from the Technical University of Denmark (DTU), Lyngby, Denmark, in 1961 and 1971, respectively. From 1961 to 1973, he was with the Electromagnetics Institute, DTU, and since 1973 he has been with Aalborg University, Aalborg, Denmark. Prof. Andersen is a former Vice-President of URSI from which he received the John Howard Dellinger Gold Medal in 2005.

Higher. Wider. Faster.
Test solutions for 5G.

The next major step beyond LTE/LTE-Advanced (4G) sets challenging requirements. Rohde&Schwarz is a world leader in all areas of RF and microwave test and measurement equipment. As a technology expert, we have been actively involved in mobile communications since the first generation. We are committed to supporting the wireless communications industry with the solutions needed to investigate, develop and standardize 5G.

Check out our test solutions at www.rohde-schwarz.com/ad/5G

Visit us at
Mobile World Congress
in Barcelona,
February 27 to March 02,
hall 6, booth 6B50 and 6C40

ROHDE&SCHWARZ

REGISTRATION OPENS FOR IEEE ICC 2017

PARIS, FRANCE, 21–25 MAY 2017

The registration process is now open for IEEE ICC 2017 to be held on 21–25 May 2017 at the Palais des Congrès–Porte Maillot in Paris, France, which is located only minutes away from the Arc de Triomphe and the legendary Avenue Champs Elysées. The conference theme, Bridging People, Communities, and Cultures, reflects the rich cultural diversity of the Paris metropolis. The technical program will include over 1100 technical papers disseminating the latest research results in communications and networking, 6 keynotes by industry leaders and eminent academic figures, 18 workshops and 24 tutorials on hot topics, as well as 18 industry sessions and panels.

"The magnificent atmosphere of Paris will offer a spectacular setting for IEEE ICC 2017 attendees," said Prof. Hikmet Sari, the Executive Chair of the conference. "In this unique environment, the technical program that we are setting up will present the latest visionary research, services, and applications that are actively reshaping our businesses, entertainment, and social practices worldwide."

"We look forward to providing everyone involved in research and in the development and deployment of future communications networks and services with an in-depth working experience and live interaction with the greatest minds in industry and academia" adds Prof. Sari.

IEEE ICC 2017 will start on Sunday, 21 May, with 12 tutorials presented by leading researchers and industry experts and 8 workshops covering diverse advanced topics. Specific areas covered include the trends and technical challenges in 5G networks, Internet of things (IoT), green technologies, software defined networks, underlying basic technologies, and future services. Later that evening, the conference will host the Welcome Reception, where attendees will have the opportunity to meet old friends and make new ones in a cozy setting.

Monday, 22 May will mark the start of the main conference with two keynotes in the Opening Ceremony. Keynote addresses will be delivered by Marcus Weldon, President of Bell Labs and Corporate CTO of Nokia, and Georges Karam, Founder, President and CEO of Sequans Communications, which is a world-leading LTE for IoT chipset solution provider for machine-to-machine (M2M) and Internet of things devices. The day will continue with technical presentations and industry panels organized in 18 parallel sessions. The Monday program will also feature the traditional Awards Luncheon during which a number of highly deserving colleagues will be recognized for their technical achievements and/or service to the society.

The program of Tuesday, 23 May will feature another keynote session with two distinguished speakers. The first speaker will be Wen Tong, CTO of Huawei, followed by Alain Aspect, a world-renowned scientist in the field of quantum optics and computing. The remainder of the day will be organized in three 90-minute time blocks during which there will be 18 parallel sessions for technical presentations and industry panels and sessions. Tuesday evening will feature the conference banquet, where attendees will enjoy traditional French cuisine along with entertainment in a very friendly and relaxed setting at the Bois de Boulogne which offers a spectacular view on the Eiffel Tower.

The program of Wednesday 24 May will closely follow that of the previous day. The plenary session of the day will feature another set of two renowned keynote speakers. The first will be Serge Willenegger, Senior Vice President, Product Management of Qualcomm Technologies, Inc., and the second is Giuseppe Caire, a world-class figure in communication theory and wireless communications. "We are honored and delighted that 6 keynote speakers of that caliber have accepted our invitations to share their vision with IEEE ICC 2017 attendees. Our industry forums and panels will feature great figures from industry worldwide." adds Prof. Sari.

In the Exhibits Hall, which will be open from Monday, 22 May to Wednesday, 24 May, a number of companies will exhibit their technologies, new products and services. Morning and afternoon networking breaks will be centrally located in that area, and industry leaders will have the opportunity to make live podium presentations here in front of a very large audience.

Finally, the program of Thursday, 25 May will include 12 tutorials and 10 workshops devoted to very hot topics for future communications networks and technologies. The Sunday, 21 May Tutorials and Workshops and the Thursday, 25 May Tutorials and Workshops are offered as separate packages with separate registrations. The registration for each one of these days gives access to any tutorial and any workshop on that day.

IEEE ICC 2017 promises to be a great event in a fantastic environment. We invite you to mark this event in your calendar and make your plans to attend and bring your family along for pre-conference and post-conference tours and visits. For any information on IEEE ICC 2017, please contact Heather Ann Sweeney of IEEE ComSoc at h.sweeney@comsoc.org, or our local PCO, C2B Congress at icc2017@c2b-congress.com. All visitors to the IEEE ICC 2017 website are also invited to reach out to friends and colleagues through the conference's Twitter, Facebook, and LinkedIn pages.

GCN GLOBAL COMMUNICATIONS NEWSLETTER

February 2017
ISSN 2374-1082

MEMBERSHIP SERVICES

Asia/Pacific Region

Interview with Takaya Yamazato, Director of the AP Region

By Stefano Bregni, Vice-President for Member and Global Activities, and Takaya Yamazato, Director of the AP Region

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

In this issue, I interview Takaya Yamazato, Director of the Asia/Pacific Region (AP). Takaya is a professor at the Institute of Liberal Arts and Sciences, Nagoya University, Japan. He received the Ph.D. degree from Keio University, Yokohama, Japan, in 1993. From 1993 to 1998 he was an assistant professor in the Department of Information Electronics, Nagoya University, Japan. From 1997 to 1998 he was a visiting researcher in the Research Group for RF Communications, Department of Electrical Engineering and Information Technology, University of Kaiserslautern, Germany. In 2006 he received the IEEE Communication Society Best Tutorial Paper Award. He served as symposium co-chair of ICC 2009 and ICC 2011. From 2008 to 2010 he was the chair of the IEEE ComSoc Satellite and Space Communications Technical Committee and the editor-in-chief of the Japanese Section of IEICE Transactions on Communications. His research interests include visible light communication, intelligent transport systems, stochastic resonance, and open educational resources.

Stefano Bregni

Takaya Yamazato

It is a pleasure for me to interview Takaya for this issue and offer him this opportunity to outline his current activities and plans as Director of the AP Region.

Bregni: Takaya, among all Regions, the AP Region is historically probably the one with the largest and most structured Board. Would you please introduce briefly the AP Region Board?

Yamazato: For over 20 years, the Asia/Pacific Board (APB) has been the premier network for ComSoc members in the Asia/Pacific (AP) Region. Our mission is to address all ComSoc activities and programs pertaining to its members and Chapters in the A/P region. In the AP Board, we have 45 officers who work tirelessly to provide value to our members by offering exceptional services. APB has five committees, and three of our vice directors take the initiative on each committee activities.

Bregni: What are the activities and scope of those five APB Committees?

The Technical Affairs Committee (TAC), which is led by vice director Prof. Saewoong Bahk and TAC co-chairs Prof. Jianwei Huang and Prof. Takahiko Saba, is responsible for two of our awards, namely the Asia Pacific Young Researcher Award and the Asia Pacific Outstanding Paper Award. This year's awardees were honored at our last APB general meeting held during Globecom2016.

The Information Services Committee (ISC), which is guided by vice director Dr. Sumei Sun and ISC co-chairs Prof. Yao-Win Peter Hong and Prof. Hung-Yun Hsieh, is responsible for our APB homepage and APB Newsletter. They launched a new homepage in September, and the page is nicely done. Please visit our homepage at http://apb.regions.comsoc.org. You will find information such as the list of awardees, APB Newsletters, our meeting slides, and more.

The Meeting & Conference Committee (MCC), which is also guided by vice director Sun and MCC chair Prof. Jiming Chen, promotes participation of the A/P region in ComSoc conferences and meetings. The MCC records conferences held in the A/P region.

The Membership Development Committee (MDC), which is led by vice director Prof. Wei Zhang and MDC co-chairs Associate Prof. Youngchul Sung and Prof. Meixia Tao, work jointly with the ComSoc chapters to promote ComSoc membership in academia and industry by better serving regional needs.

The Chapters Coordination Committee (CCC), which is also led by vice director Zhang and CCC co-chairs Prof. Eiji Oki and Prof. Miki Yamamoto, collaborates with the A/P Office to run the Distinguished Lecturer Tours (DLTs) and Distinguished Speaker Programs (DSPs). Prof. Oki is the DLT/DSP coordinator. At the time of this writing, 17 DLTs and seven DSPs have been organized. The CCC also collaborates with Sister Societies in the A/P region, and Prof. Yamamoto is in charge of the Sister Societies. Currently, we have eight Sister Societies in the A/P region.

Bregni: Now, would you give us a brief overview of the membership of the AP Region?

Yamazato: The number of ComSoc members in the A/P region is 7,733, representing 24.86% of the 31,112 ComSoc members (as of Dec. 2015). Unfortunately, the overall number of ComSoc members in the APB region decreased from 13,178 in Dec. 2014 to 7,733 in Dec. 2015. Among them, 7,430 are higher grade members (graduate students are included in higher grade members, who have voting rights) and 303 are simple student members (147 or 49.5% are from India).

Bregni: Such a decrease in ComSoc membership from 2014 to 2015 has been significant. Has this problem been particularly serious in the AP Region? Do you see any specific issue related to AP Region membership?

Yamazato: This is not just an issue in the AP Region, but also in other regions, as made evident by Table 1. Moreover, as student members are potential researchers, technologists, and educators in the communication field, we need to retain student members as higher grade members.

Another issue is that there are some members who do not belong to any chapter/council (see Table 2). There are 42 Chapters in the AP region and 396 members (corresponding to 5% of the total, from Korea, China, and Pakistan) who do not belong to any Chapter. We need to work on this issue by setting up new Chapters in Korea, China, and Pakistan.

Bregni: In 2016, the IEEE ComSoc Chapter-of-the-Year Award has

(Continued on Newsletter page 4)

CHAPTER REPORT

IEEE ComSoc Romania Chapter

Winner of the 2016 Chapter Achievement Award; Highlights of Activities in 2015

By Vasile Bota, Chair of the IEEE ComSoc Romania Chapter

The IEEE ComSoc Romania Chapter has most of its members and activities concentrated in four major academic and industrial cities of Romania: Bucharest, Cluj-Napoca, Timisoara, and Iasi. The activities of the chapter in 2015 were focused on the following main directions:

•Organizing or contributing to the organization of international conferences on communications or related topics.

•Organization of scientific symposia dedicated to students' research activities.

•Organization of presentations, short courses, and seminars offered by specialists from well known companies, which are dedicated both to students and other communications specialists.

IEEE BlackSea Communications & Networking 2015 and Other International Conferences

Chapter members contributed to the organization of international conferences on communications or to conferences organized jointly with other chapters of the IEEE Romania Section, by organizing sessions that pointed out the link of the respective field to communications.

The most important event of the chapter's activity in 2015 was the IEEE BlackSea Communications & Networking 2015 Conference (http://blackseacom2015.ieee-blackseacom.org/). Organized by IEEE ComSoc, chaired by Prof. H. Sari, vice president of IEEE ComSoc, and by dr. Al. Gelman, the conference was held at the Maritime University in the city of Constanta, the greatest Romanian city on the Black Sea coast. The conference patron was Orange Romania.

Members of the Romania IEEE ComSoc chapter had important roles in organizing this event. On the scientific side, Prof. V. Bota from Technical University of Cluj-Napoca was Technical Committee Program (TPC) co-chair, and twelve other members of the chapter were TPC members and reviewers. Moreover, Prof. O. Fratu was Operations Chair, coordinating the local organizing team from the Politehnica University of Bucharest (PUB) and Maritime University of Constanta.

The conference had three main tracks: regular technical papers, special sessions, and work-in-progress papers. Out of the 110 eligible papers submitted in the regular papers and special sessions tracks, 53 were for oral presentation, while another 15 papers were selected in the work-in-progress track.

The conference's technical program included two keynote presentations, the patron presentation, two tutorials, seven regular technical sessions, three special sessions, one poster session, and one demos session.

The keynote presentations, offered by Dr. M. Kontouris from Supelec (France) and Prof. A. Polydoros from the University of Athens (Greece), dealt with very hot research topics in the field, i.e., "Scientific Challenges of 5G" and "Opportunistically Cooperating Radios in Action", respectively.

The two half-day tutorials, namely "Green Heterogeneous Networks" offered by Dr. M. Z. Shakir, Texas A&M University at Quatar, and "Network Infrastructure for IoT and M2M Communications", offered by Dr. B. Sikdar, National University of Singapore, also covered current research topics. The keynote speeches and tutorials were highly appreciated, being watched by local and remote attendees, due to their broadcasting to PUB.

The seven regular technical sessions contained 42 high quality papers, while the three special sessions included another 11 selected papers. The poster session, and the demos session, with eight presenters, completed the program.

The TCP offered two awards: the Best Paper Award, which went to G. Montorsi and F. Kayhan from Politecnico di Torino, Italy, and the Best Student Paper Award, which was won by G. Ricciutelli, M. Baldi, N. Maturo, and Fr. Chiaraluce from the Università Politecnica delle Marche, Italy.

IEEE BlackSea Communications & Networking 2015 Chairs, prof. H. Sari and dr. A. Gelman offering prof. A. Polydoros a diploma at the end of his keynote presentation.

Dr. F. Kayhan receiving the Best Paper Award from the TPC co-chairs of the IEEE BlackSea Communications & Networking 2015.

The technical program also included two panels, discussing "5G Mobile System: Is It Just New Frequency Bands or Did We Miss Something Crucial in Earlier Generations?" moderated by Prof. M. Latva-Aho, University of Oulu (Finland), and "Internet of Things: How to Break the Barriers for Business Development?" moderated by prof. R. Verdone, University of Bologna (Italy).

The conference was attended by authors from 22 countries, from Europe, Asia, North America, Africa, and by some 150 attendees, specialists, and students, and by the remote attendees at the Politehnica University of Bucharest, which shows the interest aroused by the conference.

The chapter members also contributed to two other conferences organized jointly with other chapters of IEEE Romania Section. Chapter members from the PUB, coordinated by Prof. V. Coitoru, contributed to the "9th International Symposium on Advanced Topics in Electrical Engineering: ATEE 2015" held in Bucharest on May 7–9, 2015, by organizing a seven-paper special session on "Communications and Information Technology & Electric Circuits". They also contributed to the organization of the "8th International Conference on Speech Technology and Human-Computer Dialogue 2015", which took place at the Politehnica University of Bucharest between 14 - 16 October.

Organization of Scientific Symposia for Students

The second direction of activity was the organization of scientific symposia dedicated to the research activities of the students in telecoms. Such symposia were organized in the academic centers. As an example, the yearly one-day Symposium of Electronics and Telecommunications organized by the Communications and Applied Electronics Departments of the Technical University of Cluj-Napoca (TUCN) was an opportunity for the students to present the results of their research or of graduation theses. It had a section on communications that is structured on two levels: research results of M.Sc. and Ph.D. students, and the graduation projects for bachelor students. The 2015 edition offered best paper awards and prizes, and the best papers were published in the TUCN review "Novice Insights". The symposium was of great

CHAPTER REPORT

A Quick Look at ComSoc San Diego Chapter in 2016

By Liangping Ma, InterDigital Communications, Inc., IEEE ComSoc San Diego Chapter Chair, USA

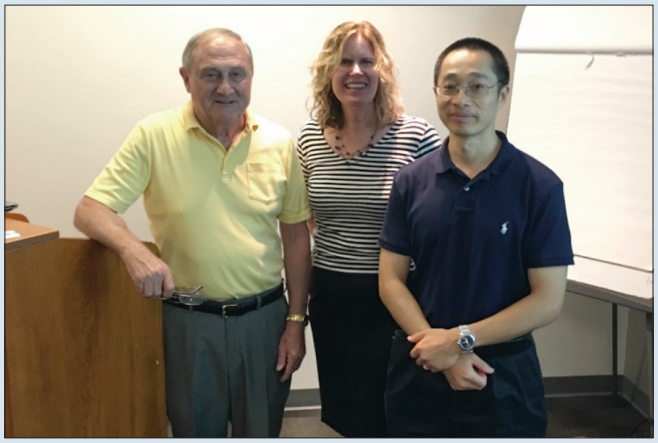
Serving the greater San Diego area, a wireless communication hub of the world, the IEEE ComSoc San Diego chapter has been very active in bringing in world-class researchers and professionals to its members. Since the start of 2016, we, in collaboration with our sister chapters, including the Vehicular Technology Society (VTS) Chapter, Signal Processing Society (SPS) Chapter, Broadcast Technology Society (BTS) Chapter, Aerospace & Electronic Systems Society (AESS) Chapter in San Diego and other ComSoc Chapters in California, have invited seven distinguished lecturers (DLs), prominent researchers, or professionals to give technical talks to our members, in an effort to help our members stay up to date with the latest advances in wireless and related technologies. We have seen excellent turnout for these talks.

On January 29, Dr. Zhensheng Zhang kicked off the technical talks for 2016 with a comprehensive overview of the latest cognitive network technologies. On February 19, Dr. S.R. Subramanya led us to dive into public key cryptography. On March 21, Mr. David Layer and Mr. Paul Shulins gave us a fresh look at broadcast radio technologies. On May 2, Dr. H. Anthony Chan gave a talk on 5G and the future wireless Internet, and a second talk on Software Defined Networking (SDN) and Network Function Virtulization (NFV). That was double service—thank you, Dr. Chan! On August 26, Dr. Touradj Ebrahimi talked about the nitty gritty of Quality of Experience (QoE) and Quality of Life (QoL). On September 21, Dr. George T. Schmidt shared his broad knowledge of navigation sensors and systems in GPS degraded or denied environments.

Success would be impossible without the help of corporate sponsors, in particular, InterDigital, whose CTO, Dr. Byung K. Yi, has been very supportive of such events, and whose employees (Dr. Tianyi Xu, Wei Chen, and Dr. Anantharaman Balasubramanian) are regular volunteers.

Moving forward, we would like to complement our technical events with additional interactive technical training sessions. To that end, one of the events that we are planning is a workshop on antenna design in collaboration with ANSYS Inc.

We are confident that 2016 is going to be another great year for the ComSoc San Diego chapter and our members!

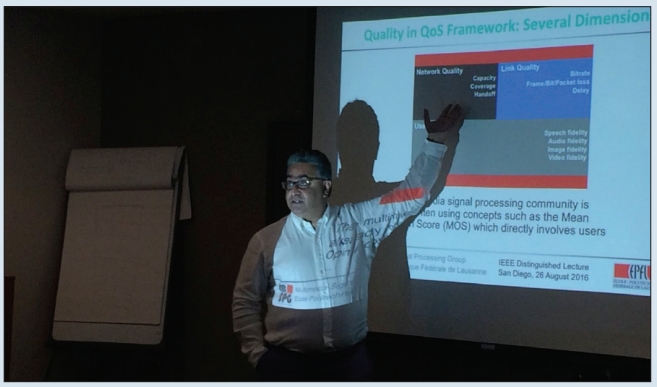


From left to right: Dr. George T. Schmidt (DL for AESS), Dr. Kathleen Kramer (AESS Chair), and Dr. Liangping Ma (ComSoc Chair).

Dr. Zhensheng Zhang was discussing with IEEE members. The left most was Dr. Anantharaman Balasubramanian, a regular volunteer.

Dr. S.R. Subramanya presenting.

Dr. H. Anthony Chan was receiving a certificate after TWO talks from Dr. Byung K. Yi (CTO of InterDigital, VTS Chapter Chair).



Dr. Touradj Ebrahimi was presenting.

Region	Higher grade 2015	Higher grade 2014	Reduction ratio	Student member 2015	Student member 2014	Reduction ratio	Total 2015	Total 2014	Reduction ratio
AP	7,430	11,168	33%	303	2,010	85%	7,733	13,178	41%
US	12,686	17,443	27%	137	1,102	88%	12,825	18,545	31%
Europe	6,270	8,539	27%	144	478	70%	6,414	9,012	29%
Total	26,386	37,150	29%	584	3,590	84%	26,972	40,735	34%

Table 1. Reduction ratio in number of members in the Asia/Pacific, US, Europe Region from Dec. 2014 to Dec. 2015.

MEMBERSHIP SERVICES/*Continued from page 1*

been assigned to a Chapter of the AP Region: Malaysia. Would you please highlight the Malaysia Chapter and its achievements in 2016?

Yamazato: The Malaysia Chapter won the IEEE ComSoc Chapter-of-the-Year Award for the second time in three years, after 2014! This is a great accomplishment, because all the ComSoc chapters worldwide are considered for this Award. The Malaysia Chapter has also been chosen as the winner of the AP Region Chapter Achievement Award in 2014 and 2016.

As the Director of the AP region, I examined questionnaires of 19 Chapters, reporting their activities in the previous year. Among those, the activities of the Malaysia Chapter were deemed excellent. They held 10 membership development seminars, 24 educational or public relations programs and student activity meetings, 12 activities for young and student members, and 13 activities that involved the local communications industry. Furthermore, they organized nine full day/half day seminars, symposia or conferences, hosted four DLT series with a total of nine talks and two DSPs, and conducted 31 technical meetings. What an accomplishment! Outstanding! The Malaysia Chapter makes us really proud.

In 2016, moreover, they planned to conduct 15 membership development programs and four paper awards: Best Paper Award, Best Ph.D. Thesis Award, Best MSc/MEng Thesis Award, and Best Undergraduate Final Year Project Award. We all know the success of ICC 2016 in Kuala Lumpur. The officers and many volunteers of the Malaysia Chapter worked tirelessly, in tight cooperation with the international Organizing Committee, to provide value to the participants by offering exceptional services.

I invited the Malaysia Chapter Chair, Prof. Fazirulhisyam Hashim, to give us a talk at our AP Board general meeting held at GLOBECOM 2016, Washington, DC, USA. I asked Prof. Hashim to enlighten us about the secret of their success and he said "learning from the best practice." He searches the homepages of other Chapters for activities that are worth doing, and he tries to organize them in their chapter as well.

Bregni: The next AP Regional Chapter Chair Congress is planned in 2017, co-located with IEEE GLOBECOM 2017 in Singapore. At that meeting, it will certainly be fruitful to share the experience of the Malaysia Chapter with all Chapters.

Area	Country	Higher grade membership	Student membership	Total	Total per country
Wuhan	China	46	2	48	48
Changwon	Korea	18	1	19	200
Daejeon	Korea	124	6	130	
Kwangju	Korea	19	1	20	
Taegu	Korea	31	0	31	
Islamabad	Pakistan	50	0	50	50
Others		18	0	18	18
Total		382	14	396	396

Table 2. List of areas WITHOUT ComSoc Chapters.

Yamazato: Yes, absolutely. We are planning to hold the 2017 AP Regional Chapter Chair Congress in coordination with the IEEE ComSoc Sister and Related Societies Summit, which is also planned to be co-located with IEEE GLOBECOM 2017 in Singapore. It will be a two-day event just before the conference. We will announce a "call for topics" to all AP chapters and coordinate interesting sessions for Chapter support and membership development. The presentation of the Malaysia Chapter will certainly be one of the highlights of the Congress.

Bregni: To conclude, is it possible to join the AP Board and collaborate on some of its activities?

Yamazato: The AP Board is open to any IEEE ComSoc member from the AP region. Please attend our APB meeting, which is usually held on the first day of the Technical Sessions of GLOBECOM/ICC. I hope that as many of our members as possible will get involved and serve on one of our Committees, as the best networking often occurs when you are working toward a common goal.

ROMANIA CHAPTER/*Continued from page 2*

interest among students, academic staff, and representatives of the local IT&C industry.

Activities Aimed at Local Industry

Another area of the chapter's activities was to build liaisons with industry. To this end, the chapter members from the technical universities of Bucharest, Cluj-Napoca, Timisoara, and Iasi, organized seminars, presentations, and intensive courses for students and staff that were offered by telecomm companies. Such a seminar on "Satellite Communications" offered by Dr. Jens Krause from SES Communications, Luxembourg, took place at PUB. A workshop on "Wireless Communications & Real Time Concept" with participation of Rohde & Schwartz Romania, was organized at the Technical University of Cluj-Napoca.

The telecom companies are also involved in the students' development by means of training courses and support for the elaboration of the graduation projects or M.Sc. dissertations. As an example, we mention the Orange Educational Program at PUB, where students in communications are offered training stages and internships to develop their graduation projects.

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 FOR MEMBER AND GLOBAL ACTIVITIES
CARLOS ANDRES LOZANO GARZON, DIRECTOR OF LA REGION
SCOTT ATKINSON, DIRECTOR OF NA REGION
ANDRZEJ JAJSZCZYK, DIRECTOR OF EMEA REGION
TAKAYA YAMAZATO, DIRECTOR OF AP REGION
CURTIS SILLER, DIRECTOR OF SISTER AND RELATED SOCIETIES

IEEE ComSoc IEEE Communications Society

www.comsoc.org/gcn
ISSN 2374-1082

UPDATED ON THE COMMUNICATIONS SOCIETY'S WEB SITE

www.comsoc.org/conferences

2017

FEBRUARY

ICACT 2017 — Int'l. Conference on Advanced Communication Technology, 19–22 Feb.

Pyeongchang, Korea
http://www.icact.org/

WONS 2017 — Wireless On-Demand Network Systems and Services Conference, 21–24 Feb.

Jackson Hole, WY
http://2017.wons-conference.org/

MARCH

NCC 2017 — Nat'l. Conference on Communications, 2–4 Mar.

Madras, India
http://ncc2017.org/

IEEE DYSPAN 2017 — IEEE Dynamic Spread Spectrum Access Symposium, 6–9 Mar.

Baltimore, MD
http://dyspan2017.ieee-dyspan.org/

ICIN 2017 — Conference on Innovations in Clouds, Internet and Networks, 7–9 Mar.

Paris, France
http://www.icin-conference.org/

NETSYS 2017 — Int'l. Conference on Networked Systems, 13–17 Mar.

Göttingen, Germany
http://netsys17.uni-goettingen.de/

IEEE WCNC 2017 — IEEE Wireless Communications and Networking Conference, 19–22 Mar.

San Francisco, CA
http://wcnc2017.ieee-wcnc.org/

OFC 2017 — Optical Fiber Conference, 19–23 Mar.

Los Angeles, CA
http://www.ofcconference.org/

IEEE CogSIMA 2017 — IEEE Conference on Cognitive and Computational Aspects of Situation Management, 27–31 Mar.

Savannah, GA
http://cogsima2017.ieee-cogsima.org/

WD 2017 — Wireless Days 2017, 29–31 Mar.

Porto, Portugal
http://www.wireless-days.com/

APRIL

IEEE ISPLC 2017 — IEEE Int'l. Symposium on Power Line Communications and its Applications, 3–5 Apr.

Madrid, Spain
http://isplc2017.ieee-isplc.org/

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

PEOPLE-CENTRIC INTERNET OF THINGS


Jorge Sá Silva


Pei Zhang


Trevor Pering


Fernando Boavida


Takahiro Hara


Nicolas C. Liebau

People can now be viewed as an integral part of the Internet of Things (IoT) ecosystem. Although considerable work has been done in the recent past regarding IoT, many challenges remain. In fact, most technologies and solutions for accessing real-world information are either closed, platform-specific, or application-specific. This Feature Topic intends to present, explore, and discuss societal aspects, scenarios, opportunities, risks, and uses of innovative people-centric IoT-based applications, which will undoubtedly be a key piece in the future and emerging people-centric society.

The emphasis of this Feature Topic was put on discussion about state-of-the-art research and development activities contributing to all aspects of people-centric IoT, which include, but are not limited to, people-IoT interactions, social network applications to mobile computing, context-aware applications and services, human in the loop, big data analysis in people-centric IoT, cloud-based people-centric IoT applications and environments, security and privacy, and prototypes, field experiments, and testbeds.

The Call for Papers resulted in 54 submitted high-quality papers, of which we have selected seven papers for publication, a 13 percent acceptance ratio. The topics addressed by these papers include aspects such as vendor-independent cross-platform architectures for mobile apps, cognitive systems for smart homes and energy management, smart city cloud platforms, user behavior classification, assessment and prediction, and human-device cooperation for security and privacy.

Currently, an increasing number of applications for personal mobile devices provide the ability for people to monitor, control, and interact with a variety of sensors and actuators. Nevertheless, the vast majority of these apps resort to vendor-specific application programming interfaces (APIs), thus preventing portability. The article "COIN: Opening the Internet of Things to People's Mobile Devices," by Maria Laura Stefanizzi, Luca Mottola, Luca Mainetti, and Luigi Patrono, proposes the COIN software architecture, an open, vendor-independent, runtime substrate that allows developers to flexibly run arbitrary IoT device tasks, implemented as loosely coupled components.

The article "Butler, Not Servants: A Human-Centric Smart Home Energy Management System," by Siyun Chen, Ting Liu, Feng Gao, Jianting Ji, Zhanbo Xu, Buyue Qian, Hongyu Wu, and Xiaohong Guan, discusses human-centric smart home energy management systems. The authors propose systems that use sensing to discover user patterns of power usage, cognitively understand the behavior of human beings, and optimally schedule home energy systems.

In the article "Smart Home: Cognitive Interactive People-Centric Internet of Things," Shuo Feng, Simon Haykin, and Peyman Setoodeh explore the cognitive IoT paradigm in the context of smart homes. The authors argue that by using the notion of cognitive dynamic systems, which build on perception-action cycles, memory, attention, intelligence, and language, it is possible to engineer IoT applications that cover a wide spectrum of tasks with minimum human intervention.

The article "The Experience of Using the IES Cities Citizen-Centric IoT Platform," by Stefanos Vatsikas, Georgios Kalogridis, Tim Lewis, and Mahesh Sooriyabandara, describes a city-scale IoT platform, covering the system's design principles and underlying capabilities. The IES Cities platform has been deployed in four cities and used for a number of different apps, yielding key insights into citizen-centric IoT application development.

In the article "Exploiting Density to Track Human Behavior in Crowded Environments," by Claudio Martella, Marco Cattani, and Maarten van Steen, the authors propose and discuss a system of mobile sensors augmented with low-cost, fixed proximity sensors to track and understand the behavior of crowds of participants. Through a deployment at a multi-story museum with 3000 daily visitors during the deployment period, the system was validated, and behavior trends of users were discovered.

The article "Gesture Detection of Passive RFID Tags to Enable People-Centric IoT Applications," by Raúl Parada Medina and Joan Melià-Seguí, presents an approach for detecting and classifying human gestures using accelerometer-enabled passive RFID tags and unsupervised machine learning. The proposed approach is expected to contribute to better authentication and personalization in IoT applications and services.

Last but not least, the article "Human Neuro-Activity for Securing Body Area Networks: Application to People-Centric Internet of Things," by J. F. Valenzuela-Valdés, M.A. López-Gordo, P. Padilla, J. L Padilla, and J. Minguillón, presents a really creative idea for securing the communication within wireless body area networks (WBANs). In IoT, applications for frequent automatic renewal of encryption keys are required. However, WBANs are typically built out of low-power-consumption and low-performance hardware. Therefore, such devices cannot generate the required secure random numbers for encryption. The novel idea presented in the article is to generate these numbers from the brain waves of the user (EEG) via a brain-computer interface.

As a final remark, the Guest Editors would like to thank all the members of the Technical Committee for their effort in putting together this Feature Topic. A special thank is due to Osman Gebizlioglu, Editor-in-Chief of *IEEE Communications Magazine*, for his support of this Feature Topic.

BIOGRAPHIES

JORGE SÁ SILVA [SM] (sasilva@dei.uc.pt) is a tenured assistant professor and a senior researcher at the Department of Informatics Engineering, Faculty of Sciences and Technology, University of Coimbra, Portugal. His main research interests are the Internet of Things, network protocols, machine-to-machine communications, and wireless sensor networks. He is a licensed Professional Engineer.

PEI ZHANG is an associate research professor in the Electrical and Computer Engineering Department at Carnegie Mellon University. Beyond research publications, his work has been featured on popular media including CNN, the Science Channel, the Discovery Channel, CBS News, CNET, Popular Science, BBC Focus, and more. He is also a cofounder of the startup Vibradotech. In addition, he has won several awards including the NSF CAREER award, and the, Edith and Martin B. Stein Solar Energy Innovation Award.

NICOLAS C. LIEBAU is chief product owner at SAP SE and responsible for the IoT big data functionality within the SAP IoT application enablement cloud service. He joined SAP six years ago, focusing on strategy and product management for Internet of Things applications. He received his Ph.D. from the Multimedia Communications Lab of the Technical University of Darmstadt in the area of accounting in peer-to-peer systems. There he also led the peer-to-peer networking group before joining SAP SE.

TAKAHIRO HARA [SM] received the B.E., M.E., and Dr.E. degrees in Information Systems Engineering from Osaka University, Osaka, Japan, in 1995, 1997, and 2000, respectively. Currently, he is a Professor of the Department of Multimedia Engineering, Osaka University. His research interests include distributed databases, mobile computing, and social computing. He is a distinguished scientist of ACM and a member of three other learned societies.

TREVOR PERING [M] (peringknife@google.com) is a senior systems software engineer at Google. His research interests include building-scale Internet of Things systems, mobile devices, and interactive experience design. He received a Ph.D. in electrical engineering and computer science from the University of California, Berkeley. He is a member of ACM.

FERNANDO BOAVIDA [SM] (boavida@dei.uc.pt) is a full professor at the Department of Informatics Engineering (DEI), Faculty of Sciences and Technology, University of Coimbra. His main research interests are people-centric Internet of Things, wireless sensor networks, mobility, and quality of service. He is a licensed Professional Engineer.

COIN: Opening the Internet of Things to People's Mobile Devices

Maria Laura Stefanizzi, Luca Mottola, Luca Mainetti, and Luigi Patrono

People's interaction with IoT devices is often mediated through personal mobile devices. Current approaches often make applications operate in separate silos, as the functionality of IoT devices is fixed by vendors and typically accessed only through low-level proprietary APIs. COIN is a system architecture that breaks this separation by allowing developers to flexibly run a slice of a mobile app's logic onto IoT devices.

ABSTRACT

People's interaction with IoT devices such as proximity beacons, body-worn sensors, and controllable light bulbs is often mediated through personal mobile devices. Current approaches usually make applications operate in separate silos, as the functionality of IoT devices is fixed by vendors and typically accessed only through low-level proprietary APIs. This limits the flexibility in designing applications and requires intense wireless interactions, which may impact energy consumption. COIN is a system architecture that breaks this separation by allowing developers to flexibly run a slice of a mobile app's logic onto IoT devices. Mobile apps can dynamically deploy arbitrary tasks implemented as loosely coupled components. The underlying runtime support takes care of the coordination across tasks and of their real-time scheduling. Our prototype indicates that COIN both enables increased flexibility and improves energy efficiency at the IoT device, compared to traditional architectures.

INTRODUCTION

The rise of smartphones and tablets makes them the established means for people to interact with Internet of Things (IoT) devices such as proximity beacons, body-worn sensors, and controllable light bulbs. A multitude of applications employ personal mobile devices to allow people to control, interact, and query sensors and actuators in the environment or on the human body. Technologies such as Bluetooth Low Energy (BLE), now commonly found on both personal mobile devices and IoT ones, are key facilitators for this trend.

Motivation. Architectures applied in these applications, however, treat the IoT devices as immutable black boxes and operate in separate silos. Mobile apps are sometimes rerouted to intermediate gateways. Whenever direct access to the IoT device is allowed, the latter typically offers application-agnostic application programming interfaces (APIs) that mainly enable extracting raw data and/or controlling basic actuator functionality. Changes to the onboard software are limited to firmware updates released by manufacturers to patch bugs or security flaws.

Such a state of affairs entails that:

- Mobile apps are developed based on *vendor-specific APIs*, preventing portability.
- Even the simplest functionality requires *intense wireless interactions*, affecting energy consumption.
- App functionality is limited to the *time of wireless connection*, that is, disconnected operations are fundamentally hampered.

Unlike current practice, many foresee the interaction between humans and IoT devices to happen over an open, vendor-independent programmable substrate that enables a multitude of apps to coexist, similar to smartphones and tablets. For example, an IoT device with body temperature and heart rate sensors may serve both fitness apps to track an individual's well being and smart health apps that communicate vital parameters to doctors. Individual building blocks necessary to realize these functionalities are often already available, but a system architecture that blends them together in a working realization is arguably still lacking.

COIN. We bridge this gap by designing a system architecture that allows IoT devices to: i) run an arbitrary slice of a mobile app's logic in an on-demand fashion, and ii) host application data according to programmer-provided criteria, independent of the connection to user devices.

The problem is unique in many respects. Unlike traditional sensor networking, for example, applications are supplied by third parties. Their characteristics, such as processing and memory requirements, are difficult to anticipate. Multiple applications may need to operate concurrently, that is, not simply run side by side, but be able to exchange data with other a priori unknown apps. Processing is also expected to be largely event-driven, for example, being dictated by connections of mobile devices, rather than occurring periodically.

COIN rests on two pillars: a custom programming model and dedicated runtime support. The former is based on a notion of a lightweight task as a programmer-defined relocatable slice of mobile app logic. In a fitness scenario, for example, programmers may define a task that computes burned calories based on the sensors available on fitness trackers. The mobile phone may opportunistically deploy such a task on the fitness tracker to limit data exchanges to a single quantity rather than raw data. Multiple tasks on an IoT device can interact in a loosely coupled manner, based on an actor-like [1] model.

COIN's runtime support accommodates existing building blocks to efficiently implement the

Digital Object Identifier:
10.1109/MCOM.2017.1600656CM

Maria Laura Stefanizzi, Luca Mainetti, and Luigi Patrono are with the University of Salento; Luca Mottola is with Politecnico di Milano and SICS Swedish ICT.

required semantics. A message broker mediates interactions across tasks, while their executions are scheduled using an Earliest Deadline First (EDF) policy. Dynamic deployment of tasks is supported using a virtual machine (VM) developers plugin. Although COIN is independent of the underlying hardware platform, our prototype targets Cortex M microcontroller units (MCUs) and BLE radios, representative of target applications where energy budgets are as small as a COIN-cell battery.

We describe the design and implementation of a pervasive game using COIN, an application otherwise infeasible with comparable features using vendor-specific architectures. We also report on the performance of COIN in energy consumption and execution times. The results indicate that the energy savings enabled by reducing wireless interactions through device-local processing overcome the cost of code interpretation, validating our design choices. The price to pay is larger execution times, but the values we obtain are not expected to impact application responsiveness.

COIN's main contributions are therefore:

- To increase the flexibility in the design of IoT applications by giving developers the ability to relocate slices of a mobile app's logic onto IoT devices
- To improve the energy efficiency at the IoT devices in applications where real-time requirements are soft or absent

The remainder of the article describes how we concretely achieve these contributions.

State of the Art

We report on application scenarios, requirements to overcome current limitations, and existing functionality.

Applications

Employing personal mobile devices as people's interface to the IoT yields applications with distinct characteristics, exemplified next.

Body Sensor Apps. Body-worn devices with physiological or inertial sensors are often used to monitor physical parameters (e.g., temperature and heart rate). The raw sensor values are streamed to a smartphone, which acts as the sole processing unit. Similar architectures tend to be inefficient. Frequent wireless interactions between smartphones and sensors are costly in energy consumption. Algorithms exist to relocate part of the processing closer to the sensors. Examples are found in activity recognition and electrocardiogram (ECG) analysis [2]. The amount of data to transmit is thus reduced, so energy consumption improves. Flexible system support on the IoT device is required to employ these algorithms in a vendor-independent and reconfigurable manner.

Immersive Computing. Interactions between people's devices and IoT ones need not be continuous, but simply occur opportunistically whenever the two are in range. Representative examples are pervasive games [3], where embedded devices are hidden in the environment to bridge the game's virtual world with the physical reality. These devices are often used as environment-immersed data stores to handle information relevant to the game plot. Access to digital information is dictated by physical location, enhancing the experience. Pervasive games are currently installed using dedicated hardware, deployed solely for running the game and later removed. Reusing already installed hardware is generally not possible, as it lacks the necessary programming facilities.

Monitoring and Tracking: Continuous interactions between mobile and IoT devices may also be broken because the latter are mobile. In supply chain applications [4], sensors are attached to packages to log information such as temperature and vibrations during transportation. When a package enters a warehouse, locally stored information is uploaded to the mobile phones of the warehouse personnel for inspection. In this scenario as well, the ability to store and process sensor data on the IoT device independent of the connection to a person's device is fundamental. Right now, such a degree of decoupling can only be realized with one-off application-specific implementations.

Requirements

System architectures at the IoT device that overcome these limitations should fulfil several requirements:

- **Device-local processing:** Running application-specific functionality on the IoT device decouples its operation from the mobile device and allows one to reduce wireless interactions, saving energy.
- **Data persistency:** The IoT device must be able to retain application data according to programmer-provided criteria independent of the connection to a mobile device, enabling disconnected operations.
- **Dynamic deployment:** The logic at the IoT device may not be known beforehand, but be provided on the fly by the mobile device; the runtime support at the IoT device must accommodate this need.
- **Real-time scheduling:** Independently developed applications may need to coexist on the same IoT device; the system must ensure that their real-time requirements are fulfilled whenever possible.
- **High-level programming and portability:** Application functionality for the IoT device must be developed using high-level languages and not require increased efforts to adapt to different hardware.

In contrast, current applications are normally developed with a "sense-and-send" design. IoT devices are employed as shipped by manufacturers, that is, with pre-loaded firmware that only enables low-level interactions. As a result, the entire application logic executes at the mobile device and is encoded in a vendor-specific manner. Besides not enabling any disconnected operation, these designs decrease portability, consequently increasing development efforts, and are detrimental to energy consumption.

Building Blocks

Approaches exist that address specific issues in the scenarios we target; for example, in the field of operating systems (OSs) for sensor nodes, VM technology for resource-constrained devices, and interoperability frameworks.

IoT devices are employed as shipped by manufacturers, that is, with pre-loaded firmwares that only enable low-level interactions. As a result, the entire application logic executes at the mobile device and is encoded in a vendor-specific manner.

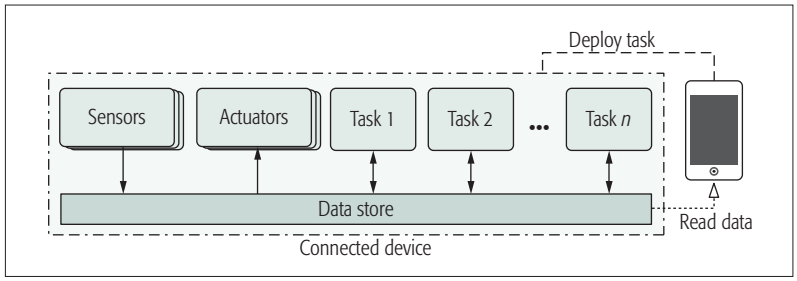


Figure 1. Tasks interacting in a loosely coupled manner.

```
1 def foo(x):
2 # do something...
3
4 def boo(x):
5 # do something else...
6
7 startTask()
8 y = foo(inputData)+boo(inputData)
9
10 output(y)
```

Figure 2. Example task code.

Sensor network OSs such as Contiki and TinyOS offer dynamic linking capabilities, which allow different applications to be added or replaced at runtime. However, the OS per se does not provide a programming model that allows dynamically deployed applications to discover each other and exchange data. Differently, components must be developed based on how they bind to already running components, which requires intimate knowledge of the latter. Most importantly, OSs in this area offer low-level programming interfaces based on languages such as C, which contrasts with the modern development tools available for mobile apps.

VMs for resource-constrained devices retain the ability of dynamic code deployment while offering hardware independence and higher-level programming languages, such as Java or Python. These aspects motivate us to base COIN on VM technology. The cost of code interpretation is the price we pay to facilitate the development process. Because of the rise of energy-efficient 32-bit MCUs, such as ARM's Cortex M series, we demonstrate that this represents an effective design point.

The design of COIN is *orthogonal* to the specific VM one plugs into the architecture. Matè [5] offers a custom language that can be tailored to specific application domains. TakaTuka [6] and Darjeeling [7] provide Java VMs for 16-bit MCUs, whereas Squawk [8] targets higher-end devices. DAViM [9] focuses on isolating multiple applications from each other. In general, the design of embedded VMs targets efficient code interpretation, without providing dedicated abstractions for coordinating concurrent third-party applications.

The heterogeneity of IoT devices motivates efforts in interoperability frameworks and reference architectures. Examples are AllJoyn and IoTivity. These ease development by defining vendor-agnostic APIs for applications to inter-operate, based on IoT devices with much greater resources compared to ours and without providing the ability to relocate parts of the application logic. Efforts such as IoT-A, SENSEI, and OpenIoT provide open architectures to facilitate application development via semantically interoperable interfaces. These are complementary to COIN, which may help realize more flexible or efficient implementations exported through the same interfaces as in these architectures.

COIN Architecture

COIN revolves around a dedicated programming model and runtime support to implement the required semantics. We only provide a few highlights here, and refer the reader to a companion technical report [10] for details.

Programming Model

We design COIN's programming model based on the characteristics of mobile apps where relocating a slice of the logic onto the IoT device may achieve benefits such as better designs or improved performance, as we discussed earlier. On the other hand, we do not target applications where IoT devices need only be equipped with application-agnostic functionality; for example, if they are used as "beacons" in the environment for indoor localization.

The key feature in COIN's programming model is the shift of interactions to data rather devices. At the core of this is a notion of a *task* as a relocatable slice of app logic.

Decoupling: To facilitate interactions among third-party functionality, tasks are fully decoupled. They cannot share global data and only interact asynchronously in an actor-like fashion [1]. Data exchanges occur through a single abstract data store, as in Fig. 1. This avoids the need to dynamically reconfigure the bindings among tasks as these come and go, and enables data-driven discovery of available functionality.

Tasks are completely defined by the data types they consume or produce. Information on these are included in a manifest deployed with the task. The data types are specified as named data structures. The manifest of the task deployed by the fitness app, for example, may indicate input types that include acceleration and blood pressure as double precision numerical values, and outputs such as burned calories as integer values. Existing sensor data models [11] can be applied to uniform naming and format.

Execution: Figure 2 shows a code snippet for a simple task using the Python syntax, but the programming model is independent of the specific language.

Task execution is reactive, and triggered only by the availability of any of the input data types. For example, we discourage the use of long-running threads whose fair scheduling may become difficult. Input data is made available using a dedicated API, which also provides operations to output the results and to indicate the start of processing, required to simplify Python's modularity model. An example of the former is the `output()` function in Fig. 2. This API is the only interface developers employ to write tasks; other than this, developers can encode arbitrary application logic.

The input data of a task may come from the sensors onboard the device, or be the result of a different task. In the former case, sensors are

automatically probed according to the required input rates specified in the manifest. For example, a health-monitoring app may employ the burned calorie information of the fitness app to augment the long-term time analysis. Such data-driven programming facilitates developing vendor-independent interaction paradigms.

Execution of tasks is also decoupled from the connection to a person's mobile device. For example, tasks may also reside on the IoT device whenever the mobile device that originally deployed them moves away. Unlike the traditional actor model [1] where data is lost if no actor immediately consumes it, COIN applies persistence to data as well. Data resides on the IoT device according to programmer-defined criteria, such as a given time interval, catering for the needs of immersive mobile computing applications [3].

Runtime Support

COIN's runtime support includes three components:

- A data broker to mediate task interactions
- A scheduler to regulate task execution
- A VM layer

Broker and Scheduler: The *broker* matches data producers and consumers based on data types. Whenever a match is identified, the consumer task is handed over to the scheduler. In our prototype, the broker maps items in the data store to BLE *characteristics* to give mobile devices standard-compliant access to data.

If multiple tasks consume the same data type, the match happens simultaneously. The *scheduler* thus implements an EDF policy. Information on the absolute deadline of a task and its expected execution time are part of the manifest. Static analysis tools and emulators can be used to estimate the latter. The scheduler also ensures that every task runs to completion; concurrent events, such as connection requests from other mobile devices, are postponed until the task finishes.

We choose EDF because of its real-time optimality: if a schedule able to meet all task deadlines exists, EDF finds one. The processing overhead of EDF is no issue in our setting, also because we do not expect a large number of tasks to be triggered simultaneously. As tasks should be short-lived, running them to completion does not pose problems, as in architectures with similar design rationales.

Virtual Machine: We port PyMite, a reduced Python interpreter, as the VM layer. COIN is independent of the language to write tasks, but we choose Python for several reasons. Compared to languages such as Java, its implementation on embedded devices is less limited; for example, PyMite retains the support to multiple programming paradigms, including object-oriented and functional. Moreover, Python directly compiles to bytecode, which reduces network traffic when deploying tasks. The most complex application we tested so far yields slightly more than 1 kB of bytecode.

We map COIN tasks to PyMite threads, which requires adapting the latter along multiple dimensions. We replace the built-in round-robin scheduler with EDF. In the original PyMite, the state of a thread is lost when the execution exits; we thus extend the VM to maintain the thread state across executions of the same task. Finally, we choose to save the precious RAM segments and store the Python bytecode on flash memory, which demands the VM to execute off the latter.

Prototype: Our prototype targets 32-bit Cortex M MCUs and BLE radios. Although COIN's programming model is independent of the underlying network technology, BLE is arguably a natural choice whenever integration with people's mobile devices is necessary and interactions may be triggered by proximity, as determined by radio connectivity [12]. We offer a primary example next. The prototype is mainly intended to provide a basis to assess the design rationale. It has a few limitations, which would require further implementation work but would *not* alter COIN's conceptual design.

PyMite does not offer resource arbitration per se, required to ensure that tasks by different parties safely share resources. This feature may be seen as desirable in any IoT VM. There is literature on the subject [13] that can be applied to address this issue. Moreover, interactions across personal mobile devices and IoT ones are currently encoded by directly accessing the APIs of the BLE stack (i.e., by reading and writing BLE characteristics). A dedicated API would, however, be needed on both sides to express such interactions at a higher level of abstraction.

Modern networking stacks, such as BLE, are already equipped with built-in security features. As a result, the main security threat for COIN is likely going to be the authenticity of the Python bytecode. Techniques such as code signing [14] exist to address this issue, and are shown to be applicable to devices even more constrained than the ones we target. For example, the digital signatures employed in the Deluge protocol [14] incur very limited processing overhead. Porting these solutions to COIN should therefore be feasible with limited effort. Most importantly, code signing would be a one-time cost at the moment of deploying a task. The performance figures we discuss later — obtained after a task is successfully deployed — would therefore retain their validity.

Our prototype targets 32-bit Cortex M MCUs and BLE radios. Although COIN's programming model is independent of the underlying network technology, BLE is arguably a natural choice whenever integration with people's mobile devices is necessary and interactions may be triggered by proximity, as determined by radio connectivity.

Using COIN

We report on the use of COIN in the design and implementation of a pervasive game [3] whose logic spans smartphone and IoT devices in the environment.

The Game. Only the bravest can become pirates! To prove their qualities, the aspiring pirates must travel to a mysterious island and overcome several challenges.

Players are divided into teams. The team that obtains the highest score becomes the pirate crew. To accumulate points, a team must collect items scattered across the island. Items are of different types: compasses, rare seeds, parrot eggs, bottles of rum, and sabers. The value of the first three kinds of items is 10 points. The latter two give 200 points, and can be obtained by paying a merchant with some of the collected items.

A team may decide to transform rare seeds and parrot eggs into plants and parrots, respectively. By doing so, the transformed item yields a score 10 times higher than does the original one. However, eggs and seeds grow only if they live at the right temperature for a sufficient time. To this end, teams must deposit the collected eggs

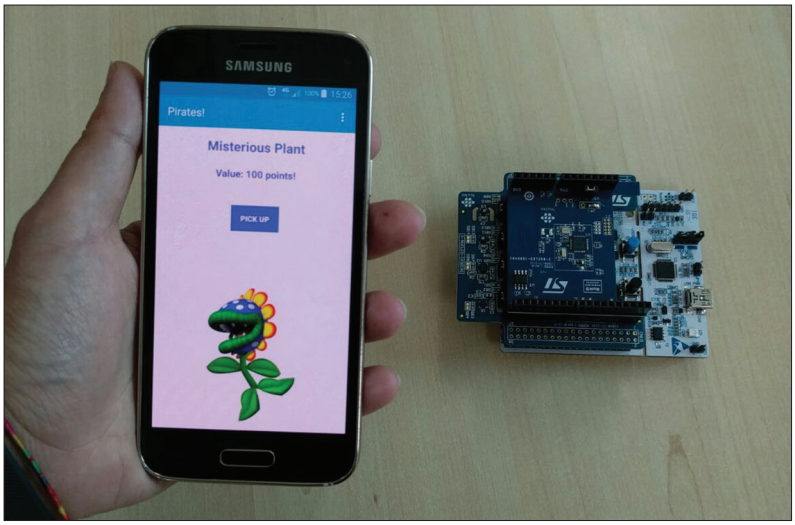


Figure 3. User interface on a person's smartphone and environment-immersed sensor device in our pervasive game.

```
1 from coin import ★
2
3 # ... initialization ...
4 startTask()
5
6 #get temperature value
7 temp = getIntInput(0)
8
9 # ...seed germination ...
10 if (count == TIME_TO_GERMINATE):
11
12 # process and output germination level
13 output(germination)
```

Figure 4. Task code to simulate seed germination over time and depending on physical temperature.

or seeds in suitable places, taking care not to be seen by opponents, who could kidnap the items after their transformation.

From virtual to physical. COIN allows us to develop the game in a way that no other existing platform would enable. Real-time user interactions happen through the touch interface of a standard smartphone. We deploy COIN on ST Nucleo-F091RC prototyping boards equipped with a BlueNRG BLE radio and an ST X-Nucleo-IKS01A1 sensor shield, as shown in Fig. 3, and install them in our department building.

We map virtual items in the game to dedicated data structures dynamically stored through COIN aboard the Nucleo boards. As a result, accessibility of items corresponds to proximity to the Nucleo board storing the corresponding data structure — as dictated by BLE communication range — and mobility on the island maps to physical mobility in our building. Implementing the collection or release of items in the game is thus as simple as reading or writing from/to COIN's data store from a player's smartphone. With this design, straightforwardly enabled by COIN, virtual and physical dimensions spontaneously blend together.

However, there is more that COIN enables. Temperature conditions that determine how eggs and seeds grow are now simple to link to temperature in the physical environment. A player's smartphone can dynamically deploy, together with the item itself, a simple COIN task that periodically probes the temperature sensor on the Nucleo board and accordingly modifies the values of the data structures representing eggs and seeds. This may happen *independent* of the connection to a player's smartphone, giving players the illusion that the game unfolds across the virtual and physical worlds. Figure 4 shows an excerpt of the task implementation that simulates the germination of seeds, which becomes as small as 320 bytes at the time of deploying the task from the smartphone.

Finally, unlike existing pervasive games [3], COIN naturally allows other apps to reuse the deployed sensing infrastructure. Say, for example, a new app is developed to control air conditioning in our offices based on temperature and individual preferences. A user's smartphone may deploy a new COIN task that computes short- and long-term trends of relevant quantities, useful as inputs for implementing the feedback loop. Provided the sampling periods are compatible, the new task may just reuse part of the sensed data that the game already requires.

Performance

Our prototype requires about 154 kB of program memory and about 10 kB of data memory. About 90 percent of this is due to PyMite and BLE drivers. However, PyMite is only meant to provide a working Python interpreter, and its memory demands could be significantly reduced by tailoring it to COIN. Even at the prototype stage, COIN fits most existing 32-bit embedded platforms. For example, the Cortex M0 core often used in SoC designs with a BLE radio provides 256 kB (16 kB) of program (data) memory.

Replacing wireless transmissions with device-local processing typically improves energy consumption. With energy-efficient protocols such as BLE, 32-bit MCUs, and the overhead of code interpretation, such a claim needs to be newly demonstrated. Therefore, we measure COIN's energy performance in a set of representative applications against a traditional "sense-and-send" design implemented in C to validate our design choices. Similarly, we compare the execution times of COIN against functionally equivalent implementations in C. In both cases, the C implementations are deployed as an immutable binary on the target platform.

Benchmarks and Metrics

We consider three applications based on the scenarios and requirements previously discussed. Each application corresponds to a COIN task.

We first consider run length encoding (RLE) compression. RLE is often advocated for applications where sensors report stable values. Next, we consider an activity detection (AD) algorithm to distinguish between *standing* or *walking* activities. The AD occurs on 5 Hz accelerometer data by computing average and standard deviation of the signal amplitude. Finally, we consider an algorithm to extract ECG information [2]. The signal is passed through multiple tap filters and then compared against a threshold to detect peaks indicating physiological issues.

We consider four prototyping platforms: a Freescale FRDM-KL46Z, a Nordic nRF51-DK, a NXP LPCXpresso1549, and a NXP LPCXpresso4337. These offer the full range of Cortex M MCUs, as well as varying amounts of program (256 to 1024 kB) and data (16 to 136 kB) memory. We attach Bluetooth extension boards where necessary. In the absence of an earth-rate sensor, we use the accelerometer; this does not impact the execution of the ECG algorithm. We disable all unnecessary peripherals.

We measure *energy consumption* and *execution times* through a Tektronix 1072B oscilloscope. The values we present are averages over at least five repetitions. The standard deviation across different runs, not shown in the charts, is always within 5 percent of the average. Detailed information on the experimental setup are found in the companion report [10].

Results

Energy Consumption: We feed data to RLE so as to achieve a 50 percent compression ratio, in fact pessimistic for RLE compression of sensor data [15]; AD reports data to a smartphone every 30 s, whereas ECG samples the sensors at 30 Hz. Results by varying these parameters are, nonetheless, available [10].

Figure 5 reports the results. The trade-off between saving transmissions by deploying COIN tasks and the additional MCU overhead due to code interpretation is in favor of COIN. In our experiments, the improvement in energy consumption is at least 25 percent, with a best case of 35 percent. This is despite the efficient energy performance of BLE radios.

The LPCXpresso4337 board shows the best performance in Fig. 5 when running the AD task. The FPU of the Cortex M4 core speeds up the execution of the floating point operations in AD. In contrast, the Cortex M0+ core on the nRF51-DK provides the best performance with sequential byte-level operations, as in RLE.

Note that the energy cost for deploying the task is a *one-time* cost. The AD task, for example, requires about 50 packets. These may be transmitted using BLE's streaming mode, which reduces efforts for packet trains. Thus, the energy overhead quickly amortizes as a task continues to run.

Execution Times: Larger execution times represent the cost for increased flexibility and better energy efficiency in COIN. For the AD and ECG tasks, we separate the case of regular local processing at every iteration from the case of data transmission that requires extra computations to prepare data for transmission.

Table 1 shows the results for the PCXpresso4337 board. The ratios are similar for the other platforms. The values are still on the same order of magnitude of packet transmissions, and should not be detrimental to the app responsiveness, including user interactions. The slowdown is vastly dominated by code interpretation, but the values in Table 1 are in line with existing literature [6, 7]. Note that PyMite is not expressly designed for the platforms we target; nor do we explicitly optimize it besides the adaptations in the previous section.

As each task takes longer to run, the slowdown may impact the overall schedulability, limiting the number of tasks concurrently executing. However, we can run up to six instances of the AD task on a Cortex M0 MCU, for example, before scheduling becomes unfeasible.

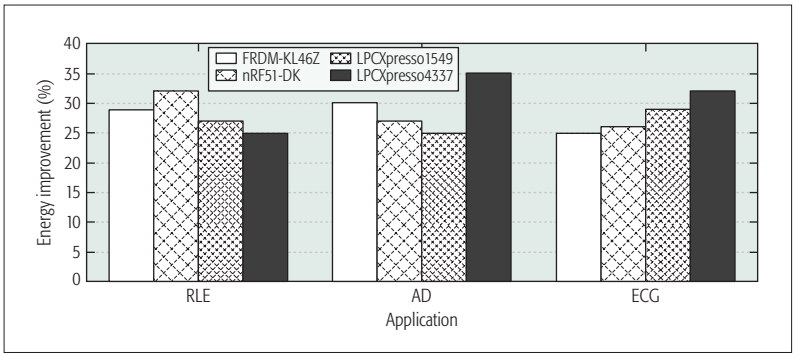


Figure 5. Energy performance of COIN compared to a sense-and-send design.

Task	Plain C (μs)	COIN (ms)	Ratio
RLE	108,37	8,63	79,93
AD-local	163,97	4,98	30,55
AD-transmission	68,31	5,22	76,76
ECG-local	127,82	3,78	29,76
ECG-transmission	89,21	4,55	51,12

Table 1. Execution times of tasks with PCXPRESSO4337.

Conclusion

We present COIN, a software architecture that provides the glue necessary to create an open vendor-independent programming substrate of IoT devices accessible from people's mobile devices. COIN offers a high-level programming model based on a lightweight notion of a task as a relocatable unit of mobile app logic. Its runtime support takes care of the tasks' dynamic deployment, real-time scheduling, and cross-task coordination. We demonstrate how COIN overcomes the limitations of traditional designs; for example, by enabling a degree of flexibility in the design of immersive computing applications that no other existing platform may similarly provide. We also show that the device-local processing COIN enables can improve a device's energy consumption up to a 35 percent factor in our tests, at the expense of larger execution times due to code interpretation.

Acknowledgments

The authors would like to thank Riccardo Paccagnella for the implementation work on the pervasive game described in this article, and Naveed Bhatti for supporting the authors during the experimental evaluation.

References

[1] G. Agha, Actors: *A Model of Concurrent Computation in Distributed Systems*, Ph.D. thesis, MIT Artificial Intelligence Lab, 1985.
[2] D. Albu, J. Lukkien, and R. Verhoeven, "On-Node Processing of ECG Signals," *IEEE CCNC*, Jan 2010.

[3] C. Magerkurth *et al.*, "Pervasive Games: Bringing Computer Entertainment Back to the Real World," *Comp. Entertainment*, vol. 3, no. 3, 2005.
[4] S. Hoppough, "Shelf Life," *Forbes Magazine*, 2006.
[5] P. Levis and D. Culler., "Matè: A Tiny Virtual Machine for Sensor Networks," *SIGOPS Oper. Sys. Rev.*, vol. 36, no. 5, 2002.
[6] F. Aslam *et al.*, "Optimized Java Binary and Virtual Machine for Tiny Motes," *IEEE DCOSS*, 2010.
[7] N. Brouwers *et al.*, "Darjeeling, A Feature-Rich VM for the Resource Poor," *ACM SENSYS*, 2009.
[8] D. Simon *et al.*, "Java™ on the Bare Metal of Wireless Sensor Devices: The Squawk Java Virtual Machine," *ACM VEE*, 2006.
[9] W. Horrè *et al.*, "DAVIM: Adaptable Middleware for Sensor Networks," *IEEE Distributed Systems Online*, vol. 9, no. 1, 2008.
[10] M. L. Stefanizzi *et al.*, "Coin: Opening the Internet of Things to People's Mobile Devices," tech. rep. 2016.17, Politecnico di Milano, http://goo.gl/oMHRRT.
[11] M. Botts and A. Robin, "OpenGIS Sensor Model Language (SensorML)," OpenGIS Implementation Specification OGC, vol. 7, 2007.
[12] L. Mottola *et al.*, "Enabling Scope-Based Interactions in Sensor Network Macroprogramming," *IEEE MASS*, 2007.
[13] A. Lachenmann *et al.*, "Meeting Lifetime Goals with Energy Levels," *ACM SENSYS*, 2007.
[14] P. K. Dutta *et al.*, "Securing the Deluge Network Programming System," *IEEE/ACM IPSN*, 2006.
[15] N. Tsiftes *et al.*, "Efficient Sensor Network Reprogramming through Compression of Executable Modules," *IEEE SECON*, 2008.

Biographies

Maria Laura Stefanizzi (laura.stefanizzi@unisalento.it) received her Ph.D. degree in computer engineering from the University of Salento, Lecce, Italy, in 2016. She is currently a postdoctoral fellow with the Innovation Engineering Department of the University of Salento. Her research interests lie broadly in the areas of embedded systems and the Internet of Things.

Luca Mottola (luca.mottola@polimi.it) is an associate professor at Politecnico di Milano, Italy, and a senior researcher at SICS Swedish ICT. His research interests focus on modern networked embedded systems. He has received numerous awards including the Google Faculty Award, the Cor Baayen Award, the ACM SigMobile Research Highlight, and Best Paper Awards at ACM MOBISYS and ACM/IEEE IPSN. He is an Associate Editor of *ACM Transactions on Sensor Networks*.

Luca Mainetti (luca.mainetti@unisalento.it) is an associate professor of software engineering and computer graphics at the University of Salento. His research interests include web design methodologies, notations and tools, services oriented architectures, and IoT applications.

Luigi Patrono (luigi.patrono@unisalento.it) is an assistant professor of computer networks at the University of Salento. His research interests include RFID, the Internet of Things, cloud, wireless sensor networks, and embedded systems. He authored almost 100 scientific papers published in international journals and conferences. He has been Organizing Chair of some international symposia and workshops, technically co-sponsored by the IEEE Communication Society, focused on RFID technologies, and the Internet of Things.

Butler, Not Servant: A Human-Centric Smart Home Energy Management System

Siyun Chen, Ting Liu, Feng Gao, Jianting Ji, Zhanbo Xu, Buyue Qian, Hongyu Wu, and Xiaohong Guan

ABSTRACT

Smart home is an emerging area that opens up a diverse set of downstream applications, such as dynamic pricing and demand response techniques, whose goals are typically to lower power consumption while providing comfortable and convenient services. Smart home has been extensively studied, and shown to be beneficial in people's real lives. Although useful, typical smart home platforms work at the "servant" level — highly dependent on user inputs with no predictions of human demands — which, if well addressed, would significantly widen their applicability. In this article, we propose a human-centric smart home energy management system (SHE) that works at the "butler" level. The system integrates ubiquitous sensing data from the physical and cyber spaces to discover the patterns of power usage and cognitively understand the behaviors of human beings. The relationship between them is established to dynamically infer users' demands for electricity, and then the optimal scheduling of the home energy system is triggered to respond to both the users' demands and electricity rates. Based on the novel framework, our SHE system provides intelligent services to satisfy the requirements of users as a butler — aiming not only to save the electricity cost or reduce the peak load, but also to predict users' demands and managing "servants."

INTRODUCTION

Smart home has been identified as one of the key techniques for the transformation of people's lifestyles, and is attracting a great deal of attention in both academia and industry. During the past decade, smart devices have gotten popular and are becoming readily available for most people. This makes smart homes more applicable and appealing to broad masses as well. However, it seems that most people have not experienced the convenience and comfort that smart homes should have brought to our life. Thus, we raise our first question:

"Is the current smart home as smart as a British butler to make life convenient and enjoyable?"

Driven by the concept of smart homes, smart appliances are developing fast and are already in service in people's homes from kitchen to garage. With the capacities of interactive communication and remote control, smart appliances, such as smart washing machines, smart heaters, and smart saucepans, intelligently perform their specific functions and tasks for occupants, just like household servants doing their own predefined tasks. Most existing smart home platforms, such as Apple's HomeKit, Google Nest, and QQ IoT, still rely on user-driven management schemes. As more smart appliances and smart grid devices are integrated into smart home systems, it will be more difficult and complicated for users to control all devices wisely. Thereby, our answer to the first question is "No." We now pose the second question:

"Could the smart home provide the same level of service as a British butler?"

A well trained British butler can provide personalized services with the particular preferences and requirements of users in mind. There are a few steps for a butler to perform his functionality:

1. Observing the daily behaviors of users
2. Predicting users' requirements in advance and making elegant decisions with professional knowledge
3. Managing the corresponding servants

Step 1 is achieved by the various sensors installed throughout the smart home system. As shown in previous related studies, machine learning tools are able to solve step 2. Moreover, based on bidirectional communication networks, smart appliances could be controlled remotely to satisfy step 3.

Along with the evolution of wearable devices, including smartphones, smart bracelets, and other widely accepted devices around people, together with the development of sensors embedded in these devices, humans are evolving into an appealing Internet of Things (IoT) application [1]. With advanced sensors and computing capability, wearable devices are responsible for monitoring human behavior related data, such as GPS for triangulating locations, accelerometers and speed sensors for detecting movements, and heart rate and blood pressure sensors for assessing sleep quality and health status. The ubiquitous sensing technologies generate a huge amount of data associated with human living behaviors in the physical world. In addition, data from high-level social network applications would reveal human

The authors propose a human-centric smart home energy management system (SHE) that works at the "butler" level. The system integrates ubiquitous sensing data from the physical and cyber spaces to discover the patterns of power usage and cognitively understand the behaviors of human beings. The relationship between them is established to dynamically infer users' demands for electricity, and then the optimal scheduling of the home energy system is triggered to respond to both the users' demands and electricity rates.

Siyun Chen, Ting Liu, Feng Gao, Jianting Ji, Buyue Qian, and Xiaohong Guan are with Xi'an Jiaotong University; Zhanbo Xu is with Berkeley Education Alliance for Research; Hongyu Wu is with Kansas State University. Ting Liu is the corresponding author.

Digital Object Identifier: 10.1109/MCOM.2017.1600699CM

Improving energy efficiency of smart homes to save energy costs is moving forward at a fast pace as an important issue in smart grid, which is commonly achieved by jointly scheduling the generation, storage devices, and loads in smart homes with electricity rate signals.

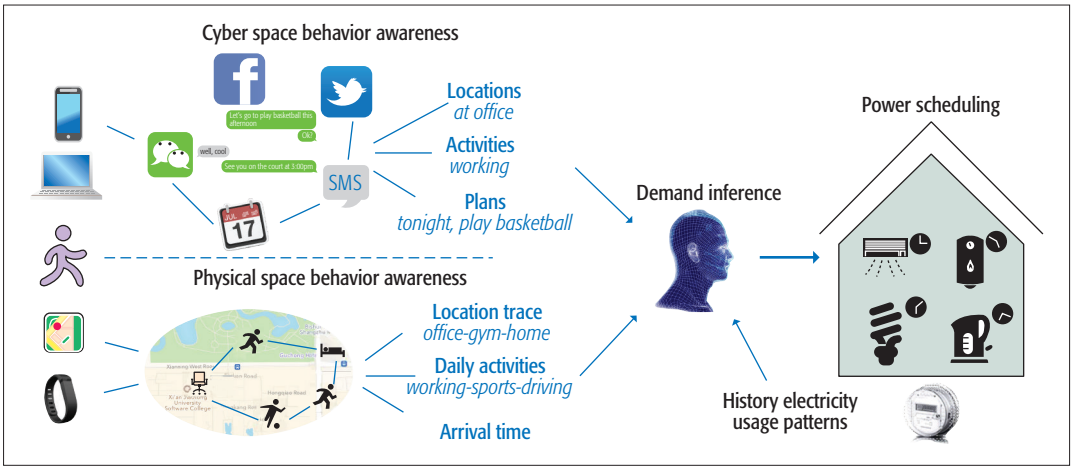


Figure 1. Overview of the SHE system.

behaviors in the cyber world. These data all together provide unprecedented opportunities to discover human behaviors, environmental factors, and social connections. With the aforementioned data and tools, we believe our answer to the second question is "Yes." We shall next expect the third question:

"Is it possible for the smart home to provide service beyond the capabilities of a British butler?"

Compared to the butler, the latest information technologies provide smart home with many advantages:

1. Ubiquitous sensing technology provides more comprehensive and complete data to cognitively understand human behaviors, especially for behaviors outside the home and unplanned sudden behaviors.
2. Detailed and accurate demand inference can be achieved based on big data analysis and artificial intelligence.
3. IoT technology provides technical support to precisely control the smart appliances in real time.

Improving energy efficiency of smart homes to save energy cost is moving forward quickly as an important issue in smart grid, which is commonly achieved by jointly scheduling the generation, storage devices, and loads in smart homes with electricity rate signals. It is believed to be a complicated and infeasible problem for butlers to come up with strategies of smart home energy systems to improve the economic benefits without sacrificing the comfort of occupants. Fortunately, based on advanced information and technologies, we can expect that the future smart home will be an omnipresent, 24/7 online, well trained and omniscient butler.

In this article, a human-centric framework of a smart home energy (SHE) management system is proposed, as shown in Fig. 1. A novel cyber-physical model is proposed to describe the human behaviors from both cyber space and physical space. Users' interaction information on social network applications, such as Facebook, Twitter, and WeChat, is monitored to understand online social behaviors in the virtual world, and the motion sensors, GPS, and biosensors on wearable devices are applied to track users' movements, locations, and sensations in the real world. Thanks to the more complete information of users, SHE can obtain more accurate inference of users' demands in smart homes and generate the optimal scheduling strategies of appliances based on comfort as well as economy. A SHE testbed is implemented in our lab to demonstrate how the SHE serves the smart home as a senior private butler.

Application Scenario

In this section, an application scenario is presented to illustrate how the SHE system works, and indicate the difference from existing smart home services.

Tom lives in a smart home with the SHE system. When he is at home, the existing context-aware smart home systems could provide personalized services to improve the comfort of his life [2, 3], such as automatically turning lights on/off, playing favorite music or TV programs wherever he goes, and pushing notifications. The SHE system mainly focuses on the major electricity consuming devices, such as the air conditioner, water heater, and washing machine. These devices are optimally controlled considering the electricity price for economy and history usage patterns for comfort. The control strategies can be adjusted based on the feedback of his physical information acquired from wearable devices.

When Tom leaves home, most existing context-aware smart home systems enter sleep mode after taking care of the home facilities, while SHE is still working. Tom often drives to work at 8:00 on weekdays, and usually comes back home at 18:00. When Tom leaves home, SHE starts to plan the optimal strategies of the focused devices to satisfy the demands when he comes back. During a break at 11:00, a basketball game invitation for 17:00 is sent to him from his friend through Facebook. SHE detects this message with Tom's confirmation and speculates that his arrival time will be delayed to 20:00. Meanwhile, Tom's demands change; for example, the hot water demand increases twice for a bath, and the washing machine is needed. Then the control strategies are updated. Due to the heat insulation capability, the water heater starts working in advance to avoid the peak load hours around 20:00. The washing machine is arranged to work at midnight at lower electricity price.

At 17:00, Tom is detected at the gym through the GPS sensor in his smartphone, and then the sport status is recognized by the motion sensors in his smart band. The event is confirmed, even if there is no appointment information at 11:00. But in this case the event can only be detected after 17:00, and the optimal working time of the water heater cannot be the lowest price time at 15:00.

The SHE system will continuously work to learn human behaviors, infer demands, and update strategies, which is the most remarkable feature compared to existing smart home systems.

Cyber-Physical Behavior Awareness

The Internet makes contact between people quick, convenient, and close. This process is further promoted by the development and popularization of smartphones and other smart devices. Various social network applications are developed to provide unprecedented services. People are enjoying a new life in cyber space, and some are even addicted. More importantly, current smart devices are commonly equipped with various sensors (i.e., location and motion sensors), which can be used to trace users' daily locations and learn the users' behavior patterns in the physical space [4]. With the integration of information in cyber space and physical space, a cyber-physical behavior awareness method is proposed in this section, as shown in Fig. 2.

Social Behavior Extraction in Cyber Space

Social networks have become a hot issue of research. A huge amount of data is generated by social behaviors in cyber space, which contain rich information on offline human behaviors, including locations, actions, and perspective. The interaction information is expressed in the forms of text, picture, voice, and video. As the most common expressions, the text content is mainly focused on in SHE system.

The attributes of interaction information in cyber space are extracted as time entity, location entity, action entity, and name entity by the named-entity recognition (NER) method [5], which is a common technique for text information processing to classify the named entities in text into predefined categories. The time entity consists of date and time, and usually follows some regular form, such as "at 9:30" or "this Friday." Location entity and action entity are presented by some related keywords with limited quantity. The name entity in social network applications is usually presented as "@user account." Without long-term training, the predefined NER method could extract the attribute vectors of social behaviors, which consist of the current time, date, location, actions, names, and so on.

Information Extraction in Physical Space

In physical space, there are various built-in sensors on a user's smartphone and wearable devices, such as GPS, motion sensors, and biosensors. The records include:

- GPS: time and locations
- Motion sensors: time, speed, tri-axial accelerometer
- Biosensors: heart rate, peripheral skin temperature

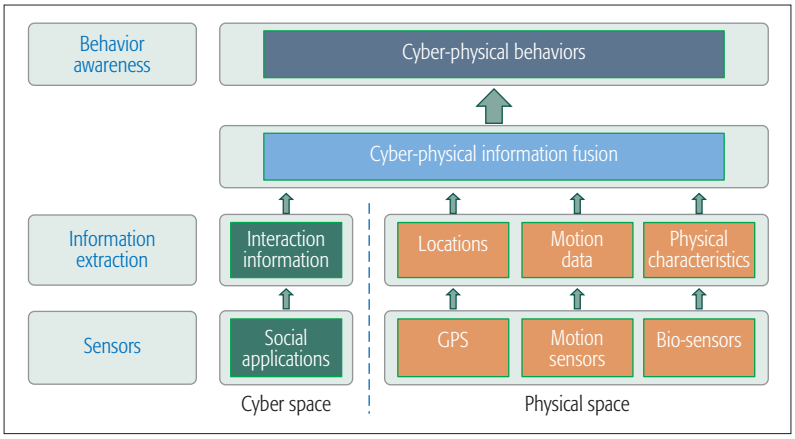


Figure 2. Cyber-physical behavior awareness.

The raw data of these sensors is collected and pre-processed on the smartphone to extract physical behavior information, including movement, motion status, physical states, and so on. The sensor data always lags behind the real behaviors, which are usually recognized from the series of records. As a result, the non-steady data is collected by a sliding window with coarse granularity to recognize the behaviors during a certain period. The Kalman filter and denoising methods [6] are employed to obtain unbiased sensor data. The GPS records are filtered to trace the valid locations. The records of motion sensors and biosensors with continuous values are discretized; for example, heart rate and speed is categorized into several levels ranging from low to high. Then the records are classified to the predefined behavior attributes through the classification decision trees. The common characteristics are sufficient to obtain the low-level motion states, such as static, walking, jogging, and running. In addition, the static characters could represent additional attributes, including the proportion, periodicity, and volatility of a certain attribute.

Cyber-Physical Behavior Awareness

The extracted behavior information in cyber-physical space is further fused to recognize the relevant behavior contexts, which include:

- Locations: where users are
- Activities: what users are doing or going to do
- Sensation: how users feel about the environment

Locations are valuable information to infer users' activities and estimate the time of arriving home. It is necessary to determine the locations of users, including home, office, supermarket, and other business and entertainment districts. They could be estimated based on the behavior information in cyber-physical space. At the initial training stage, the common places can be located with the help of the opening map services. The private places could be located through clustering the historical GPS data for several days based on the basic living habits of users. For instance, the locations of home and office can be estimated through clustering the location records at night and during work hours [7].

A user's activities are understood from the behavior information in cyber-physical space.

The relations between human behavior and consumption behavior are established by the neural network, which is trained by historical behavior and consumption patterns. Then current demands can be inferred from the users' behavior in the current time window according to their preferences for electricity consumption.

Social network information in cyber space contains rich clues of offline behaviors, which are valuable but unreliable. Whether intentional or not, the user may post inaccurate social behavior information. The behavior attributes extracted from the sensor data in physical space are reliable, able to detect and verify inaccurate behavior information in cyber space. For example, a swimming plan may be considered canceled if there is no swimming pool near the user around the date and time.

The sensations of users can be estimated from the perspective information of social behaviors and the physical data of biosensors, which provide feedback for the SHE system, especially thermal sensations. The peripheral skin temperature reflects the blood flow change related to the body temperature [8]. So the thermal sensation can be estimated through analyzing the distribution of peripheral skin temperature combined with the social behavior information. Although this process will take a relatively long time, the existing standards for comfortable living can provide tolerable settings for the SHE system.

Demand Inference Based on Human Behavior

The occupants' requirements in smart home are numerous, including indoor air cooling or heating, hot water, electric vehicle (EV) charging, and so on, which can be considered as the demands of electricity consumption of related appliances. The demands depend on many factors, including weather conditions, electricity price, and personalized behaviors and preferences of users. For instance, hot weather increases the cooling demand, sports may increase hot water usage, and the charging time of an EV depends on trip distance. With the continuous progress in artificial intelligence and deep learning [9], this kind of model can establish the complicated nonlinear relations between the demands and behavior context with the environment information.

The historical demand behaviors can be obtained through understanding the users' electricity usage patterns. Smart meters are widely deployed as an essential part of smart grid, with the capacities of communications, data storage, and calculation. The precise measurements of load profile by smart meters can be used to dig the potential information regarding the used devices and behavioral patterns of users. The key technology is load disaggregation, which aims to identify the appliances, detect the operation status, and estimate the corresponding electricity consumption through the aggregate power load [10]. In our previous work [11], the multidimensional load signatures, including active power, reactive power, and harmonic current, are extracted to detect and identify the appliances.

The characteristics of the electricity usage patterns can be extracted from the long-term disaggregated load information, such as the usage time range and distribution of each monitored appliance, the peak load, and the proportion of the electricity consumption during the peak price time. These characteristics indicate the consumption habits and preferences of users, including the preferred time range, the tendency to comfort or economy, and the responsiveness of the electricity price. These consumption behaviors of end users are meaningful not only for the SHE to generate satisfactory control strategies, but also for the utility to manage the demand side [12].

The relation between human behavior and consumption behavior are established by the neural network, which is trained by historical behavior and consumption patterns. Then current demands can be inferred from the users' behavior in the current time window according to their preferences for electricity consumption.

Optimal Home Energy Management

Future smart homes become complicated energy systems with various smart devices, energy storage (batteries of the EV), and generators (rooftop photovoltaic, PV, system). The occupants may be unavailable or not have enough electricity knowledge to manage the smart home in response to the demand response signals from the utility, such as dynamic price and load curtailment. Based on the user's demands inferred from behavior awareness, the SHE system works on providing comfortable services in smart home.

Objective: The objective of energy management in the SHE system is to minimize the electricity cost while satisfying the users' demands with the demand response signal, which is an optimization problem that includes the economic and demand objectives. On one hand, by participating in dynamic pricing-based demand response programs, occupants have incentives to manage their demand profiles to minimize energy cost in response to the varying price signal. On the other hand, the demands of occupants should be satisfied. Obviously, the above economic and demand objectives are contradictory. They can be converted into a single objective using weighting aggregation. The weight coefficients can be set according to the preferences of users.

Constraints: The operational conditions of all energy devices are formulated as the constraints based on the physical operational processes. The parameters of these models can be obtained from users or by fitting with actual data; the critical demands (so-called inelastic demands) should be satisfied during all stages according to the demand inference from the users' usage patterns. For example, the EV must be charged to meet the energy requirement for the next day's trip. In addition, the possible demand response notification of load curtailment from the utility should be considered as the constraint of the total load.

Solution: The optimization problem can be solved with several existing methods, including dynamic programming, particle swarm optimization (PSO) [13], and so on. However, since the occupants' behaviors and demands as part of objectives and constraints should be cognized and updated in every stage, which may require tremendous computational efforts. So in the SHE system, we develop an event-triggered mechanism method [14] to address this issue in a computationally efficient way. More specifically, the optimization is triggered when the behaviors are detected, and the control strategies are dynamically updated.

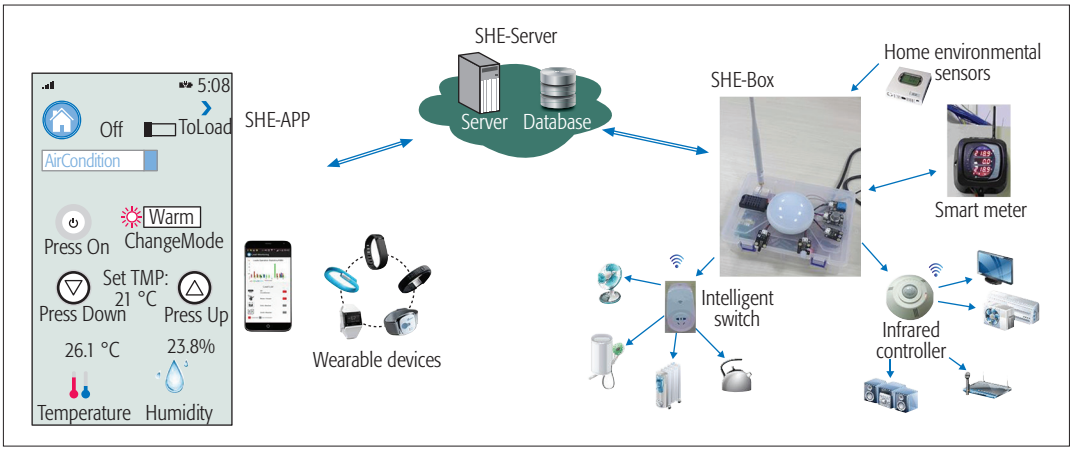


Figure 3. The testbed of SHE system.

The basic idea of SHE is to provide smart services by monitoring the user's behaviors from cyber and physical space. This framework could be extended to other user-oriented systems, such as recommendation systems, smart healthcare, and intelligent transportation systems, among others.

The Testbed of the SHE System

A testbed of the SHE system is set up on a campus with various smart devices and sensors, as shown in Fig. 3, consisting of three modules: the SHE-app, the SHE-server, and the SHE-box.

The SHE-app is an application developed on the smartphone platform, which integrates data acquisition, behavior awareness, demand inference, and data transmission. With the authorization of the user, the SHE-app monitors the social applications, the sensors of the smartphone, and the wearable devices connected to the smartphone via Bluetooth. As highly sensitive private information, the raw data and the behavior vectors are stored and analyzed on the smartphone. The sliding window of records is pre-processed to extract the behavior attributes. Then the user's demands are inferred from the behavior vectors, which contain a series of behavior attributes. As the amount of these data is small (about 18 kB for raw data, 2 kB for the behavior vectors, and 1 kB for the demands of appliances) at each period, the SHE-app consumes a small part of the total energy of the smartphone (about 10 percent), and the average computational time is about 2 s at each period. Meanwhile, the SHE-app provides the interaction with the user. The user can remotely monitor and control the smart home, and view and modify the system information.

The SHE-server is the computing center deployed on the commercial cloud computing platform. The user's demands relate to the information of required appliances with low sensitive privacy, which are anonymized and transferred to the SHE-server with encryption. The amount of the data to transfer is small (less than 1 kB). Then the major computing of decision generation is implemented on the SHE-server.

The SHE-box is the module integrated with the home gateway and the indoor environment sensors. The home gateway handles collection, storage, and transmission of data from the home sensor network. The central controller includes an infrared module and a WiFi module to control the appliances.

Case Study

Based on the testbed of the SHE system mentioned above, a case study is conducted to evaluate preference of the SHE. The participants live in a smart home, and two months of data are gathered, including the load data from the smart meter, behavior data from the wearable devices and social network appliances, and the environment data. The data of the first month are used to train the behavior awareness and demand inference models. The data of the following month are used for simulation. The human behaviors are learned to infer the demands in real time, which trigger the SHE system to update the optimal control strategies dynamically based on the real-time pricing (RTP) prices released by PJM in the United States. Here, we mainly focus on the power scheduling performance of the SHE system, and the results are shown in Fig. 4.

Figure 4a shows the power consumption profiles according to the final control decisions of a typical day. In response to the price mechanism and the detected user's demands, it is obvious that the peak load of the smart home is curtailed under the power scheduling with the SHE system, and the flexible devices are shifted to the low electric price period: the water heater works in advance at 15:00–16:00, the washing machine and clothes dryer are postponed to 4:00–5:00 the next day, and the dishwasher is scheduled to wait for an hour without changing the user's habits learned from the history usage distribution. The long-term economic performance of the power scheduling is illustrated in Fig. 4b. As the SHE system works on understanding the user's behaviors and providing more comfortable services, the user is not willing to curtail the demands. As a result, the energy consumption can hardly be reduced but can be shifted to lower price times to optimize the economic objective. In this case, the electricity cost can be reduced by 14 percent on average. To ensure the performance of cost saving, the SHE system is triggered to update the control decisions immediately after the behaviors are detected, instead of waiting until the behaviors are confirmed. As a result, the inaccurate information posted by the user may change the control decisions, and the executed parts may cause energy waste, such as on the 8th and 17th day. Although the conservative strategy that decisions are updated after the behaviors are confirmed can avoid this problem, it is unnecessary due to the infrequency of this scenario.

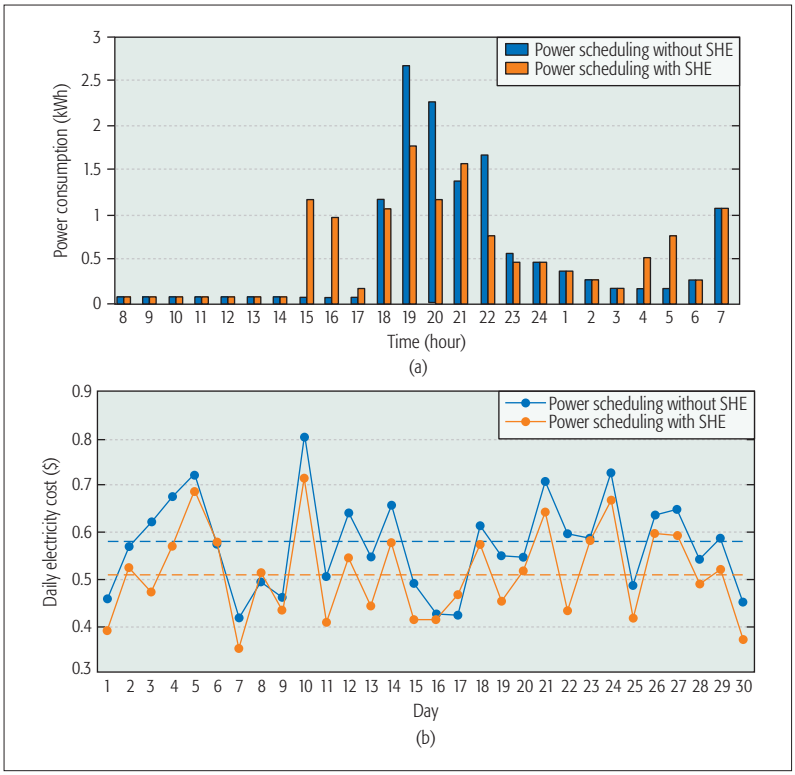


Figure 4. The results of power scheduling: a) hourly power consumption of a typical day; b) daily electricity cost of a month.

Conclusion

With the advent of smart grids, smart home technology should provide comfortable services for residential users, not only following a user's control, but also understanding and predicting the user's demand. In this article, we propose a human-centric smart home framework with the integration of understanding the behaviors of human beings, inferring the users' demands and consumption preferences, and optimally managing the energy devices in smart home. A SHE prototype is implemented to demonstrate how the SHE system, as a butler, could provide home services for the user. The SHE system can be extended for responding to various demand response signals and offering valuable support for programming and decision making at all levels of utility based on the people-centric framework.

Furthermore, the basic idea of the SHE system is to provide smart services by monitoring the user's behaviors from cyber and physical space. This framework could be extended to other user-oriented systems, such as recommendation systems, smart healthcare, and intelligent transportation systems, among others.

Acknowledgment

This work was supported by the National Key Research and Development Program of China (2016YFB0901905), the National Natural Science Foundation of China (91218301, 61473218,61304212, U1301254, 61632015), the Fok Ying-Tong Education Foundation (151067), and the Fundamental Research Funds for the Central Universities.

References

[1] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A Survey," *Comp. Net.*, vol. 54, no. 15, 2010, pp. 2787–2805.
[2] C. L. Wu and L. C. Fu, "Design and Realization of a Framework for Human_System Interaction in Smart Homes," *IEEE Trans. Systems, Man, and Cybernetics — Part A: Systems and Humans*, vol. 42, 2012, pp. 15–31.
[3] Z. Meng and J. Lu, "A Rule-Based Service Customization Strategy for Smart Home Context-Aware Automation," *IEEE Trans. Mobile Computing*, vol. 15, 2016, pp. 558–71.
[4] O. Yurur, C. H. Liu and W. Moreno, "A Survey of Context-Aware Middleware Designs for Human Activity Recognition," *IEEE Commun. Mag.*, vol. 52, 2014, pp. 24–31.
[5] D. Nadeau and S. Sekine, "A Survey of Named Entity Recognition and Classification," *Lingvisticae Investigationes*, 30.1, 2007, pp. 3–26.
[6] R. Sangeetha and Dr. B. Kalpana, "Denoising the Signals Using Kalman Filter for Target Tracking in Wireless Sensor Networks," *Proc. 2011 Int'l. Conf. Electronics Computer Technology*, vol.2, pp. 254–58.
[7] H. Liu, Y. Zhou and Y. Zhang, "Estimating Users' Home and Work Locations Leveraging Large-Scale Crowd-Sourced Smartphone Data," *IEEE Commun. Mag.*, vol. 53, 2015, pp. 71–79.
[8] K. Nakayama, T. Suzuki and K. Kameyama, "Estimation of Thermal Sensation Using Human Peripheral Skin Temperature," *IEEE Int'l. Conf. Systems, Man and Cybernetics*, 2009, San Antonio, TX, 2009, pp. 2872–77.
[9] J. Schmidhuber, "Deep Learning in Neural Networks: An Overview," *Neural Networks*, vol. 61, 2015, pp. 85–117.
[10] M. Zeifman and K. Roth, "Nonintrusive Appliance Load Monitoring: Review and Outlook," *IEEE Trans. Consumer Electronics*, vol. 57, no. 1, Feb. 2011, pp. 76–84.
[11] S. Chen *et al.*, "A Residential Load Scheduling Approach Based on Load Behavior Analysis," *2014 IEEE Int'l. Conf. Automation Science and Engineering*, Taipei, Taiwan, 2014, pp. 954–59.
[12] T. Liu *et al.*, "SHE: Smart Home Energy Management System for Appliance Identification and Personalized Scheduling," *Proc. 2014 ACM Int'l. Joint Conf. Pervasive and Ubiquitous Computing*, pp. 247–50.
[13] M. A. A. Pedrasa, T. D. Spooner, and I. F. MacGill, "Coordinated Scheduling of Residential Distributed Energy Resources to Optimize Smart Home Energy Services," *IEEE Trans. Smart Grid*, vol. 1, no. 2, Sept. 2011, pp. 134–430.
[14] S. Chen *et al.*, "SHE: Smart Home Energy Management System Based on Social and Motion Behavior Cognition," *2015 IEEE Int'l. Conf. Smart Grid Communications*, Miami, FL, 2015, pp. 859–64.

Biographies

Siyun Chen [S] (sychen@sei.xjtu.edu.cn) received his B.S. degree in control engineering and science from the North China Power Electric University in 2009. He is currently working toward a Ph.D. degree at the System Engineering Institute, Xi'an Jiaotong University, China. His research interests include smart grids, home energy management and optimization, load monitoring, and CPS.

Ting Liu [M'10] (tingliu@mail.xjtu.edu.cn) received his B.S. and Ph.D. degrees from Xi'an Jiaotong University in 2003 and 2010, respectively. He is currently an associate professor at the System Engineering Institute, Xi'an Jiaotong University. His research interests include vulnerability in smart grids, energy management, cyber-physical system, security and reliability in computer networks, and software system.

Feng Gao [M] (fgao@sei.xjtu.edu.cn) received his B.S., M.S., and Ph.D. degrees from Xi'an Jiaotong University in 1988, 1991, and 1996, respectively. He is a professor at the System Engineering Institute, Xi'an Jiaotong University. His research interests include machine learning, optimization and prediction in power system, and artificial intelligence.

Jianting Ji [S] (jtji@sei.xjtu.edu.cn) received his B.S. degree from Xi'an Jiaotong University, China, in 2014. He is currently working toward an M.S. degree at the System Engineering Institute, Xi'an Jiaotong University. His research interests include smart home systems and ubiquitous computing.

Zhanbo Xu [S'08, M'15] (zbxu@sei.xjtu.edu.cn) received his B.S. degree in electrical engineering and automation from Harbin Institute of Technology University, China, in 2008, and his Ph.D degree from the System Engineering Institute, Xi'an Jiaotong University, in 2015. Currently, he is serving as a postdoctoral researcher at Berkeley Education Alliance for Research in Singapore. His research interests are in the areas of smart grids,

building energy management and automation, and optimization of large-scale systems.

BUYUE QIAN [M] (qianbuyue@xjtu.edu.cn) received his B.S. degree in information engineering from Xi'an Jiaotong University in 2007, and his M.S. degree from Columbia University in 2009. He received his Ph.D. degree from the Department of Computer Science, University of California at Davis, in 2013. He was a research scientist at IBM T. J. Watson Research. He is currently an associate professor of computer science at Xi'an Jiaotong University. He received the Yahoo! Research Award, the IBM Eminence and Excellence Award, and the SIAM Data Mining 2013 Best Research Paper Runner Up Award.

HONGYU WU [SM'15] (hywu80@gmail.com) received his B.S. degree in energy and power engineering and Ph.D. degree in system engineering, both from Xi'an Jiaotong University, in 2003 and 2011, respectively. He is an assistant professor in the Department of Electrical and Computer Engineering at Kansas State University (KSU). He did his postdoctoral research at the Robert W. Galvin Center for Electricity Innovation, Illinois Institute of Technology, Chicago, from 2011 to 2014. Before joining KSU, he worked as a research engineer in the Power Systems Engineering Center, National Renewable Energy Laboratory (NREL), Golden, Colorado. His research interests include modeling and optimization of large-scale systems, home energy management, and renewable energy integration in smart grid.

XIAOHONG GUAN [M'93, SM'95, F'07] (xhguan@sei.xjtu.edu.cn) received his B.S. and M.S. degrees in control engineering from Tsinghua University, Beijing, China, in 1982 and 1985, respectively, and his Ph.D degree in electrical engineering from the University of Connecticut Storrs in 1993. He was a senior consulting engineer with PG&E from 1993 to 1995. Since 1995 he has been with the Systems Engineering Institute, Xi'an Jiaotong University, appointed as the Cheung Kong Professor of Systems Engineering since 1999, and the Dean of School of Electronic and Information Engineering since 2008. Since 2001 he has been the Director of the Center for Intelligent and Networked Systems, Tsinghua University, and served as the head of the Department of Automation from 2003 to 2008. His research interests include complex networked systems including smart power grids, planning and scheduling of electrical power and manufacturing systems, and electric power markets.

Smart Home: Cognitive Interactive People-Centric Internet of Things

Shuo Feng, Peyman Setoodeh, and Simon Haykin

The authors propose to integrate two entities, the Internet of Things and a cognitive dynamic system, and study a smart home scenario as the application of interest. With cognition as its foundation, the engineering paradigm of CDS provides step-by-step guidelines for systematic development of smart homes. Hence, CDS can significantly contribute to the interactive IoT ecosystem.

ABSTRACT

This article proposes to integrate two entities, the Internet of Things and a cognitive dynamic system (CDS), and studies a smart home scenario as the application of interest. As a people-centric IoT, the smart home aims to enhance the intelligence level of the living environment and improve the quality of human life. Cognition, the distinct principles of which are perception–action cycle, memory, attention, intelligence, and language, can play a key role to pave the way for building truly smart homes. With cognition as its foundation, the engineering paradigm of CDS provides step-by-step guidelines for systematic development of smart homes. Hence, CDS can significantly contribute to the interactive IoT ecosystem.

INTRODUCTION

The Internet of Things (IoT) is envisioned as a network that allows everyone and everyday objects to be connected anytime and anyplace [1]. In the past few years, IoT has gained enormous attention from the academic community as well as industrial organizations, and has been viewed as one of the main pillars of the fourth industrial revolution (Industry 4.0) [2]. Regarding the rapidly increasing number of connected devices, IoT can be viewed as a technological revolution toward ubiquitous connectivity, computing, and communications. In both the literature and real life, there are various applications of IoT, which include transportation, healthcare, agriculture, smart homes, vehicles, schools, markets, and industry, to name a few [3]. In this article, we focus our main attention on the particular scenario of interactive IoT implemented in a household, which is commonly referred to as a "smart home" or "home automation" [4]. The smart home is an interactive people-centric IoT, which is aimed at improving the quality of life. This goal is achieved via developing a more intelligent living environment.

Within the research field of IoT, efforts have been made on a particular sub-area called cognitive IoT (CIoT) [5], which aims to incorporate cognitive capability into the conventional IoT framework to some extent [6]. The notion of a cognitive dynamic system (CDS) [7, 8] provides guidelines to build cognition into IoT in a systematic way. Adopting human cognition as the frame of reference, CDS has the following five pillars: perception–action cycle, memory, attention, intelligence, and language. Introducing this framework for cognition into the engineering applications of IoT, a wide spectrum of tasks can be performed with minimal human intervention under the supervision of CDS.

This article is organized as follows. After this introductory section, the smart home scenario is presented. Then the compositional structure of CDS is discussed in detail. After that, the article will conclude in the final section.

SMART HOME SCENARIO

In general, a smart home can be viewed as an environment in which computing and communications technologies are employed for the use and control of different home appliances remotely or automatically to improve the resident's quality of life. It can be viewed as a subclass of larger categories involving smart buildings or even smart cities. In the smart home scenario, "things" in IoT refer to a set of sensors and actuators for daily use. In this context, data gathered by sensors is transferred to a decision making unit, which computes suitable control signals for achieving predefined goals. These computed control signals are then sent to the corresponding actuators. In this way, a real-time control system is built over a network. In this scenario, the IoT components are remotely controlled; therefore, in effect, the communication network plays a key role. IoT may include a diverse range of devices, such as cameras and thermometers as sensors, and electrical appliances and electronic locks as actuators. Such components can be remotely controlled by a smartphone or a computer via Bluetooth or the Internet. The smart home may also benefit from programmable devices such as light switches.

Depending on the user's personalized demands, home appliances can also be given specific instructions at any instant with various software applications (apps). The development of apps for all or some home appliances is becoming one of the most active areas in the smart home market. Currently, popular applications in smart homes include electric meters, which are viewed as indispensable devices for the realization of smart gird [9], thermostats, security systems, blinds, lighting, and door locks. An outstanding representative is the CASAS architecture [10], which facilitates the development and implementation of future smart home technologies by offering an easy-to-install lightweight design that provides smart home capabilities out of the box

This work was supported by the National Science and Engineering Research Council (NSERC) of Canada and the China Scholarship Council.

Digital Object Identifier: 10.1109/MCOM.2017.1600682CM

The authors are with McMaster University; Peyman Setoodeh is also with Shiraz University.

with no customization or training. Another representative example is the software-defined smart home platform [11], which is built on the architecture of software-defined networking (SDN) and can be connected to other platforms through an open interface.

Another promising approach is to bring the CDS into play. In this article, we mainly focus on the smart home scenario based on CIoT. The supervision of CDS over the networked components of interactive IoT in a house will make a remarkable difference in our daily lives. To be more specific, let us take a look at the following situation, which will be referred to as the "falling-asleep problem" throughout the rest of this article:

Imagine that a person is gradually falling asleep on his/her sofa while watching a TV drama in the living room on a Friday night. To ensure a good environment for sleep, the air conditioner, TV, sofa, and other appliances should be enabled to sense people's movement, gestures, body temperature, and/or voice in a coordinated manner. Then the gathered information can be used to realize whether the resident is awake, asleep, or half-asleep. Knowing the state of the person in the room, appropriate decisions can be made to further comfort her/him. For instance, the TV itself gradually lowers or turns off the voice, the sofa slowly changes its shape into a bed, and the air conditioner dynamically adjusts the temperature to be suitable for sleep based on the body and room temperature. Moreover, the washer or dryer in the basement should also stop working to avoid noise (and restart to finish the work the next morning automatically). If the lights at the front door are still on, it would be better for them to be dimmed or turned off for energy conservation. The security systems should also become alert once the resident has fallen asleep. All of these visions can be brought into reality with a CDS acting as the "brain" of the household.

Obviously, there are plenty of other situations that would also make the smart home seem very appealing. These situations share a similarity, which is that the home appliances are capable of sensing the changes in the environment and adjusting themselves to adapt to the environment cooperatively. For those applications of CIoT with similar properties, discussions here could be leveraged for their realization and improvement.

There are several questions to be answered to fully understand how a smart home will deal with this falling asleep case. Answers to these questions provide insight on how to improve the system design: How do the appliances interact with people and the environment? How do they know whether people are awake or asleep? How can they make adjustments in an effective and efficient way? How can they still perform well when there are disturbances such as a party being held in the neighborhood? The answers lie within the principles of CDS.

In the following section, we follow the well-known engineering idea of "divide and conquer." To be more precise, a complicated problem will be broken down into several simple sub-problems, which can be formulated and solved directly, and then the sub-solutions are combined to obtain a solution for the original problem.

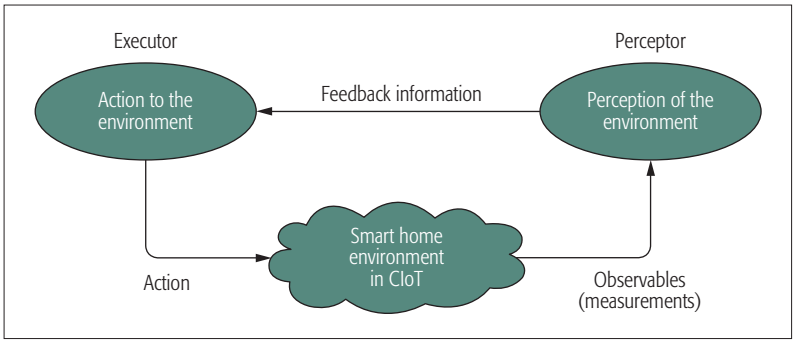


Figure 1. Perception-action cycle in a smart home supervised by the CDS.

Smart Home under the Supervision of the Cognitive Dynamic System

The functional block diagram of a smart home supervised by the CDS is built in an orderly fashion according to the principles of cognition, that is, perception–action cycle, memory, attention, intelligence, and language [7]. By building one principle over another, the performance of this typical CIoT application supervised by the CDS is improved. However, this is achieved at the price of a more complicated architecture, which demands more computational and storage resources. In practice, one of the most crucial factors that affect the system performance is the information processing capability. The required high-performance processing can be obtained by deploying more powerful processors supplemented by cloud computing. The latter relies on external networks or platforms. In the following, the role that each pillar of cognition will play in a smart home is described in detail.

Perception-Action Cycle

From the engineering perspective, the perception-action cycle is the backbone for system implementation. The cycle begins with the perceptor perceiving the environment (physical and/or social world) by processing the incoming stimuli, called observables or measurements, as shown in Fig. 1. In response to feedback information from the perceptor about the smart home environment, the executor acts, and thus, the cycle goes on [7].

The household must be equipped with an appropriate set of sensors to perceive the environment as well as the residents who live in it. The particular composition of the environmental sensors is naturally dependent on the application of interest. For instance, light sensors can be deployed on the blinds or curtains to control the light shed through the window, acoustic sensors can be installed on the doors or TV for voice control, and so on. However, it should be noted that usually the environment is not fully observable. Typically, not all households in a community are observable, and not all areas in one household are observable (due to various reasons such as privacy concerns or hardware limitations).

Among others, the sofa in the falling asleep situation is of particular importance since the sofa has direct contact with the resident. Therefore, it is reasonable to place different kinds of sensors inside the sofa such as pressure sensors and temperature transducers, hopefully without causing

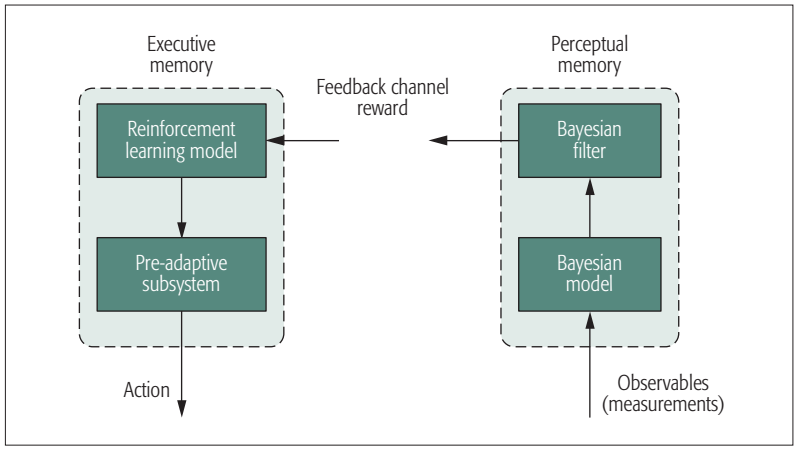


Figure 2. Functional block of memory in a smart home supervised by the CDS.

any kind of discomfort in the appearance or feel of the sofa. Moreover, different kinds of appliances have distinct working patterns. To elaborate, the pressure sensors inside the sofa should be functioning all the time, since they are responsible for not only perceiving the resident's gestures, but also determining which temperature transducers should be put online. As for the TV, the acoustic sensors installed on it could also go to sleep mode once the resident's state is classified to be in the asleep mode, since no further instructions would be given other than snores or noise. This discrimination is useful for the simplification of memory and attention design in the following subsections.

Memory

Memory is added to the two sides of the perception-action cycle as perceptual memory and executive memory, respectively. While perceptual memory enables the perceptor to interpret the observables so as to recognize their distinctive features and categorize the learned features accordingly in some statistical sense, executive memory keeps track of the chosen actions in the past and their effectiveness (Fig. 2) [7].

Typically, the environment in interactive CIoT is non-stationary, which means that the underlying behavior of the environment continually changes with time (e.g., temperature fluctuation or unconscious turning over during sleep). Given such an environment to deal with, the memory in CDS is indispensable. The function of memory, from an information processing perspective, is to learn from the environment and store the knowledge thus acquired, continually update the stored knowledge in light of environmental changes, and predict the consequences of actions taken and/or selections made by the system as a whole.

As depicted in Fig. 2, the Bayesian model of the observables lies at the bottom end of the perceptor, followed by the Bayesian filter for estimating the environment state. On top of the executor, the reinforcement learning model exploits the internal reward attributed to imperfection in the perceptor, followed by the pre-adaptive subsystem.

The Bayesian model is mainly responsible for feature extraction. Reinforcement learning is the mathematical paradigm of learning, which aims to choose the best possible action on the sole basis of environmental rewards (positive or negative). The goal of reinforcement learning is to maximize some form of rewards that will be accumulated over the course of time as the consequences of a selected action at the current cycle.

The Bayesian filter, which is the optimal recursive data-processing algorithm, serves as a general unifying framework for sequential state estimation [12]. Hence, we can easily solve the state estimation problem in the smart home scenario. As a matter of fact, a Bayesian classifier is capable of accomplishing the state estimation in simple tasks such as the falling-asleep situation with only a couple of categories. Here, we adopt the Bayesian filter for coherence.

For instance, the acoustic sensors installed on the TV can obtain the sequence of observations. Based on how long since the resident has laughed (observations) and how well he/she usually sleeps (transition-state distribution), the state estimator will be able to estimate the current state of the resident in terms of being asleep or not. In this circumstance, the function of the Bayesian filter is to estimate the physical state of the resident by monitoring his/her voice or tracking his/her motion across time, given a set of observables (measurements) obtained by the sensors.

Furthermore, more inputs/observations from pressure sensors or temperature transducers will be helpful to improve the accuracy of the estimation. For instance, if the resident has not been up for a relatively long time, or his/her breath is relatively slow, it is more likely that he/she has already fallen asleep.

Attention

Similar to memory, attention can be added to the two sides of the perception-action cycle as perceptual attention and executive attention. Attention prioritizes the allocation of available resources in such a way that the information gathering and processing power are focused on what is of critical strategic importance. While perceptual attention deals with the information overflow problem, executive attention implements a version of the principle of minimum disturbance.

A multilayer Bayesian model and a hierarchical reinforcement learning model are shown in Fig. 3. The advantages of building hierarchy in the form of layers into the design of attention are threefold. First, a deep architecture allows for learning more complicated representations. Second, the multilayer structure ensures an enlarged number of local and global feedback loops in the perception-action cycle, which in turn is a facilitator of intelligence. If exploited properly, this may lead to a more reliable decision making process in the face of environmental uncertainties. Third, a deep architecture provides the ability to learn features of features [7]. The first layer extracts relevant features contained in the observables, and subsequent layers process incoming features. Specifically, the features characterizing the observables in the perceptor or those characterizing the feedback information in the executor become increasingly more abstract and therefore easier to recognize as the hierarchical depth is increased.

In short, the importance of hierarchy can be described as improving attention and hence improving the overall performance of the CDS

over CIoT. However, there is a trade-off of increased computational complexity for system performance, particularly when the requirement is reliable decision making in the face of environmental uncertainties. Reliability can also be improved via redundancy.

In Fig. 3, the feedback within each iterative local feedback loop (as shown by the red circles) accounts for the localized attention. On the perceptor side, we have downward attention layer by layer, which means focus (i.e., the allocation of computational and storage resources) keeps improving. The modeling capability of the Bayesian model for the observables gets better and better with as the number of layers increases. On the control side, we have upward attention layer by layer, which works in a similar fashion.

Generally, the global cycle performs very well when the smart home environment in CIoT is stationary. On the other hand, the shunt cycles (i.e., local feedback loops) come into play whenever a disturbance occurs (i.e., higher risk). Here, risk is defined as the expected loss or cost, as in statistical decision theory [13].

The Bayesian model in CDS is where the uncertainty is perceived first, so it must be equipped with sensitive and effective means to detect the presence of uncertainty. With the successful detection of disturbance, we experience a change in the errors produced by the perceptor in the Bayesian filter (i.e., state estimator). That is, the error level is increased within the state estimator by virtue of the uncertainty. As a result, the entropy of the perceptor goes up, and the internal rewards in the reinforcement learning model drop to a lower level.

The manifestation of the principle of attention is not very straightforward in the smart home scenario in CIoT. To illustrate how the process of attention functions in the falling asleep situation, let us consider the following case. The resident fell asleep a few moments ago and is currently having a beautiful dream (or a nightmare). An event such as unconscious turning over, pushing the blanket away, or shaking an arm may occur at some particular point in time and specific location in the sofa, thereby resulting in a different physical state of the resident in terms of sleeping gesture and body temperature.

By virtue of built-in space-time processing, the pressure sensors and temperature transducers send information to the Bayesian filter and reinforcement learning model, identifying the precise area of the sofa where the resident is lying, and whether he/she pushes the blanket away. Furthermore, the CDS itself focuses its attention on those remaining fields/aspects that are worth monitoring. To be specific, acoustic sensors installed on the TV can be completely turned off for energy conservation. The pressure sensors inside the sofa should be functional continuously as they are responsible for detecting body movements, even tiny ones. Unlike these two types of sensors, the temperature transducers are deployed right under the resident inside the sofa. That is to say, the subset of functioning temperature transducers varies over the course of time, depending on the particular body movements of the resident. Here, the valuable computation and energy resources are allocated to that subset of sensors. This is in accordance with the design principle that attention should aim at prioritizing the allocation of available resources for what matters most.

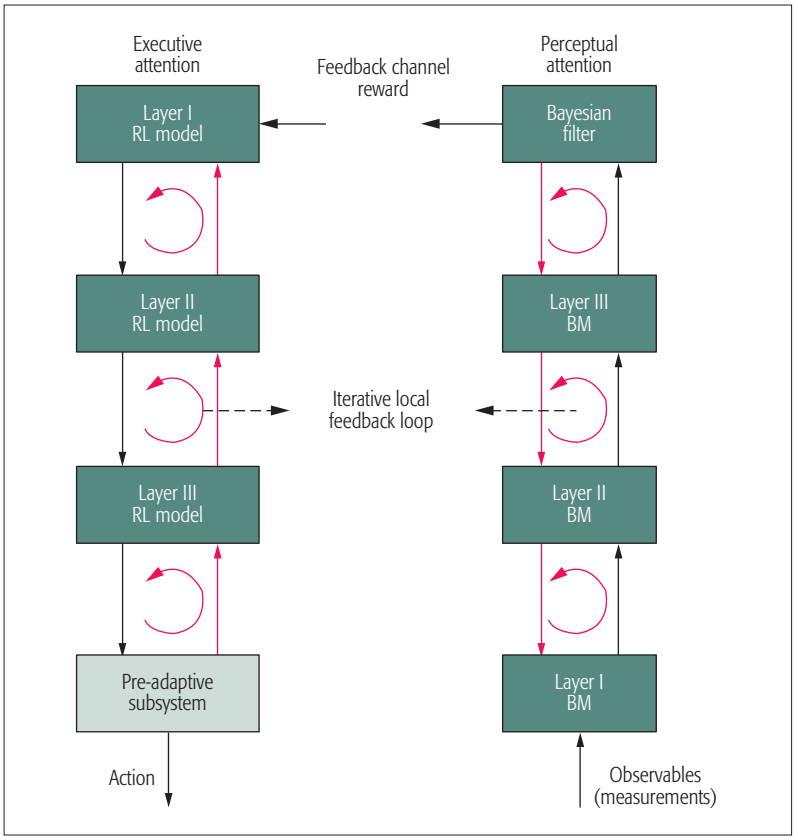


Figure 3. Attention (i.e., focusing) in a smart home supervised by the CDS.

Moreover, through its own self-organized learning process, the executive part of the CDS builds a predictive model of the resident's movement. Using this model, the CIoT is enabled to do the following: pay attention to the behavioral patterns of the resident, and predict the time interval over which the resident would not move. The latter determines the estimated valid period of the current strategy for sofa shape control and sensor scheduling.

Intelligence

Intelligence is built on the former three pillars (perception–action cycle, memory, and attention). It also brings into play the predictive pre-adaptive characteristic, which was originated in cognitive neuroscience and is a fairly new concept in engineering. The predictive subsystem of the executive part consists of two entities that are responsible for prediction and pre-adaptation, as shown in Fig. 4. Among the four principles of CDS design, intelligence is the most complicated one, which appears as an emergent behavior. Nevertheless, intelligence is the single most important function in CDS. It should be noted that the abundant presence of feedback at multiple levels, including both global and shunt cycles, facilitates intelligence. This allows the system to aim for optimal decision making in the face of inevitable uncertainties in the environment [7].

Since reinforcement learning aims to select the best possible action based on internal rewards that are the consequences of a selected action at

the current cycle, the outcome of the reinforcement learning model is naturally defined by the current cycle. Then the predictor computes the current scenario of the environment via the perceptor one cycle ahead.

In the presence of uncertainty, the probabilistic decision making block is fed with two sources of inputs, which are perturbed actions selected at the current cycle on the left (from the library of potential actions) and past experiences on the right (from the executive memory). Then the probabilistic decision-making part picks an optimal past experience based on these two inputs.

Similar to many engineering applications, the CIoT system is also subject to potential security threats. The first attack can be initiated at any time even without any sign at all. However, with the availability of the preamble, the pre-adaptation block may be able to infer that the system has been the target of a cyber-attack. This makes the system more cautious; therefore, the system will try to anticipate the next attack in terms of its initial time, target components, type of attack, and so on, thanks to the partial knowledge gained in the preamble. The defense mechanism can be improved via the learning process from one cycle to the next. In practice, there is usually a budget for the total investment on security issues. Therefore, there is naturally a cost-protection trade-off [14]. If the cost exceeds the anticipated achievement/improvement on security, the chosen strategy for protection is not justified, and we should look for other approaches.

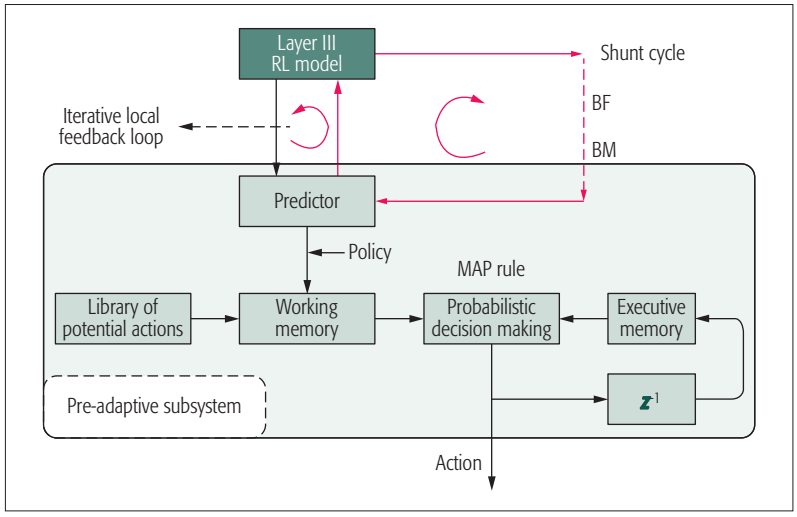


Figure 4. Pre-adaptive subsystem as a facilitator of intelligence in a smart home supervised by the CDS.

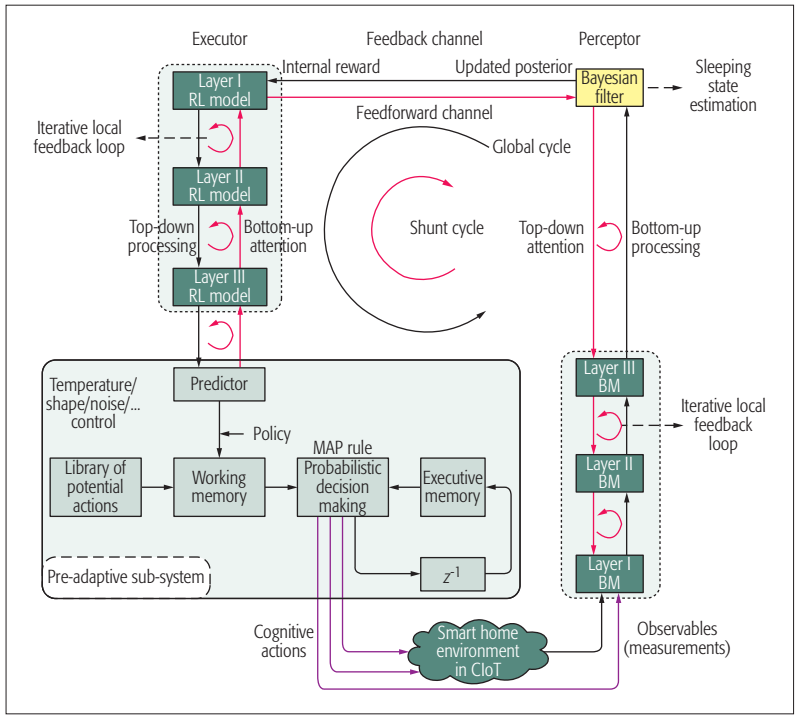


Figure 5. The complete block diagram of CDS as the supervisor of CIoT in a smart home.

Language

Just as language plays a distinctive role in the human brain, so it is in a smart home, where language paves the way for connectivity of different components in the interactive IoT via effective and efficient machine-to-machine (M2M) communications. Challenging research issues involved in enabling M2M communications include architecture, standardization, identification, addressing, security, and so on [15]. However, a detailed description regarding language is outside the scope of this article.

Compositional Structure of CDS

Finally, we provide the compositional structure of CDS as the supervisor of the interactive people-centric IoT in a smart home, which is an integration of all the previously discussed building blocks, as shown in Fig. 5. As described in [13], different kinds of cognitive actions can influence different components of the system. The next section discusses this issue for the same falling asleep scenario.

Cognitive Actions

Regarding the falling asleep situation in the smart home scenario, there are at least three kinds of cognitive actions that can be applied to the interactive system, as shown in Fig. 5.

1. This first kind of cognitive actions are those that are applied to the environment in order to indirectly affect the perception process. For example, if there is a party being held next door, the noise level and the strength of light shed through the window would naturally increase. Therefore, the controllers installed on the doors and windows will close them automatically to minimize the effect of disturbance from outside. Also, the blinds or curtains will close automatically for the night, which is part of the daily routine for smart home appliances.

2. The second kind of cognitive actions are those applied to the system itself in order to reconfigure the sensors and/or actuators. The previous discussions on how the pressure sensors determine which temperature transducers are turned on while others are turned off, and determine when to turn off the acoustic sensors, fall into this scope.

3. The third kind of cognitive actions are those applied as a part of state control actions (physical actions). In such a case, a physical action is applied

to the system but with the goal of decreasing the information gap (with or without other goals). The most direct example is the shape-changing process of the sofa. Through the shape changing, the sofa not only makes the resident more comfortable as this sofa (or sofa bed, to be more precise) takes the shape of the human body to some extent, but also gets better observables due to the closeness between the body and sensors.

Concluding Remarks

This article proposes using the cognitive dynamic system paradigm for enabling the Internet of Things with cognition, where the smart home is the application of interest. A smart home can benefit from different building blocks of the CDS:

1. The smart home must be aware of the household and its surroundings; the perceptual part of the CDS will do the job.
2. Implementing the CDS requires almost the same set of sensors/actuators, which are already used to perform daily tasks in a smart home.
3. The smart home must establish an efficient but very comfortable living environment for the resident; the executive part of the CDS equipped with attention and intelligence takes actions to achieve this goal.

This list can continue. Therefore, the combination of CDS and IoT to form a CIoT is anticipated to have great advantages compared to the existing smart home applications. With the cognitive actions being performed from one cycle to the next, the quality of service and even the quality of experience of the CIoT will be significantly improved.

Acknowledgment

The authors would like to thank Dr. Guoru Ding and Markimba Williams for their insightful suggestions and valuable comments.

References

[1] V. Ovidiu and F. Peter, *Internet of Things: Converging Technologies for Smart Environments and Integrated Ecosystems*, River Publishers, 2013.

[2] C. Perera *et al.*, "A Survey on Internet of Things: From Industrial Market Perspective," *IEEE Access*, vol. 2, Dec. 2014, pp. 1660–79.

[3] A. A.-Fuqaha *et al.*, "Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 4, 4th qtr. 2015, pp. 2347–76.

[4] A. Zanella *et al.*, "Internet of Things for Smart Cities," *IEEE Internet Things J.*, vol. 1, no. 1, Feb. 2014, pp. 22–32.

[5] Q. Wu *et al.*, "Cognitive Internet of Things: A New Paradigm Beyond Connection," *IEEE Internet Things J.*, vol. 1, no. 2, Apr. 2014, pp. 129–43.

[6] V. Foteinos *et al.*, "Cognitive Management for the Internet of Things: A Framework for Enabling Autonomous Applications," *IEEE Vehic. Tech. Mag.*, vol. 8, no. 4, Dec. 2013, pp. 90–99.

[7] S. Haykin, *Cognitive Dynamic Systems: Perception-Action Cycle, Radar, and Radio*, Cambridge Univ. Press, 2012.

[8] S. Haykin, "Cognitive Dynamic Systems: Radar, Control, and Radio," *Proc. IEEE*, vol. 100, no. 7, July 2012, pp. 2095–2103.

[9] A. Zipperer *et al.*, "Electric Energy Management in the Smart Home: Perspectives on Enabling Technologies and Consumer Behavior," *Proc. IEEE*, vol. 101, no. 7, Nov. 2013, pp. 239–408.

[10] D. J. Cook *et al.*, "CASAS: A Smart Home in a Box," *Computer*, vol. 46, no. 7, July 2013, pp. 62–69.

[11] K. Xu *et al.*, "Toward Software Defined Smart Home," *IEEE Commun. Mag.*, vol. 54, no. 5, May 2016, pp. 116–22.

[12] S. Haykin, *Neural Networks and Learning Machines*, 3rd ed., Pearson Press, 2009.

[13] S. Haykin *et al.*, "Cognitive Control," *Proc. IEEE*, vol. 100, no. 12, Dec. 2012, pp. 3156–69.

[14] L. Zhang *et al.*, "Byzantine Attack and Defense in Cognitive Radio Networks: A Survey," *IEEE Commun. Surveys & Tutorials.*, vol. 17, no. 3, 3rd qtr. 2015, pp. 1342–63.

[15] J. Kim *et al.*, "M2M Service Platforms: Survey, Issues, and Enabling Technologies," *IEEE Commun. Surveys & Tutorials.*, vol. 16, no. 1, 1st qtr. 2014, pp. 61–76.

Biographies

Shuo Feng (fengs13@mcmaster.ca) received his B.Sc. degree in electrical engineering from the University of Electronic Science and Technology of China in 2011, and his M.Sc. degree in communications and information systems from the PLA University of Science and Technology, China, in 2014. He is currently pursuing his Ph.D. degree in computational science and engineering at McMaster University, Hamilton, Ontario, Canada. His research interests include cognitive radio networks, machine learning, and cognitive dynamic system.

Peyman Setoodeh (setoodeh@ieee.org) received his B.Sc. and M.Sc. degrees in electrical engineering from Shiraz University, Shiraz, Iran, and his Ph.D. degree in computational engineering and science from McMaster University. He is currently a principal research engineer at McMaster University and an assistant professor with the School of Electrical and Computer Engineering, Shiraz University. His research interests include cognitive systems, machine learning, complex networks, and systems biology.

Simon Haykin [F] (haykin@mcmaster.ca) received his B.Sc., Ph.D., and D.Sc. degrees in electrical engineering from the University of Birmingham, United Kingdom. He is a Distinguished University Professor with the Faculty of Engineering, McMaster University. His research interests include learning from the human brain and applying it to a new generation of cognitive dynamic systems, exemplified by cognitive radar, cognitive control, and cognitive radio. He is a Fellow of the Royal Society of Canada.

The combination of CDS and IoT to form a CIoT is anticipated to provide great advantages compared to the existing smart home applications. With the cognitive actions being performed from one cycle to the next, the quality of service, or even the quality of experience of the CIoT will be significantly improved.

The Experience of Using the IES Cities Citizen-Centric IoT Platform

Stefanos Vatsikas, Georgios Kalogridis, Tim Lewis, and Mahesh Sooriyabandara

Despite people being an integral part of the IoT ecosystem, the need for people-oriented applications and features has not yet been prioritized. IES Cities aims to address this by being a new cloud-based, citizen-centric platform. It is an open-source, adaptable and easy-to-use solution that has been designed around the needs of citizens and has been extensively tested in real-world deployments in a number of European cities.

ABSTRACT

The race to create the dominant platform for the world of the Internet of Things is a crowded one, with the number of potential available solutions constantly increasing. Many of these technologies are restricted by being platform-specific, application-specific, or not being open source. In particular, despite people being an integral part of the IoT ecosystem, the need for people-oriented applications and features has not yet been prioritized. IES Cities aims to address this by being a new cloud-based, citizen-centric platform. It is an open source, adaptable, and easy-to-use solution that has been designed around the needs of citizens and has been extensively tested in real-world deployments in a number of European cities. IES Cities is a working system that allows and encourages the building and rapid deployment of people-oriented, cross-platform applications across a wide range of smart environments.

INTRODUCTION

The growth of the Internet of Things (IoT) concept continues to be explosive [1], and the number of real IoT devices is constantly growing, from wearable devices such as fitness bands to factory automation systems and infrastructure monitoring. This generates an unprecedented amount of data, the value of which will remain limited unless it can become organized in a way that provides context and makes it accessible and meaningful to people. IoT platforms can facilitate this by providing services such as data discovery, accessibility, storage, and processing.

Fortunately, there are numerous such platforms in various domains, targeted at different types of applications and users. Notable examples include Fi-WARE [2], OpenIoT [3], Libellium™ Smart Cities, IBM Watson™ IoT, Oracle™ IoT, Apple HomeKit™, Google Brillo™, and ThingWorks™. Some platforms may focus exclusively on connecting specific IoT devices to the cloud, others may focus on specific domains such as smart cities or infrastructure monitoring, while others may try to cover a wider variety of domains or topics. Not all platforms are direct competitors, as each may have its own target audience. However, an apparent common theme across most platforms is lack of people-centric features. IoT platforms should offer open, trustworthy, adaptable, and intuitive solutions with low entry barriers that enable the rapid development and deployment of socially aware, people-oriented, and cross-platform applications across a diverse range of smart environments such as smart cities, e-health, and education. Real-world data from diverse sources, such as IoT devices, open datasets, and user-generated data, should be used together to provide added value and put people in the center of the IoT ecosystem.

IES Cities is a citizen-centric platform that has been deployed in four cities across three European countries, and our experience of using it has shown that it is a strong solution for IoT usage. In the remainder of this article we present the design and features of the IES Cities platform and discuss our experience of using it for developing and deploying real-world, people-oriented applications. We also present the use case of HealthyOffice, an IoT mobile application that is built on the IES Cities platform and uses sensor and user-generated data to recognize and predict a person's mood. First, however, we present a short review of a range of IoT and smart city platforms.

REVIEW OF IOT AND SMART CITY PLATFORMS

In this section we review a selection of IoT and smart city platforms; it is worth noting that the line between smart city and IoT is often blurred, with several solutions bridging these domains. Activity in these areas has been and still remains strong, with a large number of relevant solutions having become available over the last few years. The target groups of these solutions are quite diverse; some solutions target public administrations, some target citizens and end users, while others target large enterprises, application developers, or device manufacturers. Our intention here is to provide a brief overview of the different types of platforms that exist, along with some of their defining characteristics. A more thorough survey of a large number of IoT platforms can be found in [4], while [5] discusses a range of considerations to take into account when selecting an IoT platform.

FIWARE is a European Union (EU)-funded smart city open initiative that offers a wide range of solutions; the most prominent are the FIWARE Platform itself, which offers a set of application programming interfaces (APIs) to facilitate the development of smart applications in multiple vertical sectors, and FIWARE Lab, which is a non-commercial sandbox environment where developers can test their apps and exploit open data published by cities and other organizations. CitySDK [6] is another EU-funded initiative that provides a service development kit for cities and developers, with the objective of harmo-

Digital Object Identifier:
10.1109/MCOM.2017.1600688CM

The authors are with Toshiba Research Europe Limited Bristol.

nizing application development across cities with a focus on making open and linked data available to developers. OpenIoT is an open source middleware platform for getting data from sensor clouds. Its focus is on managing cloud environments for IoT entities, such as sensors, actuators, and smart devices. In general, it aims to enable the concept of "sensing as-a service." CityZenith 5D [7] is an open city information management platform that aims to help cities better understand data coming from IoT devices and manage city resources, such as buildings and infrastructure; the current version targets both public administrations and private businesses. WSO2 is an open source middleware platform with a very wide range of components, such as the IoT Server, Dashboard Server, and Process Center. It is a complex and powerful platform that can be used across many domains and primarily targeting the enterprise world. IBM Smarter Cities™ is a suite of commercial smart city solutions primarily targeted to public administration entities, while IBM Watson IoT is a platform for device management, connectivity, real-time analytics, application development, cognitive services and storage of IoT data. Oracle Smart City Platform is another commercial solution that targets local government; this platform offers a wide range of modular services, such as data analytics, IT infrastructure, and financial administration. Oracle also offers an IoT Cloud Service, which aims to provide organizations with an easy and quick way of designing and deploying IoT applications. The Libellium Smart Cities platform is aimed primarily at hardware manufacturers and system integrators; it offers a hardware and software modular platform that allows the easy deployment of sensor-based systems and services across a city. Google and Apple are also offering their own IoT platforms; the former has been developing Brillo, an open operating system for IoT-type devices, and the latter offers a home automation framework called HomeKit. Finally, Samsung offers Artik™, an IoT solution that comprises a range of hardware modules and an open software environment, which primarily targets IoT device manufacturers.

The IES Cities Cloud Platform

The IES Cities citizen-centric IoT platform is the outcome of the EU funded project IES Cities [8] that ran between March 2013 and February 2016. The project consortium consisted of 14 partners across five countries: Italy, Spain, Slovenia, Germany, and the United Kingdom. The main goal of the project was to create an open cloud platform that would enable citizens in different cities across Europe to both provide and consume online services based on their own and external linked data related to the cities; the result of this was the successful creation and deployment of the IES Cities cloud platform. In addition, a number of citizen-centric online services (i.e., mobile apps) that utilize the platform were developed. Finally, in order to assess the platform and its impact the project also included pilot deployments in four cities: Rovereto, Italy; Bristol, United Kingdom; and Zaragoza and Majadahonda, Spain. The final outcome was successful, with more than 16,000 users in total and up to 1400 users per month actively using the apps and the platform and providing very positive feedback. The platform is available for free-of-charge use on https://cityiot.co.uk/iescities, where it will remain available for the foreseeable future. Rich documentation on how to start using the platform is also available on the same link. Finally, all the source codes for the platform and the apps are available for download on https://bitbucket.org/IES Cities/.

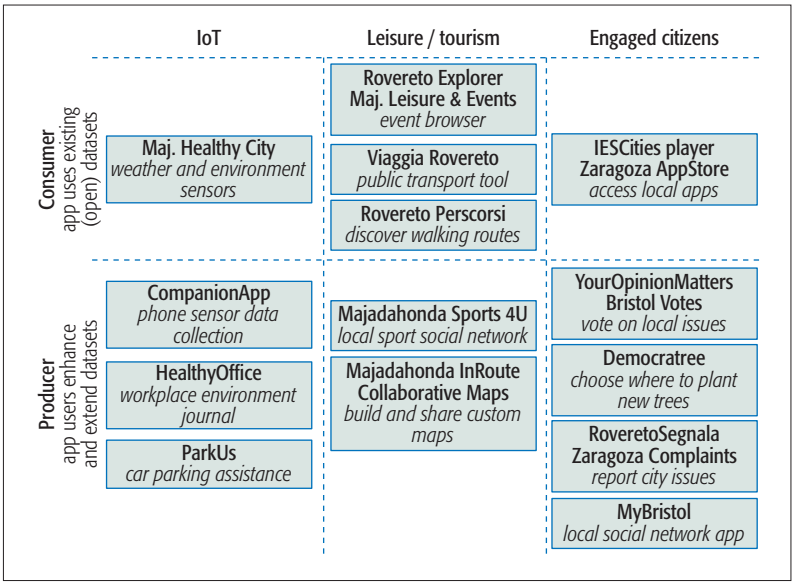


Figure 1. IES Cities app ecosystem.

What Is a Citizen-Centric IoT Platform?

A citizen-centric platform should display a range of features, such as the following, to be attractive to its intended audience:

- Free to use, with open source license
- Trustworthy, with respect to user and data privacy
- Give users full control of their data
- Easy integration of IoT devices
- Low entry barrier, with smooth learning curve
- Supported with complete and clear documentation
- Promote and include open data
- Easy integration of heterogeneous datasets
- Promote citizen-centric, socially aware services
- Provide free and easy-to-use templates
- Quick deployment of cross-platform apps
- User friendly with easy service discovery
- Easy and quick to extend, deploy, and maintain

The IES Cities platform achieves all these requirements by offering a trustworthy and secure solution with a low barrier to entry. It is open source and free for all, and it offers rich documentation, along with free (and potentially paid for) support plans. It can be deployed easily using a Docker™ container and is lightweight enough to run a basic deployment on any modern personal computer; we were even able to run it on a Raspberry Pi™ with acceptable performance. It respects user and data privacy and gives users complete control over their data. It includes an interactive and easy-to-use service discovery that is available on https://cityiot.co.uk/iescities/swagger, along with free app templates that can accelerate the creation of new services. All the socially aware, cross-platform apps, which were created by the project partners and are listed in Fig. 1, are available as open source code, so anyone can

	IES Cities	City Zenith 5D	Libellium Smart Cities	IBM Watson	Oracle IoT	FIWARE	OpenIoT	CitySDK	WSO2
Open source license	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes
Easy and interactive service discovery	Yes	N/A	No	Yes	No	No	No	Yes	No
REST interface and JSON data format	Yes	No	No	Yes	No	No	Yes	Yes	Yes
Rich documentation and support plans	Yes	No	No	Yes	Yes	No	No	Yes	Yes
App templates available for free	Yes	No	No	Yes	No	No	No	No	No
Socially aware, open source apps available	Yes	No	No	No	No	No	No	No	No
App marketplace	Yes	No	No	No	No	Yes	No	No	No
Cloud and local deployment available	Yes	No	N/A	Yes	No	Yes	Yes	N/A	Yes
Privacy, with granular control of user data	Yes	N/A	No	Yes	Yes	Yes	No	No	Yes
Open city datasets already included	Yes	Yes	No	No	No	No	Yes	Yes	No
Easy IoT device integration	Yes	Yes	No	Yes	Yes	No	Yes	No	Yes
Heterogeneous data sources supported	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes

Table 1. Citizen-centric platform aspects.

quickly build their own apps and services based on these. About a quarter of these apps were created based on user workshops and consultations, where users-citizens helped co-design apps for their city. It is also worth noting that the diversity of the apps developed highlights the flexibility and adaptability of the platform; some of the sensor data collecting apps even allowed us to use the platform as a research tool.

The platform's interfaces make extensive use of standardized web components, such as a RESTful API and JSON data format, that have been proven to facilitate the easy creation of cross-platform services and apps. The use of standardized components makes the consumption and generation of data through the apps very easy, as there is an easily accessible wealth of information and tools that enable the quick integration of REST APIs and JSON data without the need for elaborate and time-consuming coding solutions. Also, the availability of the IES Cities marketplace app where citizens can quickly discover or publish apps that were built for their city further reinforces the citizen-centric character of the IES Cities platform.

Additionally, the IES Cities project encouraged participating city councils to open up a large number of datasets for use by their citizens through the platform, bolstering further its citizen-centric credentials. It is worth noting that some of these datasets represent sensor data sourced from various locations around the cities; additionally, citizens can use IES Cities apps to share their own home-based sensor data on the platform and make them available for other citizens and app developers if they choose. Overall, the features and functionality of the IES Cities platform make it very well suited to IoT-type applications, both for storage and discovery of data coming from IoT-type devices, as well as for hosting IoT apps. This is showcased later where the HealthyOffice app is presented.

Table 1 presents a quick comparison of the IES Cities platform against a range of similar platforms. Although several platforms offer functionality similar to the IES Cities platform, none of them offer the same combination of citizen-centric and IoT characteristics. Most of these platforms target either councils or businesses, or are hardware-based platforms that only focus on IoT. Additionally, these platforms are often complex systems aimed at the most technically minded users or IT departments; their deployment and use is invariably a nontrivial technical task. In contrast to this, IES Cities is a user-friendly, people-centric platform designed with the end users, citizens, and developers in mind, in order to facilitate the quick creation and hosting of smart city and IoT apps and data.

Privacy and Control of User Data

The IES Cities platform respects user privacy in many ways; for example, it allows the registration and use of externally hosted datasets. A locally hosted database in a home server can be registered with the IES Cities cloud platform and its data accessed through it with very granular access control. The user can choose who will have access to this dataset and even create their own access policy. But even when a user dataset is hosted on the cloud platform, there is still the opportunity to set granular access policy to it. Therefore, it is possible that user generated data can always remain under the control of the individual or company providing it. This can be of great significance when potentially sensitive data is involved, such as data from personal fitness trackers. Additionally, this offers public authorities the flexibility to open up and share their data sources but still keep it on site and under their complete control while providing access to it through the platform. Also, it is worth mentioning that the platform itself is easy to deploy and manage locally, therefore providing complete control of its entire operation to the individual or organisation managing it. And of course, all information including user accounts is stored securely on the cloud platform, following all the industry standard

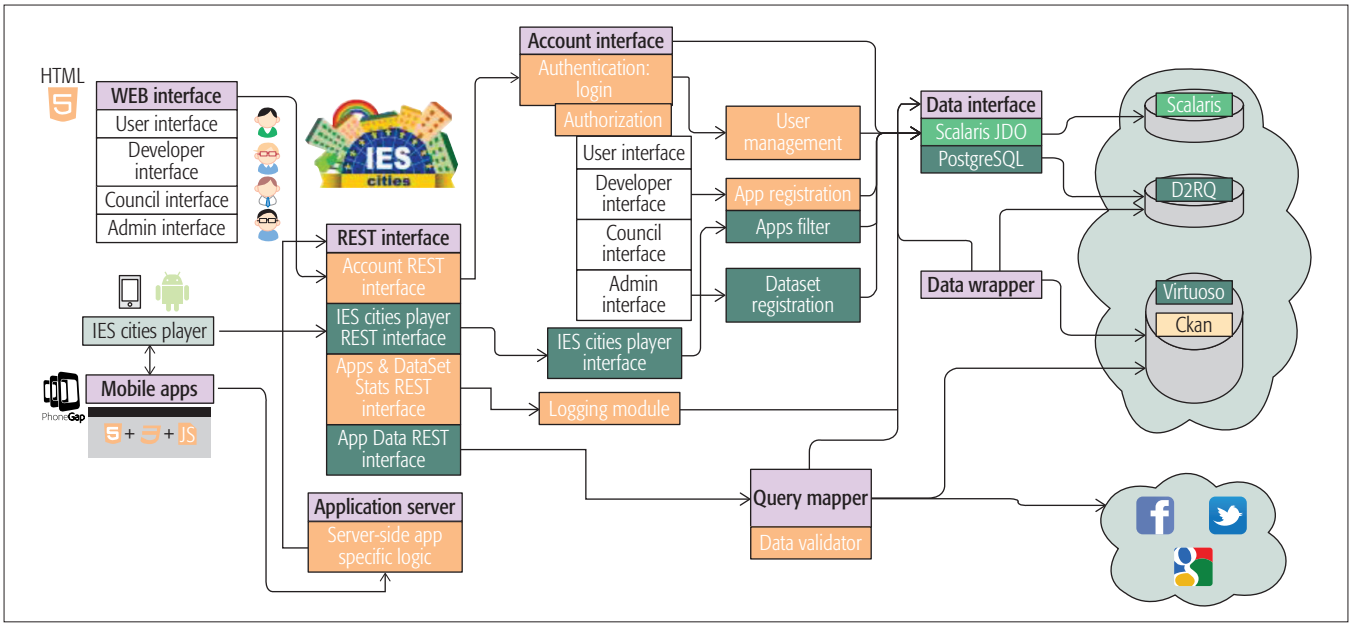


Figure 2. IES Cities platform architecture.

best practices. Finally, the adaptability of the platform allows organisations or individuals using the platform to use or integrate their own and existing user management tools that they have already developed and trust.

Design and Architecture

The architecture of the IES Cities platform is illustrated in Fig. 2 [9]; the overall structure of the platform can be viewed as a three-tier system consisting of the *client layer*, the *business layer*, and the *data layer*.

The client layer consists of the *web interface*, available on https://www.cityiot.co.uk/iescities, and the IES Cities Player marketplace app. The web interface offers a management interface for different types of users, such as end user, developer, or city council administrator. For example, it offers users the opportunity to browse and download local urban apps that use the platform or view open datasets that have been made available by the city councils or individual users. The IES Cities Player is a mobile app that functions as a marketplace for apps that are IES Cities compatible. Users can search apps by name or simply discover apps of local interest to them.

The business layer manages all the main entities of the IES Cities ecosystem: user authentication; app and dataset registration and management; event, performance, and statistics logging; as well as data querying and social media integration. All the functionalities of the platform are easily accessible through a RESTful [10] API, which groups operations into three main categories: the *account interface*, which offers Create, Read, Update, and Delete operations for all the entities handled by the platform (i.e., users, councils, apps, datasets); the *logging interface* which handles operations relevant to event, performance, and statistics logging and displaying; and finally, the *QueryMapper* [9] that offers methods to query, insert, and update data to any supported dataset through SQL commands.

The *data layer* provides read and write access to a wide range of heterogeneous datasets. A large number of different types of datasets are supported by the platform: SQL, MySQL, PostgreSQL, RDF, CSV, XML, JSON, SPARQL, even unstructured data such as Twitter and Facebook content, and also remote external data repositories such as CKAN and Socrata™. Data can be hosted either on the platform itself or in externally hosted datasets. The elegance of the data layer is that all these diverse sources of data can be queried in a unified manner, and the response that will be generated will again be formatted in a unified manner expressed in the popular JSON format, which can easily be consumed by apps.

The majority of the code for the platform has been written in JAVA™ using several relevant frameworks, such as Jersey, which facilitates the creation of RESTful APIs, and DataNucleus for data persistence. It is also worth noting that the back-end of the platform can be deployed either centrally, using a single postgreSQL node or in a distributed way, using multiple Scalaris [11] nodes. On both types of deployments, the platform displays robust and fast performance; our stress loading tests showed that a single node deployment can handle up to 2600 requests per second with an average latency of 346 ms, while the distributed deployment with 16 Scalaris nodes can handle more than 23,000 requests per second with 33 ms latency on average. More information on the IES Cities platform can be found in [9, 12], where the main focus is on the concepts of open, linked, and user-generated data; in this article we focus mostly on the experience of using the platform to develop people-oriented apps.

Developing Apps on the IES Cities Platform

In this section we discuss the development, deployment, and use of mobile apps on the IES Cities platform, as we experienced these throughout our three-year involvement with the IES Cities project.

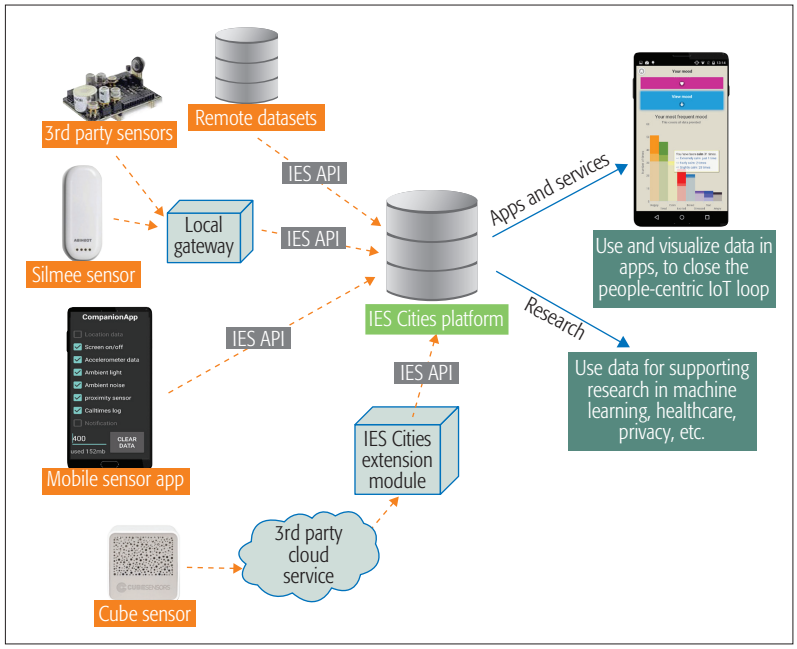


Figure 3. A representation of the IES Cities IoT ecosystem.

Mobile App Development

The platform's client, business, and data layers allow developers to easily interact with the platform via a RESTful API with simple HTTP method calls and facilitate the quick deployment of apps that use the IES Cities platform as their back-end service. As a showcase for the functionalities offered by the platform, 16 mobile apps have been developed during the IES Cities project and are available on Google Play Store; these apps range from IoT apps for sensor data collection, processing, and visualization, to socially aware apps that allow citizens to suggest locations for tree planting or to identify and report issues to the local government. A taxonomy of these 16 apps is presented in Fig. 1, where the apps are categorized according to their function and whether they primarily produce or consume data. This wide range of different apps across three countries and four cities showcases the platform's adaptability and integration capabilities, as it is able to support a wide range of requirements from apps across several domains.

Benefits for Apps Using the Platform

The benefits for developers using the IES Cities platform are outlined below:

•There is easy interaction with the platform through a universal RESTful API; additionally, easy querying of data sources through SQL queries.

•There are a wide range of available functionalities that reduce developer workload: user management, data storage, integration of diverse datasets in simple JSON format, social network integration. Instead of building these functionalities on their own, developers can simply focus on the business logic of their app.

•App templates are freely available that can significantly accelerate app development. Developers do not have to build apps from the ground app; they can quickly adapt existing app templates to suit their needs, or even reuse part of the source code of the open source apps that were developed during this project.

•Access to a wide range of published datasets from participating cities' councils; these datasets can be included in their apps to provide a richer experience to users.

•Access to a large number of existing IES Cities users who are already familiar with the IES Cities "brand." Therefore, developers can get large exposure for their app without cost in places such as the IES Cities player or IES Cities website.

•A single place for unified storage, sharing, and access to a wide range of diverse data sources. Developers can consolidate all their data in a single place, whether it is user generated data, open data, or sensor data.

•It is free to use and open source, so there is no cost to develop IES Cities compatible apps; also, they can build their own version of the platform if so required. The platform also comes with rich documentation to help developers quickly start developing IES Cities apps.

Lessons Learned

Our experience with the IES Cities platform has highlighted a few interesting points. The platform itself has proven to be reliable with good performance, and this allowed the apps to run trouble-free with a large number of users. The intuitiveness, ease of use, and provision of a wide range of functionalities and support material played a key role in minimizing development and deployment time and cost; our experience and estimations show that more than 40 percent decrease is possible due to the provided templates and open source code. Naturally, this saving becomes even larger when compared to building an app system from scratch without having access to any of the functionalities offered by this platform. Furthermore, we were able to port applications between cities and platforms with ease, thus further confirming the benefits of using the IES Cities solution. The successful development and deployment of mobile apps across different domains and cities proved the adaptability of the platform and its integration capabilities. We also learned that involving users in the design of the applications and providing intuitive and user-friendly cross-platform apps with reliable performance will ensure that users embrace them. Finally, it is worth noting that a clear and intuitive design that produces a working, reliable solution with a low barrier to entry can often be more valuable in the people-oriented IoT world than a complex solution that tries to do everything.

IES Cities IoT Ecosystem

We further discuss the potential of IES-Cities to integrate and facilitate a range of IoT networking, intelligent, and closed-loop mobile applications with efficient data fusion visualizations. These features are highlighted through the HealthyOffice and Companion mobile apps, presented below. Also, Fig. 3 illustrates a typical representation of the whole IES Cities IoT ecosystem, from data collection to sensor integration and data output and visualization.

The HealthyOffice App

HealthyOffice is an IES Cities-compatible IoT application designed to utilize user input, as well as data from wearable and environmental sensors.

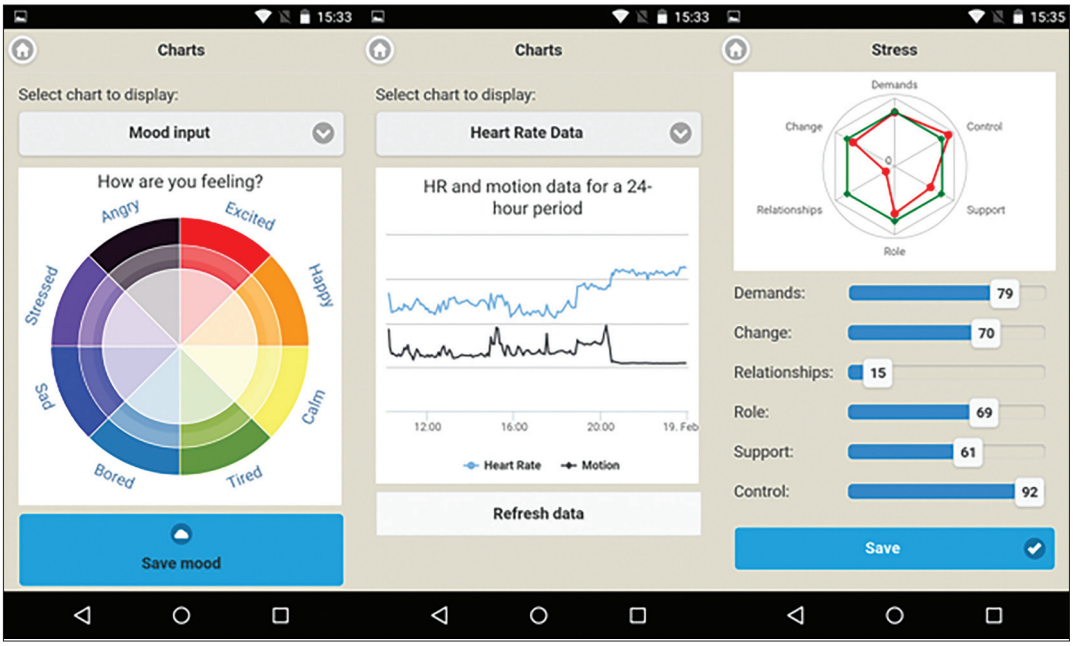


Figure 4. HealthyOffice mood input (left), Heart Rate data output (middle), stress indicators input (right).

HealthyOffice integrates machine learning modeling for the task of predicting a person's mood and analyzing well-being indicators related to their work environment.

The motivation underpinning the HeathyOffice app underlines IoT challenges ranging from:

- The secure internetworking of multimodal private sensor streams of high data rates, public ambient sensor data, or other event-driven inputs
- Coherent resolution for backend application analytics
- Effective and energy-efficient visualization of results
- A cascade architecture enabling third-party application development

The example of HeathyOffice meets these challenges, and, most importantly, it provides an API for personalized effective output that helps users visualize and manage a range of mood-causing patterns and stress levels at work.

HealthyOffice data collection includes self-reported mood input, workplace-related stress indicators [13], as well as data from wearable and environmental sensors; a few sample screenshots from the app are illustrated in Fig. 4.

The HealthyOffice app supports a range of wearables with standardized Bluetooth LE or ANT+ connectivity. At the current stage the app integrates the prototype Toshiba Silmee™ physiological multi-sensor, which includes an electrocardiogram (ECG) sensor, blood pressure from a photoplethysmogram (PPG) sensor, heart and pulse rate calculations, skin temperature, and 3-axial acceleration. Further, the app integrates the environmental Cubesensors™, which measure temperature, light, humidity, noise, and air quality. In future versions the Toshiba Wristband W20/W21 is planned to be integrated, as most users have indicated that prefer a wrist-worn device instead of the Silmee wearable, which has to be mounted on the chest using a special gel pad. The aforementioned sensors are illustrated in Fig. 5; the integration of sensor data into the app is optional.

Figure 5. Toshiba Silmee wearable (left), Toshiba Silmee Wristband (middle), and Cube sensor (right).

HealthyOffice can be used by individual users to track their own mood and well being at work. It can be used in a preventative way, to highlight potential stress factors and support staff to be proactive with managing their well being; and all data input by employees is anonymous and confidentially recorded. HealthyOffice can be used by organizations to monitor the working environment and the mood of their staff. Correlations can be drawn between staff mood and environmental stress factors, showing organizations what they do well and what they need to improve on in order to maintain good well being at work.

The Smartphone as a Sensor

CompanionApp is an IoT-type application that complements the HealthyOffice app. It is built to utilize the most versatile and widely available multi-sensor platform today: the smartphone. CompanionApp operates non-intrusively as a background service and continuously records sensor information available on the phone, includ-

HealthyOffice can be used by organizations to monitor the working environment and the mood of their staff. Correlations can be drawn between staff mood and environmental stress factors, showing organizations what they do well and what they need to improve on in order to maintain good well-being at work.

ing: location, acceleration, speed, ambient noise, ambient light, screen light, and activity recognition. It can also collect event-driven information such as the number and length of phone calls and messages. The user can choose what modalities to record and when, at which frequencies, and how to upload the data to the back-end through the IES Cities cloud platform. The app additionally offers auto-calibration of sampling thresholds and sampling rates for different sensors, so as to minimize impact on CPU usage, battery life, and data traffic, and offers compressed sensing in a data lossless manner. Our measurements show that the app generates less than 25 MB of data per day, and also that the impact on battery life is approximately 10 percent of the typical smartphone energy usage.

Data Processing and Machine Learning Approach

All the IoT data collected by HealthyOffice is uploaded to an external MySQL database for data fusion processing. In the simplest scenario, historical data and relevant mood and occupational stress trends can be visualized within the HealthyOffice app. Perhaps more interestingly, we have developed machine learning models for mood recognition that learn from and/or output user mood in five intensity levels and eight mood categories.

The machine learning process comprises four steps:

1. Preprocessing or available sensor data (noise/outlier mitigation, filtering) and data segmentation (mood prediction resolution and windowing).
2. Feature extraction (representation variables that are significant for the task, fused from available signals such as heart rate variability, acceleration, and body temperature.
3. Model training of supervised learning algorithms, based on the available mood input data (ground truth) that have been logged through the HealthyOffice application. This process is capable of building both personalized models and generalized models (by fusing all user data) as appropriate (e.g., according to the number of available instances). A model fusion process is also used to facilitate best-performing model selection, following cross-validation processes.
4. Mood prediction, where the best-performing model is used by the back-end system to record both personalized predictions as well as anonymized aggregations and statistics preparation. The IES Cities access controls and user privacy settings can be configured to allow reporting to employers or managers as appropriate.

We note that parts of the raw data processing and mood recognition model may be implemented in the back-end or the smartphone in a distributed fashion. HealthyOffice is currently available on Google Play, and more information about this work can be found in [14].

Conclusions

Currently, there is an unfulfilled need in the IoT world for open, adaptable, and easy-to-use platform solutions that will enable the building of people-oriented applications across a wide range of smart environments. Our experience with the new IES Cities platform has shown that this cloud-based, citizencentric IoT platform is a big step in the right direction. It is an open, adaptable, and already deployed system that can support and work equally well with a diverse range of people-oriented applications across different domains, such as smart cities, IoT, e-health, and even research. Its extensive integration capabilities allow for easy connection of IoT devices, as well as the inclusion of a wide gamut of heterogeneous data sources. It has proven to be reliable, with high performance in multiuser, multi-application environments. This has all allowed us to build and successfully deploy, to thousands of users across three countries, a wide range of cross-platform mobile and web apps that fuse together user generated data, open datasets, and data from wearable and environmental sensors. Through HealthyOffice we also used IES Cities as a research platform, supporting our research on machine learning algorithms for mood recognition and parking detection. Also worth noting is the intuitiveness of the IES Cities system that, along with the provided documentation and free app templates and examples, provide a low entry barrier and allow for rapid development and deployment of applications. In our experience this can mean a more than 40 percent decrease in development time and cost compared to solutions that do not offer the same level of support and user friendliness. Last but not least, IES Cities has proven to be a trustworthy platform, by providing options for fine-grained control of user data and for managing user privacy.

References

[1] Cisco, "Cisco Global Cloud Index: Forecast and Methodology, 2014–2019 White Paper," Cisco, tech. rep., 2016, http://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/Cloud Index White Paper.pdf, accessed Nov. 2016.

[2] FIWARE, "FiWARE Core Platform of the Future Internet," 2016, https://www.fiware.org, accessed Nov. 2016.

[3] OpenIoT, "Open Internet of Things," 2015, http://www.openiot.eu/, accessed Nov. 2016.

[4] J. Mineraud *et al.*, "A Gap Analysis of Internet-of-Things Platforms," *Computer Commun.*, vol. 89–90, Sept. 2016, pp. 5–16.

[5] Ayla, "How to Select the Right IoT Platform," Ayla Networks, tech. rep., 2014, http://info.aylanetworks.com/how-toselect-the-right-iot-platform, accessed Nov. 2016.

[6] CitySDK, "CitySDK," http://www.citysdk.eu/, accessed Nov. 2016.

[7] CityZenith, "City Zenith 5D," http://cityzenith.com/, accessed Nov. 2016.

[8] IESCities, "Internet Enabled Services for the Cities Across Europe," 2016, http://www.iescities.eu, accessed Nov. 2016.

[9] D. López-de Ipiñ, U. Aguilera, and J. Pérez, "Collaboration-Centred Cities through Urban Apps Based on Open and User-Generated Data," *Ubiquitous Computing and Ambient Intelligence. Sensing, Processing, and Using Environmental Information: 9th Int'l. Conf.*, Puerto Varas, Chile, Dec. 1–4, 2015, pp. 193–204.

[10] R. T. Fielding, *Architectural Styles and the Design of Network-Based Software Architectures*, Ph.D. dissertation, 2000, http://www.ics.uci.edu/ fielding/pubs/dissertation/top.htm, accessed Nov. 2016.

[11] T. Schütt, F. Schintke, and A. Reinefeld, "Scalaris," *Proc. 7th ACM SIGPLAN Wksp. ERLANG ERLANG '08*, ACM Press, 2008, p. 41, http://portal.acm.org/citation.cfm?-doid=1411273.1411280, accessed Nov. 2016.

[12] D. López-de Ipiñ *et al.*, "Citizen-Centric Linked Data Apps for Smart Cities," *Ubiquitous Computing and Ambient Intelligence. Context-Awareness and ContextDriven Interaction: 7th Int'l. Conf.*, Carrillo, Costa Rica, Dec. 2–6, 2013, pp. 70–77.

[13] HSE, "How to Tackle Work-Related Stress," tech. rep., 2009, http://www.hse.gov.uk/stress/information.htm, accessed Nov. 2016.
[14] A. Zenonos *et al.*, "HealthyOffice: Mood Recognition at work Using Smartphones and Wearable Sensors," *2016 IEEE Int'l. Conf. Pervasive Computing and Communication Wksps.*, 2016, pp. 1–6.

Biographies

Stefanos Vatsikas (stefanos.vatsikas@toshiba-trel.com) received his M.Eng. from Aristotle University, Thessaloniki, Greece. He is currently a research engineer at Toshiba TRL, where he has worked since receiving his Ph.D. from the University of Bristol, United Kingdom, in 2012. His research interests are in the areas of game theory and resource allocation in wireless networks. He also has extensive experience in developing software for mobile and connected devices.

Tim Lewis (timlewis@acm.org) is a research fellow at Toshiba TRL, where he has worked in the area of wireless network protocols since receiving a Ph.D. from the University of Edinburgh. He has published in the research areas of parallel compilers, evolutionary optimization, protocol optimization, smart grid communications, and smart city service platforms, holding patents in many of those areas.

Georgios Kalogridis (george@toshiba-trel.com) received his Diploma in electrical and communications engineering from the University of Patras, Greece, in 2000; his M.Sc. in computer science from the University of Bristol in 2001; and his Ph.D. in mathematics from Royal Holloway, University of London in 2011. He is a principal research engineer and a team leader at Toshiba Research Europe with 14+ years' experience in R&D and EU projects. He is a well recognized expert in the areas of machine learning, data privacy informatics, and IoT. He has authored more than 40 scientific publications and invented more than 30 international patents.

Mahesh Sooriyabandara [SM] (mahesh@toshiba-trel.com) received his B.Sc.Eng. (Hons) from the University of Peradeniya, Sri Lanka, and his Ph.D. from the University of Aberdeen, United Kingdom. In 2004, he joined the Telecommunications Research Laboratory of Toshiba Research Europe, where he is currently associate managing director. His current research interests include wireless networks, smart grid communications, and the Internet of Things. He is a Senior Member of the ACM and a Fellow of the IET.

Exploiting Density to Track Human Behavior in Crowded Environments

Claudio Martella, Marco Cattani, and Maarten van Steen

The authors present the design, implementation, and deployment of a positioning system based on mobile and fixed inexpensive proximity sensors that they use to track when individuals are close to an instrumented object or placed at certain points of interest. To overcome loss of data between mobile and fixed sensors due to crowd density, traditional approaches are extended with mobile-to-mobile proximity information.

ABSTRACT

For the Internet of Things to be people-centered, *things* need to identify when people and their things are nearby. In this article, we present the design, implementation, and deployment of a positioning system based on mobile and fixed inexpensive proximity sensors that we use to track when individuals are close to an instrumented object or placed at certain points of interest. To overcome loss of data between mobile and fixed sensors due to crowd density, traditional approaches are extended with mobile-to-mobile proximity information. We tested our system in a museum crowded with thousands of visitors, showing that measurement accuracy increases in the presence of more individuals wearing a proximity sensor. Furthermore, we show that density information can be leveraged to study the behavior of the visitors, for example, to track the popularity of points of interest, and the flow and distribution of visitors across floors.

INTRODUCTION

Measuring and tracking the behavior of individuals is central to the implementation of an Internet of Things (IoT) that is people-centered. To this end, it must be possible to identify when an individual is positioned in proximity to an object, at a point of interest (PoI), or in front of another person [1].

We analyze the use case of a museum. In a museum, people proximity-aware applications could enable museum staff to make sure that visitors can approach all exhibits and information without congestion and clogging, that they have access to all resources such as restaurants, restrooms, and lockers when needed, and that they do not experience queues and waits that are too long, as well as to understand at which artworks individuals stop.

A common approach to monitoring visitor behavior is to compute the *positioning* of visitors at exhibits and PoIs [2]. Here, positioning is *relative* to a point of interest, and not absolute in the coordinate space (i.e., as provided by indoor localization systems). However, as crowd density increases, sensors are known to provide more noisy, incomplete, and ambiguous data, problems that are exacerbated by complex indoor venues. Museum staff are thus left with instruments that operate unreliably in those conditions where they are most needed.

We present a mobile-nodes-assisted positioning system (MONA) that can operate in conditions of high crowd density by utilizing proximity sensing between mobile sensor nodes as well as between mobile and fixed sensor nodes. Our approach exploits the increased presence of mobile sensor nodes in the surroundings of each individual. In fact, positioning accuracy *increases* when more visitors wearing a sensor are present in the museum. MONA can position a visitor when mobile-to-anchor proximity detections are missing and even at points of interest not instrumented with anchors.

OVERVIEW

MUSEUM

The museum we study has half a million visitors per year and more than 3000 daily visitors during our experiment (we chose the day before Christmas). With a median visit duration of three hours, the museum can present crowded scenarios with peaks of nearly 3000 people visiting at the same time. Approximately 25 percent of all visitors participated in our experiment, with a peak of around 600 participants at a single moment in time.

Our museum is an open six-story area. The first four stories share a large hall in the middle connected by several stairs, with floors 3 and 4 being effectively balconies projected over the underlying stories. With this layout, network links spanning over multiple floors are common. Figure 1 shows a 3D map of the museum. Exhibits can comprise multiple items within a radius of approximately 3–4 m.

The museum curators were interested in three types of information:

- The popularity of a set of points of interest and exhibits
- The distribution of visitors across the floors
- The flows of people between floors

SENSING INFRASTRUCTURE

To collect positioning information, visitors were asked to wear a *bracelet* equipped with a 2.4 GHz transceiver and a microcontroller running a neighbor discovery protocol with a sampling rate of 1 Hz. Note that we used bracelet devices as they were the most suitable solution to our prototype environment and setup, as opposed to mobile phones, but nothing of the technique presented in this work is bound to this particular device and, in principle, it could be implemented with Bluetooth Low Energy (BLE) on mobile phones.

Digital Object Identifier:
10.1109/MCOM.2017.1600672CM

Claudio Martella is with VU University Amsterdam; Marco Cattani is with Delft University of Technology; Maarten van Steen is with the University of Twente.

Anchors: For each PoI, an anchor device was installed. These devices had the same hardware as the bracelets. They were externally powered so that it was possible to run the neighbor discovery algorithm with longer duty cycles, increasing the probability of a PoI bing discovered by bracelets. As a result we were able to discover bracelets in the range of a PoI in just a few seconds.

Every second, bracelets could receive broadcasts from anchors or other bracelets within a distance of some 5–7 m. Such broadcasts contain the unique identifier (ID) of the sender, and we consider the reception of such broadcast as proximity detection between the two nodes involved. Bracelets recorded together with the ID of the other node also signal strength information about the broadcast.

Sniffers: Once every second, bracelets reported the proximity detections collected during the previous second via a special packet sent on a dedicated channel to the backbone of the system consisting of sniffers. Like anchors, sniffers used the same hardware as bracelets in addition to a single-board computer that used either WiFi or Ethernet to commit the packet to a central database hosted on a server. The sniffers were placed uniformly in the museum to cover all areas, with some overlap.

Due to packet size limitations at the medium access control (MAC) layer, each sniffed packet could report up to three detections, favoring two anchors and one bracelet when possible. The resulting dataset contained more than 6 million anchor-to-bracelet and bracelet-to-bracelet detections, together with timestamp and signal strength information.

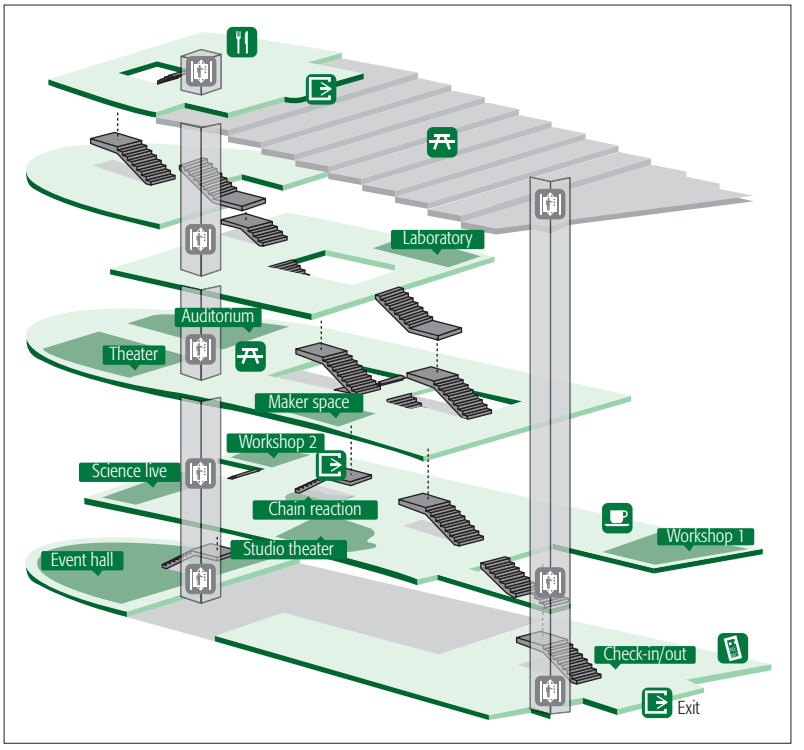


Figure 1. The map of the museum.

Related Work

There is a large body of work regarding indoor localization and positioning with mobile sensors.

In the case of museums, earlier works focus on localizing visitors at very coarse-grained levels (room level) through technologies like Bluetooth [3] to support multimedia guides [4]. More recently, proximity sensors have been used together with "physiological" sensors to classify the behavior of visitors [5].

Computer-vision techniques are an alternative approach to tracking human mobility [6] and detecting anomalies in a crowd [7]. However, cameras can suffer from poor lighting, and temporary or permanent obstructions [8]. Also, fusing the views from multiple cameras in a highly dynamic indoor scenario is challenging [9]. Finally, the privacy issue of collecting large-scale footage of visitors often restricts researchers from accessing these sources of data.

Alternatively, we can localize the absolute position of an individual. While this technique is a feasible option in outdoor scenarios — where the GPS system can be exploited — for indoor conditions, accurate localization is still an open problem.

Among the wide literature on radio-based localization techniques, only a few [10–12] are accurate enough to be employed in museum scenarios. Unfortunately, none of these techniques perform consistently throughout the museum areas as localization error increases significantly at the edges of rooms and in hallways. By installing anchors specifically at exhibits and objects of interest, we limit this problem drastically. For our tracking application, this phenomenon can be even more problematic, since even small estimation errors could lead to visitors being associated with the wrong exhibit, positioned in the wrong room, or placed on the wrong floor.

We borrow some ideas from the field of cooperative localization [13], in which nodes share measurement information in a peer-to-peer manner. We trade the advantages of deploying algorithms that can be computed in a distributed manner by the nodes "in the network" with the more global view provided by collecting and managing this data from a central repository. This aspect is particularly useful in high-density scenarios, where packet loss increases, potentially hindering cooperative approaches.

Finally, many of these approaches are usually tested outside of the extreme conditions of high mobility and density of a complex real-world museum as the one subject of our study. For a more detailed discussion of related work, also in the specific context of museums, we refer the interested reader to [2].

Model

We consider N visitors $V = \{v_1, v_2, ..., v_N\}$. While bracelets have unique IDs and were reused during the experiment, each element in V has a unique ID assigned at check-in, and we use these IDs for detection. The museum has O POIs $I = \{i_1, i_2, ..., i_O\}$ and M anchors $A = \{a_1, a_2, ..., a_M\}$. In the simple case where each anchor is assigned to a PoI, $I \equiv A$. We consider S the set of proximity sensors $S = V \cup A = \{s_1, s_2, ..., s_{N+M}\}$, as the union of detectable anchors and visitors. We define the series of proximity detections for a visitor v as an S × T matrix D_v, where $D_v(i, j) = r$ if the proximity

We designed and implemented a particle filter that takes into account the topology of the museum, the location of anchors and PoIs, and the estimated location and movement of the visitors. Particle filters have been used in indoor localization to estimate the absolute position of individuals with unreliable sensors.

sensor of visitor v detected sensor s_i at time j with signal strength r. Note that D_v has $N + M$ rows, as it comprises all proximity sensors, either used as anchors or worn by other visitors. $D_v(*, t)$ refers to all detections collected at any time t, and $D_v(i, *)$ to all detections of sensor s_i. Moreover, we define a positioning matrix M_v as the $O \times T$ matrix where $M_v(i, j) = 1$ if the visitor was at a distance shorter than d from PoI i at time j. Note that there can be times where a visitor is not positioned at any PoI, and a visitor cannot be positioned at more than one PoI at the same time (when a visitor is within d distance from multiple PoIs, we choose the closest).

Our goal is to produce the positioning matrix M_v starting from the series of detections in D_v. In principle, in the case of I ≡ A and with perfectly working sensors (i.e., with signal strength correctly mapping to distance and no inter-floor detections) we could use D_v to directly generate M_v by assigning a visitor to the anchor detected with highest signal strength, and computed to be at distance shorter than d. As the visitor walks around the museum, we would generate contiguous series of "bits" in M_v reflecting the intervals of proximity with the anchors and PoIs. However, in real-world scenarios we need to take into account *missed* as well as *spurious* detections, caused, for example, by interferences, walls, people, and the nodes themselves.

Particle Filter

We designed and implemented a particle filter that takes into account the topology of the museum, the location of anchors and PoIs, and the estimated location and movement of the visitors. Particle filters have been used in indoor localization to estimate the absolute position of individuals with unreliable sensors [14, 15]. For localization, usually a mobile sensor communicates with a few anchors installed at known locations. It is assumed that the sensor can communicate with all, or a majority of, anchors from all positions and directions, and that the sensor can measure distance from these anchors, for example, through signal strength or time of flight. Our setup is more complex, as we have a larger number of anchors that are detectable at shorter distance, and a number of detections from proximity sensors at unknown locations (i.e., the visitors), together with a multi-story venue.

The filter requires topology information about the museum, such as the sets of anchors A and PoIs I, each defined by a triple f, x, y with f being the floor number and x, y the coordinate within that floor, and a set of walls $W = \{w_1, w_2, ..., w_M\}$, each defined as a segment between two points. Every floor is modeled as a distinct two-dimensional space, with an independent origin. As floors are connected by a number of different staircases and elevators, it is difficult to model the transition spaces between floors reliably. Instead, we assume a visitor can "appear" on a certain floor and model this transition confidence in the filter based on the measurement. For each visitor we define a set of particles $P = \{p_1, p_2, ..., p_K\}$, each defined by a tuple f, x, y and a weight w that models the likelihood of the visitor to be at that coordinate. Initially, particles are spread uniformly at random across the museum floors.

For each time of the day $0 \le t < T$, the following four steps are executed for each visitor checked in at that time, given the respective detection matrix D_v.

Estimation: At each time, we estimate the floor on which the visitor is positioned, if any, by computing the floor with the largest number of particles, given enough confidence is provided by the particles. Then we compute the likelihood of each particle's estimate (i.e., its position) given the measurement at time t, that is, the set of detections in D_v contained in the tth column. For each particle p, the weight is updated using the likelihood function $\Phi(p, D_v(*, t))$. We return to details below.

Positioning: We estimate the position of v by computing the weighted average among the floor's particles (i.e., *the centroid*) and find the closest PoI i_j on the floor within distance d. We then set $M_v(i, t) = 1$, unless the confidence of the estimate is smaller than a threshold δ.

Re-sampling: We create a new set of particles by drawing with replacement from the current weighted set of particles. While drawing particles from the set, we favor particles proportionally to their weight (i.e., their likelihood). As a result, particles with higher likelihood are picked more often than particles with lower likelihood. Depending on the confidence in the current set of particles, we may choose to:

- Pick only from the particles in the current floor
- Additionally spread particles with smaller likelihood across other floors
- Re-distribute all particles across all floors (i.e., when we believe we have lost track of the visitor). More details follow below.

Movement: We move particles at walking speed in random directions, avoiding illegal moves, such as walking through walls.

Compared to previous work on positioning [2], the two steps where we focused most of our efforts to tailor the filter for our use-case are the estimation and the re-sampling steps. Those steps were redesigned to:

- Include mobile-to-mobile detections
- Consider multi-floor environments
- Support PoIs without anchors and anchors not associated with PoIs
- Omnidirectional sensing with signal strength

In general, we can compute the confidence of the particles' centroid by measuring the dispersion of the particles and the amount of time spent without collecting any detection from a sensor located on the same floor as the visitor. We consider the visitor to be on the floor with the largest number of particles, normalized by floor sizes, given enough confidence. Note that having a visitor assigned to a floor does not mean we position the visitor at any PoI, as that depends on the position of the particles within the floor, their likelihood, and dispersion.

The likelihood function $\Phi(p, D_v(*, t))$ computes the likelihood of particle p given the set of sensors detected at time t that were positioned on the same floor as p. In Φ we consider both anchors, of which we know the location, and neighbor visitors, for which we use the centroid computed at time $t - 1$. In other words, we use neighbor visitors as anchors. We consider neighbor visitors

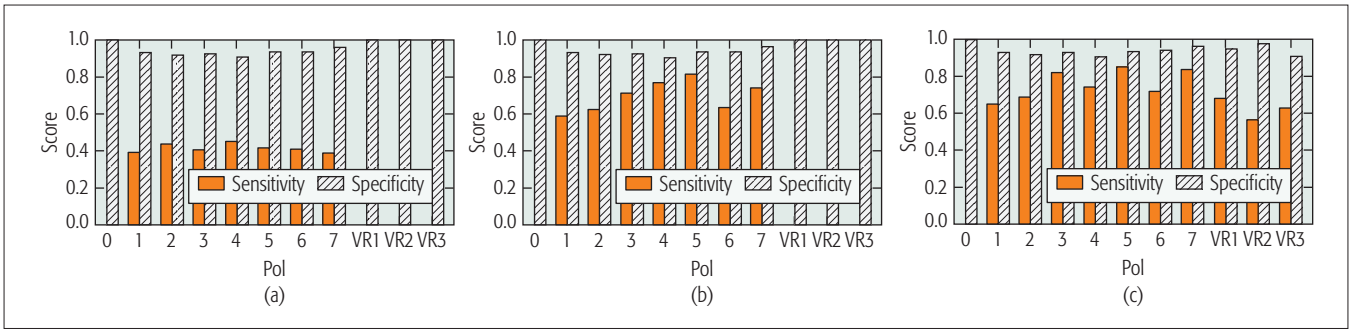


Figure 2. Positioning accuracy during the controlled experiment at the PoIs on the first flloor and three additional "virtual" PoIs not instrumented with an anchor. We compare our technique (not using mobile-to-mobile proximity) to a technique that positions the visitor at the anchor with strongest signal with and without a rolling window. The combined balanced accuracy, defined as the arithmetic mean between average sensitivity and specificity, for the two competing techniques and MONA (anchors only) were 0.62, 0.69, and 0.81 respectively: a) strongest RSSI; b) smoothened strongest RSSI; c) MONA (anchors only).

only if we are confident enough of their centroid.

Intuitively, the likelihood of a particle depends on the particle's distance from a reference point proportional to the strength of the detections. This reference point can be a single sensor or an average across sensors. We first map signal strength values to the (0, 1] continuous interval by means of a Gaussian kernel.[1] Then, if multiple sensors were detected, we compute the average coordinate across these sensors' coordinates weighted by the respective signal strength, and use the particle's distance from this average coordinate. We use a linear kernel to map distances to (0, 1].

If only one sensor is detected, instead we compute the difference between the signal strength value and the particle distance from the sensor (both mapped to (0, 1]) and use this difference.

When we update particle weights and re-sample the particles, we proceed depending on the centroid confidence c as follows (the two thresholds δ and δ' to be chosen empirically):

- $c > \delta$: This means we know the visitor is on the floor. We ignore the likelihood of particles from other floors, practically "teleporting" these particles from other floors to the current one.
- $\delta' < c \leq \delta$: This means we are starting to believe the visitor may have left the floor. We spread particles with smaller likelihood from the current floor to other floors.
- $c \leq \delta'$: This means we have lost the visitor completely. We re-distribute particles uniformly at random across floors.

As a visitor moves around a floor, particles spread toward areas that are more likely to have produced the current measurement. As the visitor moves to a new floor, absence of detections on the previous floor causes particles to disperse until we start spreading particles on the other floors, including the new one.

Finally, we apply a density-based filter and a majority-voting filter to account for particles "jumping" between between PoIs and floors [2].

Evaluation: Setup

Before the main experiment, we conducted a *controlled* experiment scripting a visit of the first floor while the complete sensing infrastructure in the museum was turned on. The script defined arrival and departure times at each PoI. Originally, the first floor had all PoIs associated with an anchor, and one additional anchor to improve positioning accuracy. To test the ability to position visitors at PoIs without anchors, we added to the script three "virtual" PoIs not associated with any particular exhibit or anchor (i.e., solely defined by a coordinate in space). In addition, one PoI with an anchor was not used in the script, but the anchor was turned on and hence could affect accuracy.

Together with the scripted visit, we distributed 15 bracelets around each interested PoI area within a radius of some 15–20 m from the respective anchor (or virtual PoI coordinate) at 1 m of height, and moved them at random across the space for the whole duration of the stop. This setup allowed us to control bracelet density and movement, and at the same time minimize external factors like body shielding effects (which were tested during the real-world experiment). Each stop lasted eight minutes, divided in four periods of two minutes. During the first periods only the visitor's bracelet was on, while during the following three periods we turned on the additional bracelets, in groups of five per period.

To evaluate real-world accuracy, for the duration of the main experiment we positioned five bracelets at known locations at PoIs, and additionally scripted a visit of the whole museum of the duration of around one and a half hours with 14 stops of duration of around five minutes each (hence, not all time was spent at PoIs). Also for this scripted visit, we added a number of "virtual" PoIs such that the visitor did not stop at PoIs associated with an anchor, except for the stops at the restaurant and the live attraction called "Chain Reaction". For both experiments, we set filter parameters to the same values.

To measure the performance of our solution in the task of positioning visitors at PoIs, we compute the number of false/true positives and negatives, and with these we compute the *sensitivity* (or true positive rate, also known as recall) and *specificity* (or true negative rate, or negative class precision) for each PoI.

We implemented the filtering pipeline with less than 1000 lines of Python code. By implementing the filters with vectorized single instruction multiple data (SIMD) operations and leveraging linear algebra libraries like numpy[2] and a just-in-time compiler with numba,[3] our pipeline can manage in real time (i.e., sub-second compute time to process 1 second worth of data) peak-time data with

[1] Signal strength decreases nonlinearly with distance.

[2] http://www.numpy.org/

[3] http://numba.pydata.org/

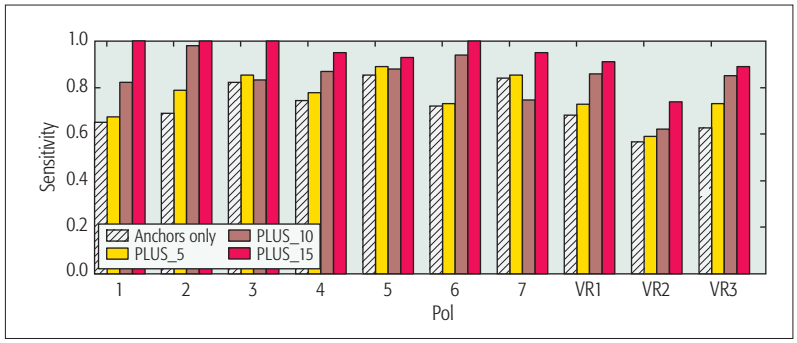


Figure 3. Impact of neighbor bracelets around a visitor on the first floor. We added five more bracelets in the surrounding of the visitor every two minutes of each stop.

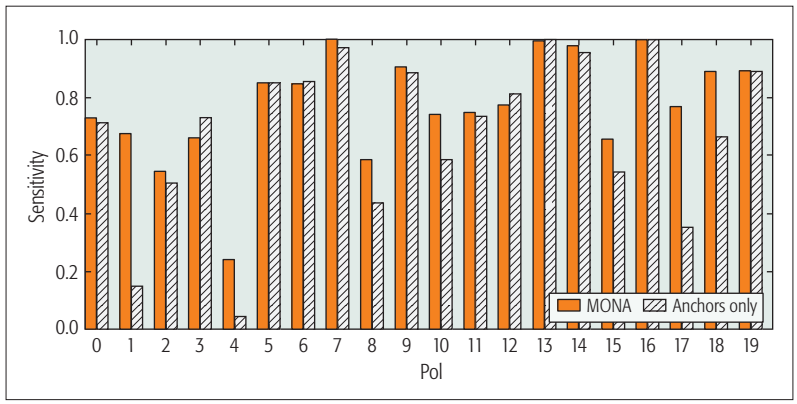


Figure 4. Positioning accuracy of our system with MONA and without it during the real-world experiment. Only PoIs 7, 9, 13, 14, 16, and 19 were instrumented with an anchor. The average sensitivity was 0:68 and 0.77, respectively.

a single thread executed on a ultra-low power laptop.[4] The pipeline can easily be ported to multicores and even GPUs for very large crowds.

Results

Controlled Experiment

We test our technique against an approach where we position the visitor at the anchor/PoI detected with strongest signal strength. As this technique is subject to missed detections and noise, we additionally extend it by making positioning decisions over a sliding window of 10 s. Figure 2 presents the sensitivity and specificity of each anchor/PoI for the two techniques as well as for MONA. Note that for this test, we compute values for MONA only for the first period, that is, we position the visitor *solely based on anchor detections*.

One can notice that raw data suffers from missed detections, yielding an average sensitivity around 40 percent of the sensitivity obtained when smoothening the decisions over a window (0.42 and 0.68). As expected, specificity is close to the maximum value of 1, as it is difficult to wrongly position a visitor at a PoI far away with a controlled transmission range. Looking at the results of MONA (anchors only), one can notice a relative improvement in average sensitivity of over 5 percent (from 0.68 std. 0.34 to 0.72 std. 0.26) compared to the smoothened technique, but the most interesting result is the impact on positioning at virtual PoIs (impossible with the other techniques). This is due to the spatial nature of the filter, and it is one of the main contributions of the filter compared to the smoothed filter.

Figure 3 presents the sensitivity results when MONA also takes into account detections of neighbor bracelets to position both the visitor and the neighbor bracelets. We do not present specificity values here as they are consistently close to 1, as in Fig. 2. One can notice that adding mobile node proximity information improves positioning ability, reaching a value of (or close to) 1 when 15 additional neighbors are used, with an average relative improvement between using only anchors and including 15 neighbors of around 30 percent (from average sensitivity of 0.72 std. 0.09 to average sensitivity of 0.94 std. 0.08). The improvement is slightly worse in the case of virtual PoIs (in particular VR2), but note that here we leveraged the anchors used for the other PoIs (except for only an additional anchor). Moreover, more data points (i.e., visits) should yield more consistent results (e.g., for PoI 7 and 3, adding 10 bracelets yields worse results than using anchors only). Finally, one would expect a slightly lower impact in the real world, where body shielding effects and other irregularities would influence mobile-to-mobile detections.

Real-World Experiment

Figure 4 presents a comparison of sensitivity values, for both the stops of the scripted visit and for the five bracelets installed at PoIs, when mobile-to-mobile detections are used and ignored for positioning. Note that 13 out of 19 PoIs are virtual; hence, they represent the hardest task for MONA (i.e., only PoIs 7, 9, 13, 14, 16, and 19 represent stops at actual PoIs instrumented with a dedicated anchor). One can notice that MONA produces a relative improvement in average sensitivity of over 13 percent (from 0.68 to 0.77), though for a few virtual PoIs the performance is lower. Again, more data points should produce more consistent results.

A main reason why we do not always see a strong impact as presented in Fig. 3 is due to the "sampling" effect of the real-world experiment. Considering that "only" 25 percent of the visitors were wearing a bracelet, the chances to leverage mobile-to-mobile proximity detections were reduced. This is aggravated by the fact that in the first and last parts of the day crowd density is lower, as fewer visitors are in the museum. Most importantly, at each second a bracelet often reports only one neighbor visitor, as bracelets favor reporting two anchors out of three detections when possible. Combined with the fact that only 25 percent of the visitors were wearing the bracelet, and all the places in the museum were not necessarily very dense at all times, we would expect a higher impact by increasing the number of reported neighbors (the current value of three was due to limitations in the packet size of the current implementation and can be increased in future versions).

Application

PoIs "Popularity" over Time

Because at each time we know which visitors are positioned at which PoI, we can estimate how many individuals are visiting a PoI. In Fig. 5a we present the number of checked-in devices during the day. One can notice that peak time

[4] Provided with a mobile 1.2Ghz Intel Core M CPU

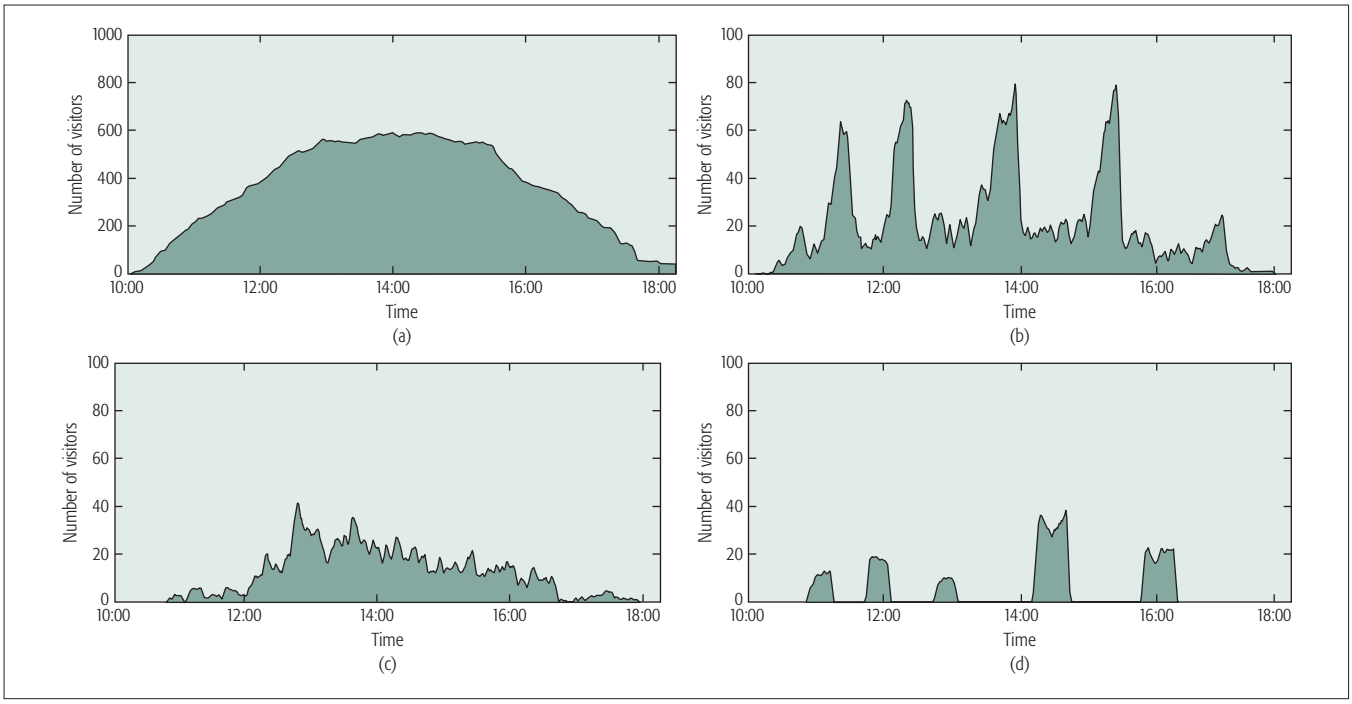


Figure 5. Number of individuals positioned at different PoIs over time: a) the number of visitors wearing a bracelet during the day (experiment participation was around 25 percent); b) the Chain Reaction PoI. Events started at 11:15 a.m., 12:15 p.m., 2:45 p.m., 3:15 p.m., and 4:45 p.m. for a duration of about 15 minutes; c) the restaurant on the top floor; d) an auditorium open to the public only on five occasions during the day.

is between 1 p.m. and 3:30 p.m., with around 600 visitors wearing a bracelet (and around 2400 individuals overall in the museum). Figure 5b shows the number of individuals spending at least two minutes at the popular Chain Reaction (CR) events throughout the day. A CR event takes place in the middle of the museum hall and is widely announced. They took place at 11:15 a.m., 12:15 p.m., 2:45 p.m., 3:15 p.m., and 4:45 p.m. for a duration of about 15 minutes. One can notice that CR events have the typical footprints of crowded events. First, in the *build-up phase*, density gradually increases during the minutes before the event, as people either stop by or approach the event location in advance. Then the event takes place, and density remains more or less constant. Finally, in the *break-up phase*, density decreases quicker, as all individuals leave the event location for another PoI.

Figure 5c shows data for the restaurant on the top floor. For this data, we aggregate positioning information for all the anchors on that floor, as the restaurant covers the whole space. Here, one can notice that lunch time peaks between 12:30 p.m. and 1 p.m., but decreases gradually, as it is used by families for breaks at the end of the visit, enjoying the view from the rooftop. Figure 5d presents data for the auditorium. The auditorium is a closed theater space that is open to the public only during scheduled events. For this reason, one can notice no visitors outside of scheduled shows and a less gradual build-up phase.

Crowd Distribution Across Floors

Figure 6 shows the number of individuals at each floor throughout the day. First, one can notice how the CR event greatly influences the distribution of visitors across the first two floors and also on the third floor. When the CR event takes place, we can see corresponding peaks on the first floor and dips on the second floor (and even smaller dips on the third).

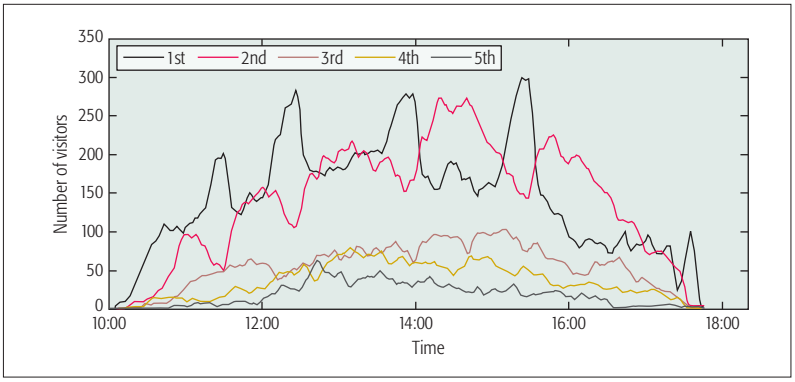


Figure 6. Distribution of visitors across floors over time.

Second, one can notice that the "popularity" peak of floors happen at different times of the day depending on the floor. The first floor hosts more people during the first half of the day, while the second floor peaks around 2 p.m. and the third floor around 3 p.m. (and the fourth floor has very few visitors at all before 12 p.m.). This is because the building is built to be visited somehow in order floor after floor, though visitors are not obliged to do so. Again, the fifth floor hosts the restaurant and has a different pattern.

Flows between Floors

Figure 7 shows the number of visitors moving per minute *from* and *to* the first floor in particular. We compute this by considering only visitors that remain on the floor for at least 10 minutes (hence filtering

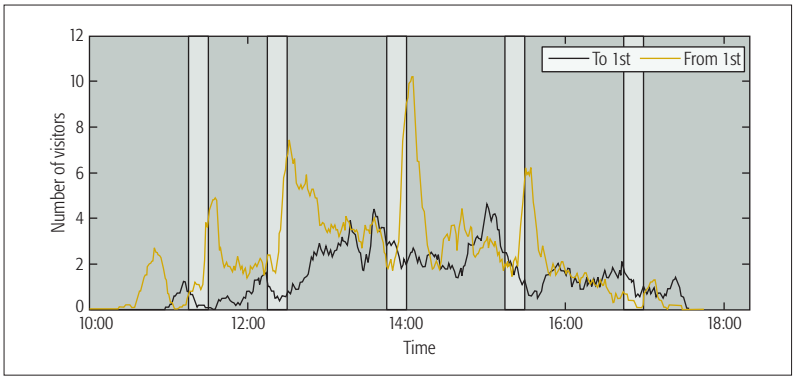


Figure 7. Flow of visitors *from* and *to* the first floor.

out visitors just passing by). One can notice again that CR events, in gray, dominate the pattern. Small peaks in the movement to the first floor appear at the minutes before CR events, while higher and more sudden peaks appear in the movement from the first floor right after the event (again, the footprints of build-up and break-up phases). As stairs to the second floor are positioned right besides the CR event location, visitors tend to move to the second floor right after the event finishes.

Conclusion

In this article, we have presented the design and evaluation of a positioning system that leverages mobile-to-mobile proximity sensing to overcome missed mobile-to-anchor detections due to high crowd density. We have shown that our approach is able to increase positioning accuracy when more visitors wearing proximity sensors are present in the museum. The museum where we conducted the study presented extreme conditions of density and challenging conditions due to a complex multi-story open space. We have tackled these challenges by tailoring a filtering pipeline to our use case. However, the approach is general and applicable to any proximity-sensing technology that is able to detect neighbor sensors with an estimate of distance.

Moreover, we have used the data to gain insights about the behavior of the visitors during the experimentation days. The insights show a clear behavioral trend, opening new questions regarding how museum staff can integrate such an approach in their work. The work we present is, however, not limited solely to the use case of a museum, but is applicable to the general problem of indoor crowd monitoring.

Acknowledgments

This research was part of Science Live, the innovative research program of Science Center NEMO that enables scientists to carry out real, publishable, peer-reviewed research using NEMO visitors as volunteers. This publication was supported by the Dutch national program COMMIT. The authors would like to thank the members of the EWiDS project and those who collaborated on this experiment, as well as the staff at the museum who provided support during the tests.

References

[1] C. Martella *et al.*, "Crowd Textures as Proximity Graphs," *IEEE Commun. Mag.*, vol. 52, no. 1, Jan. 2014, pp. 114–21.
[2] C. Martella *et al.*, "Leveraging Proximity Sensing to Mine the Behavior of Museum Visitors," *IEEE Int'l. Conf. Pervasive Computing and Commun.*, 2016.
[3] Y. Yoshimura *et al.*, "New Tools for Studying Visitor Behaviours in Museums: A Case Study at the Louvre," *Int'l. Conf. Info. and Commun. Technologies in Tourism*, 2012, pp. 391–402.
[4] E. Bruns *et al.*, "Enabling Mobile Phones to Support Large-Scale Museum Guidance," *IEEE Multimedia*, vol. 2, 2007, pp. 16–25.
[5] V. Kirchberg and M.Tröndle, "The Museum Experience: Mapping the Experience of Fine Art," *Curator: The Museum Journal*, vol. 58, no. 2, 2015, pp. 169–93.
[6] S Yaseen *et al.*, "Real-Time Crowd Density Mapping Using a Novel Sensory Fusion Model of Infrared and Visual Systems," *Safety Science*, vol. 57, 2013, pp. 313–25.
[7] N. Wijermans *et al.*, "A Landscape of Crowd-Management Support: An Integrative Approach," *Safety Science*, vol. 86, 2016, pp. 142–64.
[8] B. Zhan *et al.*, "Crowd Analysis: A Survey," *Machine Vision and Applications*, vol. 19, no. 5–6, 2008, pp. 345–57.
[9] B. Song *et al.*, , "Wide Area Tracking in Single and Multiple Views," *Visual Analysis of Humans*, Springer, 2011, pp. 91–107.
[10] R. Reimann, A. Bestmann, and M. Ernst, "Locating Technology for AAL Applications with Direction Finding and distance Measurement by Narrow Bandwidth Phase Analysis," *Evaluating AAL Systems through Competitive Benchmarking*, Springer, 2013.
[11] C. Beder and M. Klepal, "Fingerprinting Based Localisation Revisited: A Rigorous Approach for Comparing Rssi Measurements Coping with Missed Access Points and Differing Antenna Attenuations," *2012 Int'l. Conf. Indoor Positioning and Indoor Navigation*, 2012, pp. 1–7.
[12] C.-L. Li *et al.*, "Indoor Geolocation on Multi-Sensor Smartphones," *Proc. 11th ACM Annual Int'l. Conf. Mobile Systems, Applications, and Services*, 2013.
[13] M. Z. Win *et al.*, "Network Localization and Navigation via Cooperation," *IEEE Commun. Mag.*, vol. 49, no. 5, May 2011, pp. 56–62.
[14] F. Evennou, F. Marx, and E. Novakov, "Map-Aided Indoor Mobile Positioning System Using Particle Filter," *2005 IEEE Wireless Commun. and Networking Conf.*, 2005.
[15] F. Gustafsson *et al.*, "Particle Filters for Positioning, Navigation, and Tracking," *IEEE Trans. Signal Processing*, 2002.

Biographies

Claudio Martella is a Ph.D. candidate at the Large-Scale Distributed Systems group of VU University Amsterdam. His research is focused on complex networks, graph processing, and large-scale distributed systems. He is currently working on modeling collective behavior through spatio-temporal graphs by means of wearable devices and ad hoc wireless sensor networks (e.g., proximity sensors). The main use case scenario is crowd management.

Marco Cattani is a Ph.D. student at the Embedded Software group of Delft University of Technology. His research is focused on sensing and communication in extremely dense and mobile wireless networks. These networks are typical of crowd monitoring application where wearable devices must continuously exchange their status. On this topic, he is currently focusing on energy-efficient mechanisms for data collection.

Maarten Van Steen is a professor at the University of Twente, where he is scientific director of the university's ICT Research Institute CTIT. He is specialized in large-scale distributed systems, now concentrating on very large wireless distributed systems, notably in the context of crowd monitoring using gossip-based protocols for information dissemination. Next to Internet-based systems, he has published extensively on distributed protocols, wireless (sensor) networks, and gossiping solutions. He is an Associate Editor for *IEEE Internet Computing*, a Field Editor for *Springer Computing*, and a Section Editor for *Advances in Complex Systems*. He has authored and co-authored three textbooks, including a highly successful book on distributed systems, as well as an introduction to graph theory and complex networks.

IEEE WCET™

IEEE WIRELESS COMMUNICATION ENGINEERING TECHNOLOGIES CERTIFICATION

IEEE

IEEE ComSoc
IEEE Communications Society

Join an Elite Group
of Professionals Working in Wireless

IEEE WCP
IEEE Wireless Communications Professional

Globally Recognized Emblem of Achievement

WWW.IEEE-WCET.ORG

IMPORTANT DATES

SPRING APPLICATION DEADLINE
31 March 2017 by 23:59 UTC

SPRING TESTING WINDOW
17 April - 13 May 2017

FALL APPLICATION DEADLINE
8 SEPTEMBER 2017 by 23:59 UTC

FALL TESTING WINDOW
25 September - 21 October 2017

Gesture Detection Using Passive RFID Tags to Enable People-Centric IoT Applications

Raúl Parada and Joan Melià-Seguí

The authors present a solution to increase the personalization of IoT applications and services (e.g., accessing a restricted area with a contact-less card) by detecting people-object gestures with an accelerometer-enabled passive RFID tag. They demonstrate the feasibility of their proposal by achieving a precision of 85 percent in people-object gestures classification.

ABSTRACT

Our society may enhance and create new services in a people-centric IoT context through the exchange of information with sensor devices. Unfortunately, communication and services may be compromised due to a number of factors including unreliable communication, complexity, and security threats like spoofing. Within the technologies involved in the IoT paradigm, passive RFID allows the inventorying of simple objects toward wireless communication with a low-cost investment. We present a solution to increase the personalization of IoT applications and services (e.g., accessing a restricted area with a contact-less card) by detecting people-object gestures with an accelerometer-enabled passive RFID tag. We demonstrate the feasibility of our proposal by achieving a precision of 85 percent in people-object gestures classification.

INTRODUCTION

Our society has been living connected for centuries; first as individuals (i.e., face-to-face conversations) and recently through computers (e.g., on the Internet). Currently, the *object* entity has become part of the connection with people, the well-known Internet of Things (IoT) paradigm. IoT enhances and increases the interaction with entities obtaining rich information for future services where humans become the center of the interactions. From smartphones to wearables, people carry sensors that can enable people-centric services and applications within the IoT context.

Within the different commercially available IoT technologies, RF identification (RFID) enables identification and personalized services by means of a simple electronic label and a reader system. RFID, including near field communication (NFC), automatizes services such as access to restricted areas or identifying an individual in a purchase transaction, by means of simply interacting with the RFID system. The use of passive RFID brings specific benefits over other IoT technologies, like cost effectiveness and simplicity, thanks to its passive nature (i.e., no need for battery usage or replacement). The benefits of such systems may, nonetheless, come with drawbacks. These interactions may be compromised due to a number of factors like unreliable communication, the inherent complexity of IoT, and security threats like impersonation or spoofing. For instance, if an attacker obtains the IoT-enabled object identity (i.e., an RFID card identification), security will be compromised since the attacker will be able to impersonate the legitimate user.

The fact that IoT provides not only connectivity or identification, but a large range of features, may provide a solution for the above problem. For instance, RFID technology not only provides the unique identification of a given object, but also generates other relevant information such as timestamps, localization, and low-level RF indicators. For instance, the received signal strength indicator (RSSI) or RF phase (PHASE) may reveal further information such as distance, movement, and interaction with people. Moreover, RFID tags can also include sensors like temperature, pressure, and accelerometers. Thus, besides the identification code, other RFID-related features can be used to personalize IoT applications.

In this article, we present a simple method of gesture detection using passive RFID tags. The goal is to enable people-centric IoT applications and services by means of classifying specific gestures using a battery-less accelerometer sensor embedded in a passive RFID tag.

Specifically, we achieve the following contributions:

- A method to characterize the people-object gestures based on acceleration time-series information
- The implementation of an unsupervised machine learning technique to classify people-object gestures
- An empirical demonstration of the people-object gestures classification with off-the-shelf devices and equipment

The remainder of this article is organized as follows. We introduce the problem motivation and the related state of the art. The RFID-based people-object gesture detection principle is described, and we present the methodology and experimentation procedure to collect and classify the people-object gestures, which we empirically evaluate. A business model analysis of our approach is performed. Finally, the article is concluded, also pointing out future work directions.

RELATED WORK

Several authors have approached the challenge of detecting and classifying human gestures. Daniels *et al.* [1] recognized hand movements by using two cameras and middleware to extract the information of the frames. Although they obtained a high-performance solution, this approach is

The authors developed part of this work at the Department of Information and Communication Technologies, Universitat Pompeu Fabra.

Digital Object Identifier:
10.1109/MCOM.2017.1600701CM

Raúl Parada is with Università degli Studi di Padova; Joan Melià-Seguí is with Universitat Oberta de Catalunya.

expensive both economically and computationally. Ali *et al.* [2] extract biometric data thanks to a video surveillance system, and by using the Distance Based Nearest Neighbor Algorithm a given hand movement can be verified. However, it also requires expensive video surveillance systems. By using sensor equipment, Mare *et al.* [3] present a solution correctly identifying 85 percent of users. They use a bracelet with an accelerometer and gyroscope to compare the motion information with the ground truth. Regarding security and safety for people with motion diseases, Gonçalves *et al.* [4] propose two approaches to detect undesired body motions. The first approach uses the Microsoft Kinect sensor and gesture recognition algorithms, and the second approach uses a trademark device of Texas Instruments with built-in accelerometers and statistical methods to recognize stereotyped movements. A movement recognition device attached to the arm with an accelerometer and EMG sensors is implemented by Shin *et al.* [5]. Flores *et al.* [6] introduce a low-cost wireless glove controller detecting finger gestures, developed using makeshift flex sensors and a digital accelerometer. A signaling approach is presented by Björklund *et al.* [7]. They can classify human targets by comparing micro-Doppler signatures using a 77 GHz radar.

Different authors have tried to address the challenge of classifying human gestures using RFID technology. A glove equipped with accelerometer sensors and an RFID reader is presented in Hong *et al.*'s work [8]. Although these solutions show good results in detecting hand movements, they require sensors fed by a battery, besides being obstructive. Wartha and Londhe [9] introduce the topic of people verification through basic movements or by proximity with an RFID labeled object and using implanted medical devices (IMDs) or common sensor devices (i.e., car keys located in the pocket). Nevertheless, their approach requires access to IMDs, if the user has any, or additional sensor devices, thus being an intrusive solution. Parada *et al.* [10] present a method for classifying an object between being static and interactive in a context-aware smart shelf scenario by uniquely using RFID data. They defined interaction as the action of removing the labeled object from its static position from a context-aware smart shelf. However, their solution simply detects either interaction or static position of objects. Finally, Asadzadeh *et al.* [11] propose the recognition of gestures using three RFID antennas distributed within a limited matrix and classifying the movements using a hypothesis tree method. Although these approaches to gesture recognition using RFID returned promising results, they simply detect movement or require more complex equipment.

We propose a method for gesture detection using battery-less accelerometers embedded in passive RFID tags, in the context of people-centric IoT applications. The novelty of our solution in relation to to the prior work described above relies on uniquely using passive RFID technology. No vision techniques (as in [1, 2]) are required, or active sensors ([3, 6]) or unidentifiable electromagnetic signals [7]. Furthermore, our solution is non-intrusive as compared to glove utilization [8] or IMDs [9]. Finally, our proposed system is able to classify multiple gestures instead of simple interaction [10] and uses a simpler RFID setup [11].

RFID-Based People-Object Movement Detection

Within the different RFID technologies and standards, the UHF EPC Gen2 [12] RFID is a de facto standard in retail. In EPC Gen2, RFID antennas interrogate, in a time-multiplexed manner, RFID passive tags, and these RFID passive tags within the read range backscatter the signal back to the RFID reader. The RFID reader not only inventories those RFID tags within its read range, but can also record other high- and low-level indicators included in the backscattered signal.

The high-level indicators, such as identification code, timestamp, antenna port, and reader identifier, uniquely identify an object within the object population, besides providing an implicit timestamp for each sample. The low-level indicators provide an approximated measure of the RF signal from the tags as measured by the RFID antenna. For instance, the RSSI is modeled by the two-way radar equation for a monostatic transmitter, while the PHASE is approximated by the combination of the round-trip distance between the reader's antenna and the tag, plus the phase rotation introduced in the transmission and reception, and at the tag itself:

- High-level indicators
  - –Identification code (96-bit typically)
  - –Timestamp
  - –Antenna port
  - –Reader identifier
- Low-level indicators
  - –RSSI
  - –PHASE

The idea behind a people-object movement is given by a variation on the low-level RFID indicators. Detecting weaker RSSI and unstable PHASE samples may imply a long coarse-grained distance between tag and antenna. In contrast, a tag returning stronger RSSI and stable PHASE samples may imply a static tag closer to the antenna [10]. For instance, Fig. 1 shows the variation of RFID features recorded with an RFID reader and antenna while a person was taking an RFID labeled object off the shelf and later returning it back to the shelf. The interaction time of about one minute is represented with black dots as well as dashed vertical lines. It can be observed that during the interaction time, the RSSI measures a decrease of over 20 dB, and the PHASE measures deviate about 80°. On the contrary, if the object remains static (indicated with white circles), the values remain constant except for small interferences inherent to the RF scenario. Hence, if variations in the above indicators are detected using RFID equipment and the proper statistical or machine learning tools, a physical interaction with the object can be assumed (we refer the readers interested in these techniques to [10] for further information).

Nevertheless, RFID low-level signals only allow the detection of coarse movements like people-object interactions, but not fine-grained movements like users' gestures. Thus, detecting specific gestures requires further context-aware informa-

The idea behind a people-object movement is given by a variation on the low-level RFID indicators. Detecting weaker RSSI and unstable PHASE samples may imply a long coarse-grained distance between tag and antenna. In contrast, a tag returning stronger RSSI and stable PHASE samples may imply a static tag closer to the antenna.

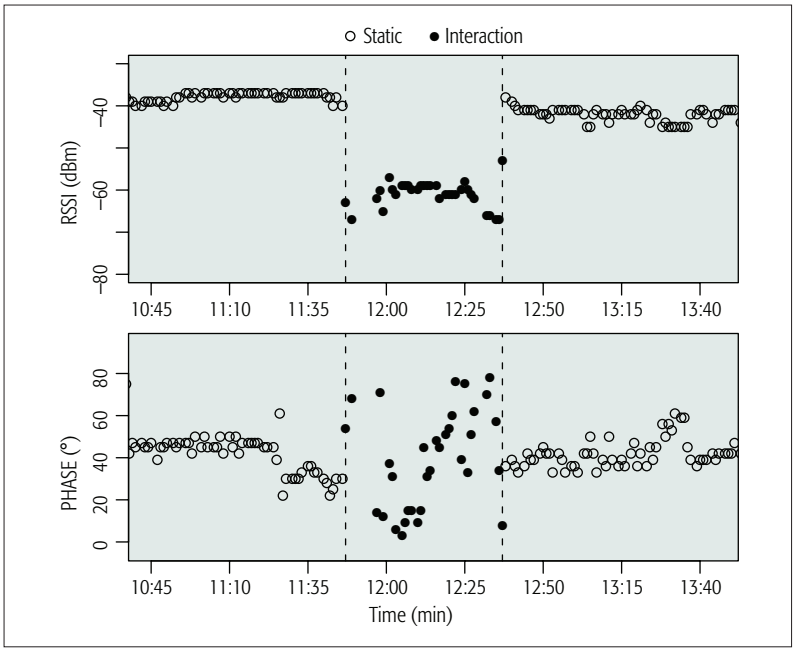


Figure 1. RFID low-level indicators like RSSI and PHASE may describe object movement, which can be inferred as interaction with persons.

tion. A solution improving movement detection accuracy within the same passive RFID technology is integrating sensors in the RFID labels. For instance, accelerometers have been demonstrated as reliable sensors to detect fine-grained movements [6, 8]. An accelerometer detects movement through spatial coordinates *x*, *y*, and *z* with respect to the gravity (measured in meters per second squared or *g*), generating a time-series of movement-related data.

Next, we detail the proposed methodology to detect specific users' gestures by using passive RFID tags with an integrated accelerometer sensor.

Gesture Detection Methodology and Experimentation

This article presents a method of gesture detection using battery-less accelerometers embedded in passive RFID tags, enabling people-centric services and applications while improving security in the IoT context. The goal is to combine the implicit ID-based RFID authentication with a specific gesture, providing two levels of authentication, and reducing security threats from third parties such as spoofing or impersonation. For instance, the proposed method could be used to improve authentication in restricted area access or authorizing payments in a commercial transaction.

Figure 2 summarizes the methodology used to enable gesture detection. Specific gestures are performed using a passive RFID tag with accelerometer sensor capabilities, together with any smartphone (from low to high range) with an integrated accelerometer for comparison and evaluation purposes (1). The gestures data is sampled by a commercial RFID reader and antenna (2) for the RFID tag using a commercial smartphone with an integrated accelerometer, and finally transferred to the same computer (3).

The data is stored as a collection of time-series information. Next, an automatic preprocessing stage is applied dividing the time series into individual gestures, also filtering the time-series static periods before and after the actual gesture (4). Feature extraction is enabled by using the Dynamic Time Warping (DTW) algorithm, which measures the similarity between two time-series, resulting in a numeric distance. The lower the numeric distance, the higher the similarity between the two time-series. DTW has already been successfully used for context recognition using accelerometers [13]. Hence, we select DTW to extract the distance between each performed gesture by either the passive RFID tag or the smartphone (5).

Finally, an unsupervised machine learning algorithm is used to analyze the time-series segment and classify the input gesture (6). Because of its simple implementation and robustness with respect to the spatial distribution of the samples, *k*-nearest neighbor (kNN) [14] is used in our experiments. In the kNN algorithm, the parameter *k* indicates the number of nearest neighbors to which a test sample is compared, classifying based on the majority of votes. For instance, if a test sample is compared with the four nearest neighbors and three of them are class A, this test sample is also considered as class A. It is worth mentioning that other unsupervised machine learning techniques could also be used. Nevertheless, our goal is to demonstrate the feasibility to classify gestures using the input accelerometer data and a simple classification technique like kNN.

The experimentation methodology is based on performing different gesture actions with both the passive RFID tag and the smartphone, following the steps described above (Fig. 2). In the experimentation stage, we used commercial state-of-the-art equipment. On one hand, we used a passive RFID tag with an integrated accelerometer from the Farsens company [15], generating around seven accelerometer samples per second. On the other hand, we used a commercial Android smartphone equipped with the Sensor Kinetic Pro app available at the Android Store [16], generating around 25 accelerometer samples per second. We defined three different gestures, which we denote as: *R*, *W*, and *C*. Figure 3 shows three images corresponding with the gestures procedure. Figure 3a–3c correspond with the movements *C*, *W*, and *R*, respectively. The idea behind the people-object gestures is given by the variation of the three time series accelerometer data associated with each of the spatial coordinates *x*, *y*, and *z*.

Figure 4 depicts the time series of the *x*, *y*, and *z* spatial coordinates while performing a gesture with a passive RFID tag (solid red line) and a smart phone (dashed black line). Figure 4a represents an example of low DTW value where both the passive RFID tag and the smartphone time series perform the same gesture *W*. Opposite, Fig. 4b shows an example of high DTW value because of the different gestures performed with the passive RFID tag (*C*) and the smartphone (*W*). It is important to stress that the DTW algorithm is able to compute the similarity between the different time series even if their lengths differ. Hence, the gestures' performance does not need to be either synchronized or restricted to any specific length.

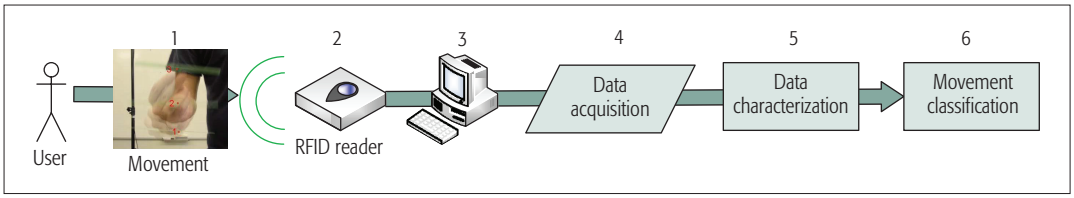


Figure 2. People-object movement classification scheme.

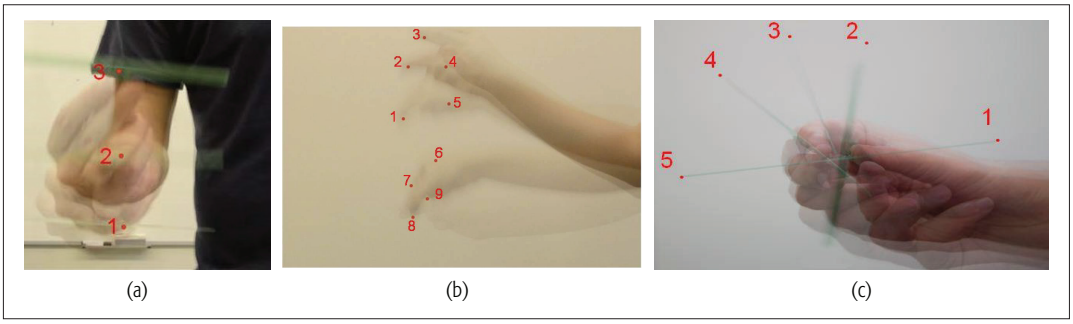


Figure 3. Specific gestures are performed during the experimentation stage: a) *C* movement; b) *W* movement; c) *R* movement.

Specifically, a total of 30 samples were performed, half with the passive RFID tag and the other half with the smartphone. With each device, volunteers executed the three predetermined gestures *R*, *W*, and *C* over an approximately 15 s period. Since each gesture is composed of three time series (one for each spatial coordinate), a total of 90 time series were generated in the experimentation stage. Therefore, a matrix of 2700 distances is generated by comparing all time sequence representations to each other.

Evaluation and Discussion

IoT-based gesture detection must ensure proper performance in the context of people-centric applications and services like authentication or intelligent systems. Hence, gesture classification performed on either passive tags or smartphones must be accurate and reliable. As described earlier, a total of 2700 time series are generated from the 30 executed tests. Ten-fold cross validation is used to evaluate the kNN clustering based on DTW features (implemented using the statistical software *R*. kNN is in turn evaluated using different values. In this article we present the results of $k = \{10, 8, 5\}$, being three of the most representative values in the experimentation phase.

Table 1 tabulates the ratio of the people-object gesture's prediction from both the passive RFID tag and the mobile device. The total number of 30 samples of movements are divided into each of the three movements *C*, *R*, and *W*, and each device. In addition, the ratio is calculated based on the number of neighbors (k). For each value of k the predicted gesture with highest ratio and its ratio are shown.

The results show how selecting a proper value of k is key for the correct performance of the kNN algorithm. Although the *C* gesture is correctly classified using all evaluated values of k, gestures *R* and *W* require fine-tuning the k parameter to 5. Not unexpectedly, the more complex the gesture, the harder it is to correctly classify. Comparing the smartphone to the passive RFID tag, the higher sampling rate of the smartphone

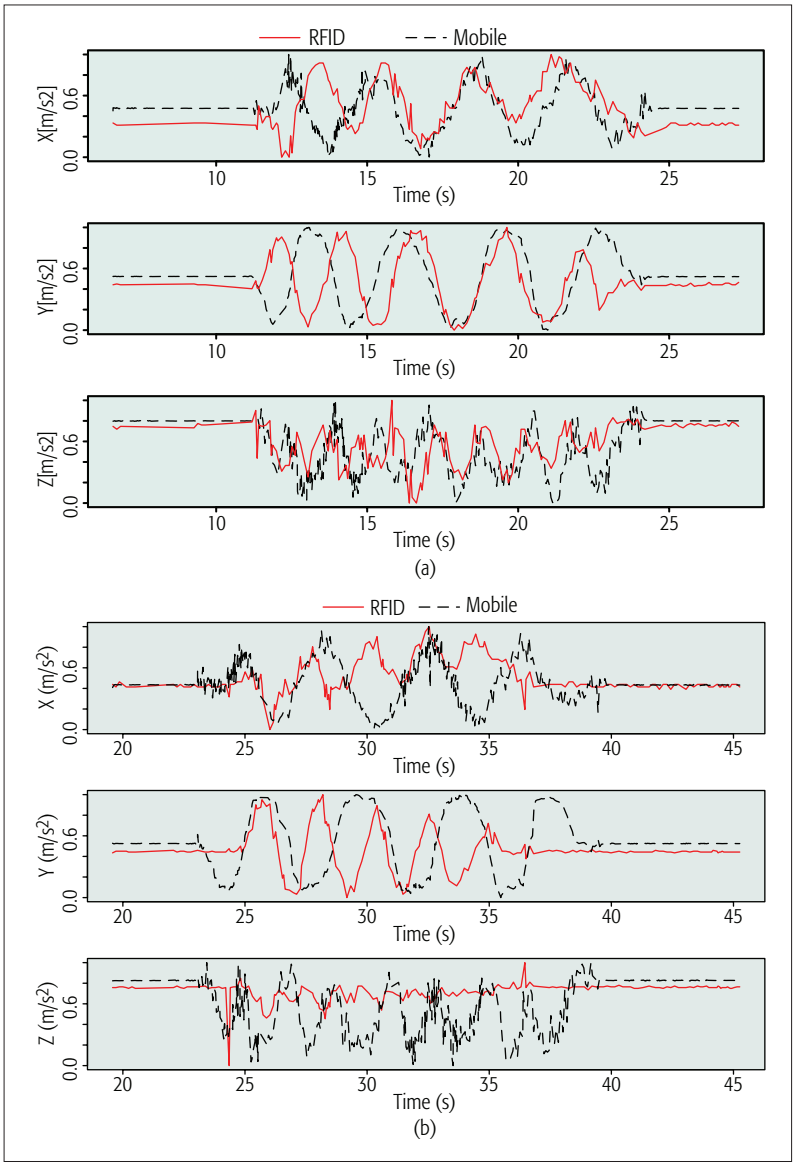


Figure 4. Each gesture generates three time series, one for each spatial coordinate.

provides more consistent results, although final classification ratios are the same for both technologies using five neighbors. Thus, we demonstrate that gesture recognition with the passive UHF tag with integrated accelerometer is able to detect specific gestures, achieving similar results as a more expensive device like a smartphone.

Besides the ratio of classification, we calculated measurement metrics — precision, recall, and accuracy — for the samples extracted from the passive RFID tag. These metric measurements differ from those of Table 1 since those represent the ratio of classification by a given gesture, while these metric measurements illustrate the overall classification procedure. Figure 5 shows the percentage (y-axis) for each metric measurement with different k values (x-axis). Notice that, for simplicity's sake, we averaged the classification metrics for all gestures *C*, *R*, and *W*, showing the confidence intervals. As we can observe in the figure, k = 5 returns the best results for all metrics, achieving around 85 percent precision and 73 percent accuracy. Focusing on the classification metrics for each gesture, the precision/recall results with k = 5 for the gestures *C*, *R*, and *W* are 56/100 percent, 100/38 percent, and 100/50 percent, respectively. In the cases of k = 8 and k = 10, these precision/recall values are 50/100, 100/29, and 100/41 percent, and 47/100, 100/21, and 100/39 percent, respectively. Notice how each gesture returns different metrics, generating either more false positives (gesture *C*) or false negatives (*R* and *W*). On average, the high precision metrics reveal good behavior of our proposed system in avoiding false positives. However, the lower recall implies a larger number of false negatives, which must be improved in future versions of the system.

Sample number	Movement	Device	Prediction/ratio		
			k = 10	k = 8	k = 5
1	C	Passive RFID tag	*C*/0.7	*C*/0.75	*C*/0.8
2			*C*/0.6	*C*/0.75	*C*/0.8
3			*C*/0.5	*C*/0.5	*C*/0.8
4			*C*/0.6	*C*/0.75	*C*/0.8
5			*C*/0.8	*C*/0.75	*C*/0.8
6		Mobile	*C*/0.6	*C*/0.625	*C*/1
7			*C*/0.8	*C*/0.875	*C*/0.8
8			*C*/0.8	*C*/0.875	*C*/1
9			*C*/0.9	*C*/0.875	*C*/1
10			*C*/0.8	*C*/0.750	*C*/1
11	R	Passive RFID tag	*C*/0.7	*C*/0.625	*R*/0.6
12			*C*/0.7	*C*/0.625	*R*/0.6
13			*C*/0.7	*C*/0.625	*R*/0.6
14			*C*/0.6	*C*/0.5	*R*/0.8
15			*C*/0.6	*C*/0.625	*R*/0.6
16		Mobile	*R*/0.5	*R*/0.625	*R*/0.8
17			*C*/0.5	*R*/0.625	*R*/0.8
18			*R*/0.5	*R*/0.625	*R*/0.8
19			*R*/0.5	*R*/0.625	*R*/0.8
20			*R*/0.5	*R*/0.625	*R*0.8
21	W	Passive RFID tag	*C*/0.5	*C*/0.5	*W*/0.8
22			*W*/0.7	*W*/0.75	*W*/0.8
23			*W*/0.7	*W*/0.625	*W*/0.8
24			*W*/0.8	*W*/0.750	*W*/0.8
25			*W* /0.9	*W*/1	*W*/0.8
26		Mobile	*W*/0.7	*W*/0.875	*W*/1
27			*W*/0.7	*W*/0.875	*W*/1
28			*W*/0.7	*W*/0.875	*W*/1
29			*W*/0.7	*W*/0.875	*W*/1
30			*W*/0.7	*W*/0.875	*W*/1

Table 1. Empirical evaluation of the gestures classification.

Nevertheless, the results confirm the feasibility of detecting people-object gestures using a passive RFID tag with an integrated accelerometer in a people-centric IoT paradigm, with similar results as performing the same gestures with a higher-range device like a smartphone. Besides passive UHF, other passive RFID technologies like near field communications (NFC) could also be suitable candidates. However, the reduced read range of NFC would make the gestures procedure harder to perform and detect.

Business Model Analysis

The proposed method could be used, for instance, within the context of intelligent systems to improve authentication in a restricted area access or authorizing payments in a commercial transaction, hence providing safer services thanks to people-centric IoT. Beyond the prototype presented in this article, in a real scenario a user would initially have performed and stored a specific gesture (i.e., in a service's signup process) to later authenticate herself in the service through the two-step process: RFID plus gesture. The authentication process could be done against a local server, the cloud, or any other storage system.

This service could be adapted to a *Physical Freemium* IoT business model, as defined by Fleisch *et al.* in [17]. The physical system (RFID reader plus tags) could be sold together with a free digital service, such as operation and maintenance. Over time, customers could switch to a premium service like cloud storage, a higher number of users, or electronic monitoring.

Conclusion and Future Work

In a people-centric IoT paradigm, objects enhance the communication between people and the Internet, improving or creating new context-aware

services. Nevertheless, this exchange of information may be compromised due to a number of factors including unreliable communications, the inherent IoT complexity, or even security threats like spoofing. Hence, new approaches are required to improve security and authentication in any IoT-related transactions.

Within the wide range of available IoT systems, passive RFID is a ubiquitous technology due to its simplicity and low cost. Besides automated identification, RFID can also extract context-aware features such as object movement. Moreover, if combined with integrated sensors like accelerometers, it may provide a finer-grained resolution of events.

In this article we propose a method for gesture detection using battery-less accelerometers embedded in passive RFID tags and unsupervised learning, intended to improve people-centric IoT applications and services like authentication and intelligent systems.

The results demonstrate the suitability of our proposal by achieving high precision in detecting specific gestures using a passive UHF RFID tag with an integrated accelerometer and the unsupervised kNN classification method. With the proposed system, we expect to contribute to a better people-centric IoT ecosystem, improving services and applications by means of simple gesture detection using context-aware technology. Our future work includes, but is not limited to:

- Improving low-level information extraction from the passive RFID tags to infer movements and gestures with higher resolution
- Expanding testing with further devices, technologies, and other machine learning techniques to improve classification metrics
- Evaluating our proposal in other people-centric IoT contexts like people with motor or cognitive disabilities

Acknowledgment

The authors of this article would like to thank Juan Carlos Horno Murillo in the laboratory tasks. We also acknowledge technical support from the Keonn and Farsens companies. This work is partially supported by the Spanish Ministry of Economy and the FEDER regional development funds under the projects SINERGIA (TEC2015-71303-R), CO-PRIVACY (TIN2011-27076-C03-02), SMART-GLACIS (TIN2014-57364-C2-2-R), and CUIDATS (IPT-2012-0972-300000).

References

[1] M. Daniels *et al.*, "Real- Time Human Motion Detection with Distributed Smart Cameras," *2007 First ACM/IEEE Int'l. Conf. Distributed Smart Cameras*, Sept. 2007, pp. 187–94.
[2] M. H. Ali *et al.*, "Automated Secure Room System," *2015 4th Int'l. Conf. Software Engineering and Computer Systems*, Aug. 2015, pp. 73–78.
[3] S. Mare *et al.*, "ZEBRA: Zero-Effort Bilateral Recurring Authentication," *2014 IEEE Symp. Security and Privacy*, May 2014, pp. 705–20.
[4] N. Gonçalves *et al.*, "Automatic Detection of Stereotyped Hand Flapping Movements: Two Different Approaches," *2012 IEEE RO-MAN: The 21st IEEE Int'l. Symp. Robot and Human Interactive Commun.*, Sept. 2012, pp. 392–97.
[5] S. w. Shin *et al.*, "Developing a Device Using Accelerometers and EMG for Hand Movement Recognition," *2013 6th Int'l. Conf. Biomedical Engineering and Informatics*, Dec. 2013, pp. 398–402.

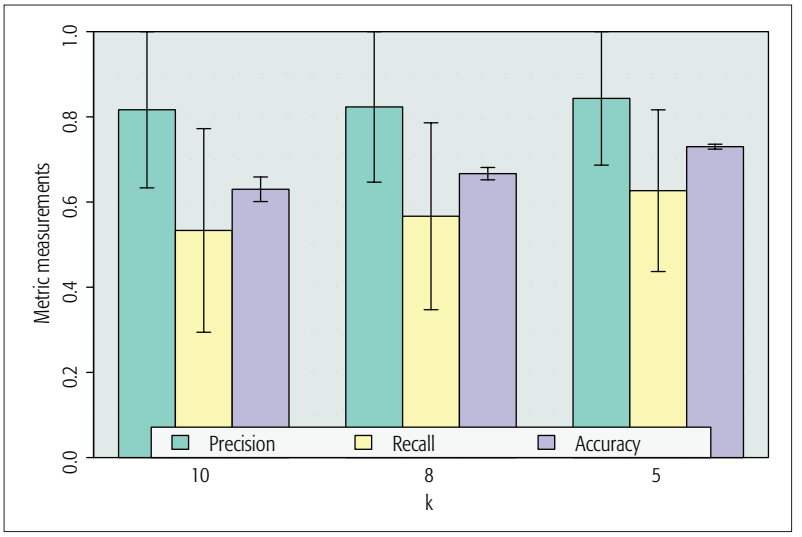


Figure 5. The highest classification metrics are achieved with $k = 5$.

[6] M. B. H. Flores *et al.*, "User-Oriented Finger-Gesture Glove Controller with Hand Movement Virtualization Using Flex Sensors and a Digital Accelerometer," *2014 Int'l. Conf. Humanoid, Nanotechnology, Info. Technology, Commun. and Control, Environment and Management*, Nov. 2014, pp. 1–4.
[7] S. Björklund *et al.*, "Millimeter-Wave Radar Micro-Doppler Signatures of Human Motion," *2011 12th Int'l. Radar Symp.*, Sept. 2011, pp. 167–74.
[8] Y. J. Hong *et al.*, "Activity Recognition Using Wearable Sensors for Elder Care," *2008 2nd Int'l. Conf. Future Generation Commun. and Net.*, vol. 2, Dec 2008, pp. 302–05.
[9] N. Wartha and V. Londhe, "Context-Aware Approach for Enhancing Security and Privacy of RFID," *Int'l. J. Engineering and Computer Science*, vol. 4, 2015, pp. 10,078–88.
[10] R. Parada *et al.*, "Using RFID to Detect Interactions in Ambient Assisted Living Environments," *IEEE Intelligent Systems*, vol. 30, no. 4, July 2015, pp. 16–22.
[11] P. Asadzadeh, L. Kulik, and E. Tanin, "Gesture Recognition Using RFID Technology," *Personal Ubiquitous Comp.*, vol. 16, no. 3, Mar. 2012, pp. 225–34; http://dx.doi.org/10.1007/s00779-011-0395-z.
[12] EPC Radio-Frequency Identity Protocols Generation-2 UHF RFID, Specification for RFID Air Interface, Protocol for Communications at 860 MHz–960 MHz, v/ 2.0.0 Ratified, EPCglobal, 2013, Nov. 2016; http://www.gs1.org/.
[13] D. Figo *et al.*, "Preprocessing Techniques for Context Recognition from Accelerometer Data," *Personal Ubiquitous Comp.*, vol. 14, no. 7, Oct. 2010, pp. 645–62; http://dx.doi.org/10.1007/s00779-010-0293-9.
[14] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, 2012, iSBN: 0262018020, 9780262018029.
[15] Farsens, "Kineo-A3DH," 2014, Nov. 2016, http://www.farsens.com.
[16] Android Play Store, "Sensor Kinetic Pro," Nov. 2016; https://play.google.com/.
[17] E. Fleisch, M. Weinberger, and F. Wortmann, *Business Models and the Internet of Things (Extended Abstract)*, Springer, 2015, pp. 6–10.

Biographies

Raúl Parada (rparada@dei.unipd.it) is a postdoctoral researcher at the Dipartimento di Ingegneria dell'Informazione (DEI), Università degli Studi di Padova and a course instructor at Universitat Oberta de Catalunya (UOC). He received his Ph.D. degree about Information and Communication Technologies at the Departament de Tecnologies de la Informació i les Comunicacions at Universitat Pompeu Fabra (UPF) in 2016. He has published several papers in the areas of wireless communications, machine learning, the Internet of Things, and antenna design.

Joan Melià-Seguí (melia@uoc.edu), Ph.D. (2011), is a lecturer at the Estudis de Informàtica, Multimèdia i Telecomunicació and a researcher at the Internet Interdisciplinary Institute, both at UOC. Before, he was a postdoctoral researcher at UPF and the Palo Alto Research Centre (Xerox's PARC). He has published more than 30 papers and one patent in the areas of the Internet of Things, intelligent systems, security, and privacy.

Human Neuro-Activity for Securing Body Area Networks: Application of Brain-Computer Interfaces to People-Centric Internet of Things

Juan F. Valenzuela-Valdés, Miguel Angel López, Pablo Padilla, José L Padilla, and Jesus Minguillon

The authors propose to use wireless brain-computer interfaces as a secure source of entropy, based on neuro-activity, capable of generating secure keys that outperforms other generation methods. In their approach, current wireless brain-computer interface technology is an attractive option to offer novel services emerging from new necessities in the context of the people-centric Internet of Things.

ABSTRACT

A former definition states that a brain-computer Interface provides a direct communication channel to the brain without the need for muscles and nerves. With the emergence of wearable and wireless brain-computer interfaces, these systems have evolved to become part of wireless body area networks, offering people-centric applications such as cognitive workload assessment and detection of selective attention. Currently, wireless body area networks are mostly integrated by low-cost devices that, because of their limited hardware resources, cannot generate secure random numbers for encryption. This is a critical issue in the context of new Internet of Things device communication and its security. Such devices require securing their communication, mostly by means of the automatic renewal of the cryptographic keys. In the domain of the people-centric Internet of Things, we propose to use wireless brain-computer interfaces as a secure source of entropy, based on neuro-activity, capable to generate secure keys that outperforms other generation methods. In our approach, current wireless brain-computer interface technology is an attractive option to offer novel services emerged from novel necessities in the context of the people-centric Internet of Things. Our proposal is an implementation of the human-in-the-loop paradigm, in which devices and humans indistinctly request and offer services to each other for mutual benefit.

INTRODUCTION

The people-centric Internet of Things (IoT), the wearable IoT (WIoT), and the health IoT are the different names that are emerging to represent the paradigm for a smart world in which ubiquitous communication occurs among heterogeneous and interconnected devices in wireless body area networks (WBANs). These networks involve a variety of low-cost devices, sensors, and gadgets with wireless communication capabilities that are placed surrounding the human body for physiological monitoring. Such WBAN devices are required to be compact, wearable, and energy-efficient in order to achieve a practical system with sufficient lifetime. These requirements impose non-negligible limitations regarding data acquisition, computation, and transmission capabilities. However, those are not the unique limitations to be considered: due to the shared wireless medium between WBAN devices, the communication security may be compromised. In this way, it is possible to have malicious attacks on body-centric systems. To avoid this, the transmitted data must be secured as it is generated, transmitted, received, stored, and analyzed within the complete system. As a consequence, WBAN security is a challenge that arises, with novel and ongoing solutions.

The WBAN nature makes it necessary to combine security with energy efficiency to provide a practical solution for wearable devices and sensors. In particular, the resources for security are very scarce; as a consequence, the solution is nontrivial. There are traditional upper layer security solutions, such as the Advanced Encryption Standard (AES), the Diffie–Hellman key algorithm, elliptic curve cryptography, and hash chains, among others, which have high computational costs. However, if lower latency or computational costs are required, it could be convenient to explore the lower layers to provide security. One of the best candidates for this task is the physical layer [1]. In this way of providing low complexity and latency, an additional approach that may provide an efficient solution is to encrypt the data prior to communication by generating random binary sequences from the communication device signals. This approach can be especially useful in the case of wireless sensors, wearable devices, and bio-signals, for which the body and its acquired signals are the essential part of the communication system. Up to now, different bio-signals from wearable devices and bio-inspired solutions have been used for securing WBANs [2]. In the last years, studies have used bio-signals such as photoplethysmogram [3], interpulse interval [4], and electrocardiogram [5], among others, to generate secure keys in the context of WBANs. In this way, other interesting signals such as EEG ones may be used for securing WBANs, which is the purpose of this work.

Digital Object Identifier:
10.1109/MCOM.2017.1600633CM

The authors are with Universidad de Granada; José L. Padilla is also with École Polytechnique Fédérale de Lausanne.

0163-6804/17/$25.00 © 2017 IEEE

In comparison with ECG, EEG signals present better characteristics for the generation of random binary sequences. For instance, unlike ECG, EEG is a non-stationary signal that, after very simple whitening processing, presents the nearly flat spectrum corresponding to an impulsive autocorrelation process. These two aspects kindly facilitate the generation of truly random and independent sequences with minimum processing and memory usage. Another aspect to consider is the bit rate. Considering the most advanced techniques in ECG, the generation of 128 random bits would take 6–10 s [5]. However, EEG signals are typically acquired at a rate of 1 ksample/s with 24 bits of resolution. This computes a binary stream of 24 kb/s. This estimation is the case for just a single EEG channel. Therefore, EEG acquisitions constituted by independent electrodes located at relatively separated positions could easily multiply this rate. In summary, and taking into account that a fraction of the total bit rate will be discarded due to existing redundancies of EEG signals, EEG has the extraordinary potential to provide very large sets of random numbers per second. This could enable the delivery of secure transmissions among the devices of a particular WBAN or to feed a repository with cryptographic passwords for a complete people-centric IoT environment (Fig. 1).

In this article, a novel approach to providing security to WBAN communications is proposed, based on secure key generation by means of EEG signal acquisition. This approach is oriented to the people-centric IoT paradigm. The article is organized as follows. We refer to brain-computer interfaces in the people-centric IoT environment. We present our experimental framework. We devote a section to the experimental results and discussion of them. Finally, conclusions are drawn.

Wireless Brain-Computer Interfaces in the People-Centric Internet of Things Ecosystem

The most relevant function of a wireless brain-computer interface (WBCI) is to establish a communication channel between the brain and other entities of the people-centric IoT or the so-called Internet of People (IoP) [6, 7]. Typically, WBCIs extract endogenous cognitive information from EEG neuro-correlates, codify this information into a binary sequence of data and stream them out [8]. The data stream is normally used either for feedback to the user, thus constituting a closed-loop communication system or as commands to computers and actuators. In the last years, WBCIs have been used for different purposes related to the human-in-the-loop paradigm, such as for the assessment of level of attention in multi-talker scenarios [9] or the cognitive workload as well as in neuro-marketing or in ambient assisted living.

The IoP paradigm enables WBCI to be part an environment in which other nodes can benefit of the generation of cognitive and electro-physiological information. Figure 1 reproduces the IoP architecture proposed in [7]. In this architecture, WBCIs are nodes of the Physical Space that upload cognitive information by means of an aggregation node towards the IoP Runtime Space. The IoP Runtime Space provides uniform access of services and applications to nodes of the physical space by an Open application programming interface (API) that abstracts their technical details. Then applications and services of different IoP scenarios could access a pool of shared resources and their data.

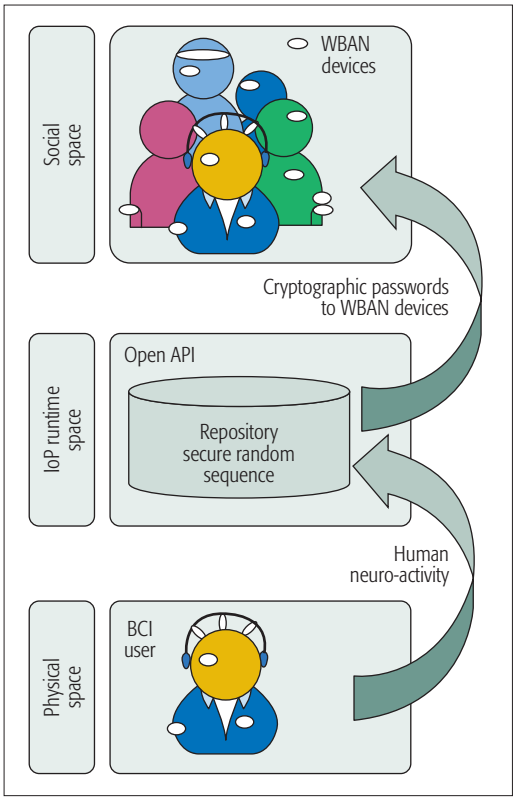


Figure 1. IoP infrastructure components.

In this context, humans are both data sources and sinks in the same way that WBAN devices of the physical space are. In our approach, WBCIs can offer to the IoP infrastructure truly random binary sequences that can be used as passwords to establish secure transactions between applications and WBAN devices that, as mentioned before, lack this capability. In this paradoxical approach, humans have the capability to offer services to applications and devices that, in turn, serve humans.

Finally, another important aspect to be considered is related to usability and wearability of the EEG acquisition systems. In recent years the level of integration, usability, and wearability of such systems has extraordinarily evolved from expensive clinical units to low-cost, mobile, head-mounted, and lightweight devices. In the future a higher level of miniaturization is expected; thus, WBCIs are becoming another integrated entity of our WBAN environment.

Experimental Frame

In this section, the experimental background is provided, in terms of description of: EEG signal acquisition, processing techniques for the experimental validation, and the statistical tests that have to be passed in order to assess the suitability of EEG signals as a source for secure communication key generation.

In recent years the level of integration, usability and wearability of such systems has extraordinary evolved from expensive clinical units up to low-cost, mobile, head-mounted and lightweight-designed devices. In the future a higher level of miniaturization is expected, thus becoming WBCIs another integrated entity of our WBAN environment.

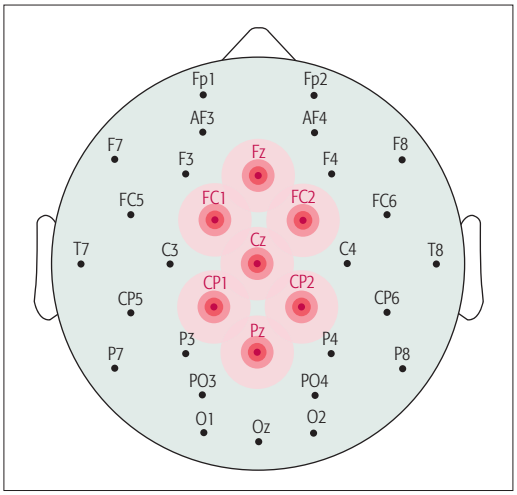


Figure 2. Electrode head map used in the experiments [10]. Red circles correspond to the electrodes in central top positions, which provide the best results in the supervised analysis.

EEG Signals for the Experimentation

The EEG datasets used to produce the results in this work have been provided by the Multimedia Signal Processing Group (MMSPG) of the Ecole Polytechnique Fédérale de Lausanne (EPFL) [10]. The acquisition system is an efficient P300-based brain-computer interface for disabled subjects. The datasets contain raw EEG data from eight subjects. Each one is formed by 32 electrode signals. The electrode positions are Fp1, AF3, F7, F3, FC1, FC5, T7, C3, CP1, CP5, P7, P3, Pz, PO3, O1, Oz, O2, PO4, P4, P8, CP6, CP2, C4, T8, FC6, FC2, F4, F8, AF4, Fp2, Fz, and Cz of the 10-20 International System (Fig. 2). The sampling rate is 2048 Hz.

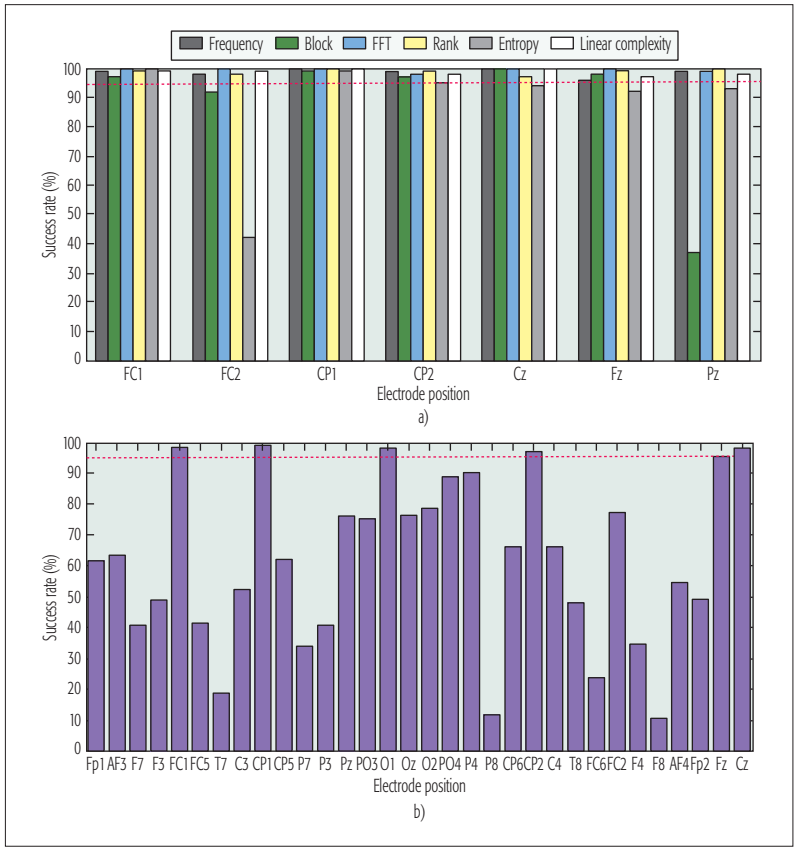


Figure 3. NIST test results: a) detailed results of the NIST tests for the electrodes in the Cz/Fz zone and surrounding area; b) mean success rate of the electrodes for the most three demanding NIST tests (frequency, block, and entropy). The red line is the NIST threshold of success.

It must be highlighted that all experiments were performed under real-world conditions. This means that the data processed in this study contain artifacts caused by eye blinks, eye movements, and muscle activity, among others, and the subjects were not always perfectly concentrated on particular tasks for the experiments.

Techniques for EEG Signal Processing

Due to their nature, the EEG data contain information mixed with artifacts and noise, typically from neuro-motor activity and electrical couplings. As a consequence, the experimentation may be affected. To overcome possible limitations, it may be recommendable to define procedures for proper EEG data processing and further analysis based on the extraction of the relevant information from the noisy captured data. Some procedures have been proposed in the literature, such as Principal Component Analysis (PCA) or Independent Component Analysis (ICA) [11, 12]. These approaches can extract the main components of a dataset by means of standard projection models so that noise and artifacts can be neglected or conveniently reduced.

The first one, PCA, is applied to find the space of maximum variance in the M-dimensional feature space of a dataset, formed by N samples of M variables each. In the case of EEG data, the M samples correspond to the different EEG acquisition channels, and the N variables are the registered signal values during the signal registration period. PCA performs a linear transformation of the original set of samples into a lower number K of uncorrelated features, called principal components (PCs), according to the computed K-subspace projection vectors. Those projection vectors are the basis for the EEG analysis in this work.

The second one, ICA, is a method of separating a signal into additive subcomponents (blind source separation). It is based on the computation of the independent vectors that compound the analyzed set of signals. ICA finds the independent components (latent sources) by maximizing the statistical independence of the estimated components. The ICA separation of mixed signals gives very good results if two assumptions are satisfied: the source signals are independent of each other, and the values in each source signal have non-Gaussian distributions, which are premises that are valid in the case of EEG.

The result in both cases, PCA and ICA, is a set of independent vectors that represent the subspace in which the signal can be represented, maximizing the independence of such vectors.

The NIST Test

The NIST Test Suite [13] is a tool for security test developed by the U.S. National Institute of Standards and Technology that is widely used for validating the performance of secure keys [14, 15]. This tool is used in this work for testing the randomness of the generated sequences in our

	Channel NIST performance (%)										
	Fp1	AF3	F7	F3	FC1	FC5	T7	C3	CP1	CP5	P7
Frequency	98	89	92	96	99	87	55	87	100	69	38
Block	67	77	16	27	97	22	0	35	99	18	26
FFT	100	100	91	100	100	100	100	98	100	98	100
Rank	98	100	99	100	99	98	92	99	100	99	98
Entropy	20	24	15	24	100	16	2	35	99	99	38
Linear complexity	99	99	97	98	99	100	97	98	100	98	99
	P3	Pz	PO3	O1	Oz	O2	PO4	P4	P8	CP6	CP2
Frequency	45	99	97	99	100	100	97	93	15	39	99
Block	38	37	92	98	35	42	83	89	7	80	97
FFT	100	99	98	100	100	97	99	99	100	100	98
Rank	99	100	100	100	99	99	100	99	99	98	99
Entropy	40	93	37	97	94	94	87	88	14	79	95
Linear complexity	100	98	99	40	99	99	99	99	100	99	98
	C4	T8	FC6	FC2	F4	F8	AF4	Fp2	Fz	Cz	
Frequency	32	88	27	98	78	25	87	98	96	100	
Block	82	30	9	92	13	4	15	28	98	100	
FFT	99	99	100	100	98	100	56	98	100	100	
Rank	99	99	94	98	96	93	96	93	99	97	
Entropy	85	26	36	42	13	3	62	21	92	94	
Linear Complexity	100	100	88	99	100	98	100	100	97	100	

Table 1. NIST test results for the 32 EEG channel signals.

In recent years the level of integration, usability and wearability of such systems has extraordinary evolved from expensive clinical units up to low-cost, mobile, head-mounted and lightweight-designed devices. In the future a higher level of miniaturization is expected, thus becoming WBCIs another integrated entity of our WBAN environment.

experimentation. For processing the generated data, it is partitioned into 100 sequences, each sequence with 20,000 bits. The NIST Test Suite provides 15 different tests. For the sake of simplicity, only the six most significant tests are reported in this article; see [14] for more information.

Results and Discussion

In this section, both the supervised and unsupervised approaches are considered. Both methods attempt to determine if the EEG signals can be used as the source signals for secure key generation, analyzing the robustness of the secured sequence by means of the NIST test.

Supervised Analysis

The first approach is the one in which the EEG signal acquisition points (EEG channels) for key generation have to be properly studied. In this case, it is necessary to identify which EEG acquisition positions are adequate to obtain the best performance in terms of key generation and its robustness. Figure 2 provides the electrode head map used in the experimentation.

All 32 channels have been analyzed to determine which ones provide good performance and pass the NIST tests (score above 95 percent). Table 1 provides the main NIST test results for the code generated by the EEG signals, according to their electrode position.

As can be noticed, the best electrode positions are CP1, FC1, Fz, CP2, and Cz, sorted in terms of performance. If their positions in the Fig. 2 map are analyzed, it clearly appears that the acquisition points on the vertex (central top of the head) are the best ones for the acquisition: the Cz and surrounding electrodes (Fig. 2). Figure 3a provides the detailed results of the electrodes in this area. This result is of importance considering the usability point of view: the best acquisition area is easily accessible and disguisable under a cap. It can be monitored with a very simple EEG acquisition system with only one central electrode or more, situated in the Cz position and surroundings, respectively.

If the different NIST tests are compared, it is noticed that the most demanding tests are the frequency, block, and entropy ones. Figure 3b provides the mean success rate of the electrodes for these three demanding tests, which ratifies the suitability of the Cz zone for the acquisition.

In comparison with other advanced proposals based on ECG [5], our EEG approach can generate binary sequences at a much faster rate. If the best five EEG channels are used (Table 1), the binary rate will be improved five times, thus

	NIST performance [%]					
	Frequency	Block	FFT	Rank	Linear. Compl.	Entropy
PCA	99	38	99	100	99	99
ICA	99	96	99	100	96	100

Table 2. NIST test results for the processed compendium signal.

	NIST performance [%]					
	Frequency	Block	FFT	Rank	Linear. Compl.	Entropy
Our best	100	99	100	100	99	100
[5]	90	90	80	100	90	–
[14]	98.9	99.4	99.4	99.4	95.6	98.9

Table 3. Comparison of the NIST test results for our best case [5, 14].

enabling the delivery of secure passwords per data flow or even transaction. In addition, the NIST test results of our approach from all the best five channels clearly exceed those of the ECG approach in [5].

Unsupervised Analysis

This second approach is the one in which the EEG electrode positions for key generation are not known a priori. In this situation, if there is no previous knowledge of the proper EEG acquisition positions to obtain the best performance, techniques such as PCA and ICA may be employed. According to the noisy nature of the EEG signals, ICA is suitable to extract different independent signals that underlie the 32-channel EEG set of signals. With these independent signals, a unique signal can be constructed as a compendium of the constituting ICA signals. Table 2 provides the main NIST test results for the code generated by this compendium signal.

As can be seen, both approaches have a significant impact in terms of NIST test performance results. In fact, ICA is the one that provides the best results, successfully fulfilling all the different tests. This case is useful in cases where the user is not familiarized with EEG signal acquisition, and the suitability of the different acquisition channels is not known in advance (e.g., BCI users with cognitive impairment or brain damage or simply due to subject inter-variability). This would avoid the need for a calibration session, thus improving the usability and plug-and-play character of our approach.

In our case, whatever the approach considered (supervised or not), the experiments reveal that EEG is a suitable source for secure communication key generation in WBANs. This demonstrates that people with very limited communication skills, such as former BCI users and the severely motor impaired, may also benefit from the generation of secure passwords for their BCI communication systems with no cognitive effort and further complexity.

Conclusion

This article presents a novel approach to provide security to wireless body area network communications based on secure key generation by means of EEG data. This proposed approach is oriented to cope with the people-centric IoT paradigm.

WBANs are mostly integrated by low-cost devices that, because of their limited hardware resources, cannot generate secure random numbers for encryption. In the context of new IoT device communication and its security, such devices require their communication to be secured, mostly by means of automatic renewing of cryptographic keys. Thus, in providing people-centric applications, security is a critical issue.

Our approach is based on brain-computer interface signal acquisition for key generation. The raw EEG signals act as the source data, based on neuro-activity, capable of generating secure keys that outperform other key generation methods. Considering the different head acquisition points available, which positions provide the best results for secure key generation must be stated. As a consequence, two cases are considered: supervised and unsupervised analysis. The first one allows us to determine which positions are the best for signal acquisition, whereas the second one is used when no previous knowledge about location suitability is available. In the case of supervised analysis, it is identified that the best acquisition points are the ones on top of the head (Cz/Fz zone and surrounding area: CP1, FC1, and CP2). In the case of unsupervised analysis, the ICA signal decomposition into independent components and compendium generation is the optimal solution for secure communication key generation.

Compared to other proposed methods in the literature such as ECG, our EEG approach generates much faster sequences with very low latency and negligible computational cost. In addition, the usability is ensured as only one channel located at the top of the head is required, thus permitting the use of a low-cost and small BCI headset with a very reduced number of channels hidden under a cap.

In an open view, our proposal can be cataloged as a particular implementation of the human-in-the-loop paradigm, in which devices and humans indistinctly request and offer services of/to each other for mutual benefit.

Acknowledgment

The authors are especially grateful to the Multimedia Signal Processing Group of the Ecole Polytechnique Fédérale de Lausanne for providing the EEG datasets used to produce the results in this work. This work has been supported by the UNGR15-CE-3311 and TIN2016-75097-P projects of the Spanish National Program of Research, Development and Innovation, the research project P11-TIC-7983 of Junta of Andalucia (Spain), and the Nicolo Association for the R+D in Neurotechnologies for disability.

References

[1] X. Wang, P. Hao and L. Hanzo, "Physical-Layer Authentication for Wireless Security Enhancement: Current Challenges and Future Developments," *IEEE Commun. Mag.*, vol. 54, no. 6, June 2016, pp. 152–58.
[2] S. Bitam, S. Zeadally, and A. Mellouk, "Bio-Inspired Cybersecurity for Wireless Sensor Networks," *IEEE Commun. Mag.*, vol. 54, no. 6, June 2016, pp. 68–74.
[3] K. K. Venkatasubramanian, A. Banerjee, and S. K. S. Gupta, "Plethysmogram-Based Secure Inter-Sensor Communication in Body Area Networks," 2008, pp. 1–7.

[4] 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.*, vol. 44, no. 4, Apr. 2006, pp. 73–81.
[5] G. Zheng *et al.*, "Multiple ECG Fiducial Points Based Random Binary Sequence Generation for Securing Wireless Body Area Networks," *IEEE J. Biomedical and Health Informatics*, 2016, pp. 1–9.
[6] F. Boavida and J. S. Silva, "IoP—Internet of People. Future Internet Networking Session," *ICT 2013*, Vilnius, Lithuania; http://ec.europa.eu/digital-agenda/events/cf/ict2013/item-display.cfm?id=10400 (2013); accessed 1 June 2015
[7] F. Boavida *et al.*, "People-Centric Internet of Things—Challenges, Approach, and Enabling Technologies," *Intelligent Distributed Computing IX*, vol. 616, P. Novais *et al.*, Eds., Springer, 2016, pp. 463–74.
[8] M. A. Lopez-Gordo and F. Pelayo Valle, "Brain-Computer Interface as Networking Entity in Body Area Networks," *Wired/Wireless Internet Commun.*, vol. 9071, M. C. Aguayo-Torres, G. Gómez, and J. Poncela, Eds., Springer, 2015, pp. 274–85.
[9] M. A. Lopez-Gordo *et al.*, "Phase-Shift Keying of EEG Signals: Application to Detect Attention in Multitalker Scenarios," *Signal Processing*, vol. 117, Dec. 2015, pp. 165–73.
[10] U. Hoffmann *et al.*, "An Efficient P300-based Brain-Computer Interface for Disabled Subjects," *J. Neuroscience Methods*, vol. 167, no. 1, 2008, pp. 115–25.
[11] J. Jackson, "A User's Guide to Principal Components," Wiley-Interscience, 2003.
[12] J. V. Stone, "Independent Component Analysis: A Tutorial Introduction," MIT Press, ISBN 0-262-69315-1, 2004.
[13] A. Rukhin *et al.*, *A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications*, NIST, Gaithersburg, MD, 2010.
[14] S. L. Hong and C. Liu, "Sensor-Based Random Number Generator Seeding," *IEEE Access*, vol. 3, 2015, pp. 562–68.
[15] G. Lo Re, F. Milazzo, and M. Ortolani, "Secure Random Number Generation in Wireless Sensor Networks" *Proc. 4th ACM Int'l. Conf. Security of Information and Networks*, 2011, pp. 175–82.

Biographies

Juan F. Valenzuela-Valdés (juanvalenzuela@ugr.es) received his degree in telecommunications engineering from the Universidad de Malaga, Spain, in 2003 and his Ph.D. from Universidad Politécnica de Cartagena, Spain, in May 2008. In 2004, he joined the Department of Information Technologies and Communications, Universidad Politécnica de Cartagena. In 2007, he joined EMITE Ing. as head of research. In 2011, he joined Universidad de Extremadura, and in 2015, he joined Universidad de Granada where he is currently an associate professor. His current research areas cover wireless communications and efficiency in wireless sensor networks. He has also been awarded several prizes, including a national prize to the Best Ph.D in Mobile Communications by Vodafone and the i-patentes award by Spanish Autonomous Region of Murcia for innovation and technology transfer excellence. He was cofounder of Emite Ing, a spin-off company. He also holds several national and international patents. His publication record is composed of more than 80 publications, including 40 JCR indexed articles, more than 30 contributions in international conferences, and 7 book chapters.

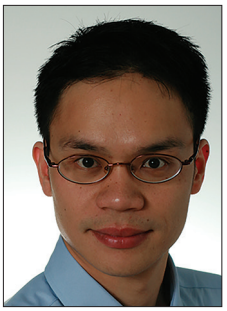
Miguel Angel Lopez received his degree in telecommunications engineering in 1998 and his Master's degree in 2011, both from the University of Malaga, Spain. He is an associate professor at the Department of Signal Theory, Telematics and Communications, University of Granada. He does research in the field of applications and wearable electronics for mobile brain-computer interfaces and bio-signal processing.

Pablo Padilla received his telecommunication engineer degree from the Technical University of Madrid (UPM), Spain, in 2005. Until September 2009, he was with the Radiation Group of the Signal, Systems and Radiocommunications Department of UPM, where he carried out his Ph.D. In 2007, he was with the Laboratory of Electromagnetics and Acoustics at Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland, as an invited Ph.D. student. In 2009 he carried out a postdoctoral stay at Helsinki University of Technology, and in September 2009, he gained an assistant professor position in the Signal Theory, Telematic and Communications Department at Universidad de Granada. Since 2012, he has been an associate professor. He is an author of more than 50 high-impact journal contributions and more than 40 contributions to international symposia. His research interests include a variety of areas of knowledge, related mainly to communication topics (radiofrequency devices, antennas, propagation, etc.), network topics (wireless communication networks), and data exploration topics.

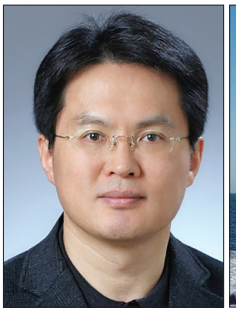
José L. Padilla received his Master of Advanced Studies degree in theoretical physics, and his Ph.D. in electronics with a Valedictorian Award Diploma from the Universidad de Granada in 2012. From 2013 to 2015, he worked as a senior researcher in the Nanoelectronic Devices Laboratory at the École Polytechnique Fédérale de Lausanne, Switzerland. He is currently a fellow of the Marie Curie Programme with the Universidad de Granada.

Jesus Minguillon received his Master's degree in telecommunication engineering in 2014 and his Master's degree in electrical and electronic engineering in 2015, both from the Universidad de Granada. He has worked in the hardware-software design of biomedical devices and 3D bio-printers. He is currently with the Department of Computer Architecture and Technology of the University of Granada, where he is a Ph.D. candidate. He does research in the fields of hardware-software implementation for biomedical instrumentation, signal processing for brain-computer interfaces, and affective computing. He is an author of several high-impact journal articles and other scientific contributions related to those fields.

PRACTICAL PERSPECTIVES ON IOT IN 5G NETWORKS: FROM THEORY TO INDUSTRIAL CHALLENGES AND BUSINESS OPPORTUNITIES


Dusit Niyato


Marco Maso


Dong In Kim


Ariton Xhafa


Michele Zorzi


Ashutosh Dutta

The Internet of Things (IoT) is a proposed development of the Internet in which everyday objects are equipped with electronics, software, sensors, and network connectivity, allowing them to send and receive data through the Internet. Formally, IEEE has described IoT as "a network of items — each embedded with sensors — which are connected to the Internet." From a practical perspective, the main difference between IoT and the existing network paradigms is that every element of IoT is able to collect and exchange data through the network without or with minimal human intervention. This opens a whole new set of interesting research challenges for future communication networks, in terms of overall system design, software implementation, data management, and service deployment.

IoT integrates different technologies and has a non-negligible impact on a variety of applications in different fields. With such capabilities, IoT has also opened interesting business opportunities in practically relevant fields such as healthcare, transportation, logistics, and manufacturing. This Feature Topic focuses on the practical aspects of IoT in 5G networks. The aim is to pave the way toward the identification of the bridge between recent theoretical findings/analytical solutions and future industrial challenges/business opportunities. Our Call for Papers attracted a number of submissions. After a thorough review process, the five papers that were best aligned with the aforementioned goal were chosen for publication.

The article "Latency Critical IoT Applications in 5G: Perspective on the Design of Radio Interface and Network Architecture" authored by P. Schulz *et al.* focuses on latency critical IoT applications, including factory automation, smart grids, and intelligent transportation systems. The article then discusses related requirements such as delay, reliability, data size, device density, and communication range. Solutions for radio interface and network architecture to meet such requirements and address the challenges, such as radio resource management, fast uplink access, transmission time shortening, and waveform design, are proposed. The article finally discusses the advantages brought by the new architecture in terms of flexibility and potential to meet all requirements.

The article "Effects of Heterogeneous Mobility on D2D- and Drone-Assisted Mission-Critical MTC in 5G" authored by A. Orsino *et al.* considers mission-critical machine type communications in LTE to support IoT applications with a variety of requirements in terms of low power, high reliability, and low latency. The article examines the impacts of mobility and heterogeneity of users and devices. Finally, the article presents numerical results to corroborate intuitions, insights, and proposals, and is concluded by a discussion on future research directions.

The article "IoT Connectivity in Radar Bands: A Shared Access Model Based on Spectrum Measurements" authored by Z. Khan *et al.* introduces a shared access framework to support IoT connectivity. The framework is developed based on the measurement results for the spectrum usage patterns and signal characteristics of ground-based fixed rotating radar systems deployed in Oulu, Finland. Subsequently, the article presents a radio environment map architecture composed of an information and measurement resource module, a database module, gateways, and device connectivity for the framework. Finally, the article highlights the benefits brought by the adoption of such a design, and outlines future research and development directions.

The article "Efficient IoT Gateway over 5G Wireless: A New Design with Prototype and Implementation Results" authored by N. Saxena *et al.* introduces the design and development of gateways to support IoT devices' communications. The gateways are able to receive IoT data traffic from the devices and provide classification and compression of delay-tolerant traffic. Then the gateways forward the traffic to a 5G cloud radio access network (C-RAN), which includes virtual base stations for processing. A prototype and testbed development based on off-the-shelf hardware and software, characterized by flexibility and suitability for real IoT applications, is presented.

The article "Wireless Caching for 5G Networks: From Theory to Implementation" authored by Y. Fadlallah *et al.* discusses the integration of wireless edge caching and coded multicast transmissions in 5G networks, with potential to support IoT applications. First, theoretical analyses of the caching-aided coded multicast technique are performed to assess

its analytical performance. Afterward, the article introduces a prototype implementation based on single-input single-output (SISO) and multiple-input multiple-output (MIMO) software defined radio platforms and a GNU Radio framework. A set of experimental results is then provided to characterize the throughput of the considered system/solution.

This Feature Topic was conceived to provide comprehensive practical perspectives on IoT in 5G networks with the focus on the transition from theory to applications. Many design and development issues for network architecture and communications technologies are reviewed. The numerous open research and development directions outlined in the five accepted papers will be useful for researchers in academia and practitioners in industry to steer the direction of their future efforts in this area.

BIOGRAPHIES

DUSIT NIYATO [M'09, SM'15, F'17] is currently an associate professor in the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He received his B.E. from King Mongkuts Institute of Technology Ladkrabang, Thailand, in 1999. He obtained his Ph.D. in electrical and computer engineering from the University of Manitoba, Canada, in 2008. His research interests are in the area of radio resource management in cognitive radio networks and energy harvesting for wireless communication.

MARCO MASO [S'08, M'14] received a Ph.D. degree in information engineering from the University of Padova as well as a Ph.D. degree in telecommunications from Supélec, France, in 2013. Since September 2014 he has been a researcher with the Mathematical and Algorithmic Sciences Lab of Huawei France Research Center. His research interests broadly span the areas of wireless communications and signal processing for the physical layer, with focus on heterogeneous networks, MIMO systems, wireless power transfer, and cognitive radios.

DONG IN KIM [S'89, M'91, SM'02] received his Ph.D. degree in electrical engineering from the University of Southern California in 1990. He was a tenured professor with the School of Engineering Science, Simon Fraser University, Canada. Since 2007, he has been with Sungkyunkwan University, Suwon, Korea, where he is currently a professor with the College of Information and Communication Engineering. Previously he served as the Founding Editor-in-Chief of *IEEE Wireless Communications Letters* from 2012 to 2015.

ARITON XHAFA [SM'10] received his Ph.D. degree from Carnegie Mellon University, Pittsburgh, Pennsylvania, in 2003. Since 2004 he has been with Texas Instruments (TI) leading projects that enable and differentiate TI's radio platforms and make them Internet of Things and Industry 4.0 ready. He is a senior member of technical staff at TI, volunteers regularly at IEEE conferences, and holds workshops for science, technology, engineering, and math (STEM) high school students in Dallas, Texas.

MICHELE ZORZI [F'07] is with the Information Engineering Department of the University of Padova. His present research interests focus on various aspects of wireless communications. He was Editor-in-Chief of *IEEE Wireless Communications* from 2003 to 2005, *IEEE Transactions on Communications* from 2008 to 2011, and, at present, *IEEE Transactions on Cognitive Communications and Networking*. He served as a Member-at-Large of the ComSoc Board of Governors from 2009 to 2011, and as Director of Education and Training from 2014 to 2015.

ASHUTOSH DUTTA [SM'03] is currently Lead Member of Technical Staff at AT&T in Middletown, New Jersey. He is co-author of the book *Mobility Protocols and Handover Optimization: Design, Evaluation and Application*, and has 30 issued patents. He serves as Director of Industry Outreach for the IEEE Communications Society and co-lead for the IEEE 5G initiative. He obtained his B.S. in electrical engineering from NIT Rourkela, his M.S. in computer science from New Jersey Institute of Technology, and his Ph.D. in electrical engineering from Columbia University.

Latency Critical IoT Applications in 5G: Perspective on the Design of Radio Interface and Network Architecture

Philipp Schulz, Maximilian Matthé, Henrik Klessig, Meryem Simsek, Gerhard Fettweis, Junaid Ansari, Shehzad Ali Ashraf, Bjoern Almeroth, Jens Voigt, Ines Riedel, Andre Puschmann, Andreas Mitschele-Thiel, Michael Müller, Thomas Elste, and Marcus Windisch

The authors focus on latency critical IoT applications and analyze their requirements. They discuss the design challenges and propose solutions for the radio interface and network architecture to fulfill these requirements, which mainly benefit from flexibility and service-centric approaches. They also discuss new business opportunities through IoT connectivity enabled by future networks.

ABSTRACT

Next generation mobile networks not only envision enhancing the traditional MBB use case but also aim to meet the requirements of new use cases, such as the IoT. This article focuses on latency critical IoT applications and analyzes their requirements. We discuss the design challenges and propose solutions for the radio interface and network architecture to fulfill these requirements, which mainly benefit from flexibility and service-centric approaches. The article also discusses new business opportunities through IoT connectivity enabled by future networks.

INTRODUCTION

Besides the traditional mobile broadband (MBB), the development of 5G networks is driven by Internet of Things (IoT) connectivity. Therefore, in addition to the classical MBB traffic demands of high throughput and capacity, new requirements of achieving low latency and high reliability for many IoT use cases are very important. In the context of new 5G use cases, IoT applications have been categorized into two classes: massive machine-type communications (mMTC) and ultra-reliable low-latency communications (URLLC). The former consists of large numbers of low-cost devices with high requirements on scalability and increased battery lifetime. In contrast, URLLC requirements relate to mission-critical applications, where uninterrupted and robust exchange of data is of the utmost importance.

In this article we focus on the latency critical IoT use cases, which are being investigated in the collaborative research project *fast wireless* (http://de.fast-zwanzig20.de/basisvorhaben/fast-wireless/). We have comprehensively analyzed such use cases and distilled their requirements. Our measurement results of the fourth generation (4G) network motivate the need for new design concepts on radio interface and network architecture in order to meet the demands of the latency critical IoT applications. In the context of the radio interface design, we discuss the latency enhancements on both the medium access control (MAC) and physical (PHY) layers. In addition, we present concepts on service-centric architecture of 5G networks. Virtualization in 5G networks leads to flexible design that enables it to shift the computing power to the edge of the network and hence reduce the latency. It also facilitates analyzing and managing the network in a service-centric fashion. This virtualization approach in the 5G network architecture allows seamless transitions between technologies or operators. Hence, service-centric management and operation disclose novel business models, from which not only network operators and service providers but also IoT customers can profit.

LATENCY CRITICAL IOT USE CASES AND REQUIREMENTS

We consider five important use cases of latency critical IoT applications and characterize them based on several different requirements as summarized in Table 1.

FACTORY AUTOMATION

Factory automation applications are typically characterized by real-time control of machines and systems in fast production and manufacturing lines, where machine parts are in motion within a limited space (e.g., a factory hall). Examples of such applications include high-speed assembly, packaging, palletizing, and so on. Factory automation applications are generally considered to be highly challenging in terms of latency and reliability demands, which also vary among different applications as given in Table 1. The reliability requirements for factory automation applications are typically 10^{-9} packet loss rate (PLR), while the latency requirements vary from 250 μs to 10 ms.

PROCESS AUTOMATION

Process automation includes applications for monitoring and diagnostics of industrial elements and processes including heating, cooling, mixing, stirring, pumping procedures, and so on. The measured values for these applications change relatively slowly. Therefore, the latency require-

Philipp Schulz, Maximilian Matthé, Henrik Klessig, Meryem Simsek, and Gerhard Fettweis are with Technische Universität Dresden; Junaid Ansari and Shehzad Ali Ashraf are with Ericsson Research, Aachen; Bjoern Almeroth is with RadioOpt GmbH; Jens Voigt and Ines Riedel are with Amdocs; Andre Puschmann and Andreas Mitschele-Thiel are with Technische Universität Ilmenau; Michael Müller is with IVM gGmbH; Thomas Elste is with IMMS GmbH; Marcus Windisch is with Freedelity GmbH.

Digital Object Identifier:
10.1109/MCOM.2017.1600435CM

	Use case	Latency (ms)	Reliability (PLR)	Update time (ms)	Data size (bytes)	Device density	Communication range (m)	Mobility (km/h)
A	**Factory automation**	0.25 to 10	10^{-9}	0.5 to 50	10 to 300	0.33 to 3 devices/m^2	50 to 100	< 30
A1	Manufacturing cell	5	10^{-9}	50	< 16	0.33 to 3 devices/m^2	50 to 100	< 30
A2	Machine tools	0.25	10^{-9}	0.5	50	0.33 to 3 devices/m^2	50 to 100	< 30
A3	Printing machines	1	10^{-9}	2	30	0.33 to 3 devices/m^2	50 to 100	< 30
A4	Packaging machines	2.5	10^{-9}	5	15	0.33 to 3 devices/m^2	50 to 100	< 30
B	**Process automation**	50 to 100	10^{-3} to 10^{-4}	100 to 5000	40 to 100	10,000 devices/plant	100 to 500	< 5
C	**Smart grids**	3 to 20	10^{-6}	10 to 100	80 to 1000	10 to 2000 devices/km^2	A few m to km	0
D	**ITS**							
D1	Road safety urban	10 to 100	10^{-3} to 10^{-5}	100	< 500	3000 /km^2	500	< 100
D2	Road safety highway	10 to 100	10^{-3} to 10^{-5}	100	< 500	500 /km^2	2000	< 500
D3	Urban intersection	< 100	10^{-5}	1000	1M / car	3000/km^2	200	< 50
D4	Traffic efficiency	< 100	10^{-3}	1000	1k	3000/km^2	2000	< 500
E	**Professional audio**	2	10^{-6}	0.01 to 0.5	3 to 1000	up to 1/m^2	100	< 5

Table 1. Communication requirements of latency critical IoT applications [1–3]. Please note that *update time* only applies to periodic traffic. The application use cases may also include sporadic or event-based traffic, but the traffic arrival distributions are not mentioned in the table.

ments for such services range from 50 to 100 ms with affordable PLR of up to 10^{-3}. The coverage area is often quite large (e.g., a power plant) and typically comprises multiple buildings and outdoor sites.

Smart Grids

Smart grid applications have relatively less stringent requirements on latency and reliability compared to factory automation applications, that is, latency and PLR requirements of up to 20 ms and 10^{-6}, respectively. However, the communication range needs to be much longer (i.e., up to a few kilometers).

Intelligent Transport Systems

Autonomous driving and the optimization of road traffic create new challenges for communications. Requirements result from different intelligent transport system (ITS) use cases such as autonomous driving, road safety, and traffic efficiency services [1].

Road safety includes warning other road devices about collisions or dangerous situations. Autonomous driving additionally requires coordination of actions, for instance, to perform overtaking or platooning. Therefore, communication systems have to operate with communication ranges of up to 500 m and latency of less than 50 ms while ensuring high reliability. However, periodic traffic consisting of small packet sizes generated at the rate of 10 Hz leads to data rates of only 2 kb/s per device.

Traffic efficiency services aim to control traffic flows. In an urban environment, these include information on the status of traffic lights and the local traffic situation to accordingly allow adapting vehicle velocities at intersections. These services require a wireless infrastructure with communication ranges of up to 2 km and high reliability, but relaxed end-to-end (E2E) latency of less than 100 ms.

Professional Audio

The majority of today's professional audio links is built based on conventional analog transmission techniques in dedicated licensed frequency bands in the VHF and UHF ranges. Compared to digital transmission, analog transmission is spectrally inefficient and requires extensive frequency

planning. Hence, it is important to treat professional audio as a part of the future 5G IoT ecosystem as well. Professional audio applications, such as live concerts, also demand extremely low latency in the transmission links. It has been observed that trained musicians find latency exceeding 4 ms between sound generation (singing voice or instrument) and tonal perception (by means of monitor speakers or in-ear-monitoring) as disturbing and thus unacceptable. In a typical stage setup, the total round-trip latency budget of 4 ms is divided into three parts: the wireless link to the central mixing desk, the tonal processing in the mixer (typically 2 ms), and the wireless link back to the musician. Each of the two wireless links must therefore add no more than 1 ms latency while providing sufficient transmission reliability.

Latency Measurements for Current 4G Networks

Dedicated E2E latency measurements in public cellular networks, such as LTE, have disclosed two key limitations: the distance to the target device and the number of active devices per cell as shown in Fig. 1. Therefore, both limitations should be considered in future 5G networks to enable the low-latency use cases discussed above.

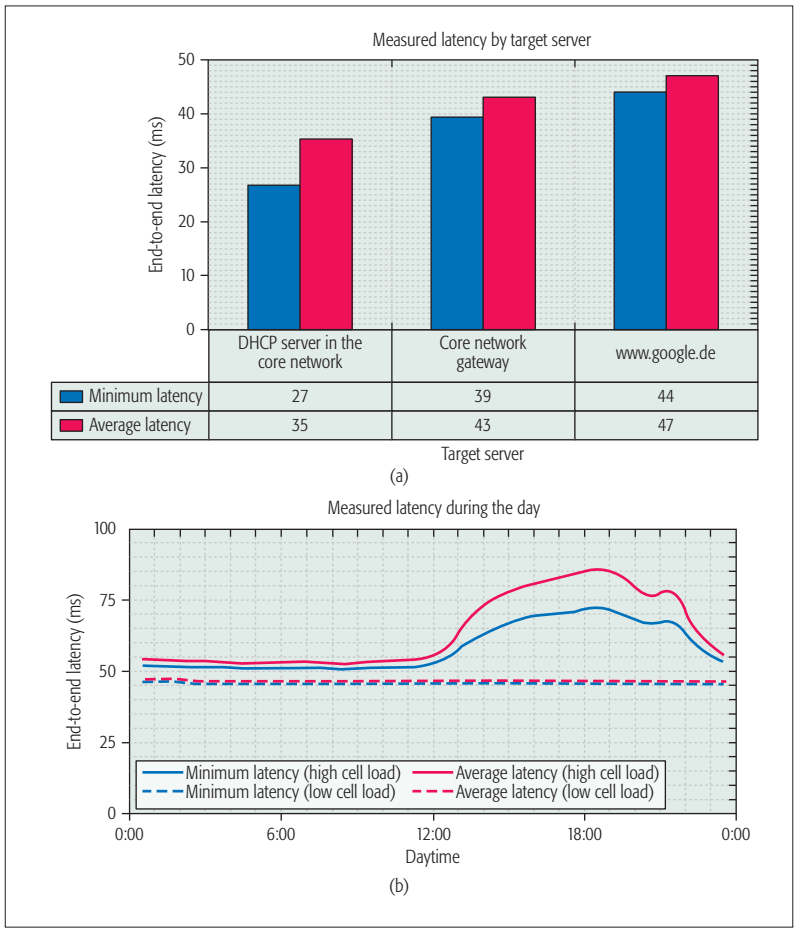


Figure 1. a) Measured minimum and average end-to-end latency (ICMP ping) for different target servers with an increasing distance from left to right; b) measured minimum and average end-to-end latency (ICMP ping) in a cell with low load (residential area) and a cell with high load (crowded market place with many active users). Both measurements have been made in a dense urban environment (Dresden, Germany, city center) for a low-mobility scenario with a proprietary application running on an Android smartphone.

First, we analyze the impact of the physical or virtual distance on the minimum achievable latency in an LTE network. As shown in Fig. 1a, E2E latency increases with a larger distance between the two endpoints. Moreover, a considerable portion of the overall latency budget is spent in the core network of the operator. For example, a minimum of 39 ms is necessary to contact the gateway of the core network toward the Internet, and only an additional 5 ms is needed to receive the reply from the Google server. This means that about 90 percent of the overall E2E latency originates from the cellular network. A second observation, made from Fig. 1b, is that the number of active devices per cell affects the achievable latency in cellular networks. In the measurement setup, two LTE cells are compared based on their minimum and mean latency during the day. The highly frequented LTE cell (high cell load) shows increased latencies, that is, the mean latency increases from 50 ms to 85 ms during the afternoon. This observation correlates with the increase in the number of active devices in the measured cell (local marketplace) during this time. For comparison, the reference cell with a low cell load is located in a residential area and shows an almost constant latency during the entire day.

These observations motivate the need for a carefully designed network architecture for latency critical IoT applications, and it is worthwhile to study the impact of placing the application close to the edge of the cellular network. Furthermore, the fundamental impact of a high number of active devices per cell on the E2E latency needs to be considered carefully.

The presented E2E latency measurements have been performed using conventional user equipment (i.e., an Android smartphone) connected to the public LTE network. The latencies are captured using the standardized ICMP procedure (layer 3 ping). Performing such dedicated ping tests is the first choice when measuring the latency of the communication link in the current systems. Nowadays, this latency measurement technique is widely used and gives valuable insights into the network performance. However, this method only gives snapshots of the actual link latency and may not represent the true latency of the communication link for a dedicated application during communication. In this regard, new solutions that enable monitoring of latency critical IoT applications need to be established. In addition, upcoming low-latency systems demand new methods to measure the latency at various levels inside the considered system, not only relying on the IP layer latency (e.g., measuring the scheduling latency of the operating system). In order to do this, timing information for events, function calls, interrupts, and so on need to be observed and assessed. Hence, the 5G network architecture enabling latency critical IoT applications has to provide interfaces to monitor the related key performance indicators (KPIs). This will allow the end user and the operator to analyze the quality of the provided service.

Concepts for Radio Interface

Latency-critical IoT applications demand changes in the current design of radio interface by schemes like resource delegation, fast uplink access, TTI shortening, and high reliability. In the Third Generation Partnership Project (3GPP), transmission time interval (TTI) is defined as the time required for the transmission of the smallest decodable data. Considering the default TTI size of 1 ms, that is, 14 orthogonal frequency-division multiplexing (OFDM) symbols, LTE Releases 8–13 do not efficiently utilize the available radio resources for small data sizes as in mission-critical IoT applications. This is primarily because the granularity of resources that can be allocated to a single device in LTE is too coarse, resulting in parts of the allocated resource being wasted. TTI duration also impacts the achievable user plane latency of the system. Therefore, changes are required in the current radio interface design to provide low user plane latency. In this section, we describe relevant enhancements for MAC and PHY layers to achieve low-latency communication for IoT applications that are not fulfilled by any of the existing wireless technology standards.

Resource Delegation Scheme

Future releases of LTE will support network assisted device-to-device (D2D) communication without directly involving the base station (BS) in data exchange between devices. The D2D communication paradigm not only allows reducing the communication latency between devices but also provides a possible solution to increase the resource utilization in the case of IoT applications. To achieve the latter, we propose a solution to delegate resources that are not needed by a device to another device that still has data to transmit. Partially or fully unused scheduling grants that were originally assigned in a dynamic or semi-persistent manner could be granted to other devices in their vicinity by leveraging from D2D communication, thus increasing the overall cell throughput. However, special care needs to be taken during D2D discovery to avoid additional access delays.

In order to allow the above mentioned secondary reuse of resources, the current LTE radio resource management (RRM) schemes need to be modified [4]. This can be achieved by splitting the RRM into a two-layer hierarchy. Thereby, the first level is managed entirely through the BS, similar to current LTE networks, whereas the second level is managed by the devices themselves. Figure 2a illustrates the approach. In particular, in the proposed RRM scheme devices are allowed to further delegate unused resources to other nearby devices, which we refer to as "sub-granting." We propose that a device which has unused resources available (i.e., the sub-grant provider) uses a small portion of its original grant to indicate to another device (i.e., the sub-grant beneficiary) that it can use the remaining portion of the grant. In order to convey the required sub-grant information, we propose to use an in-band signaling mechanism, for example, to use one OFDM symbol of a sub-frame for signaling. To minimize the overall latency of the scheme we propose to carry out the selection of sub-granting candidate pairs prior to the actual communication. For example, the BS could convey relevant information along with ordinary scheduling grants to potential sub-grant providers.

Our simulation results on this scheme show an uplink performance enhancement of 3–31 percent, depending on the number of users and sub-grant size (Fig. 2b).

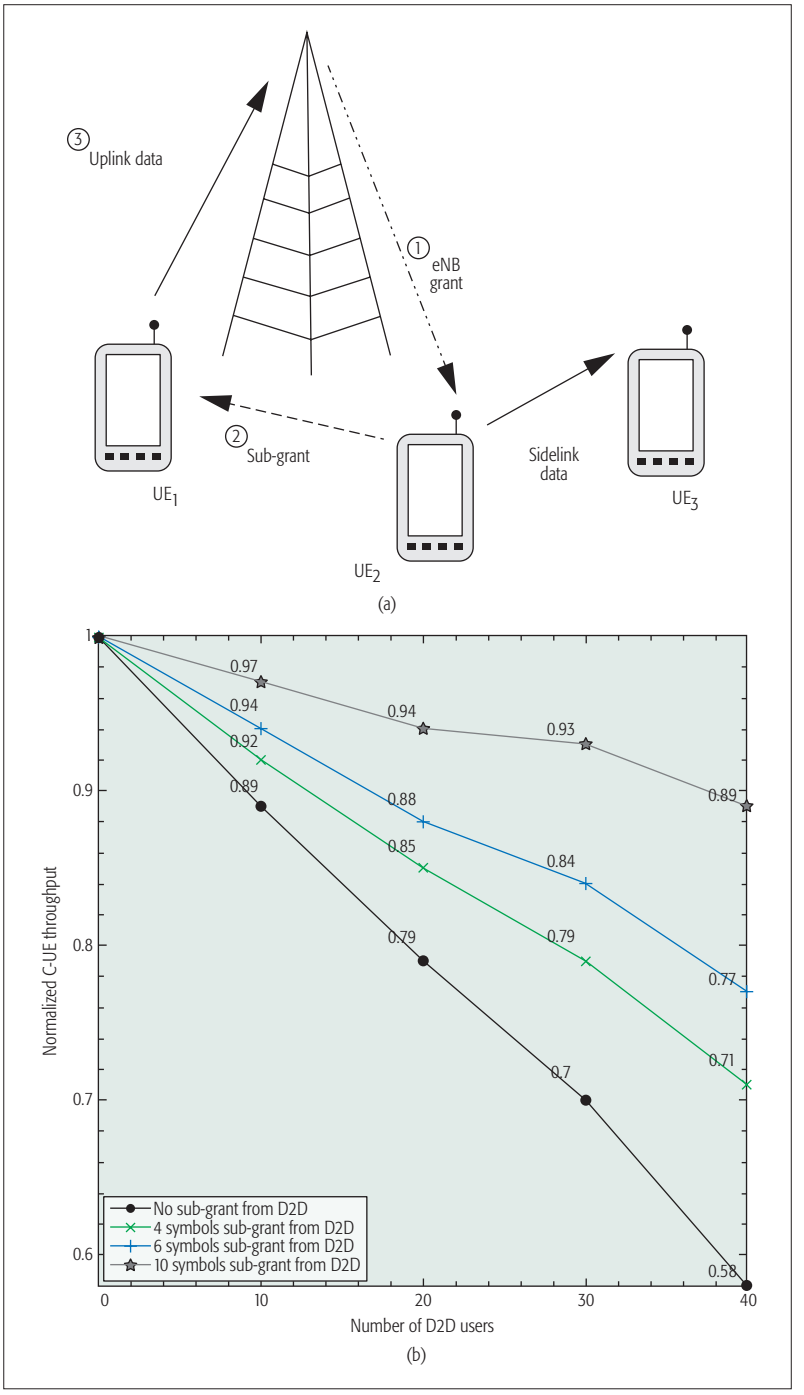


Figure 2. a) UE2 delegates its unused uplink resources to UE1; b) simulations show that in the investigated scenario with, for example, 20 devices delegating 4 OFDM symbols per sub-frame to some other devices, uplink throughput can be increased by 6 percent. More generally, the gain increases with the number of devices and the size of the sub-grant in this setting. Details may be found in [4].

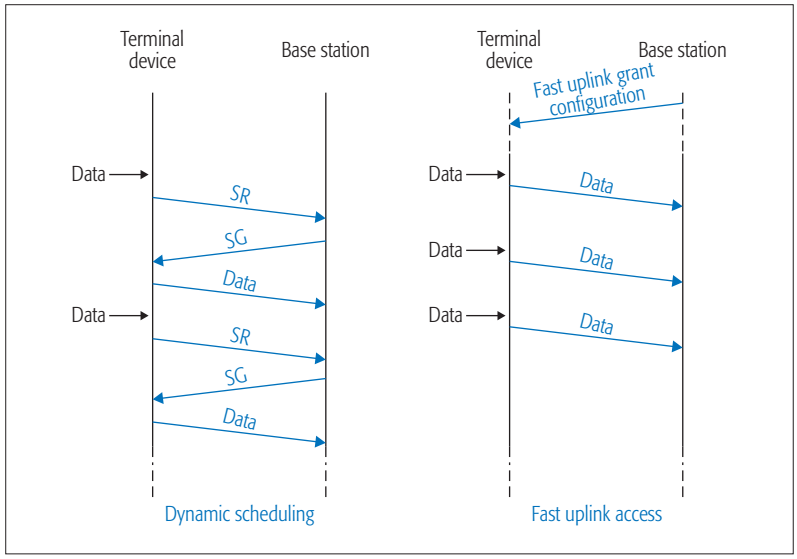


Figure 3. Message sequence chart for the dynamic scheduling and fast uplink access schemes. Dynamic scheduling imparts extra signaling delays due to the exchange of scheduling request (SR) and scheduling grant (SG) messages after data has arrived at the device. In fast uplink access, a BS configures the uplink resources in advance, and after the data arrives, it can be directly transmitted without explicit SR/SG exchange.

Fast Uplink Access

In LTE, a BS centrally coordinates channel access and RRM. The BS is able to efficiently carry out downlink transmissions as it itself manages the medium access. However, uplink transmissions using the default dynamic scheduling scheme impart extra signaling overhead, which leads to undesirable communication delays. As illustrated in Fig. 3, according to the LTE dynamic scheduling scheme, a device needs to send a scheduling request (SR) to the BS when data needs to be transmitted in uplink. The BS correspondingly allocates radio resources for the requested traffic and notifies them using a scheduling grant (SG). After receiving SG, the device is able to send its data in the assigned resources. With the default TTI size of 1 ms and the SR periodicity of 10 ms, the average uplink latency becomes 11.5 ms.

In LTE Release 13, the concept of fast uplink access has been proposed [5], which we advocate for the investigated latency critical applications (Table 1). In fast uplink access, the explicit signaling overhead of SR and SG is eliminated. Fast uplink access is based on semi-persistent scheduling [6], where resources are assigned to devices on an a priori basis. Data arriving at a device can directly be transmitted on the pre-allocated resources. When there is no data, devices do not need to send out the padding information. Using the default LTE TTI of 1 ms, fast uplink access can reduce the average communication latency to 4.5 ms, which is a significant improvement compared to the LTE dynamic scheduling. One slight drawback is lower capacity due to pre-allocation of resources. The PHY layer design features of short TTI can be complementarily applied to further minimize the overall communication delay.

TTI Shortening

Earlier, a resource delegation scheme was presented to increase the resource utilization for IoT use cases. Alternatively, TTI shortening, which not only allows low transmission latency but also increases resource utilization, is being investigated in 3GPP. Short TTI durations of 0.5 ms (7 OFDM symbols) and 0.14 ms (2 OFDM symbols) are being considered for LTE Releases 13–14 [7]. A shorter TTI duration also implies faster processing time needed for demodulation and decoding of data. Hence, we believe that a short TTI of 2 OFDM symbols is highly relevant for several latency critical IoT applications shown in Table 1, and can fulfill the latency requirements of most use cases. For instance, the average latency of using a TTI of 14 OFDM symbols along with the dynamic scheduling scheme for uplink and downlink transmission of 11.5 ms and 4.5 ms can be reduced to 2.36 ms and 0.93 ms, respectively, by restricting the TTI to only 2 OFDM symbols and assuming the reduced processing time.

However, the TTI shortening concept of Releases 13–14 is restricted by backward compatibility, which may lead to sub-optimum design for latency critical IoT applications. Most of the latency critical IoT deployments require relatively small coverage area compared to the LTE macro deployments. Therefore, LTE-based scaled numerology is being proposed for the new radio interface design of 5G [8]. Accordingly, LTE subcarrier spacing is either increased or decreased by an integer factor, which equally shortens or lengthens the OFDM symbol and cyclic prefix (CP) durations, respectively. However, a channel-dependent CP is required for robustness against inter-symbol interference (ISI) regardless of the subcarrier spacing. Especially for small-sized latency critical data with large subcarrier spacing (Table 1), the CP overhead becomes significant. It is desirable to have the least subcarrier spacing that meets the requirements of robustness against phase noise, Doppler spread, and latency without imparting unnecessarily large CP overhead.

Waveform Design

Waveform design of LTE can be enhanced to fulfill the requirements of 5G IoT use cases, where relaxed synchronization, efficiently supporting very small packet transmissions (Table 1), spectral confinement, time localization, and very low power consumption are of key importance [9]. Therefore, 5G waveform is to be chosen keeping in mind the relevant KPIs for a particular use case.

One viable option for 5G waveform design is configuring specific aspects of OFDM in order to meet the requirements of IoT use cases. Filtering or windowing adjacent bands (F/W-OFDM) leads to spectral confinement, allowing relaxed synchronization and facilitating asynchronous transmission of spectrally adjacent systems. Moreover, it increases the overall spectral efficiency by narrowing necessary guard bands.

In contrast, several new 5G waveform proposals challenge the orthogonality constraint of OFDM toward allowing relaxed synchronization and achieving spectral confinement [10]. For example, in generalized frequency-division multiplexing (GFDM) several short symbols are protected by a single CP, which keeps spectral efficiency even with very short symbols without compromising on the ISI robustness in long channels. In addi-

tion, wide subcarriers provide robustness against high Doppler spreads, and subcarrier-based filtering allows flexible spectral confinements.

Additionally, high peak-to-average power ratio (PAPR) is a common problem in multicarrier waveforms [9], which needs to be mitigated in order to achieve high power amplifier efficiency. Several techniques for PAPR reduction exist [11], but these typically reduce the overall spectral efficiency, especially in narrow-band allocations. Alternatively, allocating a single wide subcarrier to one device completely avoids the PAPR problem, but as a downside allows only low data rates per device.

Concepts for Network Architecture

Ultra-low latency cannot be achieved by improving only the radio interface design. The future 5G network will be based on software defined networking (SDN) and network functions virtualization (NFV) enabling a flexible and scalable architecture that can be adjusted to the needs of several use cases which run concurrently on the same infrastructure. Therefore, use case specific network slices [12] are introduced, which comprise appropriate subsets of network resources and settings. In particular, latency critical IoT use cases can benefit from local computational power provided by applications running in the mobile edge cloud (MEC) since this reduces the physical and virtual communication distance. This section introduces a mathematical tool to analyze such flexible network architecture and also provides insights on how IoT applications can benefit from such network architecture.

Flow-Level Modeling as a Tool to Derive Device-Centric KPIs

In addition to lower OSI-layer effects, there is another large impact on latency due to sharing of (radio) resources among a vast number of IoT devices. This impact is governed to a great extent by the spatial distribution of devices, their individual data rates, and their demand for network and radio resources. In particular, devices at the cell edge reduce network performance substantially due to their significantly lower spectral efficiency. Hence, device-centric considerations are required to guarantee a minimum latency for all devices in the network.

Flow-level models [13] are based on queuing theory and constitute useful tools to model and analyze the aforementioned effects on device-centric KPIs. Similar to the well-known SDN protocol OpenFlow, data traffic is investigated on the flow level rather than on the IP level. A data flow aggregates all information belonging to a transmitted object, which can be a sensor or control signal or a periodic message, depending on the IoT service type. Latency in this approach is then understood as sojourn time, that is, the time between the arrival of the information at the sender and when it is fully transmitted. Thus, flow-level models give a macroscopic view of the network that allows deriving statistics of device-centric KPIs, including the distribution of sojourn times, blocking probabilities, statistics on the fulfillment of service requirements, and so on.

As SDN/NFV and MECs are becoming increasingly important, investigations target evaluating flow performance at components at the edge of the network. Accurate modeling helps understand the underlying processes and evaluate existing and new concepts. Furthermore, it builds the foundation for the design of optimization algorithms, which can act on two different levels. First, there will be data and resource management within one network slice, that is, for its dedicated network elements, resource elements, and functionalities, servicing a certain application typically characterized by a particular traffic type. In addition, appropriate traffic or service models mimicking, on a macroscopic level, the applications at hand play a crucial role in appropriate algorithm design and performance evaluation. Second, on a higher level, the resources have to be allocated to each slice by a network orchestrator, ensuring the coexistence of the different applications on the same infrastructure. In the SDN/NFV context, resources are understood in a more general sense, comprising network elements, functionalities, or even radio access technologies (RATs), and thus may require more general allocation. Examples of how flow-level modeling can be applied on SDN/NFV architectures may be found in [14].

Service-Centric Analytics, Management, and Orchestration

State-of-the-art network analytics and management and orchestration (MANO) are almost exclusively deployed in cellular networks. They generally comprise multiple 3GPP-compatible radio access network (RAN) technology generations, typically incorporate a variety of equipment vendors, and go over into parts of the Evolved Packet Core (EPC) or IP Multimedia Subsystem (IMS). This includes already virtualized versions of EPC and IMS (vEPC and vIMS).

Figure 4 depicts the exemplary use case of a future integrated factory automation such as a car manufacturing plant: IoT-wearing components are manufactured at different production sites and need to be transferred between sites while being monitored at all times. In addition, the IoT components are processed by wirelessly connected sensor-actuator systems inside the production site. Heterogeneity of communication requirements, including low-latency parts all along this production cycle, is needed. This implies that services could run over a variety of radio technologies including 3GPP's cellular IoT technologies such as the future 5G RAN, Extended Coverage GSM (EC-GSM), LTE CAT-M, or narrowband IoT (NB-IoT), but also multiple WiFi flavors and non-cellular IoT technologies such as WirelessHART, ISA 100.11a, Bluetooth, Long Range WAN (LoRaWAN), and SIGFOX. Specifically, 5G RAN network slicing reveals the necessary shift from the conventional separation of the core network and RAN toward a network architecture that evolves the virtualization concept into parts of the RAN (virtualized RAN, vRAN). Initiatives such as the IEEE Next Generation Fronthaul Interface (NGFI), the Small Cell Forum, and 3GPP drive the specification of the required new fronthaul interfaces in-between. In addition, vRAN orchestration includes a dynamic and real-time capacity management, which can benefit from flow-level analysis, to follow capacity demands and traffic pattern over time

In addition to lower OSI-layer effects, there exists another large impact on latency due to sharing of (radio) resources among a vast number of IoT devices. This impact is governed to a great extent by the spatial distribution of devices, their individual data rates, and their demand for network and radio resources.

A service-centric MANO can rather be seen as an ecosystem of its own instead of just a new technology. Agility requirements suggest pure software solutions based on analytics and a full range of data science technologies as well as organically interfacing with a SDN/NFV network architecture as wide as possible.

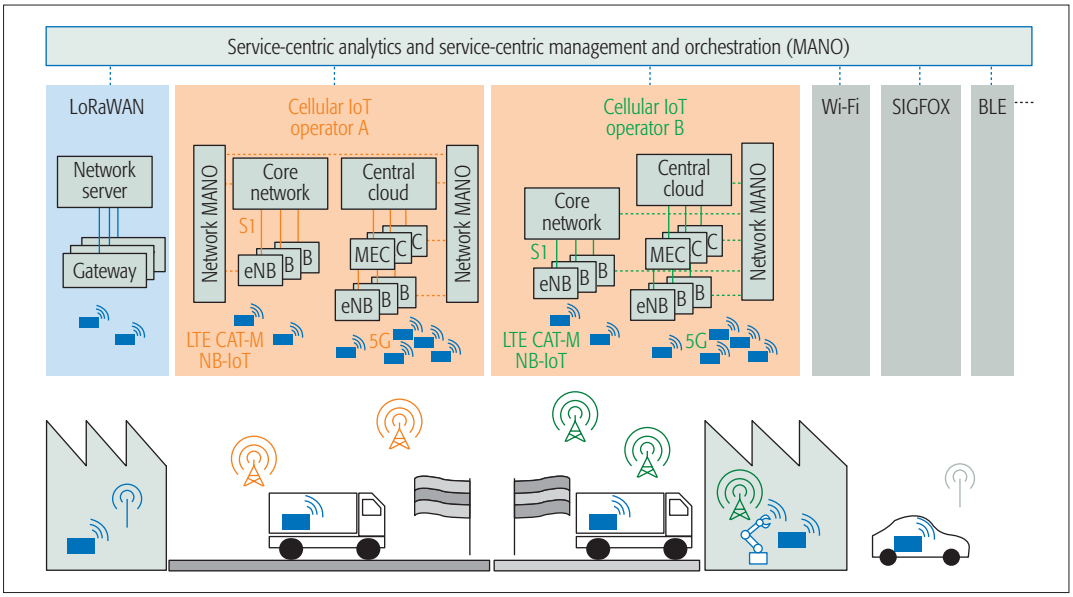


Figure 4. Service-centric analytics and service-centric management and orchestration at an exemplary IoT customer automobile manufacturer. IoT services during the production as well as over the entire life span of the end consumer product will run over a variety of radio access technologies.

and space over the various aforementioned air interface technologies as well as service-specific parameterization or vendor-agnostic orchestration of vRAN parts, at least in cellular vRAN implementations.

IoT services customers such as the exemplary car manufacturer that will have several underlying cellular operators as well as others' or its own proprietary network services under contract. Interoperability between these different networks and network technologies will be a key requirement. To support this, a future cellular network architecture should further comprise a common core network, thus enabling the revolutionary change toward service-centric analytics and MANO as well as at least common authentication, authorization, and accounting.

Considering this, a service-centric MANO will have to go far beyond a pure cellular network MANO to support a holistic view of the service on hundreds of thousands of things per service. Such a service-centric MANO will provide a vendor-agnostic view on the entire heterogeneous network and will have to cope with different life cycles of things in addition to the orchestration of the virtualized network infrastructure: The exemplary IoT customer car manufacturer may as well deploy IoT services such as ITS-related services after production, when the car is at the car dealer, and especially when the car is deployed by the end customer. Usually, network technologies develop faster than the life span of end consumer products. Consequently, a service-centric IoT SP has to maintain excellent IoT service while the underlying networks change.

Summarizing, a service-centric MANO can be seen as an ecosystem of its own rather than just a new technology. Agility requirements suggest pure software solutions based on analytics and a full range of data science technologies as well as organically interfacing with as wide an SDN/NFV network architecture as possible. Such a design would help to significantly reduce costs and to target new frontiers of integrated operational automation and agile introduction of new services in order to reduce time to market. A service-centric MANO platform needs to have capabilities that are independent of services and life cycles, as illustrated before. Furthermore, interleaving between the network and the real-time analytics could also boost further metadata-like driven business opportunities for the infrastructure owners and service providers leveraging the flow of information appearing all along the process of providing connections for an improved user experience and network efficiency.

Summary

Enabling latency critical IoT applications is one of the key targets for 5G. This article presents a comprehensive analysis of important low latency IoT use cases and their requirements on the underlying communication system. Analyzing current LTE PHY and MAC technologies and undertaking higher-layer latency measurements reveal that the requirements can only be met by introducing new radio interface design and novel network architecture concepts. In particular, we have described resource delegation schemes for D2D communication, new waveform candidates, fast uplink access schemes, and TTI shortening techniques. Moreover, a flexible network architecture incorporating SDN and NFV concepts will be able to adapt to different service requirements, where applications become less dependent on RATs or operators and follow a service-centric perspective. However, such a perspective demands concepts of network (self-) optimization and network orchestration, too, since the gained flexibility naturally comes along with increased complexity arising from the vast number of IoT use cases and their additional optimization constraints. One promising approach to efficiently control the increased complexity of service-centric management is through flow-level models. In

particular, the ability to describe networks, which serve different types of data traffic with diverse requirements, on a large scale analytically can help design and analyze SDN and NFV functionalities effectively.

From a service-centric perspective, MANO becomes increasingly important for IoT applications. Hundreds of thousands of devices will be connected through different RATs during their life cycles. Due to the prevalent environmental conditions, the application requirements, and development and changes in the underlying network, the best suitable RAT may vary over time. Therefore, interoperability between various RATs or even operators has to be guaranteed. Consequently, service-centric MANO has to provide a holistic, vendor-agnostic view of the entire network.

Acknowledgment

The work presented in this article was partly sponsored by the Federal Ministry of Education and Research within the program "Twenty20 - Partnership for Innovation" — "fast wireless."

References

[1] 5G PPP, "5G Automotive Vision," white paper, 2015.
[2] ETSI TR 102 889-2 v1.1.1, "Electromagnetic Compatibility and Radio Spectrum Matters (ERM); System Reference Document; Short Range Devices (SRD); Part 2: Technical Characteristics for SRD Equipment for Wireless Industrial Applications Using Technologies Different from Ultra-Wide Band," Aug. 2011.
[3] A. Frotzscher *et al.*, "Requirements and Current Solutions of Wireless Communication in Industrial Automation," *Proc. IEEE ICC Wksps.*, Sydney, Australia, 2014, pp. 67–72.
[4] D. M. Soleymani *et al.*, "A Hierarchical Radio Resource Management Scheme for Next Generation Cellular Networks," *Proc. IEEE WCNC Wksp. Device to Device Commun. for 5G Networks*, Doha, Qatar, 2016, pp. 416–20.
[5] 3GPP TR 36.881 v0.6.0, "Evolved Universal Terrestrial Radio Access (E-UTRA); Study on Latency Reduction Techniques for LTE (Release 13)," Feb. 2016.
[6] D. Jiang *et al.*, "Principle and Performance of Semi-Persistent Scheduling for VoIP in LTE System," *Proc. Int'l. Conf. Wireless Commun., Networking Mobile Computing*, Shanghai, China, 2007.
[7] Ericsson contribution to 3GPP TSG RAN WG1 Meeting #84, "System Level Evaluation Results for TTI Shortening Techniques," tech. rep. R1-161167, St. Julian's, Malta, Feb. 2016.
[8] Ericsson contribution to 3GPP TSG RAN WG1 Meeting #84bis, "Numerology for NR," tech. rep. R1-163227, Busan, South Korea, Apr. 2016.
[9] A. Zaidi *et al.*, "A Preliminary Study on Waveform Candidates for 5G Mobile Radio Communications above 6 GHz," *Proc. Wksp. 5G New Air Interface*, in conjunction with IEEE VTC Spring, Nanjing, China, 2016.
[10] G. Wunder *et al.*, "5GNOW: Non-Orthogonal, Asynchronous Waveforms for Future Mobile Applications," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 97–105.
[11] Y. Rahmatallah and S. Mohan, "Peak-To-Average Power Ratio Reduction in OFDM Systems: A Survey And Taxonomy," *IEEE Commun. Surveys & Tutorials*, vol. 15, no. 4, 2013, pp. 1567–92.
[12] NGMN Alliance, "5G White Paper," white paper, 2015.
[13] J. W. Roberts, "Traffic Theory and the Internet," *IEEE Commun. Mag.*, vol. 39, no. 1, Jan. 2001, pp. 94–99.
[14] K. Mahmood *et al.*, "Modelling of OpenFlow-Based Software-Defined Networks: The Multiple Node Case," *IET Networks*, vol. 4, no. 5, 2015, pp. 278–84.

Biographies

Philipp Schulz (philipp.schulz@ifn.et.tu-dresden.de) studied mathematics at Technical University (TU) Dresden, Germany, where he received his M.Sc. degree in 2014. There he also worked as a research assistant in the field of numerical mathematics, modeling, and simulation. In July 2015 he joined the Vodafone Chair Mobile Communications Systems at TU Dresden and became a member of the system-level group. Now his research focuses on flow-level modeling and the application of queuing theory on communications systems.

Maximilian Matthé (maximilian.matthe@ifn.et.tu-dresden.de) received his Dipl.-Ing. degree in electrical engineering from TU Dresden in 2013. During his studies he focused on mobile communication systems and communication theory. In his Diploma thesis he concentrated on waveform design for flexible multicarrier systems. Since 2013 he has been pursuing his Ph.D. in the Vodafone Chair Mobile Communication Systems at TU Dresden. His research focuses on the design and evaluation of MIMO architectures for future networks.

Henrik Klessig (henrik.klessig@ifn.et.tu-dresden.de) received his M.Sc. and Ph.D. degrees from the Department of Electrical and Computer Engineering of TU Dresden in 2012 and 2016, respectively. In 2011, he visited Alcatel-Lucent, Bell Labs, Germany, where he was engaged in LTE base station power modeling and energy-efficient resource management. Currently, he continues as a postdoctoral researcher at the Vodafone Chair at TU Dresden. His research interests include data traffic modeling, SON, ultra-low latency communications, and the tactile Internet.

Meryem Simsek (meryem.simsek@ifn.et.tu-dresden.de) has been a group leader at TU Dresden since 2014. She won the IEEE Communications Society Fred W. Ellersick Prize in 2015. Since June 2015, she has chaired the IEEE ComSoc Tactile Internet TSC and has initiated the IEEE P1918.1 Working Group. She joined ICSI Berkeley in October 2016. Her main research interests include fifth generation (5G) wireless systems, wireless network design and optimization, and the tactile Internet and its applications.

Gerhard Fettweis (fettweis@ifn.et.tu-dresden.de, F'09) earned his Ph.D. under H. Meyr at RWTH Aachen. After one year at IBM Research, San Jose, California, he moved to TCSI Inc., Berkeley. Since 1994 he has been Vodafone Chair Professor at TU Dresden, with 20 companies sponsoring his research on wireless transmission and chip design. He coordinates two DFG centers at TU Dresden (cfaed and HAEC), is member of the German academy acatech, and has spun out eleven start-ups.

Junaid Ansari (junaid.ansari@ericsson.com) is associated with Ericsson Research and contributes to 5G standardization. Previously, he worked as a postdoctoral researcher and research assistant at the Institute for Networked Systems at RWTH Aachen University. He received his Ph.D. (2012) and M.Sc. (2006) degrees from the same university. He has actively contributed to several collaborative national and European Union funded research projects. His research interests include embedded intelligence and system architecture design for the next generation of wireless networks.

Shehzad Ali Ashraf (shehzad.ali.ashraf@ericsson.com) is an experienced researcher at Ericsson Research, which he joined in 2013. He holds an M.Sc. in electrical engineering from RWTH Aachen University. Since joining Ericsson, he has been deeply involved in European Union and German government funded research projects related to the development of 5G concepts. Currently, he is also involved in 3GPP standardization. His research interests include 4G and 5G radio access technologies and machine-type communications.

Bjoern Almeroth (bjoern.almeroth@radioopt.com) is a data scientist at RadioOpt GmbH. He received his Ph.D. in electrical engineering in 2015 from TU Dresden. During his studies, he investigated the subject of analog-to-digital conversion in the context of multi-band signal reception. His current research focus is on the topic of quality-of-service and quality-of-experience monitoring in today's and upcoming 5G networks using crowd-sourced mobile customer experience measurements.

Jens Voigt (jens.voigt@amdocs.com) is currently with Amdocs Network Solutions in Dresden, Germany. His professional background includes radio access network analytics and optimization. He is passionate about technology, which he has proven in long-term university collaboration, successfully managed research projects, and product innovation. He holds a diploma (1995) and a doctoral (2001) degree in electrical and computer engineering from TU Dresden and is the co-author of 40+ scientific publications and multiple international patent families.

Ines Riedel (ines.riedel@amdocs.com) received her Dipl-Ing. (2006) and Dr.-Ing. (2014) degrees from TU Dresden. During internships at Sony, Germany, and Telefónica R&D, Spain, she was involved in software defined radio developments. From 2006 to 2014 she worked as a research associate and senior research associate at the Vodafone Chair at TU Dresden and joined Amdocs in 2014. Her research interests include radio access network analytics and optimization, and future network technologies.

Due to the prevalent environmental conditions, the application requirements, and development and changes in the underlying network, the best suitable RAT may vary over time. Therefore, interoperability between various RATs or even operators has to be guaranteed. Consequently, service-centric MANO has to provide a holistic, vendor-agnostic view on the entire network.

ANDRE PUSCHMANN (andre.puschmann@softwareradiosystems.com) received his Dipl.-Inf. and Ph.D. degrees in computer engineering from Technische Universitt Ilmenau, Germany, in 2009 and 2015, respectively. He is now with the CONNECT Centre for Future Networks and Communications in Dublin, Ireland, and also a senior engineer with Software Radio Systems Ltd. His research focuses on lower-layer radio protocols and resource management strategies for safety-critical applications, including vehicular networks and machine-to-machine communication.

ANDREAS MITSCHELE-THIEL (andreas.mitschele-thiel@tu-ilmenau.de) is a full professor and head of the Integrated Communication Systems group at the Ilmenau University of Technology, Germany. He received an M.S. in computer and information science from the Ohio State University (1989), and a doctoral (1994) and Habilitation degree (2000) in engineering from the University of Erlangen. He has held various positions in the telecommunication industry including at Lucent Bell Labs and Alcatel. His research focuses on the engineering of telecommunication systems.

MICHAEL MÜLLER (michael.mueller@ivm-sachsen.de) is the head of research of the Institut für Vernetzte Mobilitt gGmbH (IVM) and was the former chief of research and development of the MUGLER AG. His research interests are (mobile) communication networks and interconnected mobility. He has long research experience in physics and wireless networks as well as in the interdisciplinary study of the mobility sector.

THOMAS ELSTE (thomas.elste@imms.de) has studied computer science at Ilmenau University of Technology. Since 2005 he has been working at the Institut for Microelectronics and Mechatronics Systems GmbH (IMMS), Ilmenau, Germany, in the System Design Department, where his work is focused on kernel-level programming and application development for embedded systems running linux, real-time operating systems, and real-time network technologies for automation applications.

MARCUS WINDISCH (marcus.windisch@freedelity.com) is co-founder and CEO of Freedelity, a leading specialist for low-latency wireless communications systems with focus on professional audio applications. After Master's studies at the University of Madison, Wisconsin, he received his Dipl.-Ing. degree (2002) and his Ph.D. (2007) in electrical engineering from TU Dresden. He co-founded two start-up companies and holds several patents.

Effects of Heterogeneous Mobility on D2D- and Drone-Assisted Mission-Critical MTC in 5G

Antonino Orsino, Aleksandr Ometov, Gabor Fodor, Dmitri Moltchanov, Leonardo Militano, Sergey Andreev, Osman N. C. Yilmaz, Tuomas Tirronen, Johan Torsner, Giuseppe Araniti, Antonio Iera, Mischa Dohler, and Yevgeni Koucheryavy

ABSTRACT

mcMTC is starting to play a central role in the industrial Internet of Things ecosystem and have the potential to create high-revenue businesses, including intelligent transportation systems, energy/smart grid control, public safety services, and high-end wearable applications. Consequently, in the 5G of wireless networks, mcMTC have imposed a wide range of requirements on the enabling technology, such as low power, high reliability, and low latency connectivity. Recognizing these challenges, the recent and ongoing releases of LTE systems incorporate support for low-cost and enhanced coverage, reduced latency, and high reliability for devices at varying levels of mobility. In this article, we examine the effects of heterogeneous user and device mobility — produced by a mixture of various mobility patterns — on the performance of mcMTC across three representative scenarios within a multi-connectivity 5G network. We establish that the availability of alternative connectivity options, such as D2D links and drone-assisted access, helps meet the requirements of mcMTC applications in a wide range of scenarios, including industrial automation, vehicular connectivity, and urban communications. In particular, we confirm improvements of up to 40 percent in link availability and reliability with the use of proximate connections on top of the cellular-only baseline.

EMERGING INTERNET OF MOBILE, RELIABLE THINGS

The number of connected machine-type devices is expected to exceed 28 billion by 2021, thereby surpassing the number of human-centric connections significantly [1]. This fascinating development is a driving force behind the convergence of the physical and digital worlds that promises to create an unprecedented Internet of Things (IoT) market of US$19 trillion over the next decade [2]. Currently, a diverse range of IoT use cases includes intelligent transportation systems, smart grid automation, remote health care, smart metering, industrial automation and control, remote manufacturing, public safety surveillance, and numerous other applications [3].

When it comes to technical requirements, these diverse use cases are expanding the market to comprise "low-end" massive IoT applications as well as significantly more complex "high-end" solutions that may be labeled as critical IoT. At one end of the scale, in high-volume IoT deployments, there are smart sensors that report on a regular basis to the cloud infrastructure. At the other end of the scale are advanced critical IoT applications that have stringent requirements in terms of communications reliability, availability, and latency [4]. Accounting for the rapid growth pace of various IoT applications, the ultimate objective of enabling machine-type communications (MTC) technology is to construct comprehensive connections among diverse stationary and mobile devices, as well as other *things* across extensive coverage areas.

Today, this construction is being decisively attempted by the fifth generation (5G) of mobile networks, and the relevant standardization has recently started. Seamless connectivity support for mobility is particularly important for mission-critical MTC (mcMTC) devices moving at various speeds over a certain geographical area. Ironically, while mobility models have been routinely used in the evaluation of human-centric communications technologies, such as mobile ad hoc and legacy cellular networks, the effects of mobility in *mobile* 5G systems that are targeting the mcMTC market are much less understood [5].

In the context of ad hoc wireless networks, the impact of mobility on per-user throughput has been comprehensively characterized by [6]. For applications with loose delay constraints, where network topology may change over the timescale of single-packet delivery, the per-user throughput can increase dramatically when nodes are mobile rather than static. However, this important result may not be applicable in 5G networks that have strict latency budgets and where an infrastructure node arbitrates access to licensed spectrum resources.

Another line of research has established that employing multiple radio access technologies

The authors examine the effects of heterogeneous user and device mobility — produced by a mixture of various mobility patterns — on the performance of mcMTC across three representative scenarios within a multi-connectivity 5G network. They establish that the availability of alternative connectivity options, such as D2D links and drone-assisted access, helps meet the requirements of mcMTC applications in a wide range of scenarios.

Antonino Orsino is with University Mediterranea of Reggio Calabria and Tampere University of Technology; Aleksandr Ometov, Dmitri Moltchanov, Sergey Andreev, and Yevgeni Koucheryavy are with Tampere University of Technology; Leonardo Militano, Giuseppe Araniti, and Antonio Iera are with University Mediterranea of Reggio Calabria; Gabor Fodor is with Ericsson Research and Wireless@KTH; Osman N. C. Yilmaz, Tuomas Tirronen, and Johan Torsner are with Ericsson Research; Mischa Dohler is with King's College London and Worldsensing.

Digital Object Identifier:
10.1109/MCOM.2017.1600443CM

Given the decisive past progress in 3GPP to support MTC requirements and the ongoing efforts to define the NB-IoT radio interface, many massive MTC usage scenarios can already be accommodated by existing LTE releases. However, the enhancements promised by the new radio are needed in order to enable large-scale deployments of mcMTC applications.

and multi-access networks can drastically improve connection reliability, robustness to link failures, as well as utilization of spectrum resources in environments housing devices with diverse capabilities and quality of service (QoS) requirements [7]. Most recently, the integration of ad hoc and cellular technologies [8] facilitated novel applications of user-deployed and provisional wireless access points in the form of flying robots, or *drones*. This development has sparked new interest in understanding the benefits of multi-connectivity in the context of truly mobile access for IoT applications.

In this article, we aim to thoroughly quantify the impact of heterogeneous mobility in 5G networks that support mcMTC to enable advanced IoT applications. To this end, we consider a multi-connectivity system where devices can utilize cellular, direct device-to-device (D2D), and drone-assisted connections to communicate and access information. In particular, we focus on system-level performance characterization in representative 5G-grade IoT scenarios featuring mixed mobility patterns. The legacy LTE cellular systems are adopted as a benchmark where the devices are served over the infrastructure-based connections. These are compared to proximity services (ProSe)-based LTE solutions, where additional connectivity is made available by employing both D2D links between proximal devices and communications via "mobile" access points (i.e., drones).

Toward a Converged 5G-IoT Ecosystem

MTC Requirements and Challenges in 5G Standardization

The rapid proliferation in numbers and functionalities of IoT devdices has meant that the standards community is decisively advancing to outline the novel 5G mobile technology. To this end, the vision for the future development of international mobile telecommunications (IMT) and beyond was published [9], which presents the overall objectives and requirements for such next-generation systems. That document introduces three broad classes of usage scenarios with very different performance requirements:

- Enhanced mobile broadband
- Massive MTC
- Ultra-reliable and low-latency communications

The Third Generation Partnership Program (3GPP) is eagerly responding to this initiative by starting to ratify a new, non-backward-compatible radio technology in centimeter- and millimeter-wave spectra, and the early commercial deployments of this *new radio* technology are planned for 2018–2020. It is expected that 3GPP's *new radio* will be accompanied by further LTE evolution in parallel. Recognizing the benefits of cellular networks built around a global standards suite, the work in 3GPP includes technology components such as LTE Wi-Fi link aggregation (LWA), licensed assisted access (LAA), D2D communications to support smartphone relaying for wearables, power saving for MTC devices, MTC service enabling layers (oneM2M; M2M: machine-to-machine), as well as support for low-throughput and low-complexity MTC devices, realized both as a new LTE user equipment (UE) category (Cat-M1) and a new narrowband IoT (NB-IoT) radio interface in LTE Release 13 [10]. For MTC, these developments primarily mean a clear distinction between "massive" and "critical" usage scenarios, even though certain IoT applications may simultaneously belong to both categories (e.g., critical industrial alarms).

Given the decisive past progress in 3GPP to support MTC requirements (which started as early as 2005) and the ongoing efforts to define the NB-IoT radio interface, many massive MTC usage scenarios can already be accommodated by existing LTE releases. However, the enhancements promised by the *new radio* are needed in order to enable large-scale deployments of mcMTC applications, as their main demands are aligned along the lines of mobility (with speeds of up to 500 km/h) and latency (with end-to-end delays of under 1 ms) [9]. This support is crucial in order to promptly leverage the rich business opportunities around mcMTC as part of the 5G landscape, but it also poses many important system design questions.

In order to support more reliable consumer and industrial IoT applications [11], we envision that leveraging and integrating across the available heterogeneous access options, such as multi-radio uplink, downlink, and direct D2D link, will be crucial. The latter is particularly attractive, as the distinction between the network and the UE is becoming blurred, which offers excellent opportunities to utilize specialized UE as part of increasingly complex network tasks. This trend goes hand in hand with improved degrees of intelligence in the networked devices, from sensors, wearables, and UE to connected cars and mobile robots, which require very different levels of support for mobility, reliability, and spectrum management [12]. As cooperation between network and UE is becoming essential to improve performance of future IoT applications, the impact of more frequent handovers becomes a growing concern [13].

The promising market of connected — and soon self-driven — cars imposes unprecedented requirements in the form of extreme latency and reliability of data delivery at very high-speed mobility. This is particularly challenging given the fundamental fact that higher mobility naturally contradicts better reliability. As the respective operational models in the emerging vehicle-to-everything (V2X) business are taking shape, we need a comprehensive set of tools to handle the unconstrained mobility. This demand is particularly pressing since the impact of mobility has not been revisited in network architectures for a decade or so, and now, as we are entering a new era of converged 5G-IoT, is an appropriate time to understand and analyze the various implications of mobility on system performance, as well as to possibly rethink the ways of managing it in 5G networks.

Representative 5G-Grade mcMTC Scenarios

Today, the landscape of the global consumer and industrial IoT business is already extremely broad, stretching from wearable fitness trackers and health care devices to consumer electronics and connected cars. The most challenging study cases emerge in the form of crowded urban sce-

narios with very high connectivity demands, possibly under unreliable network coverage [14]. In addition to this, in environments with high-speed unrestricted mobility, the availability and reliability of wireless links are of primary importance to ensure strict service-level agreements in new markets around 5G-grade applications and services.

To address the performance of these connected machine-centric networks, we evaluate the relevant mixtures of realistic mobility models and study their effects on the availability and reliability metrics under partial cellular network coverage. In particular, we focus on three reference study cases that constitute representative 5G-grade mcMTC scenarios with very diverse application requirements (Fig. 1).

Industrial Automation (CASE A): Factories of the future will be something more than standalone "connectivity islands." There is, in fact, an ongoing trend to connect them as part of a broader industrial ecosystem. Accordingly, we consider the typical mobility aspects related to the supply chain processes within the factory itself or in proximity to its buildings. We focus on time constrained communications for the management of assets and goods as part of the on-site production and logistics sectors. This scenario becomes of interest since reliable management of the entire supply chain is crucial to avoid faults and improve the overall factory automation efficiency.[1]

Vehicular Connectivity (CASE B): Communications in the V2X study case comprise data exchange between a connected vehicle and:

- Another vehicle (i.e., vehicle-to-vehicle, V2V)
- A road infrastructure (i.e., vehicle-to-infrastructure, V2I)
- A personal device moving at pedestrian speeds (i.e., vehicle-to-pedestrian, V2P)

Here, the transmitted information can be periodic messages such as speed, positioning, and time related data needed to support critical safety and best effort entertainment applications, as well as offer efficient and comfortable driving experience. In this context, D2D interaction and "mobile" access points, including drones, may be of particular interest to achieve higher communications reliability and improved connection availability.

Urban Communications (CASE C): This study case covers a set of practical situations where a very large number of mobile end users, potentially carrying several wearable devices, are crowded together in locations including stadiums, shopping malls, open air festivals, and other public areas. The network infrastructure should in these cases be ready to accommodate high densities of active users and their connected devices with large amounts of aggregated traffic. Consequently, the key challenge in these environments is the provision of relatively reliable and available wireless connections to people moving according to certain pedestrian patterns, likely crossing areas with partial connectivity from the pre-deployed infrastructure network.

Mobility-Centric Perspective on mcMTC

Multi-Connectivity System Setup

Available Connectivity Options: In the subsequent evaluation, we adopt the legacy LTE solution as our benchmark where cellular infrastructure serves the mcMTC devices of interest. In addition, proximity-based D2D communications between the involved devices and drone-assisted mobile access are considered to augment the connectivity experience. This set of technologies, referred to as the ProSe-based LTE solution, leverages on D2D links whenever mcMTC devices have an opportunity to establish them in proximity (assuming a partner with the desired content), to improve the chances of reliably acquiring the relevant data.

In particular, drones that carry radio transceiv-

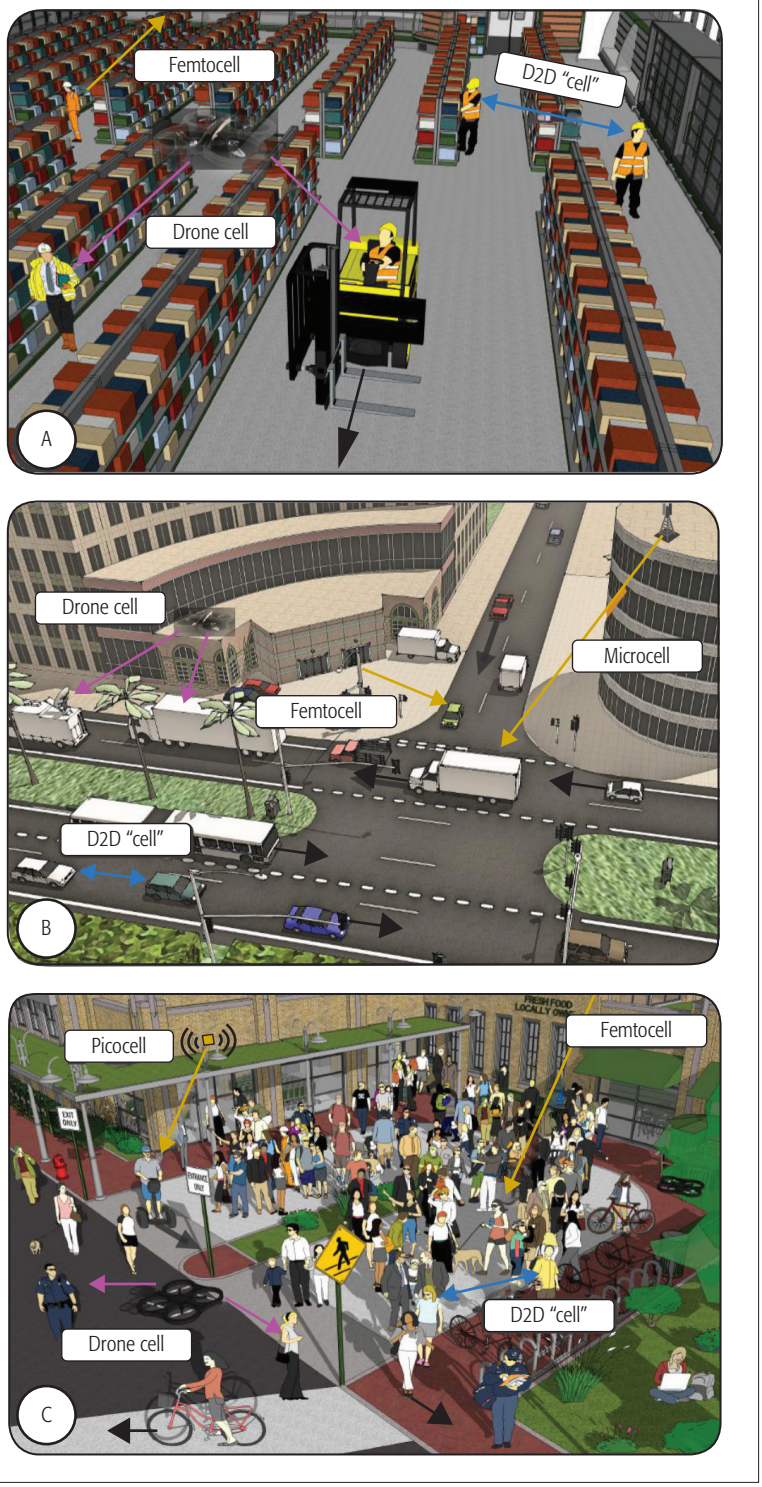


Figure 1. Characteristic 5G-grade IoT study cases.

[1] Note that connectivity between factory entities (e.g., robots, sensors, vehicles, workers) does not necessarily require ultra-low latencies, as response times are typically less constrained for humans than for machines.

Mobility model	Corresponding application	Brief description
Random walk	Short timescale movement of humans and vehicles	The random travel direction is uniformly distributed in [0, 2π]. The speed follows the distribution between the boundary values. After a constant time interval, a node computes new direction and speed for future movement.
Levy flight	Long timescale movement of humans and vehicles	Multiple short "runs" within a restricted area are interchanged with long-distance travel in a random direction.
Manhattan	Movement of vehicles and pedestrians in an urban environment	At each cross-road intersection, a node chooses to continue in the same direction with probability of 0.5, while turning left or right with equal probabilities of 0.25.
Reference point group	Mobility of drones	Each node follows the logical center (identified by the group leader). The nodes additionally have their own short timescale random walk mobility within the group.

Table 1. Utilized mobility models.

ers (i.e., drone small cells) are essentially *mobile* access points that provide better network coverage and bring higher data rates to challenging locations where LTE layout may be underprovisioned. Further, we assume that the D2D connection setup is managed by the LTE infrastructure for device discovery, session continuity, and security arbitration, whereas WiFi Direct links in unlicensed spectrum are selected as the actual D2D technology. Ultimately, mcMTC connectivity is considered to be *unreliable* in situations when:

- The device is outside of cellular LTE coverage.
- It has no opportunity to establish a D2D link with a relevant partner.
- The device cannot be served by a neighboring drone small cell.

Characteristic Mobility Models: To comprehensively assess the effects of realistic mobility in our three mcMTC study cases, we consider four mobility models and their heterogeneous combinations. These models have been carefully selected to capture both short- and long-term timescales of mobility as appropriate for the chosen study cases. While some of the models originally come from the realm of human mobility, we flexibly adopt them to become representative of mcMTC moving patterns. We briefly summarize the selected models in Table 1 and highlight their main features.

First, with the random walk (RW) model, the devices move from their current to the new target position randomly by appropriately choosing their speed and travel direction; this behavior aims to capture the short-term mobility on the scale of tens of minutes. Further, the Levy flight (LF) model is able to mimic movement patterns over a larger time span where mixed effects may be experienced. Another consideration is the Manhattan model, which is widely used to follow the mobility of vehicles in urban settings. Finally, the reference point group (RPG) model is particularly suitable to track the mobility of drones. To make it compliant with our scenarios, we assume that the drones follow a reference point, which is identified by the

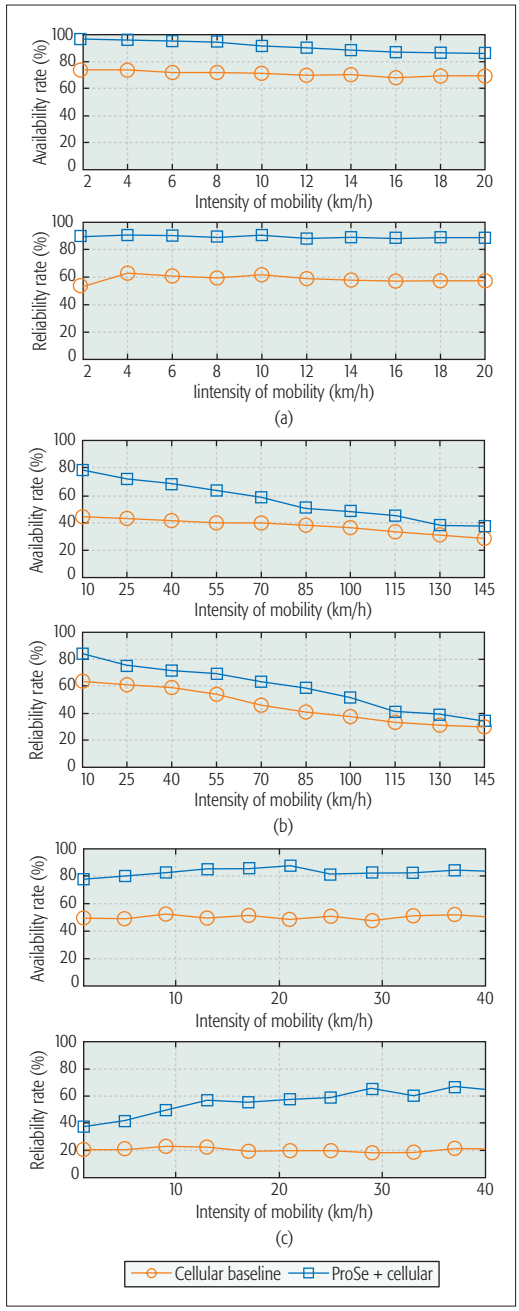


Figure 2. Analysis of system performance in terms of availability and reliability rate as a function of the average device speed in the considered study cases: a) industrial automation; b) vehicular connectivity; c) urban communications.

zone within the area of interest where the density of users is the highest. This setup allows the drone small cell to provide additional capacity and coverage in locations where large user densities may cause congestion and network overloads.

Deployment Parameters: We consider three mcMTC scenarios that reflect the industrial automation (A), vehicular connectivity (B), and urban communications (C) applications. In all study cases, the concerned devices acquire information over the link that offers them the highest data rate. We compare the following access technologies: legacy LTE cellular, WiFi Direct for the D2D links, and millimeter-wave (mmWave) over

licensed operator bands for drone small cells. For the mmWave technology, we select 28 GHz frequencies as a viable candidate for the 5G *new radio*, where the channel propagation, building penetration, and reflection parameters are adopted from [15].

In our three scenarios, we assume that the LTE coverage is *partial* within the modeled area, which corresponds to when the network is either underprovisioned (e.g., in rural regions) or serves challenging environments (e.g., with obstacles for signal propagation, e.g., walls or basement locations). We therefore consider that reliable cellular connectivity for the mcMTC devices in all study cases is only available over about 70 percent of the total area of interest based on deterministic modeling, since network coverage may be intermittent at the cell edges and beyond.

Further, the human users and networked mcMTC devices are allowed to move freely within the considered location according to their specific mobility patterns. For CASE A, the setup is represented as an indoor/outdoor area of [200,200] m where industrial robots and machines (i.e., in the indoor part) are first deployed uniformly and then move around within a range of 2 m at low speeds (i.e., around 1 km/h). The logistics related procedures are carried out by humans and vehicles where the corresponding mobility is modeled according to the RW model.

For CASE B, our setting is the area of [500,500] m where connected vehicles drive according to the Manhattan model. For more realistic simulations, we also add some background data traffic from pedestrian users. The latter are characterized by the LF mobility with α factor of 1.5. Finally, CASE C represents a crowded urban scenario where vehicles and users that carry a number of wearable devices are initially deployed within the area of [1000,1000] m and then move around. The respective mobility models are the Manhattan model for vehicles and the LF or RW model for humans (i.e., people prefer LF or RW pattern probabilistically) where the maximum speed of the nodes is limited by 20 km/h. For further information and details on the simulation settings, refer to Table 2.

		CASE A	CASE B	CASE C
Application parameter	Amount of data Inter-arrival time	300 kB 10 s	1500 B 100 ms	20 MB 1 s
System parameter	Cell radius	100 m	250 m	500 m
	Number of nodes*	100 M/30 H/20 V	450 V/50 H	300 H/650 M/50 V
	Density of nodes	0.75 node/m^2	1.0 node/m^2	1.0 node/m^2
	Mobility model	RW/Manhattan	Manhattan/LF	Manhattan/LF/RW
	Number of drones	5	10	10
	D2D range		50 m	
	D2D link setup		1 s	
	D2D target data rate		40 Mb/s	
	Drone altitude		[10–20] m	
	Drone speed		10 km/h	
	Drone mobility		RPG	
	Simulation time		30 min	
	Number of simulation runs		1000	

* M: machines, H: humans, V: vehicles.

Table 2. Simulation setup and parameters.

Selected Numerical Results

The reported performance assessment has been conducted with our custom-made simulator, named WINTERsim (WINTERsim system-level simulator: http://winter-group.net/downloads/, accessed August 2016). The main objective of this system-level analysis is to reveal the effects of heterogeneous device mobility in the 5G-grade mcMTC scenarios outlined above, as well as to quantify the contributions of various multi-connectivity options to the overall communications reliability.

Hence, the output metrics of our evaluation are:

- The *availability rat*e, that is, the proportion of users that experience certain connectivity, even though successful acquisition of all the desired data may not be guaranteed
- The *reliability rate*, defined as the actual data acquisition probability with which the mcMTC devices are able to successfully receive their data of interest
- The *impact of connectivity* options characterizing the relative shares of different multi-connectivity links, including cellular-, D2D-, and drone-based alternatives

With our system-level analysis, we are also able to evaluate other metrics of interest, such as the number of handovers between the available connectivity options, the handover delay, and the signaling load caused by unnecessary handovers.

The availability and reliability rates in the three scenarios under investigation have been simulated over a period of 30 min, and are summarized in Fig. 2. As we learn from these curves, for CASE B higher mobility speeds affect the system-wide performance considerably. By contrast, in the case of low mobility, the results do not vary dramatically, which holds for CASES A and C. For the vehicular scenario, D2D communications and drone small cells demonstrate diminishing benefits with the growing intensity of mobility (i.e., 100 km/h and beyond). However, the ProSe-based solution still offers consistent improvements on top of the legacy LTE baseline in all of the study cases. In particular, the gains in terms of the data acquisition rate vary from 25 to 35 percent for CASES A and C, as well as from 5 to 40 percent for CASE B. With respect to the reliability rate, we see an increase of 25 and 20 percent on average when considering CASES A and B, whereas the improvements for CASE C reach 40 percent.

The impact of alternative connectivity options in the studied scenarios is reported in Fig. 3. Interestingly, we conclude that D2D connections are utilized the most for mcMTC data acquisition. The explanation behind this fact is in the large number of potential contact opportunities for proximate users. However, with the growing intensity of

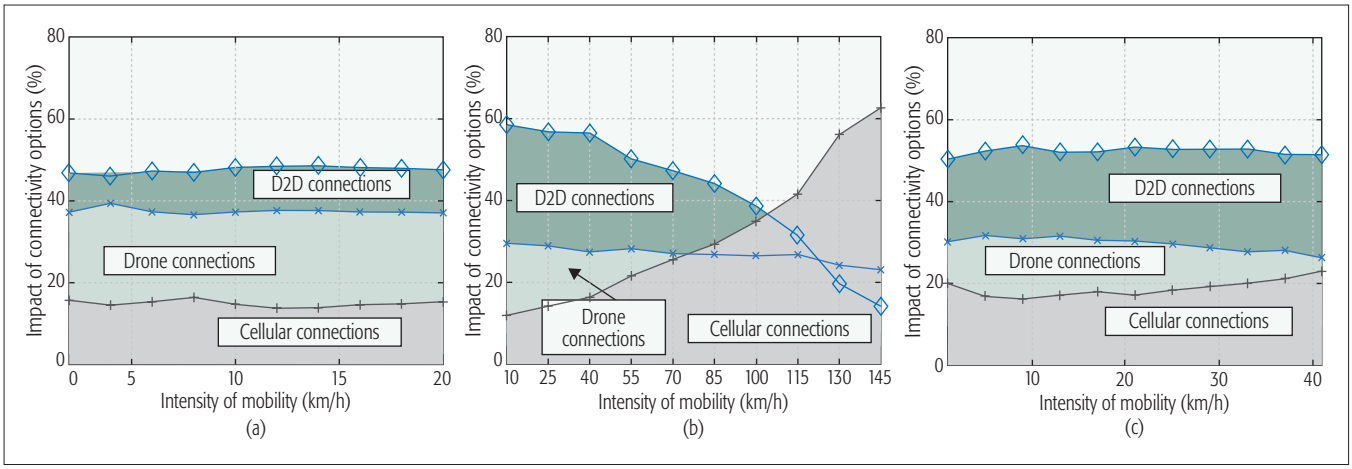


Figure 3. Impact of available radio access technologies on overall connectivity. The vertical axes display the contribution of each connectivity option: a) industrial automation; b) vehicular connectivity; c) urban communications.

mobility, the number of feasible contacts drops, and hence contact duration becomes the dominant factor that determines the chances of receiving the relevant content successfully. A similar trend is observed in Fig. 3b, where at speeds of above 85 km/h the devices prefer — by attempting to maximize their throughput — the more stable cellular LTE connections to any proximate links. In contrast, the impact of mobility is not as severe in Figs. 3a and 3c, where D2D- and drone-based links are used more often than the infrastructure-based connections.

In summary, for the scenarios with low (CASE A) and limited (CASE C) mobility, link availability and reliability may not be affected dramatically by the device mobility. However, this situation could change for other types of similar mcMTC applications having different packet sizes [16]. In these study cases, exploiting D2D- and drone-assisted communications leads to a significant improvement in the data acquisition rates as well as bringing along higher reliability. On the contrary, in the very different vehicular scenario where the intensity of mobility is typically higher (CASE B), we observe a considerable system-level performance degradation at speeds above 85 km/h. This is because proximity-based communications and drone small cells gradually lose their efficiency to provide additional capacity and coverage, while the only viable alternative remains to acquire data through the cellular infrastructure.

Toward Future Highly Reliable mcMTC

Conventional Performance Assessment in Wireless Systems

The three mcMTC applications considered, although remaining fully suitable for the purposes of our performance evaluation, are only examples of the potential wide diversity of emerging critical IoT scenarios. Broadly, focusing on their reliability and sustainable operation requires careful consideration and subsequent handling of many potential disruptions, such as interruptions in public services, data losses, remote control faults, and device malfunctions, among others. To analyze the occurrence probabilities of these unwanted events, the research community will need to develop a comprehensive set of tools suitable for monitoring and minimizing (or at least controlling) the probabilities of such rare events.

Today, a thorough performance assessment of modern wireless systems is per se a highly demanding task, which involves deep understanding of complex, cross-layer protocol structure as well as multi-layer composition of contemporary access networks. While analytical models tailored to a specific use case or functionality are somewhat successful in characterizing the system behavior in abstraction of many real-world factors, the resulting insights may appear to be rather limited in practice. As a consequence, analytical modeling of such complex networks often becomes overly complicated and feasible only in selected special cases.

In contrast, system-level simulations become the de facto method to assess the performance of complex multi-layer and heterogeneous wireless networks. Indeed, modern system-level simulator (SLS) tools require less abstraction work to capture the functionalities of the involved components compared to analytical assessment, while remaining efficient enough to deliver the results of required precision within a reasonable timeframe. Today, the latter is the main reason for the popularity of various SLS methodologies in both academia and industry to assess system key performance indicators (KPIs). However, even advanced SLS tools may have difficulty in quantifying the probabilities of rare events, which are essentially (short-lived) situations that occur infrequently.

Therefore, aiming to assess the system-wide performance of future mcMTC applications, additional efforts are needed to optimize the respective system-level simulations for modeling the targeted rare events. In particular, for 5G-grade IoT services, the levels of availability of a network node are required to be on the order of 10^{-8}–10^{-12}, while the packet loss probability has to be on the scale of 10^{-9}–10^{-12}. An adequate analysis of these KPIs in conventional SLS tools still remains an extremely challenging task due to prohibitive runtime.

Applicable Techniques for Modeling Rare Events

To obtain the probabilistic reliability and availability parameters while utilizing SLS methods within a reasonable time span, we propose to rely on

rare event simulation (RES) techniques. Depending on a particular RES implementation as part of the overall modeling framework, we differentiate between:

- Importance sampling (IS)
- Variance reduction techniques (VRT)
- Trajectory splitting (TS)

These methods have their roots in mathematical physics, where they have been successfully applied to characterize the events that occur within the range of 10^{-10}–10^{-20} [17]. Indeed, for the envisioned mcMTC requirement of "5 nines" (i.e., 99.999 percent) reliability, the number of events to be modeled has to be at least 10^7 or 10^9 for point and interval estimates, respectively.

Going further toward the next-generation ultra-reliable mcMTC systems, we anticipate the need to offer availability on the order of "9 nines" to control the probabilities of the underlying rare events. Accordingly, the number of modeling samples to reach the desired accuracy levels would increase to 10^{11} and 10^{13}, which is not feasible for modern SLS tools. Based on our literature analysis across various applied fields, the use of RES methods may decrease the modeling times by 2 to 6 orders of magnitude [17], as illustrated in Fig. 4. However, the application efficiency of all the RES techniques depends significantly on the type of the metric of interest as well as the complexity of the simulation scenario in question. For these reasons, significant modifications to the simulation logic have to be introduced in order to make the entire process transparent for a system designer, which needs to receive prompt research attention.

For completeness, even though the use of RES techniques requires considerable integration efforts to support them within the existing simulation pipelines, here we provide useful recommendations for including these methods as part of large-scale SLS methodologies:

- Large-scale SLS campaigns must be carefully designed first by applying simplified analytical or simulation models to understand the key qualitative trade-offs between the involved variables when identifying the most appropriate RES method to be used in final simulations.
- Conducting practical test trials is critical for the efficient implementation of RES techniques within a large-scale SLS as this should allow for assessing the performance of the chosen method as well as the expected time to complete complex simulations.
- A specialized, single-purpose SLS tailored to a certain set of target applications may not only be much more efficient, but also necessary for the performance assessment of future IoT systems, since reduced complexity also simplifies implementation of RES methods.

Conclusions, Key Lessons, and Perspective

In this article, we study the impact of heterogeneous mobility on connection availability and reliability in mcMTC scenarios when devices can use multiple connectivity options to establish wireless links. Since mcMTC are becoming an enabling technology in scenarios as diverse as industrial automation, vehicular connectivity, and urban communications, we utilize four appropriate mobility models to construct a realistic simulation environment around these scenarios. Our evaluation results demonstrate that with the increasing speed of movement, the availability of D2D links and drone small cells provide diminishing benefit over the LTE cellular-only baseline case. On the other hand, D2D connections and drone-assisted links are highly utilized and improve the availability and reliability of mcMTC data acquisition at low and moderate device speeds. In the vehicular scenario, the cellular infrastructure becomes the only viable communications alternative as speeds grow beyond 60 km/h.

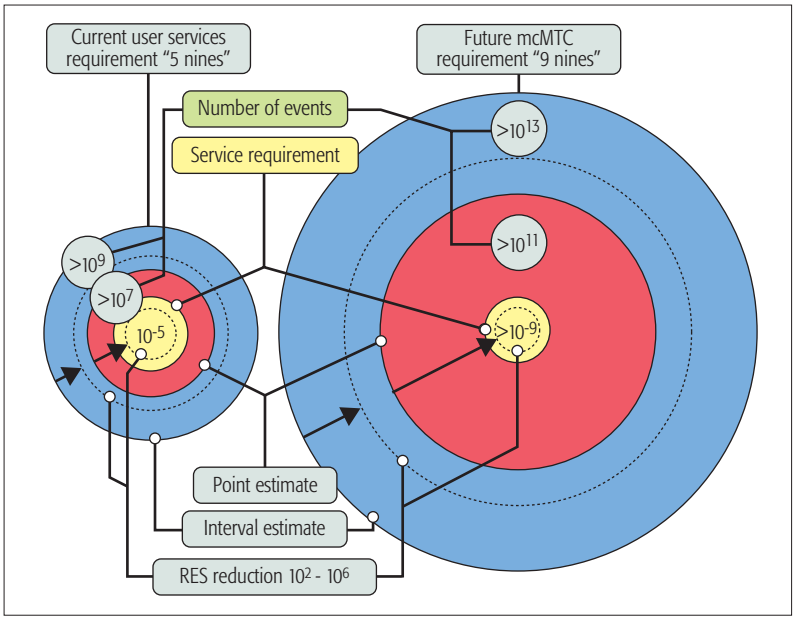


Figure 4. Target requirements for current and future mcMTC applications.

As a summary, the main contributions of this work are:

- Analyzing the impact of mixed mobility on system-level performance in three characteristic mcMTC scenarios
- Quantifying the contributions of various connectivity options for a realistic mix of mobility models
- For rare events, a review of relevant modeling techniques that are suitable for large-scale mcMTC simulations

Finally, the findings of this article have a strong impact on the landscape of IoT business, where it offers opportunities but also poses formidable challenges to overcome. Notably, it has become evident that the advocated heterogeneous multi-connectivity approach, including the use of D2D- and drone-assisted links, allows for meeting the stringent industrial control requirements at a relatively moderate capital expenditure (CAPEX). On the other hand, the operational expediture (OPEX) are likely going to be higher than in a traditional business. Furthermore, the heterogeneity of the equipment used, that is, fiber, masts, drones, licensed and license-exempt spectrum, and so on, poses operational challenges that will need to be addressed by the rapidly growing industrial IoT ecosystem.

The advocated heterogeneous multi-connectivity approach, including the use of D2D- and drone-assisted links, allows for meeting the stringent industrial control requirements at a relatively moderate capital expenditure (CAPEX) cost. On the other hand, the operational costs (OPEX) are likely going to be higher than in a traditional business.

Acknowledgment

This work was supported by the Academy of Finland and by the project TAKE-5: The 5th Evolution Take of Wireless Communication Networks, funded by Tekes.

References

[1] A. Ali, W. Hamouda, and M. Uysal, "Next Generation M2M Cellular Networks: Challenges and Practical Considerations," *IEEE Commun. Mag.*, vol. 53, no. 9, Sept. 2015, pp. 18–24.

[2] J. Bradley, J. Barbier, and D. Handler, "Embracing the Internet of Everything," Cisco White Paper; https://www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoE_Economy.pdf; accessed Aug. 2016.

[3] Ericsson, "Cellular Networks for Massive IoT," White Paper; https://www.ericsson.com/res/docs/whitepapers/wp_iot.pdf; accessed Aug. 2016.

[4] O. N. C. Yilmaz *et al.*, "Analysis of Ultra-Reliable and Low-Latency 5G Communication for a Factory Automation use Case," P*roc. 2015 IEEE Int'l. Conf. Commun. Wksp.*, June 2015, pp. 1190–95.

[5] J. Wu and P. Fan, "A Survey on High Mobility Wireless Communications: Challenges, Opportunities and Solutions," *IEEE Access*, vol. 4, 2016, pp. 450–76.

[6] M. Grossglauser and D. N. C. Tse, "Mobility Increases the Capacity of Ad Hoc Wireless Networks," *IEEE/ACM Trans. Net.*, vol. 10, Aug. 2002, pp. 477–86.

[7] G. Fodor, A. Eriksson, and A. Tuorinioemi, "Providing QoS in Always Best Connected Networks," *IEEE Commun. Mag.*, vol. 41, no. 7, July 2003, pp. 154–63.

[8] S. Andreev *et al.*, "Understanding the IoT Connectivity Landscape: A Contemporary M2M Radio Technology Roadmap," *IEEE Commun. Mag.*, vol. 53, no. 9, Sept. 2015, pp. 32–40.

[9] ITU-R M.2083-0, "IMT Vision — Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond," tech. rep., Sept. 2015.

[10] H. Shariatmadari *et al.*, "Machine-Type Communications: Current Status and Future Perspectives toward 5G Systems," *IEEE Commun. Mag.*, vol. 53, no. 9, Sept. 2015, pp. 10–17.

[11] G. Fodor *et al.*, "Device-to-Device Communications for National Security and Public Safety," *IEEE Access*, 2014, pp. 1510–20.

[12] L. Song *et al.*, "Game-Theoretic Resource Allocation Methods for Device-to-Device Communication," *IEEE Wireless Commun.*, vol. 21, no. 3, June 2014, pp. 136–44.

[13] F. Guidolin *et al.*, "Context-Aware Handover Policies in HetNets," *IEEE Trans. Wireless Commun.*, vol. 15, Mar. 2016, pp. 1895–1906.

[14] METIS II Deliv. D1.1, "Refined Scenarios and Requirements, Consolidated Use Cases, and Qualitative Techno-Economic Feasibility Assessment," tech. rep., Jan. 2016.

[15] T. S. Rappaport *et al.*, "Millimeter Wave Mobile Communications for 5G Cellular: It Will Work!," *IEEE Access*, vol. 1, 2013, pp. 335–49.

[16] A. Orsino *et al.*, "Direct Connection on the Move: Characterization of User Mobility in Cellular-Assisted D2D Systems," *IEEE Vehic. Tech. Mag.*, vol. 11, Sept. 2016, pp. 38–48.

[17] G. Rubino and B. Tuffin, *Rare Event Simulation Using Monte Carlo Methods*, Wiley, 2009.

Biographies

Antonino Orsino (antonino.orsino@unirc.it) received his B.Sc. degree in telecommunication engineering from University Mediterranea of Reggio Calabria, Italy, in 2009 and his M.Sc. from the University of Padova, Italy, in 2012. Currently, he is a Ph.D. student at the DIIES Department, University Mediterranea of Reggio Calabria, and a visiting researcher at Tampere University of Technology, Finland. His current research interests include device-to-device and machine-to-machine communications in 4G/5G cellular systems. He has served as a reviewer for several major IEEE conferences and journals.

Aleksandr Ometov (aleksandr.ometov@tut.fi) received his specialist degree in information security from St. Petersburg State University of Aerospace Instrumentation, Russia, in 2013. He has been a research assistant at the Department of Electronics and Communications Engineering of Tampere University of Technology since 2013. Currently, his major research interests are wireless communications, information security, heterogeneous networking, cooperative communications, and machine-to-machine applications.

Gabor Fodor (gabor.fodor@ericsson.com) received his Ph.D. degree in teletraffic theory from Budapest University of Technology and Economics, Hungary, in 1998. Since then, he has been with Ericsson Research, Stockholm, Sweden. He is currently a master researcher with specialization in modeling, performance analysis, and protocol development for wireless access networks. He has authored around 50 papers in reviewed conference proceedings and journals, and holds over 40 patents (granted or pending).

Dmitri Moltchanov (dmitri.moltchanov@tut.fi) is a senior research scientist in the Department of Electronics and Communications Engineering, Tampere University of Technology. He received his M.Sc. and Cand.Sc. degrees from Saint Petersburg State University of Telecommunications in 2000 and 2002, respectively, and his Ph.D. degree from Tampere University of Technology in 2006. His research interests include performance evaluation and optimization issues in wired and wireless IP networks, Internet traffic dynamics, quality of user experience of real-time applications, and traffic localization in P2P networks. He has served as a TPC member for a number of international conferences and has authored more than 50 publications.

Leonardo Militano (leonardo.militano@unirc.it) is currently an assistant professor at University Mediterranea of Reggio Calabria. He received his M.Sc. degree in telecommunications engineering in 2006 and his Ph.D in telecommunications engineering in 2010 from the University of Reggio Calabria. He was a visiting Ph.D student at the Mobile Device group at the University of Aalborg, Denmark. His major areas of research are wireless networks optimization, user and network cooperation, device-to-device communications, and game theory.

Sergey Andreev (sergey.andreev@tut.fi) is a senior research scientist in the Department of Electronics and Communications Engineering at Tampere University of Technology. He received his Specialist degree (2006) and Cand.Sc. degree (2009) from St. Petersburg State University of Aerospace Instrumentation, as well as his Ph.D. degree (2012) from Tampere University of Technology. He has (co-)authored more than 100 published research works on wireless communications, energy efficiency, heterogeneous networking, cooperative communications, and machine-to-machine applications.

Osman N.C. Yilmaz (osman.yilmaz@ericsson.com) has been working in global radio access system research and standardization projects since 2008, at NSN Research, Nokia Research Center, and Ericsson Research, respectively. In parallel with his research period in industry, he received his M.Sc. degree in telecommunications from Aalto University in 2010 and has been pursuing his Ph.D. degree since then. His research interests include self-organizing networks, heterogeneous networks, device-to-device communications, machine-type communications, and 5G in general. He is the inventor/co-inventor of 50+ patent families, as well as the author/co-author of 20+ international scientific publications and numerous standardization contributions in the field of wireless networks. He received the Nokia Top Inventor Award in 2013. He is currently working as a senior researcher at Ericsson Research Finland.

Tuomas Tirronen (tuomas.tirronen@ericsson.com) is a senior researcher at Ericsson Research, which he joined in 2012. He received his M.Sc. in teletraffic theory in 2006 from Helsinki University of Technology and his D.Sc. in communications engineering in 2010 from Aalto University. He is currently working on developing concepts and standards for 4G and 5G wireless technologies with focus on machine-type communications and the Internet of Things, performance evaluation, and energy efficiency. He has (co-)authored several conference and journal papers, and is involved in innovation work and patenting.

Johan Torsner (johan.torsner@ericsson.com) is a research manager in Ericsson Research and is currently leading Ericsson's research activities in Finland. He joined Ericsson in 1998, and has held several positions within research and R&D. He has been deeply involved in the development and standardization of 3G and 4G systems, and has filed over 100 patent applications. His current research interests include 4G evolution, 5G, and machine-type communication.

Giuseppe Araniti (araniti@unirc.it) is an assistant professor of telecommunications at the University Mediterranea of Reggio Calabria. From the same university he received his Laurea (2000) and Ph.D. degree (2004) in electronic engineering. His major areas of research include personal communications systems, enhanced wireless and satellite systems, traffic and radio resource management, multicast and broadcast services, and device-to-device and machine-type communications over 4G/5G cellular networks.

ANTONIO IERA [SM'07] (antonio.iera@unirc.it) graduated in computer engineering at the University of Calabria, Italy, in 1991, and received a Master diploma in information technology from CEFRIEL/Politecnico di Milano, Italy, in 1992 and a Ph.D. degree from the University of Calabria in 1996. Since 1997 he has been with the University of Reggio Calabria and currently holds the position of full professor of telecommunications and director of the Laboratory for Advanced Research into Telecommunication Systems. His research interests include next generation mobile and wireless systems, RFID systems, and the Internet of Things.

MISCHA DOHLER (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 *Wiley Transactions on Emerging Telecommunications Technologies* and *EAI Transactions on the Internet of Things*. He is a frequent keynote, panel, and tutorial speaker. He has pioneered several research fields, contributed to numerous wireless broadband, IoT/M2M, and cyber security standards, holds 12 patents, has organized and chaired numerous conferences, has more than 200 publications, and has authored several books. He acts as a policy, technology, and entrepreneurship adviser. He has talked at TEDx, and had coverage on TV and radio.

YEVGENI KOUCHERYAVY (yk@cs.tut.fi) is a professor and lab director in the Department of Electronics and Communications Engineering of Tampere University of Technology. He received his Ph.D. degree (2004) from the same university. He is the author of numerous publications in the field of advanced wired and wireless networking and communications. His current research interests include various aspects of heterogeneous wireless communication networks and systems, the Internet of Things and its standardization, as well as nanocommunications. He is an Associate Technical Editor of *IEEE Communications Magazine* and an Editor of *IEEE Communications Surveys & Tutorials*.

IoT Connectivity in Radar Bands: A Shared Access Model Based on Spectrum Measurements

Zaheer Khan, Janne J. Lehtomäki, Stefano Iellamo, Risto Vuohtoniemi, Ekram Hossain, and Zhu Han

To address the challenge of more spectrum for IoT connectivity, the authors propose an SA framework with rotating radars. The proposed framework is based on the results of the authors' measurement campaign in which they measured spectrum usage patterns and signal characteristics of three different ground-based fixed rotating radar systems near Oulu, Finland.

ABSTRACT

To address the challenge of more spectrum for IoT connectivity, this article proposes an SA framework with rotating radars. The proposed framework is based on the results of our measurement campaign in which we measured spectrum usage patterns and signal characteristics of three different ground-based fixed rotating radar systems near Oulu, Finland. In our work, we review different IoT protocols and their use of licensed or unlicensed spectrum. We make the case that IoT systems generate much data that cannot be accommodated with licensed/unlicensed spectrum, which already suffer from congestion. We identify the suitability of shared access between different rotating radars and IoT networks. We then present a zone-based SA framework in rotating radar spectrum for the operators providing IoT services, highlight its benefits, and also specify challenges in its implementation. To fully develop the considered zone-based SA method that ensures coexistence of IoT devices with no harmful interference to the rotating radars, we propose an REM-enabled architecture for the SA. The proposed architecture provides principles and rules for using the SA for the IoT, and it does not require modifications in the incumbent radar systems.

INTRODUCTION

The Internet of Things (IoT) is regarded as the next stage in digital communications with a wide range of applications, such as tasked sensors, controllers, smart metering, security systems, and industrial control [1, 2]. Communication is the "glue" that binds all the sensors, actuators, management platforms, and databases together to form the IoT. Wireless communications are the key to providing connectivity in the IoT, and as a result IoT can further congest wireless networks. For the regulators, this means freeing up more spectrum for wireless communications at a time when we are already running out of frequency spectrum [1, 2].

There is a plethora of new wireless technologies for IoT connectivity currently being developed. However, there is much uncertainty as to where spectrum might come from to efficiently support millions of connected devices once these technologies are deployed globally. The problem of spectrum scarcity in the wireless world has triggered regulators' interest in novel spectrum sharing mechanisms, which enable coexistence between distinct radio technologies and services. In terms of new spectrum sharing models, the potential use of shared access (SA) between radar and wireless communications systems has generated particular interest [3]. One reason for this interest is the fact that communications systems and radar systems jointly consume most of the highly desirable spectrum below 6 GHz [4, 5]. However, different works and reports have shown that existing sharing models either do not take into account the real spectrum usage of radar systems or are often counter-productive to the goals of spectrum sharing in the radar bands [5].

To address the challenge of more spectrum for the IoT, in this article, we present results of our measurement campaign and identify the suitability of frequency spectrum used by the rotating radars for providing connectivity to the IoT (sensors, actuators, and gateways) on the basis of an SA framework. Our contributions in this article include the following:

- First, we present results of our measurement campaign in which we measured spectrum usage patterns and signal characteristics of different ground-based fixed rotating radar systems in Finland.
- Based on the measured/analyzed features of three different radar systems, we identify the suitability of measured frequency spectrum for providing connectivity to the IoT on the basis of SA. To the best of our knowledge, our study represents the first evaluation of more spectrum for the IoT under SA in the radar spectrum.
- We propose the use of a radio environment map (REM) architecture as an enabler to provide SA to the IoT networks in frequency channels used by different rotating radar systems. REM is a cognitive tool that can be utilized to enhance the awareness of IoT entities of their operational radio environment [6].

It is important to note that our work in [7]

Digital Object Identifier:
10.1109/MCOM.2017.1600444CM

Zaheer Khan is with the University of Liverpool and was also with the University of Oulu; Janne J. Lehtomäki and Risto Vuohtoniemi are with the University of Oulu; Stefano Iellamo is with the Institute of Computer Science at FORTH; Ekram Hossain is with the University of Manitoba; Zhu Han is with the University of Houston.

0163-6804/17/$25.00 © 2017 IEEE

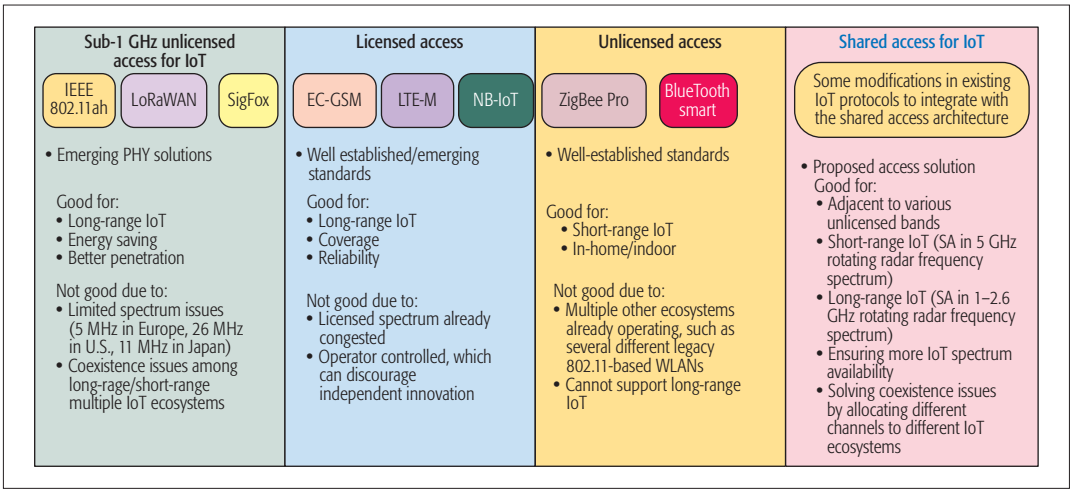


Figure 1. Comparison of different wireless connectivity schemes for IoT and their spectrum usage.

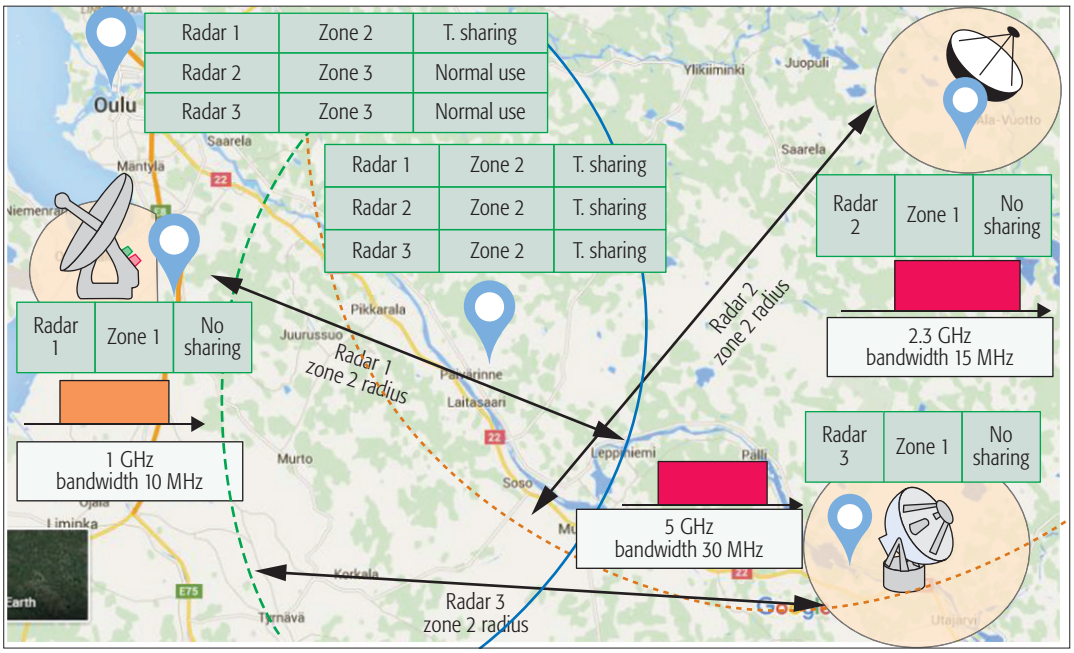


Figure 2. Approximate locations, spectrum band utilization, and bandwidth of each of the utilized channels, and different sharing zones at five different locations (blue location markers) for each of the three different radar systems measured by the authors. "T. Sharing" means temporal sharing.

presented measurement results for the spectrum usage of a weather radar in the 5 GHz band. Different from [7], in this work we present results for three different rotating radar systems. Each of the measured radars is used for a different application, operates in a different spectrum band, and has channel bandwidth utilization between 10 to 30 MHz; see Fig. 1 for illustration. Moreover, different from [7], we identify the suitability of providing SA for IoT in the frequency channels used by rotating radar systems, and also propose the use of an REM architecture as an enabler to provide SA.

The rest of the article is organized as follows. We overview different IoT wireless technologies and their use of radio spectrum. Then we present results of our measurement campaign and also provide the reasons for the suitability of SA for IoT. To this end, we present a REM-based SA framework before we conclude our work.

Different IoT Wireless Technologies and Their Use of Radio Spectrum

Typically, IoT can generate different spectrum demands. In terms of spectrum usage, in general, there are three alternative tracks for IoT services:

- Licensed spectrum
- Unlicensed spectrum
- SA spectrum

In Fig. 2, we highlight and compare the different spectrum usage approaches for IoT that are either currently being used or under consideration for use to meet the needs of different types of IoT services.

Licensed Spectrum and IoT

Cellular networks operate on licensed spectrum and are rapidly evolving with new functionalities to form an attractive solution for emerging low-power wide-area IoT applications. NB-IoT is a narrowband radio technology specially designed

Cellular networks operate on licensed spectrum and are rapidly evolving with new functionalities to form an attractive solution for emerging low power wide area IoT applications. NB-IoT is a narrowband radio technology specially designed for the Internet of Things (IoT) and can be deployed in GSM and LTE licensed spectrum.

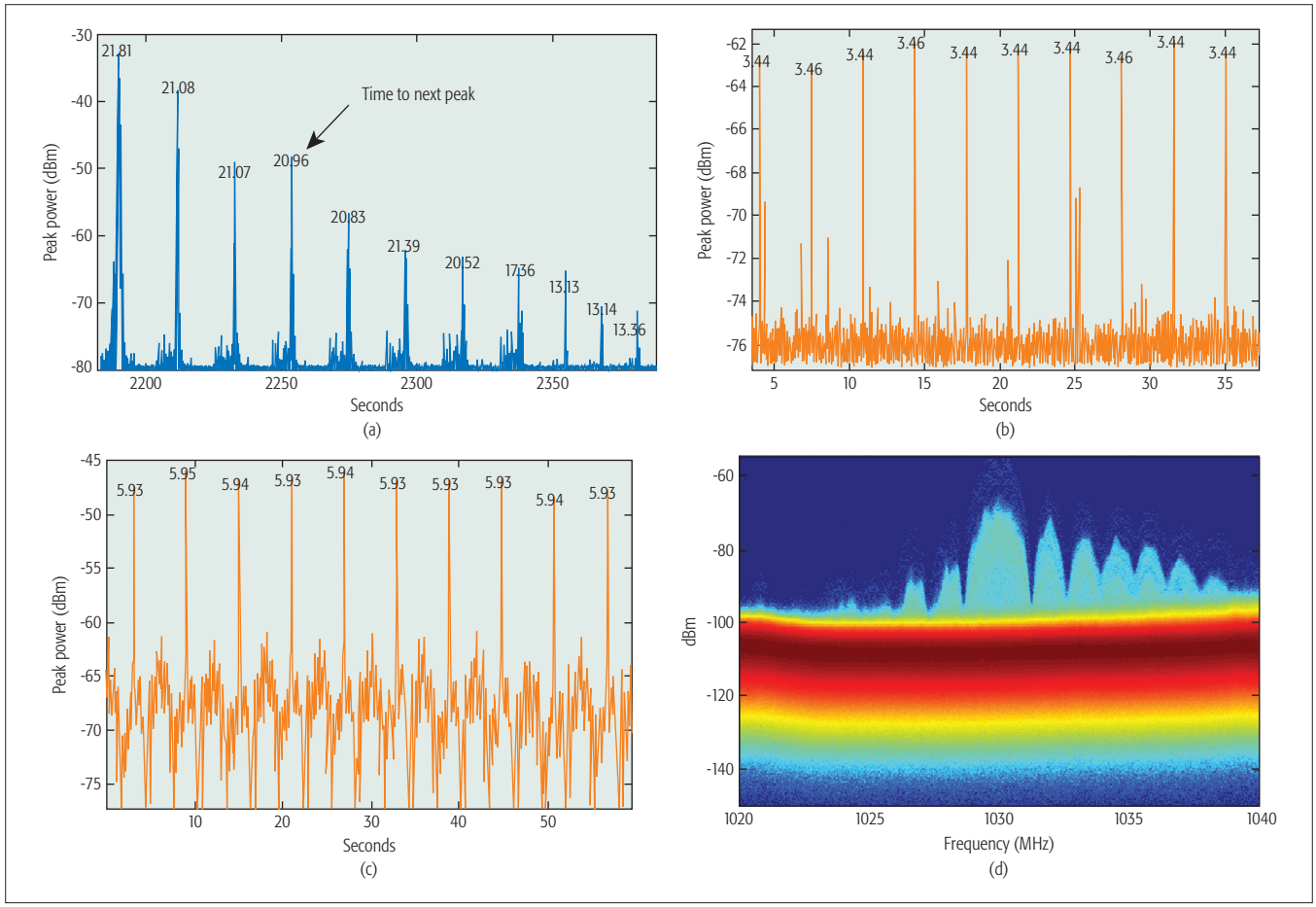


Figure 3. Example measurement results showing the measured times between the main beam peaks of the three rotating radar systems, and the logarithmic two-dimensional spectrograms of the recorded power values of the airport surveillance radar signals.

for IoT and can be deployed in GSM and LTE licensed spectrum. Ericsson and Orange are testing EC-GSM (Extended Coverage GSM) using the 900 MHz licensed band, with the aim to enhance device reachability by up to 20 dB or seven-fold improvement in the range of low-rate applications. This extends the coverage of GSM to reach challenging locations such as deep indoor basements, where many smart meters are installed. LTE for machine-to-machine (LTE-M) is another cellular IoT solution that utilizes the licensed spectrum and is based on LTE [8].

Unlicensed/Sub-GHz Unlicensed Spectrum and IoT

In 2.4 GHz and 5 GHz unlicensed spectrum, Bluetooth Smart is a modified Bluetooth-based protocol for IoT applications that require low-power connectivity over short ranges of typically within 200 m. ZigBee PRO and ZigBee Remote Control are based on the IEEE802.15.4 protocol and use unlicensed spectrum for access. It target applications that require relatively infrequent data exchanges, low data rates, and coverage of within a 100 m range such as in a home or building.

The WiFi alliance is working on a new IEEE 802.11ah standard that can manage low-power wide area network (LPWAN) connectivity for IoT devices. IEEE 802.11ah intends to operate over a set of unlicensed radio bands in the sub-1 GHz unlicensed band. Some of the prominent features of the new IEEE 802.11ah are its energy saving mechanisms; its use of spectrum below 1 GHZ ensures wider coverage for LPWAN IoT. To power IoT with new communication solutions, independent IoT network groups have devised two different solutions for LPWANs, which are called SigFox and LoRaWAN. SigFox is a narrowband technology and uses a standard radio transmission method called binary phase shift keying (BPSK). LoRaWAN looks at a wider amount of spectrum than SigFox [9]. Both LoRa and SigFox are planned to share spectrum with other solutions in the sub-1 GHz license-exempt bands.

Shared Access Spectrum and IoT

The total demand of thousands of IoT devices in a given area using heterogeneous access protocols can have a significant effect on future radio spectrum use. The number of IoT devices and the nature of traffic will thus require far more frequency spectrum than is commercially available for them today. Given that radar bands are now also a potential candidate for sharing between wireless communication systems and radar systems [1, 4, 5], in the context of making more spectrum available, SA in radar spectrum for IoT is an important and useful idea.

In the next section, based on measurements of spectrum usage of different rotating radars, we describe the suitability of rotating radar channels for wireless connectivity of IoT devices.

Measurement Results, Zone-Based SA, and the Suitability of SA for IoT

Measurement Strategy, Setup, and Results

The rotating radars that operate in different bands have highly directional rotating antennas and provide coverage of applications over a large area (e.g., they can have a range of 150–200 km). The presented measurement results in this section include spectrum usage behavior of three ground-based fixed rotating radar systems in Finland: a weather radar system in the 5.6 GHz band, uplink of an airport aircraft surveillance radar system in the 1.03 GHz band, and a surveillance radar in the 2.3 GHz band. These radars transmit a narrow beam and perform more listening than talking. For example, a weather radar may emit a pulse for 2 μs, then listen for approximately 2 ms. They rotate to horizontally scan 360°, and some of them also tilt vertically. In Fig. 1, we illustrate approximate locations, spectrum band utilization, and the utilized channel bandwidths for each of the three different radar systems measured by the authors near the city of Oulu. Measurements were performed with an Agilent N6841A RF sensor connected to a wideband, omnidirectional antenna (ARA CMA-118/A) [10]. For both surveillance radars, the measurements were based on recording continuous (no time domain gaps) stream of I/Q samples. The sampling rate was (depending on the scenario) 2 MHz or 10 MHz, leading to the minimum time resolution of 0.5 ms. For the weather radar case, measurements were based on recording a continuous stream of fast Fourier transform (FFT) processed outputs (with 20 MHz sampling rate). Each measurement duration was more than 50 min at each location.

In Figs. 3a–3c, we present the measured times between main beam peaks of the three rotating radar systems operating in the three different spectrum bands. The three figures also illustrate the received peak power as a function of time in seconds. It can be seen in the three figures that there are pauses in the received signal from the radar due to its antenna rotation. When the rotating radar's main scan beam points to the measurement locations, a signal peak is received. It can also be seen from the figures that while the radar's pulse interval (i.e., the time between two consecutive pulses received at the same location) are constant (Figs. 3b and 3c) for the measured surveillance radars, they are not constant for the weather radar. The pulse intervals of the surveillance radars in Figs. 3b and 3c are periodic with pauses of 3.44 and 5.93 s between the scan pulses; however, the pulse intervals of weather radar are quasi-periodic with pauses between the scan pulses that vary from 13.1 to 21.1 s. This is due to the fact that the measured radar has two scanning modes: the normal-mode with pulse repetition frequency (PRF) 570 Hz, pulse duration 2 ms, rotation speed 16.9°/s, lowest elevation angle 0.3°; and the dual-mode with dual-PRF 900/1200 Hz, pulse duration 0.8 ms, rotation speed 26.7°/s, lowest elevation angle 0.4°.

Figures 3a–3c also show that while the received peak power for the two surveillance radars does not vary significantly, for the weather radar, the received power varies over a period of time. The reason for this received peak power variation is that unlike the two surveillance radars, the weather radar horizontally scans 360° at different vertical angles. For the weather radar, the highest received peak power in Fig. 3a is obtained when the radar directs its beam downward to the measurement location. In Fig. 3d, we present logarithmic two-dimensional spectrograms of the recorded power values of the airport surveillance radar signals.

SA in Radar Spectrum

Our measurement results show that there are pauses in the received signal from each of the three rotating radars due to their antenna rotation (Figs. 3a–3c). This offers the potential of low-power IoT devices to use zone-based SA in the radar bands [7, 11]. Different from this work, the potential of zone-based SA in the context of small cell networks has been explored by [11, 12].

Previous works [4, 5] have proposed the use of large exclusion zones in which the radii of exclusion zones vary, depending on the specific site, between 72 and 121 km. This model may guarantee 100 percent protection to the radars; however, different works and reports have shown that large geographic exclusion zones are unnecessary and counter-productive to the goals of spectrum sharing in the radar bands. In our work, the three different zones around a radar are modeled as follows. At a distance of a few kilometers from a rotating radar system, a network of sensor devices are deployed in a circle around the radar station. Our measurement results show that for the measured radars, approximately within 2 to 3 km, even the sidelobe signal from the radar can be strong enough to interfere with the wireless communications. Thus, it is suitable for both the radar system and the wireless communication systems to have the radius of Zone 1 of approximately 3 to 4 km. Due to this reason we call Zone 1 the exclusion zone (Zone 1) as any secondary transmissions are forbidden in this zone. The minimum distance from which the aggregate received signal strength at the radar does not exceed a minimum threshold value (defined by a regulatory body) can be used to establish the starting point for the Zone 3 region. This distance needs to be calculated by a regulatory body using extensive measurement campaigns. In Zone 3, the users are free to use the spectrum as they are outside the interference area of the radar. Finally, between Zone 1 and Zone 3 is Zone 2. In Zone 2, only temporal sharing is allowed in which the network is not allowed to transmit during the time when the radar's main beam is pointing at it, and is also not allowed during the guard interval before and after that time period. To avoid any possible interference with the side lobes of a radar, when sensor devices detect aggregate received signal strength exceeding a critical threshold value (defined by a regulatory body), they notify the REM repository, which in turn instructs the IoT gateways to move some of their users to another channel to avoid any possibility of interference. It is easy to see that compared to the GEZ models

The minimum distance from which the aggregate received signal strength at the radar does not exceed a minimum threshold value can be used to establish the starting point for the Zone 3 region. This distance needs to be calculated by a regulatory body using extensive measurement campaigns.

Features	Benefits for the operators providing IoT services	Challenges in implementation
Rotating radars operate in various higher and lower frequency spectrum.	To avoid inter-system interference. operators can connect long-range networks for IoT in lower frequency spectrum and short-range IoT systems in higher frequency spectrum.	Careful design of SA connectivity for the multitude of IoT ecosystems and their appropriate frequency spectrum selection.
Heterogeneous sharing zones due to distinct locations of radar systems.	An operator can allocate networks of delay-tolerant IoT devices to the shared spectrum of Zone 2 radar and delay-intolerant IoT devices to the shared spectrum of Zone 3 radar.	Appropriate SA IoT allocation design that takes into account delay tolerance/delay intolerance of IoT applications.
Surveillance radars with periodic scanning period	Periodic SA for the gateways and the IoT devices	• Database assistance for any change in scanning pattern over longer periods. • Design of appropriate guard intervals for radar's protection.
Radars with quasi-periodic scanning periods	Quasi-periodic scanning weather radars have longer scan pulse intervals (between 13 to 22 s) and can provide longer communication intervals.	Regular database assistance for any change in scanning and careful design of transmission/quiet periods.
In general, radio navigation frequency reservations are almost similar across the globe.	Operators can have the possibility of designing unified standards under SA for IoT devices.	Coordination across different regulatory bodies across the globe.
Typically, single radar system per channel, wide coverage area, and co-located transmitter/receiver.	Database-assisted SA systems require less interaction with the IoT networks.	Design of appropriate database technology.
Type of IoT application.	Goodness metric.	Explanation.
Delay-tolerant.	$B_f 1_{[f_{min} \le f_o \le f_{max}]}$.	Amount of radar channel bandwidth B_f, and its frequency f_o is within radio range requirements of a particular IoT.
Delay-sensitive (time important but not critical).	$B_f 1_{[f_{min} \le f_o \le f_{max}]}(d_o/d)$.	Along with bandwidth B_f and the frequency range, one needs to also take into account the desired timescale of packet arrival d_o and its actual delay d.
Delay-intolerant.	Bf $1_{[f_{min} \le f_o \le f_{max}]} 1_{[d < d_{max}]}$.	Along with bandwidth B_f and the frequency range, actual packet delays d not exceed the defined maximum delay d_{max}.

Table 1. Six factors in zone-based SA's suitability in frequency channels used by different rotating radars, implementation challenges, and spectrum goodness metrics for IoT applications.

of [4, 5], the use of opportunistic temporal spectrum in our zone-based model allows a higher number of secondary user transmissions with the same level of interference protection to the radar system.

In the next section, based on the analyzed features of different rotating radar systems, we present a REM architecture as an enabler to provide shared wireless access for IoT devices.

Reasons for Suitability of Zone-Based SA, Implementation Challenges, and Spectrum Goodness

In Table 1, we provide six factors in the suitability of zone-based SA for IoT devices, and also present challenges involved in the implementation of SA in the radar channels. It is also important to identify which rotating radar channels are suitable for which IoT applications. For example, in a given area, a Zone 3 radar channel can be more suitable for applications that are intolerant to delays, whereas delay-tolerant applications can use a Zone 2 radar channel with little or no degradation in performance. In Table I, we also present spectrum goodness metrics that can be utilized for finding a suitable SA channel for an IoT application.

Enabling IoT Connectivity through REMs

The general concept of REM was first introduced by [6]. In [6], REM is defined as a network entity that enhances the awareness of cognitive radios by providing them information about their radio environment. The provided information includes device locations and activities, policies and regulation on spectrum access, and other information.

Role of the Gateways

In the proposed REM architecture (Fig. 4), the IoT gateways are used to act as a transparent bridge relaying messages between end devices and a REM repository server in the back-end. Although Internet-connected smartphones and tablets can be used as gateways to collect/transfer data from/to IoT devices, for the IoT to encompass millions of devices, the gateways would be required to operate on a much larger scale. The gateways would require less human intervention to collect and transmit data. To this end, the gateways will be included in hubs for smart homes, and in industrial equipment for purposes of tracking and asset management. In general, on one

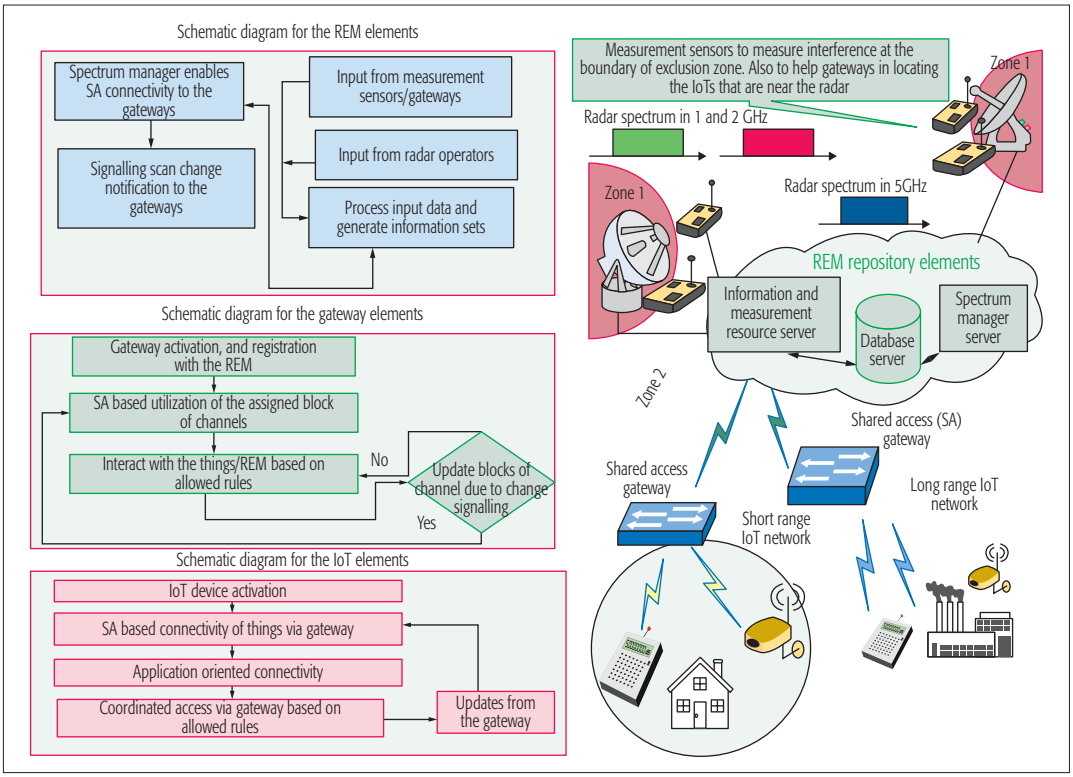


Figure 4. Different components of the proposed SA architecture and the simplified high-level schematic diagrams for different components involved in the proposed architecture.

side, the gateways will communicate via wireless technology downstream and upstream with the small IoT devices. On the other side, the gateways will be wirelessly connected further upward to the REM server.

The Proposed Architecture

The novel aspects of our proposed SA architecture relate to the access and operation in both various radar channels and adjacent unlicensed bands to these channels. It is good for our proposal that the unlicensed bands are close to/adjacent to several radar bands. There are already IoT devices that are operating in different unlicensed frequency spectrum. This means that basically the same front-end with some modifications can be used for our proposed framework.

The proposed architecture can be divided into four components (Fig. 4):

- REM repository.
- Different radar operators.
- Measurement capable devices (MCDs), such as a network of interference measurement and location estimation sensors, which are deployed at the boundary of a radar's exclusion zone.
- IoT network entities, such as gateways and IoT devices.

In our proposed framework, the REM repository is a collection of resources that can be accessed by the IoT gateways. The REM repository consists of:

- An information and measurement resource module (IMRM)
- A database module (DM)
- A spectrum manager (SM)

In Fig. 4 we present simplified high-level block diagrams for different components involved in the proposed REM-based SA architecture for IoT. Next, we explain the components of the proposed architecture.

The REM Repository Elements

Information and Measurement Resource Module

(Input from the Radar Operators): The IMRM module of the REM repository takes low-overhead static and dynamic information from different radar operators as input. The *static* (one time) information includes:

- Location of a radar system
- A particular radar system, which allows temporal sharing or not, and an exclusion zone established by a regulatory body to prohibit secondary transmissions in a specific area around a radar
- A reference power threshold to ensure that a secondary network entity does not fall into the exclusion zone
- The rotation rate of a radar

Also, the time radar's rotating main beam spends at a reference point. The *dynamic* information includes:

- Any change in scan speed of radar systems that are periodic/quasi-periodic rotating radars

A low-overhead message using a few bits can be utilized by a radar system operator to provide information about scan change notifications. Note that this does not require any changes in the operation of a radar system itself.

(Input from Measurement Sensors/Gateways): Information about the radio environment: The sensors collect information about the interfer-

It is good for our proposal that the unlicensed bands are close to/adjacent to several radar bands. There are already IoT devices that are operating in different unlicensed frequency spectrum. This means that basically the same front end with some modifications can be used for our proposed framework.

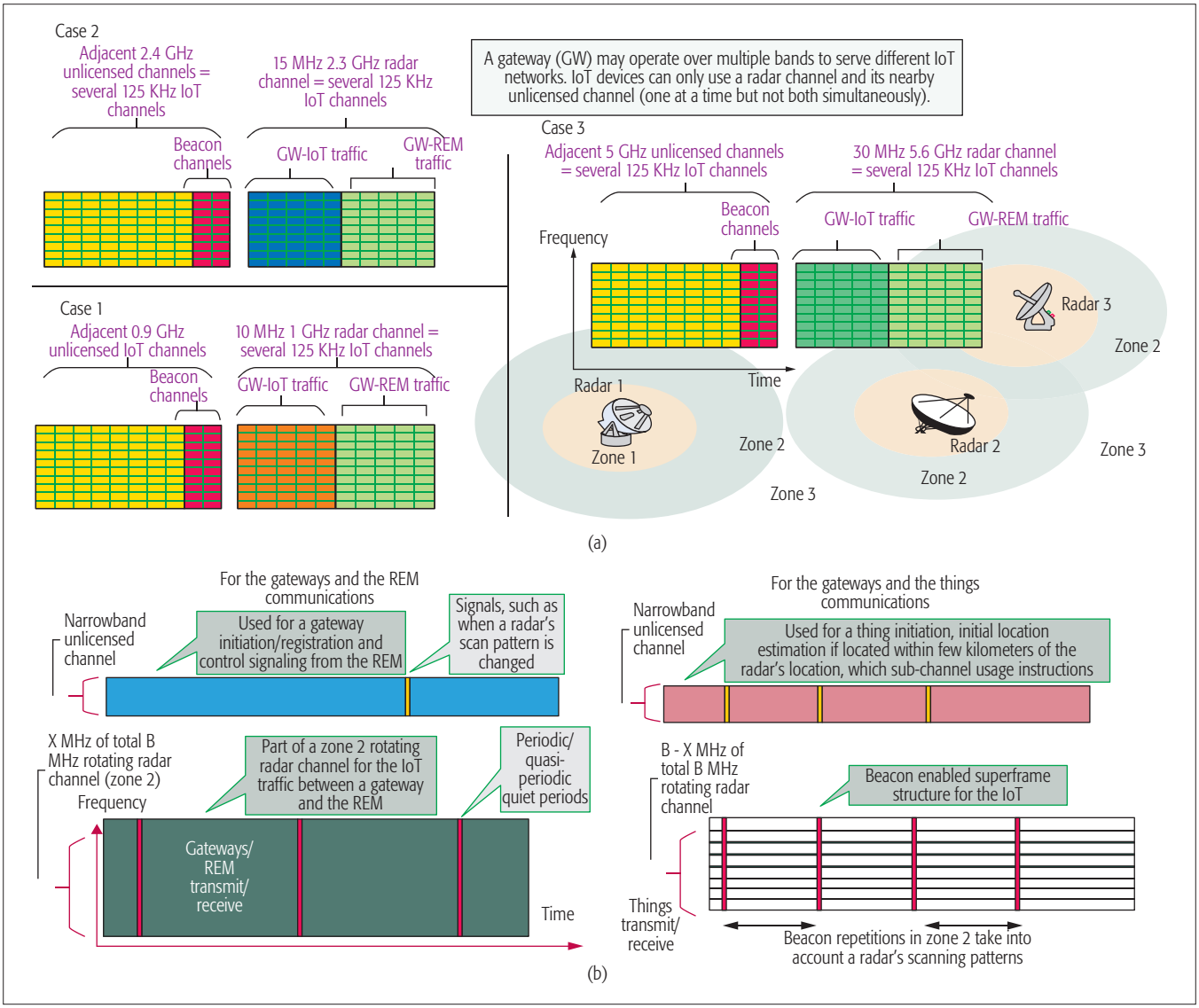


Figure 5. Examples of beacon and traffic channel blocks for the proposed SA for IoT devices.

ence environment. For example, a sensor network deployed at the boundary of Zone 1 of a radar can particularly facilitate interference-free temporal sharing in Zone 2 with the radar. By deploying interference measurement sensors, an operator can know when and where the reference power threshold (defined by a regulatory body) is exceeded, if at all, due to aggregate transmissions of the IoT entities. If the aggregate power received at the sensors exceeds the threshold, the REM repository instructs the gateways to move some of its users to another channel. The REM also uses the Zone 1 sensors and the gateways (that are located near Zone 1) to perform *sensor-gateway triangulation* for the location estimation of the IoT devices. The location estimation near Zone 1 helps determine whether an IoT device is within or outside the exclusion zone. If the device is within Zone 1, it can only use unlicensed channels; otherwise, it can use the rotating radar channels adjacent to the unlicensed channels.

Database Module: The DM processes information from the IMRM and generates instruction sets for the IoT gateways operating in the area. Based on the processed information from the IMRM, it lists channels that are available in an area for sharing, and also lists rules of sharing for a particular channel.

The instruction set generation at the DM provides a secure way of ensuring sharing with such radar systems, as the access is controlled and managed by a trustworthy authority (cellular/IoT operator), which is authorized to operate in a given area by an official regulatory body.

Spectrum Manager: The SM on one side interacts with different gateways, for example, it interacts with a gateway when it is activated to collect its location information and transmission power characteristics, and on the other side, it collects generated instructions from the DM. It then processes the obtained two-sided information and notifies the gateways about which portion of spectrum is available to them for utilization.

The Gateway Elements

Gateway Activation/Registration: Depending on how many different applications it can serve, a gateway can operate over multiple bands or a single band. For example, if a gateway is deployed

in a residential home, it may require only short-range IoT connectivity, and hence may operate only in the higher 5 GHz bands. On the other hand, if a gateway is deployed by an operator to provide connectivity in a given area, it may be required to provide long-range and/or short-range connectivity to a variety of different IoT applications. Hence, it may be required to operate over multiple bands.

When activated, a gateway, in order to obtain channel access authorization, needs first to register with the SM. This procedure can be carried out as follows: An unlicensed channel adjacent to a rotating radar channel is partitioned into several subchannels of 125 kHz bandwidth. A small set of these subchannels, called beacon channels, is used for the beacon transmissions (Fig. 5). On activation, a gateway first listens to one of these beacon channels and registers with the REM repository. The registration of a gateway involves providing its location information and transmission power characteristics in order for the list of available/forbidden channels to be computed.

SA-Based Utilization of the Channels: The gateway obtains a list of available channels from the REM repository, which is a set of unlicensed channels and available radar channels, and also obtains the rules for sharing in each of the available channels. Each of the available unlicensed/radar channels, whose bandwidth may vary between 10–30 MHz, is partitioned into several subchannels of 125 kHz bandwidth. A set of these subchannels is utilized by the gateway to communicate with the REM repository, and the other subchannels are utilized to communicate with the IoT devices; see Fig. 5 for illustrative examples.

The Things Elements: At a given time, an IoT device can operate only over a single radar channel or an adjacent unlicensed channel, but not both. Each gateway continuously transmits information, such as its identification number (ID), on the beacon unlicensed channels adjacent to the radar channels on which it can operate. When the signal is picked up by a nearby IoT device, which is just initiated, it responds to the signal. The gateway selects a set of subchannels for exchanging traffic with the IoT device. When the radar channel is not available in the area, the gateway selects this subchannel set from the unlicensed channel adjacent to the radar channel; otherwise, it selects the set from the radar channel.

Superframe-Based Communication Format

On a radar channel, the communication format consists of a periodic/quasi-periodic superframe. The superframe starts with a beacon signal transmitted by a gateway (see the beacon signals on the right of Fig. 5b). More than one gateway in an area can transmit beacon messages at the same time and avoid interference by using spread spectrum radio modulation, used in existing IoT protocols like LoRaWAN [9]. The beacon signal notifies the IoT devices about the beacon repetition rate (i.e., when to listen for the next beacon) and the communication period message, which reports the length of the period after the beacon signal during which the devices can transmit/receive their traffic (for the illustration, see Fig. 5b).

In a Zone 3 radar channel, the superframe duration can be adjusted to any length suitable for the network. In a Zone 2 periodic radar channel, such as the radars at 1 GHz and 2.3 GHz, the beacon can be transmitted to the devices after the radar's main beam leaves the slice in which the network is located. The network stays quiet during the time the radar's main beam spends on the slice T_s and also during the guard interval times T_g before and after that slice. This means that if the radar's main beam points every T_r s at the slice, a superframe of length $T_r - T_s - 2T_g$ can be utilized for the IoT traffic. For example, with $T_s = T_g = 0.5$ s this can be equal to $T_r - 1.5$ s.

In a Zone 2 quasi-periodic radar channel, such as the radar in 5 GHz (Fig. 3a), the time T_r can vary over different periods. To take into account this quasi-periodicity due to the slow scan mode and fast scan modes of the radar, the length of the superframe can be set to be $min(T_r) - T_s - 2T_g$ for the IoT traffic.

Conclusions and Future Directions

Radar bands are a potential candidate for spectrum sharing between wireless communications and incumbent systems. To better understand the operating principles of various rotating radars which operate in different spectrum bands, and to determine their spectrum usage patterns, we ran an extensive measurement campaign near the city of Oulu, Finland. During the campaign, the spectrum usage behavior of three ground-based fixed rotating radar systems at different locations was measured. Based on the measurement results, in this article, we identify the suitability of the rotating radar spectrum for the IoT shared spectrum access. We present reasons for the proposed SA suitability, identify related implementation challenges, and discuss spectrum goodness metrics for IoT applications. We also propose a framework that enables SA for IoT devices through REMs. For potential future work, this research can be extended to explore the challenges in the implementation of the proposed REM-based SA in the rotating radar's channels, and challenges such as the required number of measurement sensors to support the REM, update rate of the REM, its algorithmic complexity, and security issues. The prototype can also be developed to enable SA through REMs for IoT connectivity.

Acknowledgment

This work was funded by the Academy of Finland under grant number 26687 and in part by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC). It was also funded by U.S. grants NSF CPS-1646607, ECCS-1547201, CCF-1456921, CNS-1443917, ECCS-1405121, and NSFC61428101.

References

[1] H. R. Schindler *et al.*, "Europe's Policy Options for a Dynamic and Trustworthy Development of the Internet of Things," tech. rep., 2013; accessed 17 Oct. 2016; http://www.rand.org/pubs/research_reports/RR356.html

[2] S. Forge, "Radio spectrum for the Internet of Things," tech. rep., 2016, Accessed 17 Oct. 2016; http://www.emeraldinsight.com/doi/abs/10.1108/info-11-2015-0050.

[3] FCC, "Enabling Innovative Small Cell Use In 3.5 GHz Band NPRM & Order," Docket 12-148, 2012; accessed 17 Oct. 2016; https://www.fcc.gov/document/enabling-innovative-small-cell-use-35-ghz-band-nprm-order.

The beacon signal notifies the IoT devices about the beacon repetition rate, that is, when to listen for the next beacon, communication period message which notifies the length of the period after the beacon signal during which the devices can transmit/receive their traffic.

For potential future work, this research can be extended to explore the challenges in the implementation of the proposed REM-based SA in the rotating radar's channels, challenges such as the required number of measurement sensors to support the REM, update rate of the REM, its algorithmic complexity, and security issues.

[4] NTIA CSMAC Committee, "Interference and Dynamic Spectrum Access," tech. rep., Nov. 2010; accessed 17 Oct. 2016; https://www.ntia.doc.gov/files/ntia/ publications/ csmac_interferencecommitteereport_01102011.pdf.

[5] M. Cotton *et al.*, "Developing Forward Thinking Rules and Processes to Fully Exploit Spectrum Resources: An Evaluation of Radar Spectrum Use and Management," Mar. 2012; 17 Oct. 2016; http://www.its.bldrdoc.gov/publications/2669.aspx.

[6] Y. Zhao, *Enabling Cognitive Radios through Radio Environment Maps*, Ph.D. dissertation, Virginia Tech, 2007; accessed 17 Oct. 2016; https://theses.lib.vt.edu/theses/available/etd-05212007-162735/

[7] Z. Khan *et al.*, "On Opportunistic Spectrum Access in Radar Bands: Lessons Learned from Measurement of Weather Radar Signals," *IEEE Wireless Commun.*, vol. 23, no. 3, June 2016; accessed 17 Oct. 2016, pp. 40–48.

[8] Nokia, "LTE-M: Optimizing LTE for the Internet of Things," white paper, tech. rep., 2015; accessed 17 Oct. 2016; https://iotfuse.com/wp-content/uploads/2016/02/nokia_lte-m_-_optimizing_lte_for_the_internet_of_things_white_paper.pdf.

[9] Technical Marketing Workgroup, "LoRaWAN: What Is It?" LoRa Alliance, tech. rep., 2015; accessed 17 Oct. 2016; https://www.lora-alliance.org/portals/0/documents/whitepapers/LoRaWAN101.pdf.

[10] Agilent Technologies, "Agilent Radar Measurements," app. note, tech. rep., 2014; accessed 17 Oct. 2016; http://cp.literature.agilent.com/litweb/pdf/5989-7575EN.pdf

[11] M. Tercero, K. Sung, and J. Zander, "Temporal Secondary Access Opportunities for WLAN in Radar Bands," *Proc. 14th Int'l. Symp. Wireless Personal Multimedia Commun.*, 2011; accessed 17 Oct. 2016, pp. 1–5.

[12] F. Hessar and S. Roy, "Spectrum Sharing between a Surveillance Radar and Secondary WiFi Networks," *IEEE Trans. Aerospace Electronic Systems*, vol. 52, no. 3, June 2016; accessed 17 Oct. 2016, pp. 1434–48

Biographies

Zaheer Khan received his Dr.Sc. in electrical engineering from the University of Oulu, Finland, and his M.Sc degree in electrical engineering from University College Boras, Sweden, in 2011 and 2007, respectively. Currently, he has a tenure track position at the University of Liverpool, United Kingdom. He worked as a research fellow/principal investigator at the University of Oulu. He was the recipient of the Marie Curie fellowship for 2007–2008. His research interests include application of game theory to model distributed wireless networks, prototyping access protocols for wireless networks, IoT location tracking systems, cognitive and cooperative communications, and wireless signal design.

Janne J. Lehtomäki graduated with an M. Sc. and a Ph. D. in telecommunications from the University of Oulu in 1999 and 2005, respectively. Currently, he is an adjunct professor at the University of Oulu, Centre for Wireless Communications (CWC). His research interests are in nanonetworks, spectrum measurements, energy detection, and cognitive radio networks. He co-authored the winner of the Best Paper Award at IEEE WCNC 2012.

Stefano Iellamo works as an experienced researcher at ICS-FORTH, Heraklion, Greece, since 2015. He obtained his M.Sc. from Politecnico di Milano, Italy, in 2009 and his Ph.D. from Telecom ParisTech, France, in 2014. His research interests include 5G wireless networks, IoT connectivity, spectrum management, and machine learning.

Risto Vuohtoniemi received his M.Sc. (Tech) and Licentiate of Technology degrees in telecommunications from the University of Oulu. He is now working as a university teacher in the Department of Communications Engineering and as a research scientist in CWC at the University of Oulu. His research interests are in RF technology, especially in future wireless communication systems, cognitive radio, power line communication, and passive radar applications.

Ekram Hossain [F'15] is a professor (since March 2010) in the Department of Electrical and Computer Engineering at the University of Manitoba, Winnipeg, Canada. He is a Member (Class of 2016) of the College of the Royal Society of Canada. He received his Ph.D. in electrical engineering from the University of Victoria, Canada, in 2001. His current research interests include design, analysis, and optimization of wireless/mobile communications networks, cognitive radio systems, and network economics. He has authored/edited several books in these areas (http://home.cc.umanitoba.ca/~hossaina). He serves as the Editor-in-Chief of *IEEE Communications Surveys & Tutorials* and an Editor for *IEEE Wireless Communications*. Also, he is a member of the IEEE Press Editorial Board. Previously, he served as the Area Editor for *IEEE Transactions on Wireless Communications* in Resource Management and Multiple Access from 2009 to 2011, an Editor for *IEEE Transactions on Mobile Computing* from 2007 to 2012, and an Editor for the *IEEE Journal on Selected Areas in Communications* Cognitive Radio Series from 2011 to 2014. He has won several research awards including the IEEE VTC 2016 (Fall) Best Student Paper Award, the IEEE Communications Society Transmission, Access, and Optical Systems Technical Committee's Best Paper Award at IEEE GLOBECOM 2015, the University of Manitoba Merit Award in 2010, 2014, and 2015 (for Research and Scholarly Activities), the 2011 IEEE ComSoc Fred Ellersick Prize Paper Award, and the IEEE Wireless Communications and Networking Conference 2012 Best Paper Award. He was elevated to an IEEE Fellow "for spectrum management and resource allocation in cognitive and cellular radio networks." He was a Distinguished Lecturer of IEEE ComSoc, 2012–2015. Currently he is a Distinguished Lecturer of the IEEE Vehicular Technology Society. He is a registered Professional Engineer in the province of Manitoba, Canada.

Zhu Han [S'01, M'04, SM'09, F'14] received his B.S. degree in electronic engineering from Tsinghua University in 1997 and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Maryland, College Park, in 1999 and 2003, respectively. From 2000 to 2002, he was an R&D engineer of JDSU, Germantown, Maryland. From 2003 to 2006, he was a research associate at the University of Maryland. From 2006 to 2008, he was an assistant professor at Boise State University, Idaho. Currently, he is a professor in the Electrical and Computer Engineering Department as well as in the Computer Science Department at the University of Houston, Texas. His research interests include wireless resource allocation and management, wireless communications and networking, game theory, big data analysis, security, and smart grid. He received an NSF Career Award in 2010, the Fred W. Ellersick Prize of IEEE ComSoc in 2011, the EURASIP Best Paper Award for the *Journal on Advances in Signal Processing* in 2015, the IEEE Leonard G. Abraham Prize in the field of communications systems (best paper award in *IEEE JSAC*) in 2016, and several best paper awards at IEEE conferences. Currently, he is an IEEE ComSoc Distinguished Lecturer.

Efficient IoT Gateway over 5G Wireless: A New Design with Prototype and Implementation Results

Navrati Saxena, Abhishek Roy, Bharat J. R. Sahu, and HanSeok Kim

ABSTRACT

The growth of IoT is expected to herald a better quality of life and open up new industrial opportunities. However, it is also raising new challenges in resource constrained wireless networks. Originally designed for human-to-human wireless communications, present wireless networks fail to satisfy the huge connectivity and varying traffic requirements of IoT. Next generation 5G wireless networks are expected to exploit emerging features, like mmWave, massive MIMO, and C-RAN, for providing the enormous connectivity, resource pooling, and energy efficiency required to support commercial rollout of IoT. Exploring these emerging features, we develop 5G-enabled IoT gateways to communicate with the RRHs of 5G C-RAN. The gateways are endowed with efficient compression schemes to significantly improve uplink resource utilization. Our 5G C-RAN prototyping and laboratory experimental results point out a huge opportunity for supporting a massive number of IoT-enabled devices by sophisticated 5G C-RAN deployment and efficient IoT gateway development.

INTRODUCTION

As coined by the famous British entrepreneur Kevin Ashton, the Internet of Things (IoT) [1] refers to a network of physical objects, or "things," embedded with electronics, software, sensors, and network connectivity, for automatic information exchange. Interestingly, from an IoT's perspective, "things" can refer to a wide variety of devices, including surveillance cameras, environment monitoring sensors, healthcare equipment, biochip transponders on farm animals, and sensor-equipped automobiles. Increasing penetration of smart devices and the introduction of ultra-modern communication platforms, imbued with ubiquitous computing, are steadily contributing toward the ultimate realization of IoT. Continuing in this trend, in the near future, IoT is expected to have significant impact on home and business applications, thereby ushering in a better quality of life and new business opportunities for global economic growth.

Figure 1a shows an example of IoT architecture and connectivity, involving a myriad of IoT-enabled devices, connected to the Internet for supporting a wide variety of applications. The huge connectivity and large number of devices mandate the use of some gateway or another for Internet connectivity. Naturally, there are different candidate access technologies for this gateway design. While WiFi could be a good solution for many applications, it suffers from increasing packet collisions with increasing amounts of uplink access. Moreover, traditional WiFi is also not suitable for quality of service (QoS) support and consumes a fair amount of power. Similarly, the wireless access solutions in unlicensed frequency bands require separate deployment from existing cellular networks, thereby resulting in additional capital expenditure (CAPEX) and operating expenditure (OPEX). As the proprietary and unlicensed solutions are not optimized for spectral efficiency, with exponential increase in IoT deployment, these solutions are very likely to congest the unlicensed bands and trigger complaints from existing users. At the same time, a variety of new wide area applications (e.g. smart cities, traffic monitoring, and smart grids) for IoT are offering new markets to wireless operators for enhancing their revenues. As a result, the Third Generation Partnership Project (3GPP) has recently decided to include narrowband IoT (NB-IoT) in Release 13 standards. However, as mentioned below, present fourth generation (4G) macrocellular cellular networks (e.g. LTE, LTE-Advanced) are facing significant new challenges in supporting large-scale IoT applications.

Massive rollout of IoT involve a wide number of devices with automated connectivity. Unfortunately, current commercial 4G systems mostly support lower than 1000 connected users or devices per cell [2]. With increasing numbers of users, an LTE radio resource control (RRC) block starts rejecting these connections [2]. Emerging IoT will typically involve many thousands of connected devices. Figure 1b (blue bars) demonstrates Cisco's recent forecast of almost exponential increase in number of connected devices (Cisco VNI: Global Mobile Data Traffic Forecast Update; http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html [Oct-22-2016]).

Moreover, the sporadic communications from a myriad of heterogenous IoT devices might also impose huge traffic bursts on wireless cellular

Next generation 5G wireless networks are expected to exploit emerging features, like mmWave, massive MIMO, and C-RAN, for providing the enormous connectivity, resource pooling, and energy efficiency required to support commercial rollout of IoT. Exploring these emerging features, we develop 5G-enabled IoT-gateways to communicate with the RRHs of 5G C-RAN.

Navrati Saxena and Bharat J. R. Sahu are with Sungkyunkwan University; AbhishekRoy and HanSeok Kim are with Samsung Electronics.

Digital Object Identifier:
10.1109/MCOM.2017.1600437CM

IoT-enabled devices are likely to involve a wide variety of application traffic, ranging from static, intermittent, delay tolerant, small-sized packets to mobile, frequent, delay sensitive, large packets. The heterogeneity of traffic and applications make the already complicated network support even more challenging.

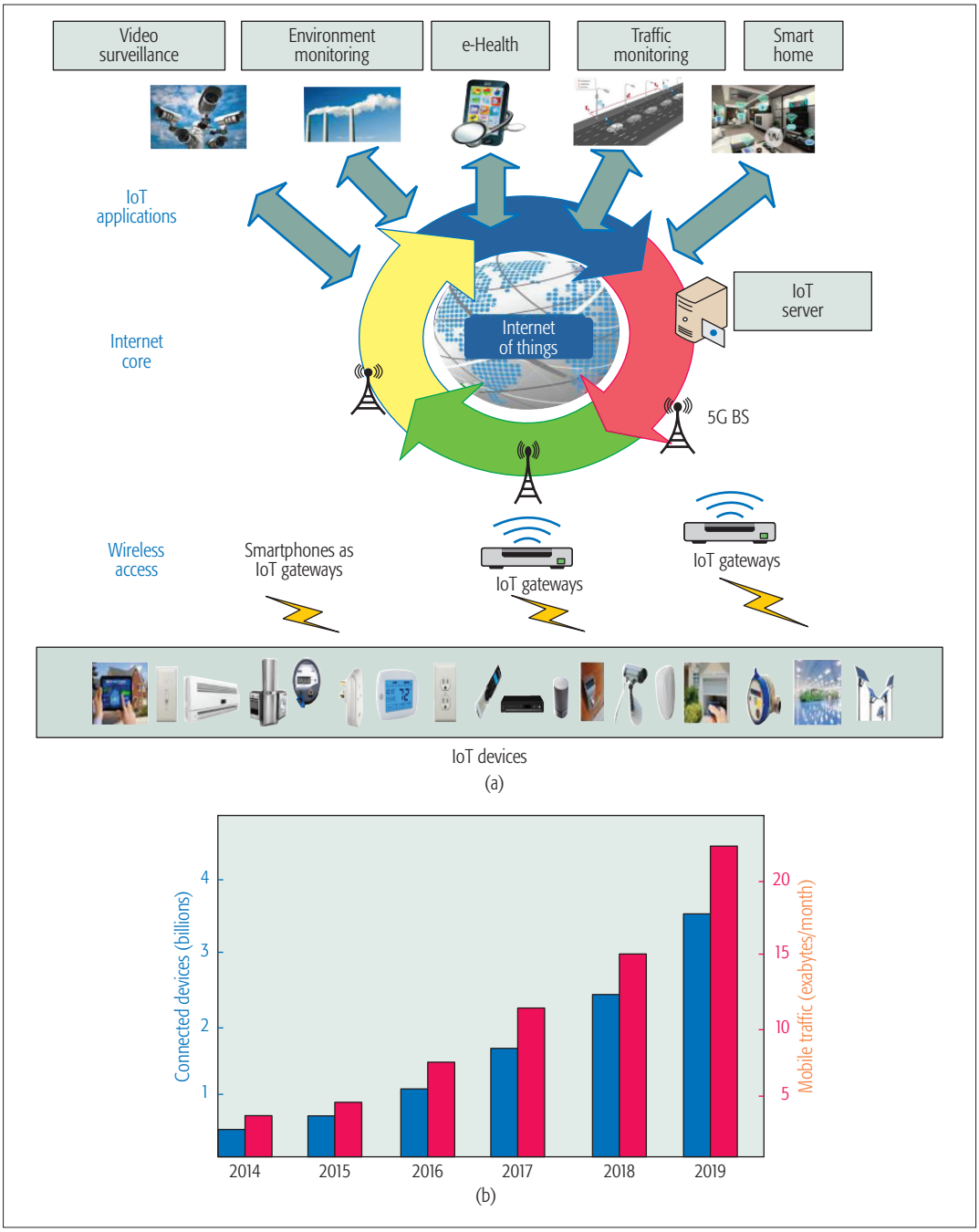


Figure 1. IoT architecture with connectivity and traffic forecast: a) IoT architecture involving IoT-gateways and 5G BS; b) forecast of mobile devices and traffic.

networks. As shown in Fig. 1b (red bars), Cisco's recent traffic forecast in commercial cellular networks has pointed out a whopping five times increase in cellular traffic volume over the next four years. It is less likely that existing 4G wireless networks will be capable of sustaining this ever increasing traffic demand in the near future.

Last but not least, IoT-enabled devices are likely to involve a wide variety of application traffic, ranging from static, intermittent, delay-tolerant, small-sized packets to mobile, frequent, delay-sensitive, large packets. The heterogeneity of traffic and applications make the already complicated network support even more challenging.

The above-mentioned challenges have triggered new research investigations across different industries as well as academia. The exhaustive survey in [3] has identified the requirements, challenges, plausible architectures, and applications of machine type communications (MTC) over cellular LTE-Advanced networks. A rule-based intelligent gateway to enable the efficient integration of horizontal IoT services with QoS features is discussed in [4]. The work in [5] uses smart IoT devices with gateways to develop a real-time challenge response for advanced security measures. Existing research works also include simplified air interface protocols, with low-energy, low-complexity IoT modules for cellular licensed bands [6]. Next generation 5G wireless networks [7, 8] are expected to explore innovative radio technologies for providing better coverage, higher data

rates, dense connectivity [9] and superior user experience at a relatively lower cost than existing wireless systems. However, the introduction of ultra-dense 5G wireless networks also threatens to increase a network's energy consumption, thereby aggravating the detrimental greenhouse (CO_2) gas emissions [7]. Cloud radio access networks (C-RANs) [10] are rapidly coming up to improve resource management, energy efficiency, environmental sustainability, and OPEX. C-RAN [10] is a cellular architectural evolution, based on distributed base stations (BSs), where remote radio heads (RRHs) are separated from the digital baseband unit (BBU) by high-speed (typically several gigabits per second) fiber optical fronthaul cables. C-RAN exploits real-time virtualization techniques to dynamically allocate resources from the entire resource pool to the software stacks of BBUs, according to network load. We believe that in the future, C-RAN will gradually evolve toward supporting massive IoT-enabled devices and applications. This motivates us to design and implement 5G-enabled IoT gateways and 5G C-RAN for providing efficient connectivity to IoT devices. More precisely, our contributions in this article are as follows.

•We begin with a brief discussion of the major emerging features of 5G wireless, like millimeter-wave (mmWave) communications, massive multiple-input multiple-output (MIMO) and C-RAN. We also mention how these emerging new features will help in satisfying the massive connectivity and diverse traffic generated by a huge number of IoT devices.

•Subsequently, we explain the design and implementation of new 5G-enabled IoT gateways. The novelty of these gateways lies in efficient uplink IoT traffic classification and optimal uplink data (traffic) compression strategies. This helps in relaxation of uplink traffic burden and results in efficient utilization of uplink wireless resources.

•Next, we discuss our prototype development of 5G C-RAN for providing connectivity to the IoT devices through 5G-enabled IoT gateways.

•Our testbed implementation and laboratory experiments demonstrate the efficiency of 5G C-RAN and IoT gateways for providing massive IoT connectivity with reduced energy consumption and negligible overheads.

5G Wireless: The Harbinger of Industrial IoT

Supporting almost pervasive device connectivity and explosive growth in data usage is an extremely daunting task in present 4G (LTE) cellular wireless networks. The concept of 5G wireless communications, with mmWave channel and sub-millisecond (i.e., < 1 ms) medium access control (MAC) [7, 8], are emerging for alleviating the constraints of current 4G cellular networks. Naturally, large-scale research and 5G prototype development [7, 8] activities are gearing up across academia and industries. For example, 3GPP Mobile and wireless communications Enablers for the Twenty-twenty Information Society (METIS) project and the HORIZON 2020 program (HORIZON 2020, The EU Framework Programme for Research and Innovation"; http://ec.europa.eu/programmes/horizon2020/ [October 22, 2016]) are some of the major research programs and initiatives for 5G wireless.

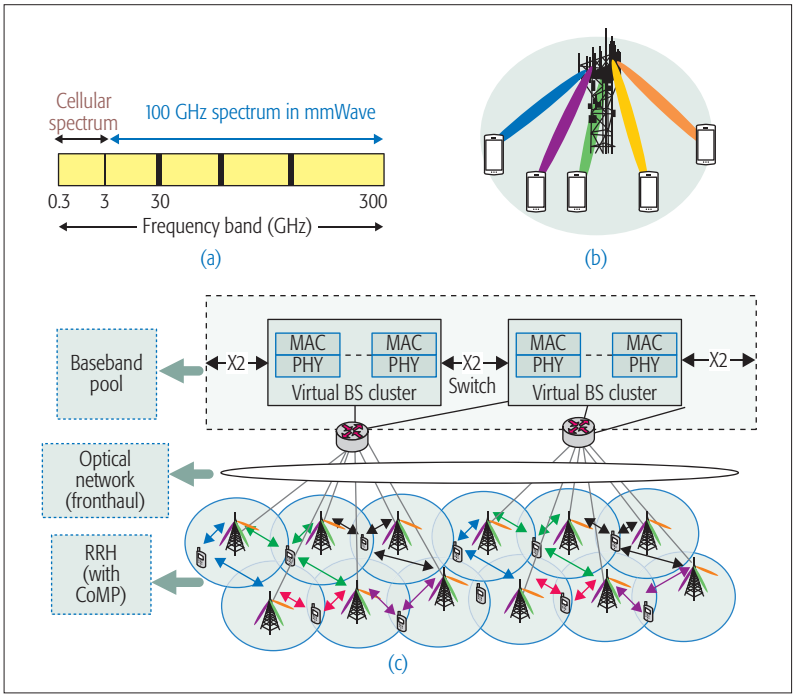


Figure 2. 5G C-RAN with mmWave and massive MIMO.

mmWave and Massive MIMO: New Frontiers in Wireless

The user capacity (i.e., number of users supported) of any wireless communication system depends on spectral efficiency and bandwidth. With the ongoing trends of reduction in cell size and increasingly aggressive modulation and coding schemes (MCSs), it is the system bandwidth that needs to be explored. Figure 2a points out that presently almost all wireless communications use 300 MHz to 3 GHz spectrum band, often referred to as "sweet spot" or "beachfront" spectrum [7, 8]. However, emerging 5G wireless networks envision usage of unused high-frequency mmWave bands, ranging from 3~300 GHz. Interestingly, even a small fraction of available mmWave spectrum can support hundreds of times more data rate and user capacity over the current cellular spectrum [7, 8]. Thus, the availability of a big chunk of mmWave spectrum is opening up a completely new horizon to support spectrum constrained wireless networks.

Unfortunately, mmWave communications suffer from path loss and line of sight (LOS). Recent research [11] related to path loss shows directional transmissions of narrow mmWave beams can reduce interference and increase spatial multiplexing capabilities [11]. LOS as well as non-LOS (NLOS) propagation characteristics of reflected, scattered, and diffracted mmWave signals are under investigation for achieving 5G channel models. The propagation results [11] also demonstrate very little signal penetration through glass doors and windows, thus pointing out the need for different small-sized cells to serve indoor and outdoor coverage areas. Interestingly, such small cells, with only 200 m radius, are already deployed in dense urban areas of Japan and

The development software is modular, allowing easy customization, including changes to the application software. There are multiple wireless interfaces: in the downlink, it can connect with the IoT devices using ZigBee or near field communications (NFC); and in the uplink, it uses our 5G-wireless protocol stack to communicate with the RRHs of 5G C-RAN.

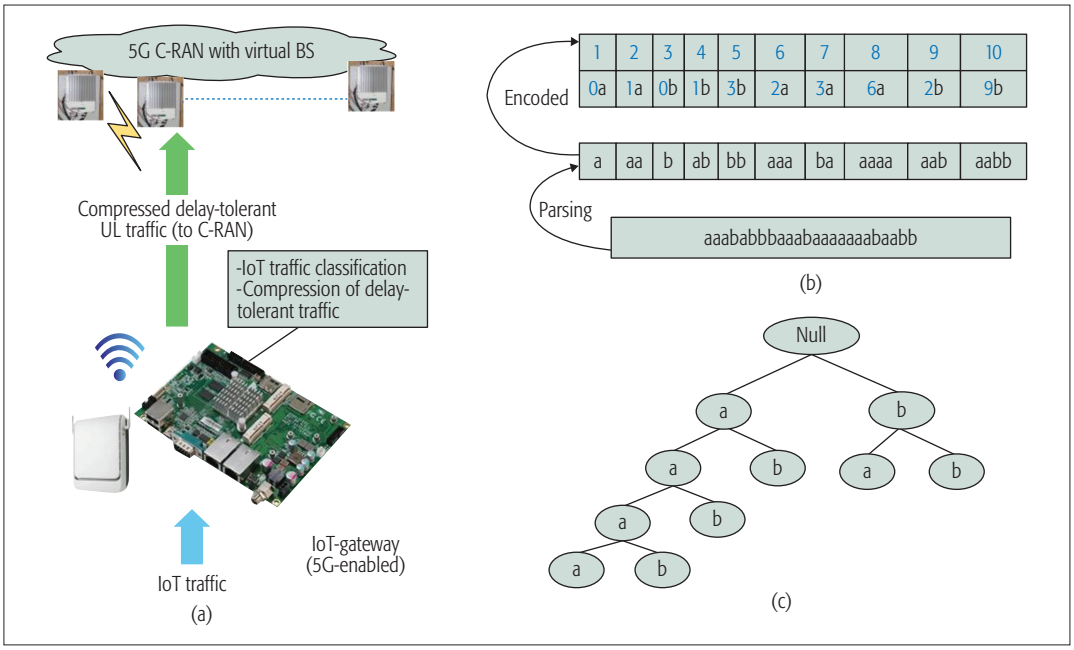


Figure 3. 5G-enabled IoT-gateway design with uplink traffic encoding implemented as a trie: a) 5G-enabled IoT gateways; b) uplink traffic encoding in IoT-gateways; and c) encoding dictionary implemented as a trie.

Korea [8]. Adaptive and hybrid mmWave beamforming are explored for reducing the interference [12]. The beams are aligned at such angles that the emitted signal components are added constructively only at the target (reference) locations and reduced in most other locations. Figure 2b depicts that massive MIMO [13] provides the BS with a huge number of directional antennas for significantly enhancing spectral and energy efficiency. A suitable combination of mmWave communications and massive MIMO is likely to form the foundation of 5G wireless required for providing native support to massive IoT connectivity.

Emergence of 5G HetNets and Cloud RAN

5G wireless networks [7, 8] are expected to be made up of ultra-dense heterogeneous networks (HetNets), comprising small low-power cells, besides the legacy macrocells. Joint research works by Qualcomm and Samsung Electronics [13] have pointed out that commercial deployment of HetNets calls for sophisticated resource and interference management. Figure 2c delineates C-RAN [10] — an aspiring cellular architectural evolution for significant improvement in wireless resource management, OPEX, and energy usage. C-RAN is a cellular architectural evolution based on distributed BSs, where RRHs are separated from digital the BBU by high-speed (typically several gigabits per second) fiber optical fronthaul cables. In next generation C-RAN, hundreds of RRHs will be connected to a centralized BBU pool, with high-speed switches for load balancing.

Noticeably, as only the antenna is needed at the cell sites, the usage of expensive processing equipment in every site is not needed and results in unnecessary increase in network energy consumption and OPEX. Another tenet of C-RAN is virtualization, which aims at reducing CAPEX by applying real-time virtualization techniques to dynamic resources allocation from the pool to the software BBU stacks according to the network load. Hence, the individual BBUs in the C-RAN are often called as virtual BSs (VBSs), and the entire pool is called a VBS cluster (VBSC). Wireless operators are realizing significant cost reduction from pooling and virtualization of BBU processing, as there is no longer a need for per-site peak capacity provisioning. A quick survey [10] of recent research works indicates the impending necessity and major global research works related to C-RAN. Major mobile operators, like China Mobile, Ericsson, NEC, ZTE, and Korea Telecom (KT), have already started deploying C-RAN (http://labs.chinamobile.com/cran/ [October 22, 2016]) for prospective energy savings. Thus, while 5G wireless, with mmWave and massive MIMO, can provide native support to dense, heterogenous connectivity, C-RAN helps reduce network energy consumption and CAPEX. Based on these basic foundations, in the remainder of this article, we illustrate the development of our 5G-enabled IoT gateway for supporting enormous IoT connectivity.

5G-Enabled IoT-Gateway Development

Figure 3a demonstrates our 5G-enabled IoT gateway development platform for supporting massive IoT connections. The development platform supports enhanced computing and graphics capabilities. The gateway provides intensive data aggregation, data analysis, and information visualization for connecting everyday wireless-enabled devices to the Internet using 5G wireless. The development software is modular, allowing easy customization, including changes to the application software. There are multiple wireless interfaces: in the downlink, where it can connect with the IoT devices using ZigBee or near field communications (NFC); and in the uplink, it uses our 5G wireless protocol stack to communicate with the RRHs of 5G C-RAN.

Referring to Fig. 1b, Cisco's recent cellular traffic forecast points out the impending connectivity and traffic burden in commercial cellular networks in the near future. A significant fraction of the uplink traffic will be generated by several thousand connected IoT devices. Hence, optimal data handling and data compression are required for efficient wireless resource management. Data compression [14] transforms data into a compressed form, such that the original data could be completely or partially recovered, with (lossy compression) or without (lossless compression) any information loss. Among the major lossless text compression schemes [14] (e.g., Huffman coding, run length coding, arithmetic coding, Lempel Ziv coding), Lempel Ziv (LZ) coding is an optimal text (data) compression algorithm, capable of working on almost any type of data.

As many IoT devices often generate delay-tolerant data traffic, we introduce new uplink data classification, buffering, and compression (encoding) in 5G-enabled IoT gateways. Our designed 5G-enabled IoT gateways first classify the uplink data packets into two categories: delay-sensitive packets and delay-tolerant packets. While delay-sensitive packets are immediately transmitted over the 5G radio interface, delay-tolerant packets are buffered in the uplink MAC buffers of IoT gateways until one of the following conditions hold true:

- A delay-sensitive packet arrives.
- The periodic uplink timer threshold expires.
- The maximum data size threshold expires.

The 5G-enabled IoT gateways now use optimal data compression schemes [14] to encode the uplink data traffic generated by the devices and transmit it to the RRH of 5G C-RAN. LZ-based text compression algorithms [14] typically replace a string of characters with a reference to a dictionary (codebook) location of the same string. The dictionary reference is represented by a (*Len*, *Dist*) tuple, where *Len* and *Dist* represent the length and offset of the string, respectively. Among the many variants of optimal LZ coding, we use Lempel-Ziv-Storer-Szymanski (LZSS) for the following fundamental reasons:

- LZSS [14] is an optimal lossless compression technique, where any dictionary reference longer than the substring is omitted if the substring length is less than a "break even" point [14].
- Compared to most other optimal text compression schemes, LZSS is easier to implement [14], with relatively lower storage and less computation overhead.
- Moreover, unlike most other compression schemes, LZSS does not require an external dictionary [14]. Instead, it uses a sliding window as a dictionary. This prevents frequent dictionary updates during uplink transmissions.

The sliding window dictionary in LZSS compression contains recently encoded uplink data, represented as a sequence of characters. The larger the size of the dictionary, the longer it takes to find a match to replace. The entire sequence of data withheld since the last transmission is transmitted in the uplink 5G radio interface in an encoded form. As the dictionary is maintained as a sliding window, once it is full, the oldest characters are replaced by new characters. A dictionary search is needed to find a match for the sequence of characters being encoded. The un-encoded (un-compressed) characters are written without any modification. A flag bit is used to distinguish between un-encoded and encoded sequences of characters. Encoding starts from left to right for a fixed length of F characters. At each encoding step, a portion of the input characters is read into a window of size N characters. The first $(N - F)$ characters have already been encoded, and the remaining F characters are in the look-ahead buffer. The window is searched to find the longest match for the look-ahead buffer. The longest match is encoded and transmitted in the uplink by the 5G-enabled IoT gateways. Thus, the traffic "$\tau_1\tau_2\tau_3$ ... " received by the IoT gateway is transmitted in the uplink as a sequence "$C(w1)C(w2)$ $C(w3)$ ...," where w_is are non-overlapping distinct segments of data $\tau_1\tau_2\tau_3$ ..., and $C(w)$ is the encoding for w. For example, as shown in Fig. 3b, on receiving the uplink IoT data string *aaababbbbaaabaaaaaaabaabb* ..., our 5G-enabled IoT gateway first parses the data as *a*, *aa*, *b*, *ab*, *bb*, *aaa*, *ba*, *aaaa*, *aab*, *aabb*, ... Now, assuming encoding index (code-book index) as simple integers (i.e., encoding an index of NULL, *a*, ... , *aabb*, as 0, 1, ... , 10, respectively), the IoT gateway now encodes the segments as 0*a*, 1*a*, 0*b*, 1*b*, 3*b*, 2*a* 3*a*, 6*a*, 2*b*, 9*b*, ..., and transmits it in the uplink using 5G radio interface. Such a symbol-wise traffic model can be efficiently stored in a dictionary, implemented as a search trie. Figure 3c demonstrates the corresponding trie, implemented as a heap in our IoT gateway.

The corresponding decompression (decoding), a relatively simple and straightforward process, is performed at the C-RAN. The decoding requires reading the dictionary offset and copying a specified number of characters. We now mention the memory and complexity of the compression scheme:

- The incremental parsing accumulates larger and larger data in the dictionary, thereby estimating higher-order traffic probabilities. Essentially, the strategy approaches optimality for ergodic traffic sources, with the improvement bounded by $\Omega(\log_2 N \log_2 \log_2 N)$ for N symbols.
- Including a flag bit, the pointer pair could be represented using $(1 + \log_2 N + \log_2 F)$ bits. Naturally, the performance depends on the values of N and F.
- For a sequence of characters, the time taken for each encoding step is bounded by $(N - F)$ comparisons, which could be bounded by $\log_2 N$ using a binary search. Thus, for encoding N characters, the comparison operation is upper bounded by $O(N)$.

Further research and development activities are currently underway to enhance the processing capabilities of 5G-enabled IoT gateways for supporting local processing and localized applications. Initial investigations have demonstrated that 5G-enabled gateways have the potential to reduce latency and increase reliability by offloading some popular applications from the server and C-RAN. Moreover, 5G wireless is expected to provide native support for device-to-device (D2D) communications. Our 5G-enabled IoT

Further research and development activities are currently underway to enhance the processing capabilities of 5G-enabled IoT gateways for supporting local processing and localized applications. Initial investigations have demonstrated that the 5G-enabled gateways have the potential to reduce latency and increase reliability.

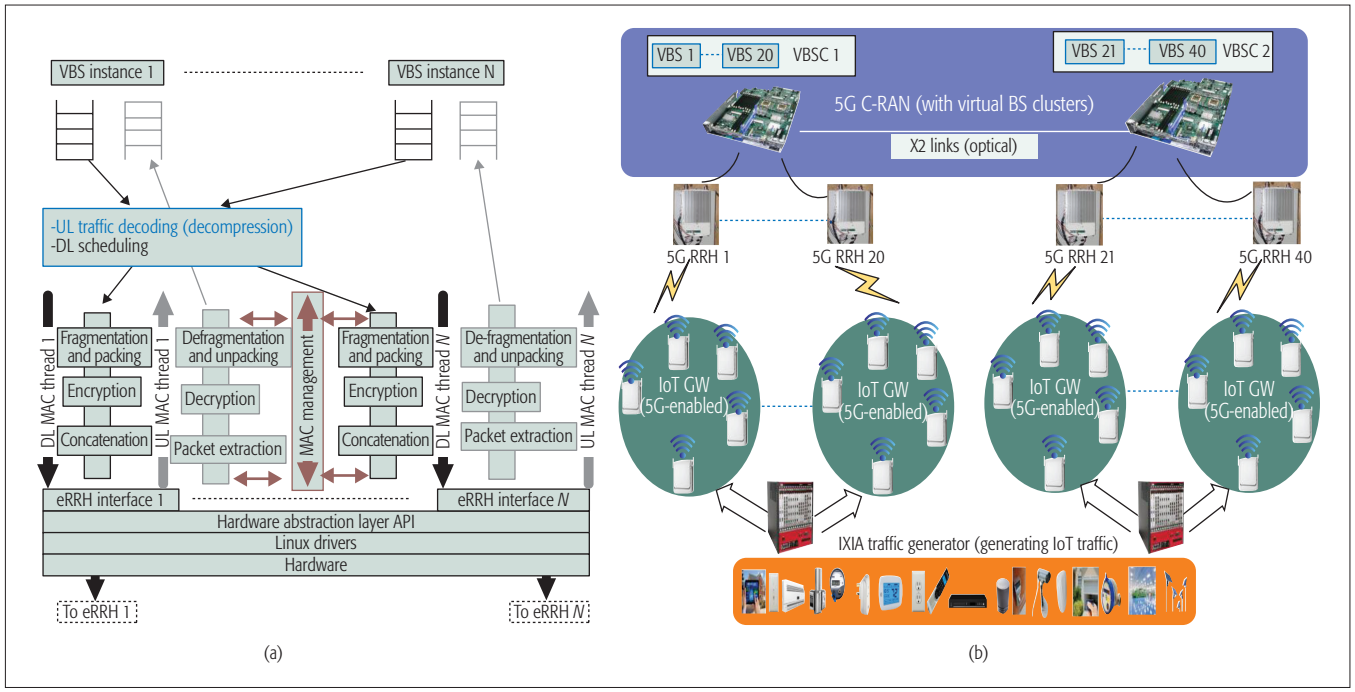


Figure 4. C-RAN prototype and testbed: a) C-RAN prototype with software VBS instances; b) 5G CRAN testbed setup for experiments.

gateways are now being equipped with proximity-based D2D communications for supporting direct inter-gateway connectivity bypassing the C-RAN. Once the D2D connectivity is set up, using the gateways, the IoT devices can communicate with other IoT devices under the coverage of a neighboring gateway (in close proximity). This will enhance the IoT service quality and reduce the burden on the 5G-enabled C-RAN. In the subsequent section, we explain the 5G C-RAN prototypes and testbed for providing connectivity to these IoT-enabled 5G gateways.

5G C-RAN Prototype and Testbed Development for IoT Support

We now briefly describe our 5G C-RAN prototype containing multiple VBS software instances and then discuss the testbed setup for experimental evaluation.

Prototype Implementation

Our VBSC prototype is built on IBM X3650 servers, having dual-core Intel X5355 processors over a Linux kernel. High-speed 10 Gb/s optical fibers are used for fronthaul communication between VBSC and Ethernet RRHs (eRRHs). Multiple VBSCs (server nodes) are interconnected by 10 Gb/s X2 interface. As shown in Fig. 4a, multiple VBSs in a single server operate as separate processing channels, linking to their dedicated data interfaces on that eRRH. Separate eRRH interfaces are provided to each VBS instance by using a hardware abstraction layer (HAL) application programming interface (API). To avoid resource conflicts, the resource management component provides the physical resource allocation to each VBS in the server. The VBSC supports the 5G wireless physical layer [12] and its associated system parameters.

For efficient design of MAC and upper layers, VBS instances use:

- Multi-threading to parallelize the workload
- Concurrent data structures to improve thread scalability
- Minimum memory copy to reduce memory overhead
- Thread pooling to reduce thread creation and destruction overhead

Each software MAC has its own self-configured threading model, driven by the actual data traffic experienced. Figure 4a shows this multithreaded design of uplink and downlink MAC layer, where one MAC process handles multiple logical MAC instances. The MAC processing divides the tasks into two groups based on their real-time requirements. Tasks requiring strict deadlines, like scheduling, packing, fragmentation, and encryption, are allocated more working threads, higher priority, and dedicated hardware resources. On the other hand, tasks with more relaxed deadlines, like automatic repeat request (ARQ) and MAC management, have limited working threads and hardware resources. The decompression of uplink traffic is incorporated in the VBS MAC instances, working across multiple threads. Every VBS MAC instance can communicate with other instances in the same cluster, using fast inter-process communication (IPC) for efficient resource pooling. The VBS MAC instances also communicate with other instances in the different cluster (residing in different servers) using gigabit X2 interfaces. The MAC management and control signals are communicated to the associated eRRHs by using the HAL API and fronthaul optical fibers.

Testbed Setup

Figure 4b shows the dense 5G C-RAN testbed setup for validating our C-RAN prototype. The testbed consists of two VBSCs, implemented in two IBM X3650 servers, each consisting of 20 VBS instances (i.e., 40 VBS instances total). Every VBS instance is connected to an eRRH using 10 Gb/s optical fronthaul cables. The two servers are

5G radio access network models	
Frequency	5G mmWave: 27.925 GHz [12] 4G: 2.3 GHz
Channel bandwidth	5G: 200 MHz [11], 4G: 20 MHz
Channel model	Urban channel [11]
5G RRH's Tx power	31 dBm
IoT gateway radius	100 m (4G and 5G)
MAC frame	5G: 0.2 ms, 4G: 1 ms
Antenna gain	9 dBi
Penetration loss	20 dB
Path loss compensation	3.8
Number of IoT gateways	50–200
IoT gateway's Tx power	10 dBm
Uplink scheduler	Proportional fair
Max. time threshold	1 s
Packet inter-arrival	[0.1–1] s
Packet size	100 bytes [15]

Table 1. 5G radio and system parameters.

also connected using gigabit backhaul interfaces. We have collected a month's actual IoT traffic data from *Fiesta-IoT* (http://fiesta-iot.eu/fiesta-experiments/ [October 22, 2016]) — a global federation of IoT testbeds and data sets, comprising thousands of IoT devices. Statistical analysis of the *Fiesta-IoT* traffic [15] reveals that most of the packets are small-sized (typically within 100 bytes) with widely varying inter-arrival rates of 0.1–1 s [15]. The IXIA traffic generator (https://www.ixiacom.com/sites/default/files/resources [October 22, 2016]) is used to emulate and generate this real IoT traffic for input into 5G-enabled IoT gateways. Due to the limitation in number of IoT gateways, a set of 5 IoT gateways are kept in the coverage area of every RRH, that is, a total of 200 gateways are used for communicating with 40 RRHs. We have used a dense urban 5G channel model [11] and RF parameters [12], which mentions 31 dBm RRH transmission power, 5 dBm idle power, 20 dB penetration loss, 0.8 shadowing deviation, 10 dBm IoT gateway maximum power in 27.925 GHz frequency band, and 200 MHz bandwidth. Table 1 highlights the major parameters related to 5G physical channels, IoT gateways, and IoT devices used for all testbed and simulation experiments. Over the next section, we perform the series of experiments and discuss the corresponding results.

Results and Discussions

Figure 5a shows the reduction in uplink radio access distribution experienced. While existing IoT gateways, over 4G wireless networks, experience up to 120 uplink accesses per second, the introduction of 5G-enabled gateways with packet classification and delayed uplink transmission reduces the uplink access to around 60 (i.e., almost half the original accesses). The efficient compression schemes further reduce the average uplink access to only ≤ 25, that is, less than half of the uplink access achieved by only packet classification and delayed access. Thus, the overall uplink access reduction, achieved by combination of both packet classification and compression, is almost 80 percent (from 120 to 25).

As 5G BSs are expected to support many human-to-human (H2H) users, only a small fraction of the uplink resources will be available for the uplink IoT traffic. Figure 5b depicts the improvement in device capacity of 5G wireless for different fractions of uplink radio resource availability. For 15 percent radio resource availability, exploring wider mmWave bandwidth, legacy 5G-enabled gateways (i.e., without classification and any compression) can support up to 110 IoT gateways (i.e., almost 9 times over legacy 4G wireless). Our proposed 5G-enabled IoT gateway with LZSS-based traffic compression further doubles the capacity to support more than 210. Thus, the combination of wider mmWave bandwidth and our proposed compression schemes results in a whopping 20 times total capacity gain over existing 4G wireless IoT gateways. For a comparative capacity gain study, we have also evaluated the IoT gateway capacity with two major alternative compression algorithms: LZW compression and Huffman coding. Figure 5b demonstrates that all three compression schemes achieve similar capacity gains, with LZW and LZSS compressions achieving slightly better performance over Huffman coding.

As mentioned before, a wide variety of IoT applications are expected to be delay-tolerant. These applications generally issue sporadic, intermittent uplink transmissions. Our 5G-enabled IoT gateway exploits these delay-tolerant applications to efficiently compress the uplink data generated by IoT-enabled devices. This grouping and compression introduce some additional latency in the 5G IoT gateways. For practical purposes, the additional latency should not be significant enough to have a negative impact on service quality. Figure 5c illustrates the average packet delay distribution. The sub-millisecond 5G MAC reduces the round-trip latency (downlink + uplink) from 50 ms to around 10 ms and 5 ms with and without uplink data compression, respectively. The increase in IoT capacity is achieved at the cost of a 5 ms increase in latency, which is negligible for delay-tolerant applications. Note that the delay-sensitive packets generated from the real-time applications do not incur any additional delay.

Figure 5d points out that our 5G C-RAN prototype can achieve almost 60 percent lower total power consumption in comparison to existing 5G wireless. The software stack implementations of VBS in the cloud, with the resource virtualization and pooling, help in reducing this energy consumption. Furthermore, in order to capture the effects of both traffic and energy dynamics, we have also used the energy efficiency metric, measured as traffic transmitted by unit energy. Figure 5e depicts that our proposed scheme offers almost 50 percent improvement in the overall

A wide variety of IoT applications are expected to be delay-tolerant. These applications generally issue sporadic, intermittent uplink transmissions. Our 5G-enabled IoT gateway exploits these delay-tolerant applications to efficiently compress the uplink data generated by IoT-enabled devices.

energy efficiency over traditional 5G wireless. The improvement is even more during off-peak periods (before 9 a.m. and after 8 p.m.).

Finally, it should be noted that our proposed 5G-enabled IoT gateways buffer the delay-tolerant IoT packets to perform compression before the uplink transmissions. This requires some additional storage overhead in the IoT gateways. Using the IXIA traffic generator, we emulate the uplink traffic generated by up to one million IoT devices. Figure 5f shows the additional uplink storage and CPU overhead associated major data compression algorithms. It points out that even with 1-million-device capacity, the average additional storage overhead experienced by the IoT gateways, equipped with our LZSS compression, is

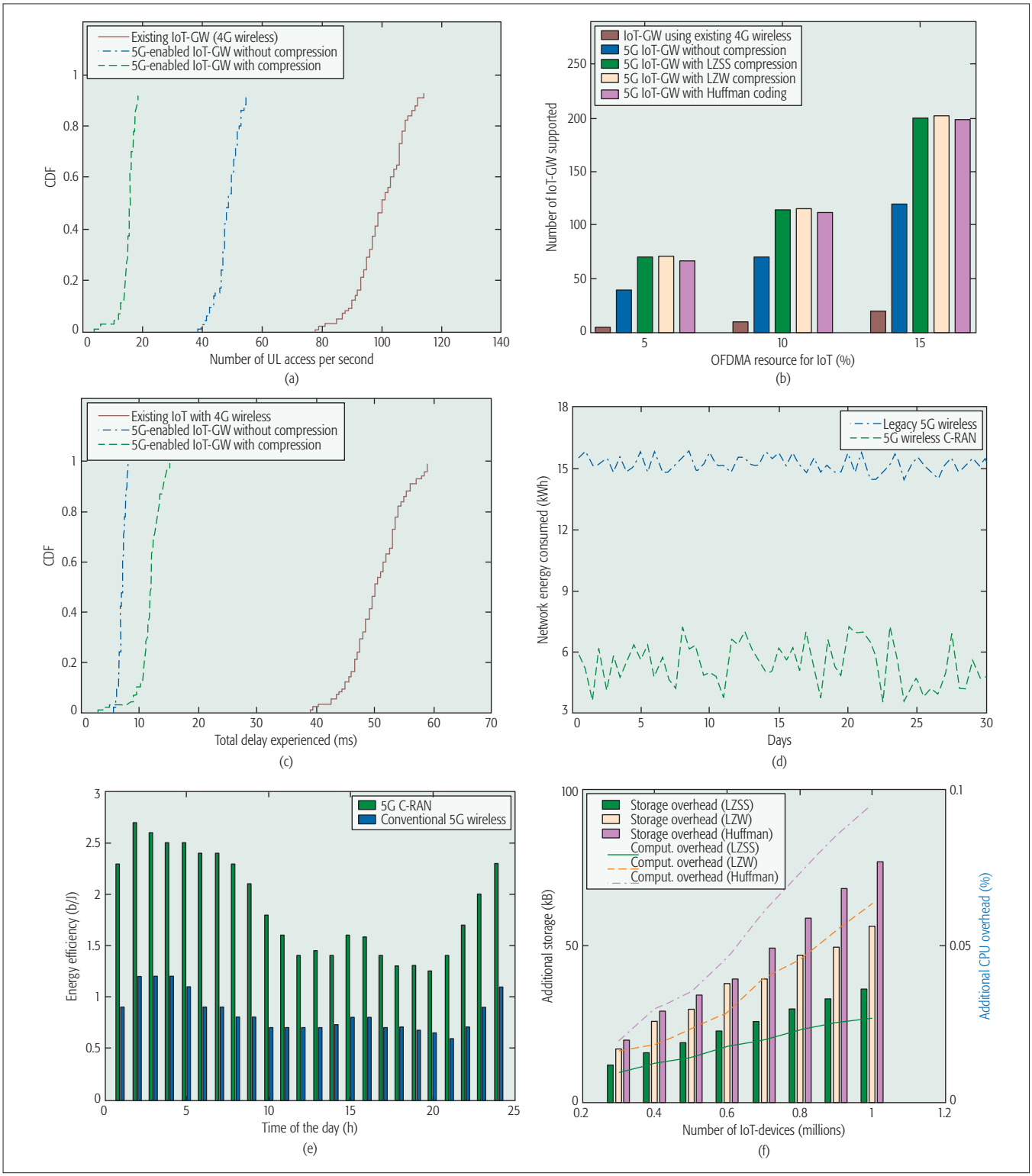


Figure 5. Simulation results showing uplink access, IoT GW capacity, latency dynamics, energy consumption, energy dynamics, and overhead: a) uplink access distribution; b) device capacity of cells; c) delay experienced by IoT-enabled devices; d) improvement in energy consumption; e) energy efficiency using C-RAN; f) traffic and computation overhead.

only 40 kbytes. On the other hand, complex compression schemes like Lempel-Ziv-Welch (LZW) and Huffman coding involve almost twice as much additional memory storage. The sliding-window-based codebook of LZSS is responsible for these gains in storage requirements. With present multi-core embedded processors, the additional computation overhead of our LZSS-based compression scheme demands only 0.04 percent extra CPU usage. On the other hand, the more complex LZW and Huffman codings involve almost three and five times more additional CPU usage compared to our proposed LZSS compression. Note that this additional storage and computation overhead represents extra memory and processing (besides the existing uplink buffer required for storing the uplink packets) associated with compression and codebook management. Thus, while LZW and Huffman coding achieve similar capacity gains, the complex algorithms impose more storage and CPU requirements on the IoT gateways. This points out the effectiveness of relatively simple LZSS compression schemes for IoT gateway design.

Conclusion

The increasing penetration of smart IoT devices is raising significant new challenges in already resource-constrained wireless networks. Existing wireless access, like WiFi and unlicensed bands, often suffer from heavy collisions and spectrum congestion with exponential increase in IoT devices. Next generation 5G wireless networks are expected to explore a set of new, emerging features, like mmWave, C-RAN, HetNets, and massive MIMO for providing enormous connectivity and data for massive rollout of IoT. Exploring these new features, we develop a 5G-enabled IoT gateway and introduce optimal uplink traffic compression schemes for efficient uplink resource utilization. Our 5G-enabled IoT gateway design, 5G C-RAN prototype, and laboratory experiments point out that huge improvement in IoT-device capacity is possible by using sophisticated uplink data compression and efficient wireless resource virtualization.

Acknowledgments

This research is supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF- 2016R1D1A1B03935633).

References

[1] A. Al-Fuqaha *et al.*, "Internet of Things: A Survey on Enabling Technologies, Protocols and Applications," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 4, 2015, pp. 2347–76.

[2] H. Holma, A. Toskala and J. Reunanen, *LTE Small Cell Optimization: 3GPP Evolution to Release 13*, Wiley, ISBN: 978-1-118-91257-7, 2015, pp. 1–462.

[3] F. Ghavimi and H-H Chen, "M2M Communications in 3GPP LTE/LTE-A Networks: Architectures, Service Requirements, Challenges, and Applications," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 2, 2015, pp. 525–49.

[4] A. Al-Fuqaha *et al.*, "Toward Better Horizontal Integration among IoT Services," *IEEE Commun. Mag.*, vol. 53, no. 9, Sept. 2015, pp. 72–79.

[5] M. W. Condry and C. B. Nelson, "Using Smart Edge IoT Devices for Safer, Rapid Response with Industry IoT Control Operations," *Proc. IEEE*, vol. 104, no. 5, 2016, pp. 938–46.

[6] C. S. Bontu, S. Periyalwar, and M. Pecen, "Wireless Wide-Area Networks for Internet of Things", *IEEE Vehic. Tech. Mag.*, vol. 9, no. 1, 2014, pp. 54–63.

[7] M. Agiwal, A. Roy, and N. Saxena, "Next Generation 5G Wireless Networks: A Comprehensive Survey," *IEEE Commun. Surveys & Tutorial,s* vol. 18, no. 3, 2016, pp. 1617–55.

[8] J. G. Andrews *et al.*, "What Will 5G Be?" *IEEE JSAC*, vol. 32, no. 6, 2014, pp. 1065–82.

[9] 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.

[10] A. Checko *et al.*, "Cloud RAN for Mobile Networks — A Technology Overview," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 1, 2015, pp. 405–26.

[11] T. S. 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.

[12] 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.

[13] F. Boccardi *et al.*, "Five Disruptive Technology Directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 74–80.

[14] T. Bell, J. Cleary, and I. Witten, *Text Compression*, Prentice-Hall, 1990, pp. 1–318.

[15] M. Shafiq *et al.*, "A First Look at Cellular Machine-to-Machine Traffic: Large Scale Measurement and Characterization," *Proc. Int'l. Conf. Measurement and Modeling of Computer Systems*, 2012, pp. 65–76.

Biographies

Navrati Saxena [M] (navrati@skku.edu) is an associate professor in the Electrical Engineering Department, Sungkyunkwan University (SKKU), South Korea. She was an assistant professor in Amity University India and a visiting researcher at the University of Texas at Arlington. She completed her PhD from the Department of Information and Telecommunication, University of Trento, Italy. Her research interests involve 4G/5G wireless, IoT and smart environments. She directs the Mobile Ubiquitous System Information Center (MUSIC) at SKKU and serves as a Guest Editor and Technical Program Committee member of international journals and conferences. She has co-authored one book (Taylor & Francis) and published in 40 international journals.

Abhishek Roy [M] (abhishek.roy@samsung.com) is currently working in the Networks Division, Samsung Electronics South Korea. He received his PhD in 2010 from SKKU and his M.S. in 2002 from the University of Texas at Arlington. His research interests include mobility and resource management in 4G/5G wireless systems, IoT, and smart grids. He serves as a Guest Edito and Technical Program Committee member of many international journals and conferences. He has co-authored one book (Taylor & Francis) and published in 40 international journals.

Bharat J. R. Sahu [StM] (bjrsahu@skku.edu) is currently a Ph.D. candidate in the Electrical and Computer Engineering Department, SKKU. Prior to joining SKKU, he worked as a research fellow at LNMIT, India. He received his B.S. and M.S. degrees from Sambalpur University, India. His research interests include 4G/5G wireless systems, IoT, M2M communications, and smart grids.

HanSeok Kim (hs365.kim@samsung.com) received his B.S. and M.S. degrees in electronics engineering from Seoul National University, Korea, in 1990 and 1992, respectively, and received his Ph.D. degree in electrical and computer engineering from Purdue University in 2003. He worked for LG Electronics developing office automation products from 1992 to 1997. In 2003, he joined Samsung Electronics, where he is presently a Samsung Master in the Network Division. His research interests include traffic modeling, flow control, resource management, and performance analysis in wireless networks.

Our 5G-enabled IoT gateway design, 5G C-RAN prototype and laboratory experiments point out that huge improvement in IoT-device capacity is possible by using sophisticated uplink data compression and efficient wireless resource virtualization.

Coding for Caching in 5G Networks

Yasser Fadlallah, Antonia M. Tulino, Dario Barone, Giuseppe Vettigli, Jaime Llorca, and Jean-Marie Gorce

The authors provide an overview of the emerging caching-aided coded multicast technique, including state-of-the-art schemes and their theoretical performance. Then they focus on the most competitive scheme proposed to date and describe a fully working prototype implementation in CorteXlab, one of the few experimental facilities where wireless multiuser communication scenarios can be evaluated in a reproducible environment.

ABSTRACT

One of the major goals of the 5G technology roadmap is to create disruptive innovation for the efficient use of the radio spectrum to enable rapid access to bandwidth-intensive multimedia services over wireless networks. The biggest challenge toward this goal lies in the difficulty in exploiting the multicast nature of the wireless channel in the presence of wireless users that rarely access the same content at the same time. Recently, the combined use of wireless edge caching and coded multicasting has been shown to be a promising approach to simultaneously serve multiple unicast demands via coded multicast transmissions, leading to order-of-magnitude bandwidth efficiency gains. However, a crucial open question is how these theoretically proven throughput gains translate in the context of a practical implementation that accounts for all the required coding and protocol overheads. In this article, we first provide an overview of the emerging caching-aided coded multicast technique, including state-of-the-art schemes and their theoretical performance. We then focus on the most competitive scheme proposed to date and describe a fully working prototype implementation in CorteXlab, one of the few experimental facilities where wireless multiuser communication scenarios can be evaluated in a reproducible environment. We use our prototype implementation to evaluate the experimental performance of state-of-the-art caching-aided coded multicast schemes compared to state-of-the-art uncoded schemes, with special focus on the impact of coding computation and communication overhead on the overall bandwidth efficiency performance. Our experimental results show that coding overhead does not significantly affect the promising performance gains of coded multicasting in small-scale real-world scenarios, practically validating its potential to become a key next generation 5G technology.

INTRODUCTION

Along with the Internet revolution, IP traffic is growing at a tremendous pace and is expected to reach two zetabytes per year by 2019. Mobile data networks are envisioned to support up to 14 percent of this global data traffic coming from a plethora of different market segments. Among these segments, multimedia streaming is the service with the highest penetration rate, having a major impact on the overall traffic increase. On the other hand, current mobile network generations cannot cope with this explosive traffic growth due to the capacity limitations of radio access, backhaul, and core mobile networks, and the increasingly unicast and on-demand nature of users' content demands. In order to support this traffic expansion, the fifth generation (5G) of mobile networks is under preparation. Among the key performance challenges that 5G needs to address are throughput, latency, and energy efficiency. That is, 5G is expected to provide 1000× higher throughput, sub-millisecond service latencies, and up to 90 percent overall energy savings [1]. Despite the myriad of technological advances at the physical (PHY) and medium access control (MAC) layers, such as inter-cell interference coordination (ICIC), massive multiple-input-multiple-output (MIMO), and carrier aggregation, targeted data rates are still significantly out of reach. To this end, 5G envisions novel architectural components for the next generation radio access network (RAN), including small cell densification, efficient wireless backhauling, and network self-organization [1]. In this context, the use of inexpensive storage resources within the RAN is emerging as a promising approach to reduce network load and effectively increase network capacity.

PROMINENCE OF WIRELESS CACHING IN 5G

Wireless caching (i.e., caching content within the wireless access network) is gaining interest, especially in ultra-dense networks where many connected devices try to access various network services under latency, energy efficiency, and/or bandwidth limitation constraints [1]. Proactively caching content items at the network edge (e.g., at the RAN) helps relieve backhaul congestion and meet peak traffic demands with lower service latency, as Fig. 1 illustrates. For maximum benefits, network operators can intelligently exploit users' context information, classify content by popularity, and improve predictability of future demands to proactively cache the most popular content before being requested by end users. Such a strategy is able to fulfill the quality of service (QoS) requirements while significantly reducing the use of bandwidth resources and the associated energy consumption. Content items can be cached at different locations of the mobile network. Within the RAN, base stations (or small base stations), user equipment (UE) devices, and access points (APs) can be enhanced with additional memory for content caching. While caching can also hap-

Yasser Fadlallah and Jean-Marie Gorce are with INSA-Lyon; Yasser Fadlallah is also with the University of Sciences and Arts in Lebanon; Antonia M. Tulino, Dario Barone, and Giuseppe Vettigli are with the University of Naples Federico II; Antonia M. Tulino is also with Nokia Bell Labs; Jaime Llorca is with Nokia Bell Labs.

Digital Object Identifier:
10.1109/MCOM.2017.1600449CM

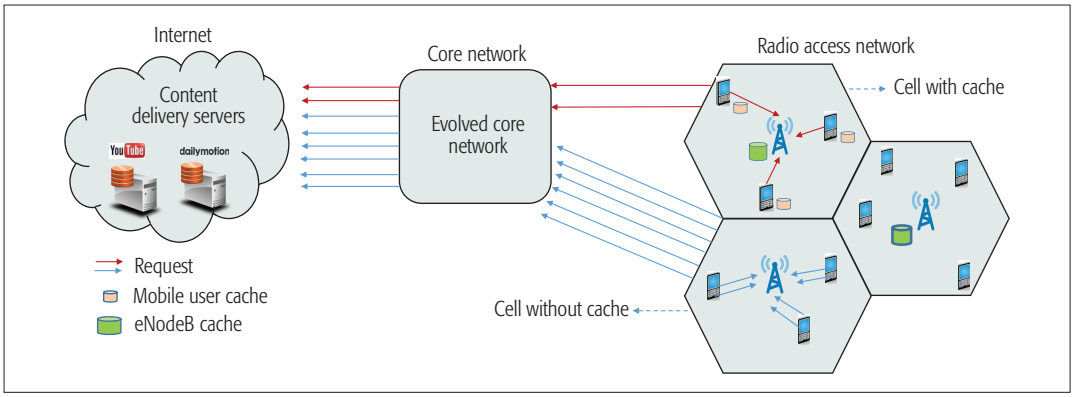


Figure 1. Caching within the radio access network: impact on network load and traffic congestion.

Wireless caching (i.e., caching content within the wireless access network) is gaining interest, especially in ultra-dense networks where many connected devices try to access various network services under latency, energy efficiency, and/or bandwidth limitation constraints.

pen within the evolved packet core (EPC), the main benefit of caching at the EPC is to reduce peering traffic between Internet service providers (ISPs). It is the additional deployment of cache memories within the RAN that can crucially help minimize intra-ISP traffic, relieving backhaul load and reducing service latency [2].

From Uncoded to Coded Content Distribution

A substantial number of recent studies have analyzed the use of wireless caching as a promising solution for 5G. Among these studies, [3] introduced the idea of femtocaching and addressed the question of which files should be assigned to which helper nodes (femtocell-like base stations), while [4, references therein] considered the improvement in caching efficiency that can be obtained by dynamically learning content popularity and updating cache contents at the network edge. Despite considerable interest, such studies focus on the data placement problem in isolation, assuming the use of unicasting or naive (uncoded)[1] multicasting during transmission, and hence ignore the potential benefits of joint placement and transmission code design.

In [5], the data placement problem is generalized to the coded content distribution problem where the goal is to jointly determine the placement and routing of (possibly coded) information over the network, showing that joint code design significantly increases multicast efficiency, leading to substantial improvements in reducing network load and access latencies. A number of information-theoretic studies have then characterized the order-optimal performance of a caching network of special practical interest, the shared link caching network, formed by a single source node (e.g., base station) with access to a library of content files connected via a shared multicast link to multiple user nodes (e.g., end devices or APs), each with caching capabilities. In this context, the work in [6] showed that under worst case demands, caching portions of each file uniformly at random and using index coding (IC) [7] during transmission yield an overall load reduction that is proportional to the aggregate cache size. In [8], the authors analyzed the case in which user demands follow a Zipf popularity distribution, designing order-optimal achievable schemes[2] that adjust the caching distribution as a function of the system parameters to balance the gains from local cache hits and coded multicasting. Shortly after, [9] showed that the gains achieved by these schemes require a number of chunks per requested item that grows exponentially with the number of caches in the system, leading to codes of exponential complexity that compromise their theoretical gains. Efficient polynomial-time schemes (e.g., [10]) have then been proposed to recover a significant part of the promising multiplicative caching gain.

In terms of practical implementations, the work in [4] provided a big data platform where learning algorithms can be used to predict content popularity and drive caching decisions, but the benefit of the learning techniques for improving caching efficiency is evaluated via numerical simulations. In addition, only conventional uncoded schemes are considered, and aspects related to advanced coding techniques such as caching-aided coded multicasting that can potentially provide much larger gains are largely overlooked. It is also important to note that so far, only information-theoretic studies have shown the potential gains of such schemes, and the emulation work in [11] only considers two to four users and three files, a very limited scenario that does not allow showing the real impact of computational complexity and coding overhead. Moreover, it is unclear whether existing schemes meet the requirements of current technologies, thus leaving plenty of open questions regarding practical performance benefits.

The Need for Experimental Validation

This article aims to bridge the gap between theory and practice in order to validate the benefits of caching-aided coded multicasting by designing a fully working prototype implementation and testing it in a large network testbed. Such a testbed and prototype implementation provide a cornerstone for the evaluation of future schemes with more advanced wireless caching protocols and cache-enabled PHY layer techniques such as joint source-channel coding.

We first provide an overview of the caching and coded multicasting framework, and discuss the key concepts behind the ability to provide load reductions that are proportional to the aggregate cache size. We then introduce a new frame structure that includes specific fields to account for all the practical aspects required for a fully working real-world implementation. The primary role of the newly designed frame structure is to allow decoding of coded data at each receiver. Our MAC layer frame design is combined with an orthogonal frequency-division multiplexing

[1] The term uncoded is used to refer to a scheme in which at each use of the channel, the transmission is composed of packets that belong to the same file, while the term coded refers to a scheme in which transmissions can be composed of a mixture of packets from different files.

[2] An achievable scheme is said to be order-optimal if, as the file size goes to infinity, the number of transmissions needed to satisfy the user demands scales as the information-theoretic optimal number of transmissions needed to satisfy the user demands; that is, the ratio between the achievable and optimal number of transmissions is upper bounded by a constant independent of all the system parameters.

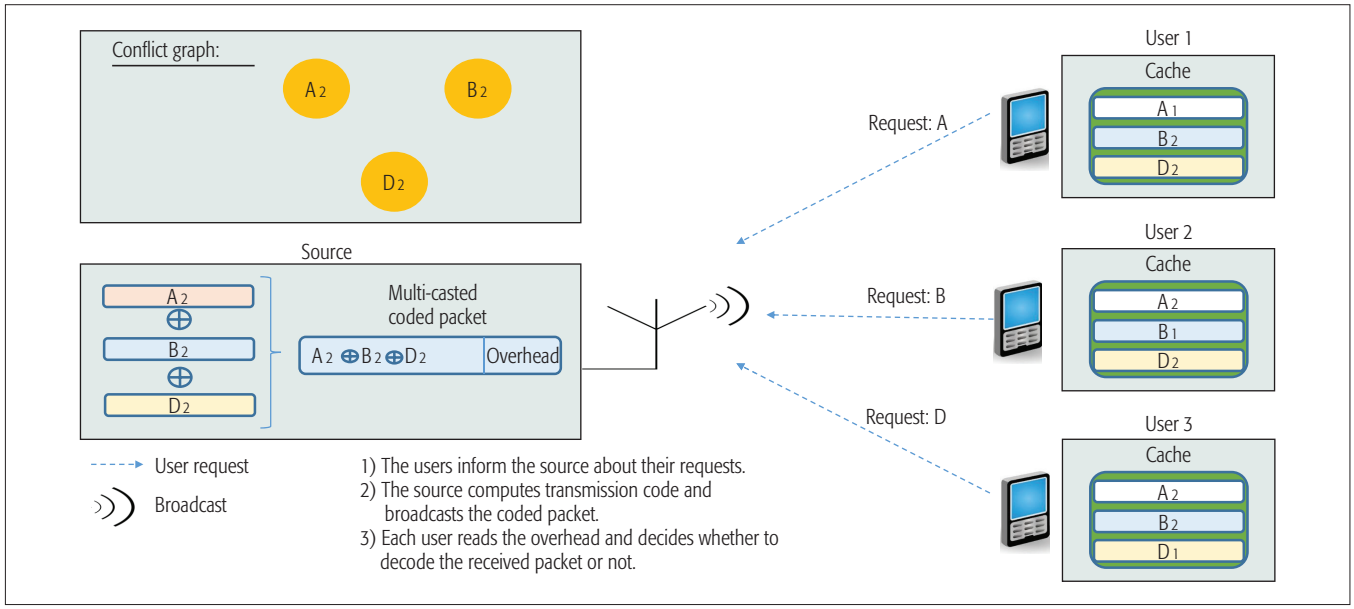


Figure 2. Caching-aided coded multicast design in a three-user SLCN where each user is equipped with a cache of storage capacity of one and a half files and requests one file from a library of four binary files.

(OFDM) PHY layer, which makes it compatible with LTE-Advanced mobile networks or further PHY layer standards. The resulting fully working prototype is implemented in a large-scale testbed facility, CorteXlab [12], composed of tens of highly flexible radio nodes deployed in a controlled and reproducible environment. We present experimental results in the context of key 5G performance metrics related to transmission delay, bandwidth usage, and energy efficiency. Our experimentation validates the fact that memory can be effectively turned into bandwidth, leading to substantial network throughput gains.

Caching-Aided Coded Multicasting

As previously stated, the use of caching together with smart offloading strategies in a RAN composed of evolved NodeBs (eNBs), APs (e.g., WiFi), and UEs can significantly reduce backhaul traffic and service latency. In this context, a shared link caching network (SLCN) topology can be identified at different levels of the mobile network. Indeed, a radio cell constitutes an SLCN where the eNB acts as the source node connected to the UEs via a shared multicast link. In addition, an SLCN can also be formed by a core network (CN) server (source node) connected to a set of eNBS via a shared wireless backhaul. In both cases, user nodes are equipped with storage resources for content caching. Accordingly, we focus on the analysis and implementation of an SLCN composed of a source node, with access to a library $\mathcal{F}$ of m binary files, connected to n user nodes via a shared multicast link. Each user node is equipped with a cache of storage capacity equivalent to M files, and can make up to L file requests according to a Zipf demand distribution. A multicast link is a shared channel in which any transmission can be overheard by all receivers.

A caching-aided coded multicast scheme is performed over two phases:

- The caching phase, where the source node populates the user caches with appropriate functions of the content library
- The delivery phase, where the source forms a multicast codeword to be transmitted over the shared link in order to meet the users' content demands

These phases are generic for both coded and uncoded schemes, but naively performed in the uncoded case. In fact, when relying on uncoded or naive multicasting during the delivery phase, it is well known that the optimal caching strategy is to cache the top M most popular files at each user cache. However, in general, this is far from optimal when coding can be used in the delivery phase [8]. In the following, we discuss the potential of caching-aided code design and illustrate its major advantages compared to the optimal caching policy under uncoded (naive) multicasting.

Random Fractional Caching

Each binary file $f \in \mathcal{F}$ is divided into B_f equal-size chunks. Given the *caching distribution* $\{pf\}$, with $\Sigma_{f=1}^{m} p_f = 1$, each user caches chunks of file f with probability p_f. That is, each user caches a number of chunks $p_f MB_f$ of file f chosen uniformly at random. It is important to note that the randomized nature of the selection process allows users to cache different sets of chunks of the same file, shown to be key in creating coded multicast opportunities during the delivery phase. In [8], the authors showed that the optimal caching distribution can be approximated by a truncated uniform distribution $p_f = 1/\tilde{m}$, $\forall \leq \tilde{m}$ and $p_f = 0$, $\forall f > \tilde{m}$, without affecting order optimality,[3] and referred to this caching policy as random least frequently used (RLFU). Compared to the least frequently used (LFU) caching policy (the best option under naive multicasting), where the same most popular files are entirely cached at each user, RLFU maximizes the amount of distinct chunks collectively cached by the network.

Coded Multicasting

A simple example in Fig. 2 illustrates the key benefits of coded multicasting during the delivery phase. The network is composed of a source and

[3] Details about the selection of the optimal $\tilde{m}$ are given in [7].

3 user nodes requesting files from a library of $m = 4$ binary files $\mathcal{F} = \{A, B, C, D\}$. Each file (e.g., video segment) is divided into 2 chunks, yielding a library of chunks $\mathcal{C} = \{A_1, A_2, B_1, B_2, C_1, C_2, D_1, D_2\}$. During the caching phase, users 1, 2, and 3 randomly fill their caches with chunks $\{A_1, B_2, D_2\}$, $\{A_2, B_1, D_2\}$, and $\{A_2, B_2, D_1\}$, respectively. During the delivery phase, at a given request round, users 1, 2, and 3 make requests for video segments A, B, and D, respectively. Under an uncoded naive multicasting transmission scheme, the source needs to transmit the missing chunks A_2, B_2, and D_2 over the shared multicast link using three time slots. In contrast, by employing coded multicasting, the source can mix the three chunks A_2, B_2, and D_2 via an XOR operation (binary addition) and multicast the coded chunk $A_2 \oplus B_2 \oplus D_2$ using only one time slot. Clearly, in this case, coded multicasting reduces the number of transmissions (and hence the number of delivery time slots) by a factor of three.

As illustrated in the above example, a given user is able to decode its requested chunk from a mixture of combined chunks if and only if it has knowledge of all other combined chunks. Such a problem can be seen as an IC problem [7], and can be described by what is referred to as the *conflict graph* [8]. The conflict graph is constructed such that each graph vertex corresponds to one requested chunk, and an edge between two vertices is created if they correspond to different requested chunks, and for each vertex, the associated chunk is not included in the cache of the user requesting the chunk associated with the other vertex. Notice that an edge between two vertices indicates that their associated chunks must be separately transmitted, while non-connected vertices can be modulo summed via XOR operation [7]. The goal is to find the best chunk combinations such that the total number of transmissions is minimized. A common approach, referred to as chromatic index coding (CIC) [8], is to compute a minimum *graph coloring* of the IC conflict graph, where the goal is to find an assignment of colors to the vertices of the graph such that no two connected vertices have the same color, and the total number of colors is minimized. The multicast codeword is constructed by generating sub-codewords obtained by XORing the chunks with the same color and then concatenating the resulting sub-codewords. The conflict graph of the example given in Fig. 2 is illustrated in the top left corner of the figure. The graph consists of three vertices corresponding to the three requested chunks A2, B2, and D2. There are no edges between the vertices of the graph since, for each vertex, the associated chunk is included in the cache of the users associated with the other vertices. Therefore, all vertices can be assigned the same color and binary added into a single coded transmission, as shown in Fig. 2.

The work in [8] showed that the combined use of RLFU caching and CIC coded multicasting is order-optimal[4] under any Zipf demand distribution, and that RLFU-CIC provides multiplicative caching gains, that is, the per-user throughput scales linearly or super-linearly with the cache size. In order to prove this result, the authors resort to a polynomial-time approximation of CIC, referred to as greedy constrained coloring (GCC). While GCC exhibits polynomial complexity in the number of users and chunks, both CIC and GCC can only guarantee the promising multiplicative caching gain when the number of chunks per file grows exponentially with the number of users, significantly limiting their practical performance [9]. Subsequently, the works in [10] and [13] extended the RLFU-CIC and RLFU-GCC schemes to the non-homogeneous SLCN and proposed two improved coded multicasting algorithms: the greedy randomized algorithm search procedure (GRASP) based on a greedy randomized approach; and the hierarchical greedy coloring (HGC). These algorithms have been shown to recover a significant part of the multiplicative caching gain, while incurring a complexity at most quadratic in the number of requested chunks.

Decoding Phase

From the observation of the received multicast codeword and its cached content, each user has to decode its intended chunks via its own decoding function. In order to guarantee decoding, the receiver needs to be informed (e.g., via a packet header that carries all necessary information, as shown in Fig. 3a) of the sub-codewords in the concatenated multicast codeword that contain any of its intended chunks. For each of the identified sub-codewords, the receiver obtains its intended chunks by performing simple binary addition.

In the next section, we describe a fully working prototype implementation that includes the design of the required packet header to ensure full decodability.

Implementation of Caching-Aided Coded Multicasting

While we previously described state-of-the-art wireless caching and transmission code design, the impact of real protocol overheads on the multiplicative caching gain remains an open question that we address via a real prototype implementation in the following.

Our prototype implementation is based on the following components: a simplified application layer for generating and combining the requested chunks; a MAC layer extended with additional header fields to allow decoding of coded chunks; and a PHY layer compliant with LTE standards. Our basic MAC layer frame implementation does not account for a complete standardized frame structure, and the generated data is not encapsulated through the protocol stack, since our main goal is a proof of concept of caching-aided coded multicasting and its real-time feasibility. In the following, we describe in detail the MAC layer frame structure.

Frame Structure

For a clear understanding of the implementation process, the basic frame structure is given in Fig. 3. Every packet is composed of two parts: header and payload. The payload represents the coded chunks (divided as: $payload_1$, ..., $payload_K$), a mixture of original data chunks with elements in the Galois Field of order two GF(2), making it easy to encode and decode with a simple XOR operation. The header illustrated in Fig. 3c contains

Our prototype implementation is based on the following components: a simplified application layer for generating and combining the requested chunks; a MAC layer extended with additional header fields to allow decoding of coded chunks; and a PHY layer compliant with LTE standards.

[4] Order-optimal in the sense that the number of transmissions needed to satisfy the user demands scales (in number of users, number of files, and memory size) as the optimal scheme.

The fact that the header information in the coded scheme depends on the number of served users implies a variable header length.

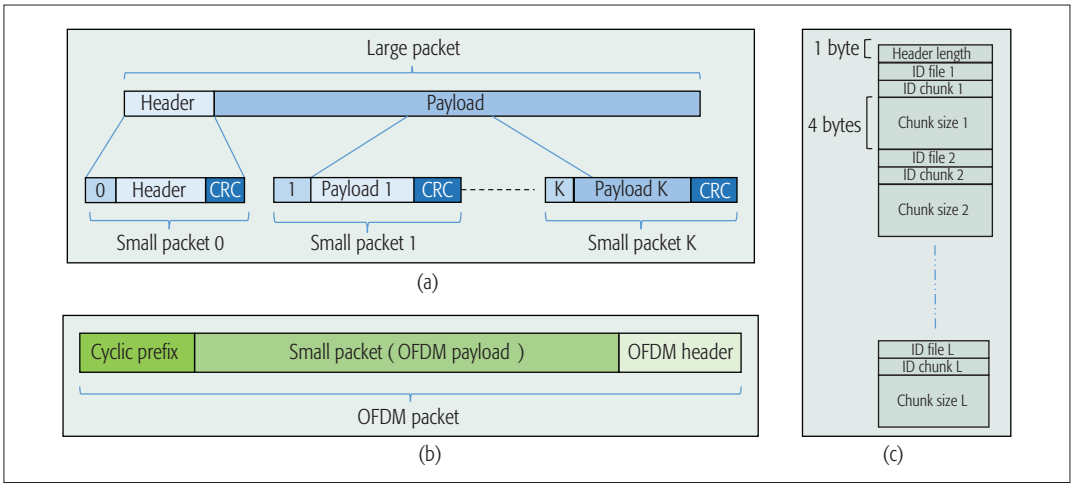


Figure 3. An example of the proposed frame structure: a) MAC layer frame; b) PHY layer frame; c) header structure.

the minimal information required for a successful extraction of each individual chunk. It carries the combined chunks identities and consists of the following information:

- Header length: This is the first element of the header, and its size is fixed to 1 byte. It carries the number of header bytes.
- File IDs: These are the IDs of the files to which the combined chunks (payload) belong. Each ID requires a multiple of 1 byte, depending on the number of content files in the library.
- Chunk IDs: These are the IDs of the combined chunks within the transmitted packet. Each ID requires a multiple of 1 byte, depending on how many chunks a file is partitioned into.
- Chunk sizes: These are the sizes of the combined chunks, and are encoded with a multiple of 4 bytes to make the receiver able to recognize the size of each chunk.

In a practical network scenario, it is unusual to have a header length exceeding 1 byte since the number of requests and the number of simultaneously served users is generally limited. Notice that the uncoded design will necessitate the same header structure but only for one target user (unless the same chunk is destined to multiple users) because no chunk combination is performed. This means that in an uncoded scheme each packet is separately transmitted with its associated header, without the need for additional overhead information. This is due to the fact that we are assuming multicast transmission over the downlink shared channel (DLSCH). This LTE physical layer transport channel is the main channel for downlink data transfer, and it is used by many logical channels. The fact that the header information in the coded scheme depends on the number of served users implies a variable header length. An example of the header decomposition is illustrated in Fig. 3c, where the number of files and chunks are assumed to not exceed 1 byte each, and the maximum size of a chunk is limited to 4 bytes. The payload length of a coded packet is equal to the largest chunk's length among the combined ones. Before being transmitted, the coded packet is partitioned into small packets and numbered such that the receiver can rebuild the original coded packet. Each coded packet (Fig. 2) is dedicated to users with IDs indicated in the header information. Aiming to decrease the packet error probability (PER), the first small packet will be limited to the header data, and the others are charged with the payload. A 32-bit cyclic redundancy check (CRC) is appended to each small packet for error detection. In the header information, if the CRC detects some errors, the whole coded packet is lost, and the user drops all related small packets. Otherwise, each user checks whether concerned or not. If so, the user proceeds to the small packet decoding, based on its cached data and the reported files and chunks IDs. Conversely, if the user is not concerned, the packet is dropped, and the user waits for the next header to check out its affiliation. In case of channel erasure, the small packets are replaced with dummy bytes.

CorteXlab Platform

The resulting fully working prototype is implemented in a large-scale testbed facility, CorteXlab [12] which is a testbed for cutting edge radio experimentation, composed of a mix of radio nodes, including single-input single-output (SISO) and multiple-input multiple-output (MIMO) software defined radio (SDR) nodes. The testbed shown in Fig. 4 is installed in a large (180 m^2) shielded room partly covered with electromagnetic wave absorbing material. User nodes are placed over a regular grid with an inter-node distance of 1.8 m, and accept any PHY layer implementations on both hardware and software. A unified server is available for starting, coordinating, and collecting the results of experiments. As a development tool, the GNU Radio software is employed for real-time experimentation.

End-to-End Performance Results and Perspectives

Setup Environment

Our SLN experimentation consists of one radio source node and n = 10 radio user nodes. Each user requests L = 10 files from a library $\mathcal{F}$ of m = 20 binary files, each of size 2.8 Mb. A cache of size M files is deployed at every user. Such a scenario can be seen as if the users are APs carry-

ing multiple requests from different UEs, and the source is the eNB having access to the content library. The file request distribution is drawn from a Zipf distribution with Zipf parameter α: $\alpha = 0$ returns a uniform request distribution; the higher the Zipf parameter α, the more skewed the request distribution. The binary files are partitioned into equally sized $B = 100$ chunks yielding a library of $m_b = 2000$ chunks. Both RLFU with $\tilde{m}$ optimized as in [8] and LFU caching policies are adopted for the coded and naive multicasting delivery schemes, respectively.

The output of the multicast encoder goes into an OFDM modulator with the following transmission parameters. Each PHY frame is decomposed into small packets of size 100 bytes to which a cyclic redundancy check (CRC-32) and an OFDM header are appended for error detection and OFDM demodulation, respectively. The OFDM header and payload data are mapped into binary phase shift keying (BPSK) and quadrature PSK (QPSK), respectively, and each symbol is transmitted over a sample duration of $T_s = 1$ µs. The data is carried over $L_f = 48$ subcarriers spaced by $\Delta_f = 15$ kHz, and the central frequency is set to 2.45 GHz.

Experimentation Results

The focus herein is on the gain at the MAC layer that is based on counting the total number of required bytes to serve all UEs. Assuming the same number of requests L from all users, the normalized minimum rate (NMR) is defined as $R_t/(L \times \text{file size})$, where R_t is the total number of required bytes at the MAC layer to satisfy all user demands. Note that NMR is in general a non-decreasing function of the number of users and a decreasing function of cache size, M; in particular, for $M = 0$, the NMR is equal to the total number of distinct user requests. In the following, we provide a numerical validation of the prototype performance in terms of NMR. Specifically, we analyze the performance of our prototype solutions prot-HGC and prot-GRASP in terms of NMR compared to:

- HGC and GRASP for finite file chunking simulated in the Matlab environment without taking implementation overhead into account
- Naive multicasting with LFU caching policy at the rate of the worst channel receiver
- The benchmark upper bound GCC when $B = \infty$ [8]

The trend in terms of NMR demonstrated by the prototype confirms the gains predicted by the theory. Figures 5a and 5b show the NMR as a function of the cache size and the Zipf parameter α, respectively. This metric is specially illustrative of the amount of bandwidth resources the wireless operator needs to provide in order to meet the receiver demands. In Fig. 5a, we assume a Zipf parameter $\alpha = 0$. Observe first the performance of naive multicast. As expected, the load reduces approximately linearly with the cache size M. Now observe how the significant multiplicative caching gains (w.r.t. naive multicast) quantified by the upper bound (RLFU-GCC with $B = \infty$) are remarkably preserved by the prototype solutions (prot-HGC and prot-GRASP), which achieve an NMR almost indistinguishable from the corresponding schemes implemented in the Matlab environment without taking into account the encoding and decoding overheard. Figure 5a clearly shows the effectiveness of the proposed implementation in allowing receivers to decode their requested files at an NMR very close to the theoretically optimal NMR. From Fig. 5a, it is also apparent that the two coloring algorithms have similar performance for $\alpha = 0$. The effectiveness of coded multicasting is highly influenced by the Zipf parameter α, as illustrated in Fig. 5b for cache size $M = 2$ files. Observe how the reduction in NMR enabled via coded multicasting is much more attractive in the region of $\alpha \leq 1$.

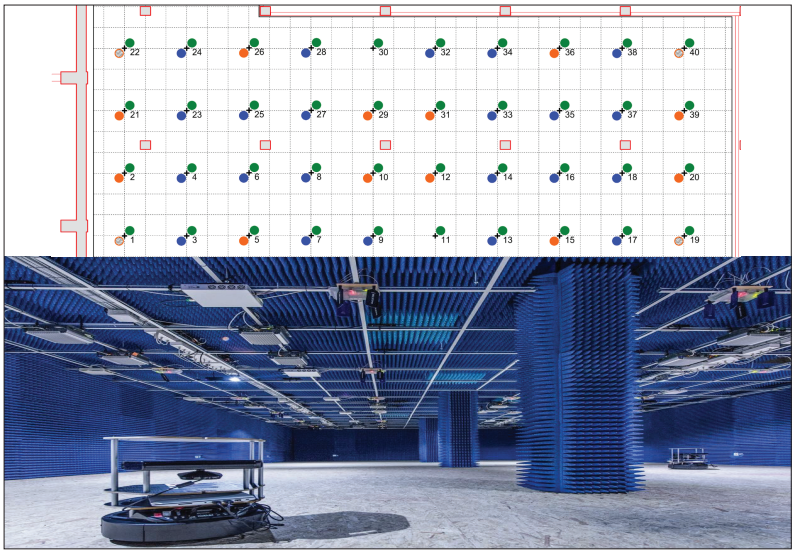


Figure 4. CorteXlab platform and the nodes placement map.

In order to illustrate the behavior of the multiplicative gains, Figs. 5c and 5d show the prototype NMR gains as a function of the cache size M and the Zipf parameter α, respectively. The gain of a given scheme is defined as the ratio between the NMR achieved by naive multicasting with LFU caching policy and the NMR achieved by that scheme. In particular, when the scheme is a prototype implementation, the NMR of naive multicasting is computed with its associated overhead. From Fig 5c, we can observe that the gain is a monotonic non-decreasing function of the cache size. In particular, for GCC-$B = \infty$ and large memory, such gain admits the following analytical expression, $M(1 - (1 - q_m)^n)$, where q_m denotes the probability of requesting file m. Note that we do not plot the point at $M = 20$ since it is well known that the NMR is zero for all the schemes, and hence the gain is given by an indeterminate form of type 0/0. Figure 5c shows that the gains achieved by prot-GRASP and prot-HgC are very close to the gains achieved by the corresponding MATLAB simulated schemes, confirming the little impact of the implementation overhead on the overall performance. Furthermore, it is worth noticing that due to the reduced number of transmitted coded packets compared to the number of uncoded packets transmitted by naive multicasting, the total overhead size is also smaller. That is, even though each packet header length is larger, the total number of header bytes over all transmissions is also reduced.

In terms of the Zipf parameter α, Fig. 5d shows that for $M = 2$ files, a gain close to 1.3 is obtained

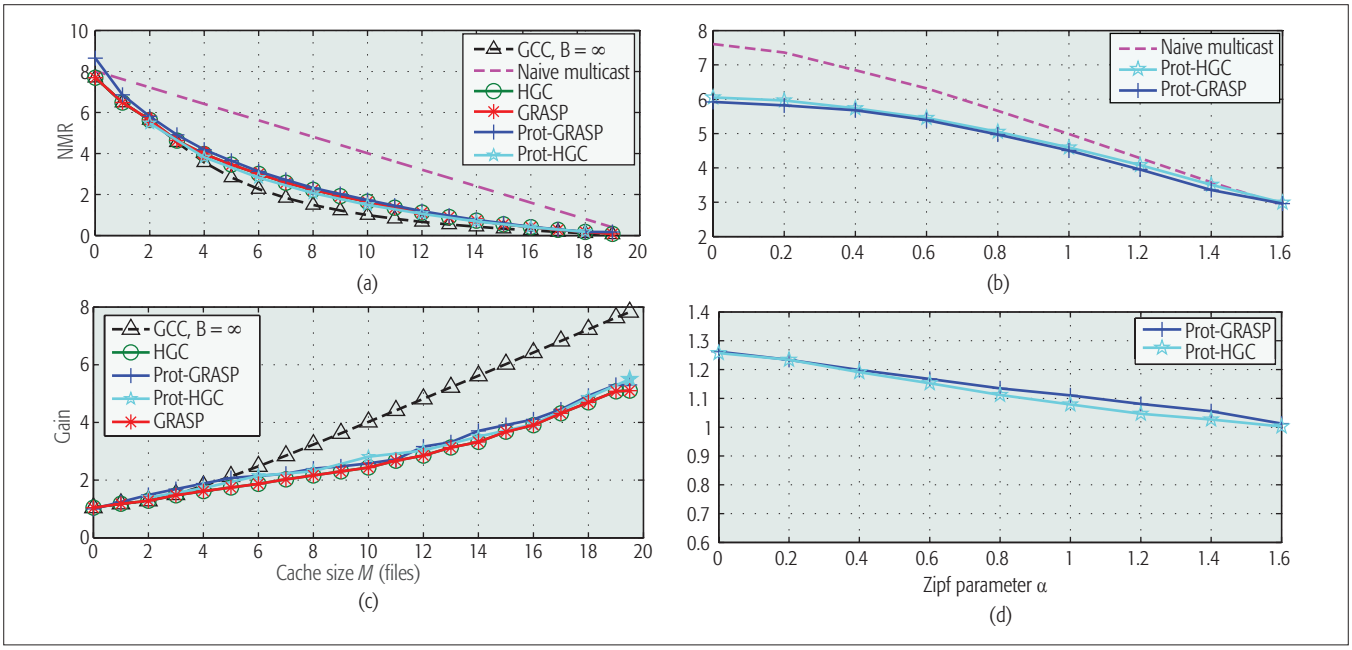


Figure 5. Comparison of the total minimum rate normalized by the one-user network minimum rate with respect to: a) the cache size with $\alpha = 0$; b) the Zipf α parameter with $M = 2$ files; and the gain of the coded over the uncoded scheme with respect to: c) the cache size with $\alpha = 0$, and d) the Zipf α parameter with $M = 2$ files.

under uniform popularity ($\alpha = 0$), and this gain tends to 1 as the popularity distribution becomes more skewed.

Turning Memory into Bandwidth

In this section, we evaluate the physical layer bandwidth gains enabled by coded multicasting. To do so, we assume a fixed video transmission delay (e.g., according to users' QoS) and evaluate the bandwidth required to serve the video segment requests of all users. Figure 6 illustrates the bandwidth gain (BG) evolution at the PHY layer with respect to the number of users for different cache sizes. We define the PHY BG as the bandwidth required to serve all requests via naive multicasting over the bandwidth required via the use of coded multicasting. The increase in BG can be clearly observed with respect to both the cache size and the number of users. For instance, assuming a cache size $M = 10$ percent of the library size, the gain starts with a value around 1.1 for 5 users and goes up to 1.31 for 40 users. Similarly, at $M = 30$ percent, the gain increases from around 1.4 for 5 users and reaches around 1.68 for 40 users. The increase of the BG with respect to the number of users is especially relevant, as it illustrates the scalability benefits of coded multicasting.

Future Directions

In the above discussion, coding overhead and computational complexity have been proven not to limit the performance gain of caching-aided coded multicasting. However, several open problems related to PHY layer protocols are currently under investigation, among which we cite the following.

Variation of the Channel Characteristics: Regarding the variations of channel statistics across users (e.g., different signal-to-noise ratios), the work in [14] provided a theoretical analysis that takes into account the wireless channel characteristics in the presence of any combination of unicast/multicast transmission and wireless edge caching. They proposed a channel-aware caching-aided coded multicast scheme based on joint source-channel coding with side information. Such a scheme is able to guarantee a rate to each receiver that is within a constant factor of the optimal rate they would obtain if the remaining users experience their same channel conditions — avoiding throughput penalizations from the presence of receiver with worse propagation channel. The implementation of this scheme in CorteXlab is part of our next steps for future work. As opposed to network emulation platforms such as in [11], CorteXlab will allow properly testing user mobility and realistic channel degradation across wireless endpoints.

Combination with MIMO Schemes: The use of MIMO schemes is an interesting topic with significant active research. Undergoing studies such as [15] have shown that coded multicasting is indeed complementary to MIMO, and the combination of both provides cumulative gains in most practical scenarios. This is also an interesting topic for future work, where CorteXlab can again provide a key advantage in order to easily include next generation radio technologies.

Dynamic Scenarios: Our current implementation setting is limited to static scenarios with respect to file popularity and number of users. Ideas related to cache adaptation with respect to dynamic popularity distributions and varying number of users are of interest for future work and currently under investigation.

Conclusion

This article discusses the potential of caching-aided coded multicasting for improving bandwidth efficiency in next generation wireless access networks. A real-time implementation for performance evaluation in real-world environments

has been presented. On the way from theory to practical evaluation, a complete frame structure for the transmitter and the receiver has been proposed. Our prototype implementation integrates coded multicasting into an OFDM-based PHY layer, and has been deployed in CorteXlab, a shielded experimentation facility with multiple radio nodes. Interestingly, the additional coding overhead does not compromise performance and leads to an overall positive multicasting gain, reducing bandwidth requirements and transmission delay when compared to the best uncoded schemes. We have integrated the coded multicast design in an OFDM-based PHY layer, and deployed the scenario in CorteXlab, a shielded experimentation room, using radio nodes. Our work also shows the potential of CorteXlab to validate new concepts relative to advanced radio technologies for 5G networks.

References

[1] A. Osseiran *et al.*, "Scenarios for 5G Mobile and Wireless Communications: The Vision of the METIS Project," *IEEE Commun. Mag.*, vol. 52, no. 5, May 2014, pp. 26–35.

[2] X Wang *et al.*, "Cache in the Air: Exploiting Content Caching and Delivery Techniques for 5G Systems." *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 131–39.

[3] N. Golrezaei *et al.*, "Femtocaching: Wireless Video Content Delivery through Distributed Caching Helpers," *Proc. IEEE INFOCOM*, Orlando, FL, 2012, pp. 1107–15.

[4] E. Zeydan *et al.*, "Big Data Caching for Networking: Moving from Cloud to Edge," *IEEE Commun. Mag.*, vol. 54, no. 9, Sept. 2016, pp. 36–42.

[5] J. Llorca *et al.*, "Network-Coded Caching-Aided Multicast for Efficient Content Delivery," *Proc. IEEE ICC*, Budapest, Hungary, 2013, pp. 3557–62.

[6] M. Maddah-Ali and U. Niesen, "Decentralized Coded Caching Attains Order-Optimal Memory-Rate Trade-Off," *IEEE/ACM Trans. Net.* , vol. 23, no 4, 2015, pp. 1029-–40.

[7] A. Blasiak, R. Kleinberg, and E. Lubetzky, "Index Coding via Linear Programming," preprint arXiv:1004.1379, 2010; http://arxiv.org/abs/1004.1379.

[8] M. Ji *et al.*, "Order-Optimal Rate of Caching and Coded Multicasting with Random Demands," preprint arXiv:1502.03124, Feb. 2015 ; https://arxiv.org/abs/1502.03124.

[9] K. Shanmugam *et al.*, "Finite Length Analysis of Caching-Aided Coded Multicasting," *IEEE Trans. Info. Theory*, vol. 62, no. 10, Oct. 2016, pp. 5524–37.

[10] G. Vettigli *et al.*, "An Efficient Coded Multicasting Scheme Preserving the Multiplicative Caching Gain," *Proc. IEEE INFOCOM Wksps.*, Hong Kong, 2015, pp. 251–56.

[11] U. Niesen, and M. A. Maddah-Ali, "Coded Caching for Delay-Sensitive Content," *Proc. IEEE ICC*, London, U.K., 2015, pp. 5559–64.

[12] L. S. Cardoso *et al.*, "CorteXlab: A Facility for Testing Cognitive Radio Networks in a Reproducible Environment," *Proc. EAI CROWNCOM*, Oulu, Finland, 2014, pp. 503–07.

[13] M. Ji *et al.*, "An Efficient Multiple-Groupcast Coded Multicasting Scheme for Finite Fractional Caching," *Proc. IEEE ICC*, London, U.K., 2015, p. 3801–06.

[14] A. S. Cacciapuoti *et al.*, "Speeding up Future Video Distribution via Channel-Aware Caching-Aided Coded Multicast," *IEEE JSAC*, vol. 34, no. 8, Aug. 2016, pp. 2207–18.

[15] S. Yang, K. Ngo, and M. Kobayashi, "Content Delivery with Coded Caching and Massive MIMO in 5G," *Proc. IEEE ISTC*, Brest, France, 2016, pp. 370–74.

Biographies

Yasser Fadlallah [S'10, M' 14] received his Telecommunication Engineering diploma from the Faculty of Engineering, Lebanese University, Lebanon, in 2009, his M.S. degree from the Université de Bretagne Occidentale, France, in 2010, and his Ph.D. degree from Télécom Bretagne, France, in 2013. In 2012, he was a visiting Ph.D. student at the Coding and Signal Transmission Laboratory, University of Waterloo, Canada. Between 2013 and 2014 he was an R&D engineer at Orange Labs, Paris. In 2015, he held a post-doctoral research position at INRIA until September 2016, when he joined the University of Sciences and Arts in Lebanon (USAL), where he is currently an assistant professor. His research interests lie in the wireless communications area, focusing on interference management, advanced and low-complexity receivers, wireless caching, and multiple-antenna systems.

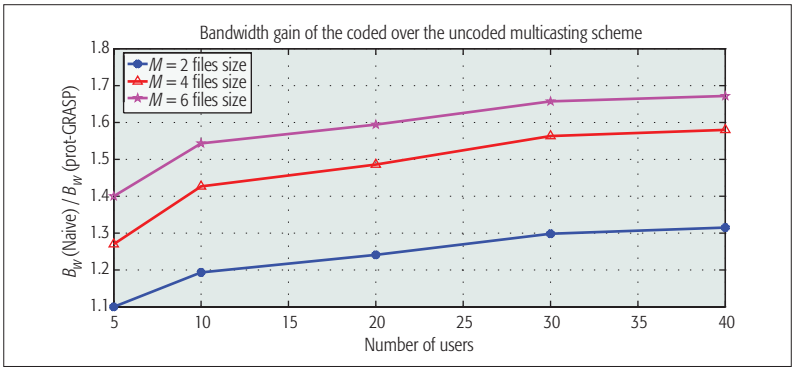


Figure 6. The bandwidth gain of coded multicasting over naive multicasting for different memory cache sizes.

Antonia M. Tulino [S'00, M '03, SM '05, F '13] received her Ph.D. degree in electrical engineering from Seconda Universita' degli Studi di Napoli, Italy, in 1999. She held research positions at Princeton University, at the Center for Wireless Communications, Oulu, Finland, and at Universita' degli Studi del Sannio, Benevento, Italy. From 2002 she has joined the Faculty of the Universita' degli Studi di Napoli "Federico II," and in 2009 she joined Bell Labs. Since 2011, she has been a member of the Editorial Board of *IEEE Transactions on Information Theory*. She has received several paper awards, among the others the 2009 Stephen O. Rice Prize in the Field of Communications Theory for the best paper published in *IEEE Transactions on Communications* in 2008. She has been principal investigator of several research projects sponsored by the European Union and the Italian National Council, and was selected by the National Academy of Engineering for the Frontiers of Engineering program in 2013. Her research interests lie in the area of communication systems approached with the complementary tools provided by signal processing, information, theory and random matrix theory.

Dario Barone received his B.E. in telecommunications engineering from Universita' degli Studi di Napoli "Federico II," Italy, in 2016. He developed his thesis working on this project where he makes his first contribution in research and development. He is currently studying for his Master's in telecommunications engineering at the same university.

Giuseppe Vettigli received a Master's degree in computer science from the Université Federico II di Napoli in 2014. He is now a Ph.D. student in computer science at the same university.

Jaime Llorca [S'03, M'09] received his B.E. degree in electrical engineering from the Universidad Politecnica de Catalunya, Barcelona, Spain, in 2001, and his M.S. and Ph.D. degrees in electrical and computer engineering from the University of Maryland, College Park, in 2003 and 2008, respectively. He held a postdoctoral position at the Center for Networking of Infrastructure Sensors (CNIS), College Park, Maryland, from 2008 to 2010. He joined Nokia Bell Labs at Holmdel, New Jersey, in 2010, where he is currently a research scientist in the Network Algorithms Group. His research interests include energy-efficient networks, distributed cloud networking, content distribution, resource allocation, network information theory, and network optimization. He was a recipient of the 2007 Best Paper Award at the IEEE International Conference on Sensors, Sensor Networks and Information Processing, the 2016 Best Paper Award at IEEE ICC, and the 2015 Jimmy H.C. Lin Award for Innovation.

Jean-Marie Gorce [M'07, SM'14] is a member of the scientific committee of the joint lab INRIA Nokia Bell Labs. He received his Ph.D. degree in electrical engineering from the National Institute of Applied Sciences (INSA), Lyon, France, in 1998. He held a research position at Bracco Research, S.A. Geneva, and was recruited by INSA Lyon in 1999. He was a co-founder of the Centre for Innovation in Telecommunications and Integration of Services Laboratory, of which he was the director from 2009 to 2014. He has been a researcher associated with INRIA since 2003, and he was visiting scholar at Princeton University from September 2013 to August 2014. He is the director of the Telecommunications Department of INSA Lyon and the holder of the SPIE ICS Industrial Chair on the Internet of Things. He has been the principal investigator of several research projects sponsored by the French government or the European Union. He is an Associate Editor of the *EURASIP journal of Wireless Communications and Networking* (Springer). His research interests lie in wireless networking, focusing on realistic modeling, wireless system optimization, and performance assessment considering both infrastructure-based and ad hoc networks. He is a scientific coordinator of the experimental facility FIT-CorteXlab.

INTERNET OF THINGS: PART 2


Christos Verikoukis


Roberto Minerva


Mohsen Guizani


Soumya Kanti Datta


Yen-Kuang Chen


Hausi A. Muller

The Internet of Things (IoT) is seen as a set of vertical application domains that share a limited number of common basic functionalities. In this view, consumer-centric solutions, platforms, data management, and business models have to be developed and consolidated in order to deploy effective solutions in the specific fields. The availability of low-cost general-purpose processing and storage systems with sensing/actuation capabilities coupled with communication capabilities are broadening the possibilities of IoT, leading to open systems that will be highly programmable and virtualized, and will support large numbers of application programming interfaces (APIs). IoT emerges as a set of integrated technologies — new exciting solutions and services that are set to change the way people live and produce goods. IoT is viewed by many as a fruitful technological sector in order to generate revenues. IoT covers a large wealth of consumer-centric technologies, and it is applicable to an even larger set of application domains. Innovation will be nurtured and driven by the possibilities offered by the combination of increased technological capabilities, new business models, and the rise of new ecosystems.

This Feature Topic (FT) issue addresses several promising approaches to sensors, actuators, and new consumer devices. New communication capabilities (from short-range to LPWAN to 4G and 5G networks, with NB-IoT). In addition, there are new communication protocols and the exploitation of NFV/SDN for better communications; new solutions for large distributed systems (e.g., combination of cloud, grid, and edge/fog computing); new vusiness models and ecosystems; and consumer-centric aspects including IoT application development, utilization of semantics, and security, privacy, and trust.

This timely FT has gathered articles from a wide range of perspectives in different industrial and research communities of IoT. In response to the Call for Papers, 103 high-quality manuscripts were received, and after a very careful review process, four outstanding papers have been selected for Part 2 of this FT, giving an overview of recent developments in quantum-resistant cryptosystems for securing the IoT, game theoretic models for resource management for IoT, the potential of UAVs equipped with IoT devices, and new business models for IoT.

In the first article, "Securing Internet of Things in a Quantum World" by C. Cheng, R. Lu, A. Petzoldt, and T. Takagi, the authors present the concept of quantum-resistant cryptosystems for securing the IoT as well as the existing implementations on constrained devices suitable for the IoT. They also present ongoing projects that will help develop the future security solutions for the IoT.

In the second article, "Management in Massive Wireless IoT Systems" by P. Semasinghe, S. Maghsudi, and E. Hossain, some non-conventional game theoretic models that fit the inherent characteristics of large-scale IoT networks are explored. The present five type of games: evolutionary games, mean field games, minority games, mean field bandit games, and mean field auctions, and discuss the potential IoT-related resource management problems that can be solved by using these models.

In the third article, "UAV-Based IoT Platform: A Crowd Surveillance Use Case" by N. Hossein Motlagh, M.Bagaa, and T.Taleb, the potential of UAVs, equipped with IoT devices, in delivering IoT services from height is discussed. As a use case the article demonstrates, by using a developed platform, how UAVs can be used for crowd surveillance based on face recognition.

In the fourth article, "Business Development in the Internet of Things: A Matter of Vertical Cooperation" by A. Ghanbari, A. Laya, J. Alonso-Zarate and J. Markendahl, the relevance of vertical cooperation in the IoT ecosystem is discussed, and the need to develop new value networks that leverage this cooperation and enable the creation of new business models is highlighted. The authors use the examples of two major building blocks of smart cities: intelligent transport systems and health and well being services.

BIOGRAPHIES

CHRISTOS VERIKOUKIS [S'95, M'04, SM'07] (cveri@cttc.es) got his Ph.D. from Universidad Politècnica de Catalunya (UPC) in 2000. He is currently a Fellow Researcher at CTTC, head of the SMARTECH Department, and an adjunct associate professor at the University of Barcelona. He has published 100 journal papers and over 170 conference papers. He is also a co-author of three books, 14 chapters in other books, and two patents. He is currently Chair of the IEEE ComSoc CSIM Technical Committee.

ROBERTO MINERVA holds a Ph.D. in computer science and telecommunications from Telecom Sud Paris, France, and a Master's degree in computer science from Bari University, Italy. He is the Chairman of the IEEE IoT Initiative, an effort to nurture a technical community and to foster research in IoT. He is at TIMLab, involved in activities on SDN/NFV, 5G, big data, and architectures for IoT. He is the author of papers published in international conferences, books, and magazines.

MOHSEN GUIZANI [S'85, M'89, SM'99, F'09] received his B.S. (with distinction) and M.S. degrees in electrical engineering, and M.S. and Ph.D. degrees in computer engineering from Syracuse University in 1984, 1986, 1987, and 1990, respectively. He is currently a professor and the Electrical and Computer Engineering Department Chair at the University of Idaho. He currently serves on the Editorial Boards of several international technical journals. He is the author of nine books and more than 450 publications in refereed journals and conferences.

SOUMYA KANTI DATTA is a research engineer at EURECOM and a co-founder of an IoT startup, Future Tech Lab. His research focuses on innovation, standardization, and development of next-generation technologies in mobile computing, IoT,

M2M communication, and security. He is an active member of the IEEE Consumer Electronics Society and W3C. He has published more than 40 papers in top IEEE conferences and journals. Currently he is involved in oneM2M and the W3C Web of Things Group.

YEN-KUANG CHEN [F'12] received his Ph.D. degree from Princeton University. He is a principal engineer at Intel Corporation, Santa Clara, California. His research areas span from emerging applications that can utilize the true potential of IoT to computer architecture that can embrace emerging applications. He has 50+ U.S. patents, 20+ pending patent applications, and 90+ publications. He is the Editor-in-Chief of the *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*. He is a Distinguished Lecturer of the IEEE Circuits and Systems Society, 2016–2017.

HAUSI A. MULLER is a professor in the Department of Computer Science and associate dean of research in the Faculty of Engineering at the University of Victoria. He is a member of the IEEE Computer Society Board of Governors and the 2016–2017 Vice-President of the IEEE CS Technical and Conference Activities Board. His research interests include software engineering, software evolution, IoT, smart cyber physical systems, and self-adaptive systems. He is a Fellow of the Canadian Academy of Engineering.

Securing the Internet of Things in a Quantum World

Chi Cheng, Rongxing Lu, Albrecht Petzoldt, and Tsuyoshi Takagi

Currently, we rely on cryptographic algorithms such as elliptic curve cryptosystems (ECCs) as basic building blocks to secure the communication in the Internet of Things. However, public key schemes like ECC can be easily broken by the upcoming quantum computers. Due to recent advances in quantum computing, we should act now to make the IoT be prepared for the quantum world.

ABSTRACT

Currently, we rely on cryptographic algorithms such as elliptic curve cryptosystems (ECCs) as basic building blocks to secure the communication in the IoT. However, public key schemes like ECC can easily be broken by the upcoming quantum computers. Due to recent advances in quantum computing, we should act now to prepare the IoT for the quantum world. In this article, we focus on the current state of the art and recent developments in the area of quantum-resistant cryptosystems for securing the IoT. We first demonstrate the impacts of quantum computers on the security of the cryptographic schemes used today, and then give an overview of the recommendations for cryptographic schemes that can be secure under the attacks of both classical and quantum computers. After that, we present the existing implementations of quantum-resistant cryptographic schemes on constrained devices suitable for the IoT. Finally, we give an introduction to ongoing projects for quantum-resistant schemes that will help develop future security solutions for the IoT.

INTRODUCTION

The past decade has witnessed the steady development of the Internet of Things (IoT). As illustrated in Fig. 1, Gartner has estimated that by 2016 there will be 6.4 billion connected devices in use, and this number is further expected to hit 20.8 billion by 2020. The world population is believed to reach 7.6 billion by 2020, which means that on average each person in the world will have nearly 3 connected devices. Since these connected things, such as implantable medical devices and vehicles, play vital roles in our daily lives, strong security requirements for the IoT have become a must.

Generally, the main security goals for the IoT are confidentiality, integrity, and authentication [1]. Confidentiality guarantees that sensitive information cannot be leaked to unauthorized entities, while integrity prevents information from being modified en route, and authentication ensures that the communicating entities are indeed those they declare to be. As shown in Fig. 2, important communication protocols for the IoT include the IEEE 802.15.4 standard, the Constrained Application Protocol (CoAP), and the IPv6 over Low-power Wireless Personal Area Networks (6LoWPAN) standard. To achieve the aforementioned security goals for the IoT, these protocols use cryptographic primitives such as the Advanced Encryption Standard (AES) for confidentiality and integrity, and elliptic curve cryptosystems (ECCs), which include the Elliptic Curve Digital Signature Algorithm (ECDSA) for integrity and authentication and the Elliptic Curve Diffie-Hellman (ECDH) algorithm for exchanging keys used in AES [2].

However, recent advances in quantum computing threaten the security of the current IoT using these cryptographic schemes. Just as the security of Rivest, Shamir, and Adleman (RSA) and Diffie-Hellman (DH) key exchange schemes are based on the difficulty of solving some number-theoretic problems such as integer factorization and discrete logarithms, the security of the ECC is based on the difficulty of solving the elliptic curve discrete logarithm problem. As early as 1994, mathematician Peter Shor of Bell Laboratories showed that quantum computers can solve the integer factorization problem and the (elliptic curve) discrete logarithm problems in an efficient way, sparking great research interest in quantum computing. Since then, quantum algorithms like Grover's search algorithm have been proposed, which provide significant speedup for many problems. Other examples include the quantum algorithms using the quantum Fourier transform, the quantum walk for solving searching problems, and adiabatic quantum computing for optimization problems. Besides that, much research is performed on how to design and build more powerful quantum computers with less resources to implement these algorithms [3].

It is still unclear when large-scale quantum computers will come into existence, but more and more scientists believe that we only need to overcome significant engineering obstacles. Based on recent advances in quantum computing, some scientists even claim that within 20 years our currently used public key infrastructures will become insecure because of the availability of large-scale quantum computers [4].

Even though there are quantum secure replacements for the cryptographic standards in use today, it will take a long time for the transition from currently used IoT systems to their quantum-resistant counterparts. Regarding the fact that we are at the very beginning of the standardization process for quantum resistant algorithms, and research on their application in the IoT is limited,

Digital Object Identifier:
10.1109/MCOM.2017.1600522CM

Chi Cheng is with China University of Geosciences (Wuhan) and Kyushu University; Rongxing Lu is with the University of New Brunswick; Albrecht Petzoldt and Tsuyoshi Takagi are with Kyushu University.

0163-6804/17/$25.00 © 2017 IEEE

it is urgent to make significant efforts in securing IoT systems against possible attacks by quantum computers. Therefore, no matter whether we can predict the exact arrival time of large-scale quantum computers, we should act now to prepare IoT systems for the quantum world.

In this article, we focus on the current state of the art and recent developments in the area of quantum-resistant cryptosystems for securing the IoT. The structure of this article is as follows. In the next section we demonstrate the impacts of large-scale quantum computers on the security of the cryptographic schemes used today, and then give an overview of the recommendations for cryptographic schemes that can be secure under attacks of both classical and quantum computers. After that, we consider the implementations of quantum-resistant cryptographic schemes on constrained devices for the IoT. We give an introduction to ongoing projects and developments for post quantum cryptography that will help develop the future security solutions for the IoT, and conclude this article.

Impact of Quantum Computers on Current Cryptographic Algorithms

The existing cryptosystems used for securing the IoT can be divided into two groups: symmetric and asymmetric (or public key) cryptosystems. In a symmetric cryptosystem two parties share a common secret key, which is then used to encrypt and decrypt messages. On the other hand, an asymmetric cryptosystem makes use of two keys: a private key and a public key. Everybody can use the public key to encrypt messages, but only the owner of the private key can decrypt the ciphertexts. In the context of signature schemes, the private key is used to generate a signature for a document, while everyone can use the public key to check the validity of the signature.

Currently, the most well-known example of a symmetric cipher is the AES, which was selected and standardized in 2001 by the National Institute of Standards and Technology (NIST) via a public competition. AES allows messages to be encrypted with secret keys of length 128, 196, and 256 bits, which are denoted as AES-128, AES-196, and AES-256, respectively. Among them, AES-128 is the most widely deployed in securing the IoT. To date, the best known attack against AES is a brute force search covering all possible keys. Since Grover's algorithm speeds up this process dramatically using quantum computers, the key size of AES needs to be doubled. That is, in order to achieve a security level of 128 bits against attacks with quantum computers, we need an AES key size of 256 bits.

For a public key environment, hash functions, public key encryption schemes, signature schemes, and key exchange protocols are the basic building blocks. A hash function is a map that transforms data of arbitrary length to a hash value of small fixed length. Hereby, it should be difficult to find two different messages that map to the same hash value (collision resistance). Today, the most widely used hash functions are SHA-2 and SHA-3, which are members of the Secure Hash Algorithm (SHA) family selected by NIST. Depending on the output length, SHA-2 can be further divided into SHA-256, SHA-384, and SHA-512. According to NIST's recommendation, we also need to enlarge the output of hash functions to prevent attacks using Grover's algorithm.

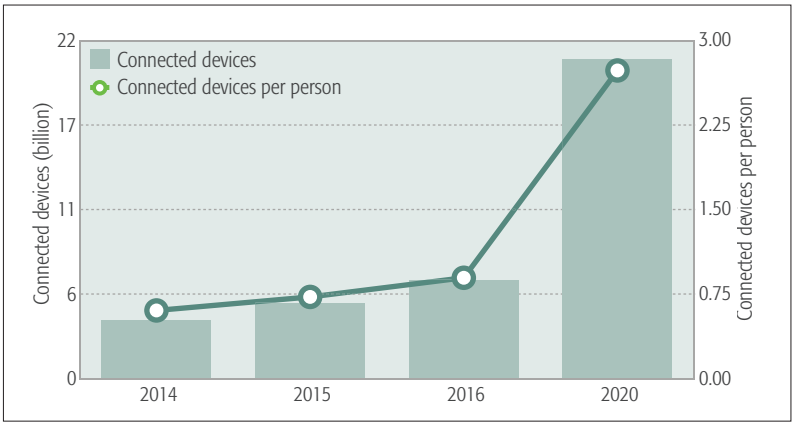


Figure 1. Number of connected devices in the IoT.

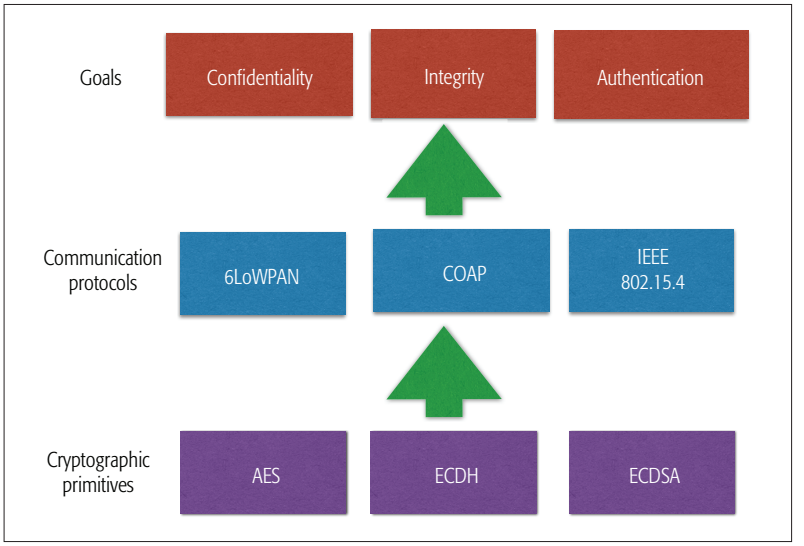


Figure 2. Current cryptographic primitives for securing communication in the IoT.

The impact of quantum attacks on the existing public key encryption and digital signature schemes is even more dramatic. The currently used cryptographic schemes for these purposes include RSA, the Digital Signature Algorithm (DSA), DH key exchange, and ECC, whose security is based on the hardness of certain number theoretic problems such as integer factorization and solving (elliptic curve) discrete logarithms. However, Shor's algorithm can solve these problems very efficiently on a quantum computer, which makes all these classical schemes insecure as soon as large quantum computers arrive.

To summarize, quantum computers have a great impact on the security of all cryptographic schemes used today. While for symmetric schemes and hash functions, it is relatively easy to prevent quantum attacks (increase key and output sizes respectively), public key schemes like RSA and ECC are completely broken (Table 1).

Therefore, we need to develop new schemes for public key encryption and signatures whose security is based on mathematical problems not affected by attacks using quantum computers. In

Algorithms	Purpose	Impact
AES	Symmetric encryption	Double the key size
SHA-2, SHA-3	Hash functions	Enlarge the output
RSA, ECC	Public key encryption and signature	Insecure
DH, ECDH	Key exchange	Insecure

Table 1. Impact of large-scale quantum computers.

Purpose	Type	Candidate algorithms
Symmetric encryption	Symmetric ciphers	AES-256, Salsa20
Public key encryption	Code-based	McEliece with binary Goppa
	Lattice-based	NTRUEncrypt
Public-key signature	Hash-based	XMSS, SPHINCS-256
	Multivariate-based	Rainbow, TTS, HFEv-
	Lattice-based	GPV, GLP, BLISS

Table 2. Initial recommendations for quantum-resistant algorithms.

the next section we give an overview of the existing candidates for this purpose.

Initial Recommendations for Quantum-Resistant Algorithms

To address the challenges in securing the IoT in the quantum world, we first need to know which kind of cryptographic primitives can be secure under the attacks of both classical and large-scale quantum computers. According to NIST [4], widely accepted quantum-resistant public key cryptosystems include hash-based signatures, code-based cryptosystems, multivariate polynomial-based cryptosystems, and lattice-based cryptosystems. The other recommendations given in [4] are based on the difficulty of the isogenies problem over supersingular elliptic curves and the conjugacy search problem in braid groups.

The first code-based cryptosystem was proposed by McEliece in 1978 and is a public key encryption scheme based on an error correcting code called Goppa code. The basic idea of the McEliece scheme can be described as follows: A message is encrypted into a codeword with some added errors, and only the private key holder can remove the errors and recover the original message. After nearly four decades, the McEliece scheme has withstood all proposed attacks [5, 6]. In particular, there is no quantum attack known that breaks the McEliece cryptosystem.

The construction of hash-based signatures employs only hash functions, and therefore minimizes the security requirements for building digital signature schemes. The first hash-based signature scheme was proposed by Merkle, who used a binary hash tree to construct the signatures. The Extended Merkle Signature Scheme (XMSS) is an improved version of Merkle's signature scheme, which reduces the signature size and requires weaker security assumptions [7]. A common requirement of the hash-based signature schemes is the need to record information about previously signed messages, which is called "state." This can lead to problems when signatures are generated on several devices since these devices have to be synchronized after each signature generation. To avoid this, a stateless hash-based signature scheme called SPHINCS has been proposed, which can be described as a multi-tree version of XMSS [8].

The European research group PQCRYPTO has given initial recommendations with specific parameters for quantum-resistant schemes, and we summarize their results in Table 2.

The security of multivariate polynomial-based cryptosystems is based on the difficulty of solving a system of multivariate quadratic (degree 2) equations over a finite field, which is proved to be an NP-hard problem. Depending on the field size used in the system, the multivariate polynomial-based schemes can be divided into small field ones, which include signature schemes such as Unbalanced Oil and Vinegar (UOV), Rainbow, and TTS, and big field ones such as Hidden Field Equations (HFE) [6]. As a variant of HFE, the HFEv- scheme is very useful due to its efficiency and ability to produce the shortest signatures among all existing multivariate polynomial-based schemes.

Previously, lattices were regarded as an important tool in breaking cryptographic schemes. However, starting with Ajtai's pioneering work on using lattices to construct cryptographic systems, numerous works have been done in this area [9]. In 1998, Hoffstein, Pipher, and Silverman proposed NTRUEncrypt (also known as NTRU), a lattice-based public key encryption algorithm that has attracted a lot of attention due to its efficiency and compact keys. Currently, the security of lattice-based cryptosystems mainly depends on the hardness of two problems: the short integer solution (SIS) problem and the learning with errors (LWE) problem, as well as their corresponding variants over rings, the ring-SIS problem and the ring-LWE problem. The advantage of cryptosystems based on the ring-SIS problem and ring-LWE problem is that they are more efficient and significantly reduce the key size compared to schemes based on the non-ring versions of the corresponding problems. Stele and and Steinfeld have proposed a variant of NTRUEncrypt, which can be proven to be secure under the ring-LWE assumption. Another hot topic in lattice-based cryptography is the design of lattice-based signature schemes, which include schemes based on preimage sampleable functions such as GPV, schemes based on the decisional ring-LWE problem such as GLP, and schemes based on the ring-SIS problem such as BLISS.

Quantum-Resistant Cryptographic Schemes on Constrained Devices and Networks

The IoT cannot become reality without the help of various kinds of constrained devices, which not only help us collect and gather information from nature, our households, and factories, but also process and even act on this information. As defined in [10], constrained devices refer to small

devices with limited resources in CPU, memory, and power. These limited resources bring special challenges for the cryptographic schemes used to secure constrained devices in the IoT. Since some of these devices may be used for decades, we should make them secure against long-term attacks. ECC with appropriate parameters is regarded as a solution to this problem. However, devices using ECC become insecure as soon as quantum computers appear. Therefore, the design and implementation of quantum-resistant cryptographic algorithms for constrained IoT devices are of vital importance.

Lattice-based and multivariate polynomial-based algorithms have shown their efficiency in providing quantum-resistant security for constrained devices. In [11] the signature scheme BLISS is implemented on a 32-bit ARM Cortex-M4F microcontroller with 1024 kB flash memory, taking 35.3 ms for signing and 6 ms for verification to achieve 128-bit security. In [12], the implementations of a ring-LWE-based encryption scheme, RLWEenc, and BLISS are conducted on an Atmel ATxmega128A1 microcontroller, which is equipped with an 8-bit CPU running at 32 MHz and a 128-kB flash memory. Specifically, in order to achieve security levels higher than 156 bits, it takes 68 ms for Ring-LWE encryption and 18.8 ms for decryption. For 128-bit security, BLISS needs 329 ms for signing and 88 ms for verification.

For multivariate polynomial-based cryptosystems, in [13] implementations of enhanced TTS (enTTS) and Rainbow are also done on an 8-bit Atmel ATxmega128A1 microcontroller. It is shown that the enTTS needs 66.9 ms for signing and 962.2 ms for verification, respectively, for a 128-bit security level. At the same time, for Rainbow it costs 257.1 ms for signing and 288.0 ms for verification. Since the two implementations in [12, 13] are done on the same 8-bit microcontroller, we list their results in Table 3, which compares the different implementations regarding key and signature sizes as well as the running times for signature generation and verification (for a security level of 128 bits).

The Transport Layer Security (TLS) protocol provides a good solution for Internet security, achieving both confidentiality and authentication. Meanwhile, CoAP, which is safeguarded by the Datagram Transport Layer Security (DTLS) protocol, has been designed for the IoT, especially for constrained devices. Just as TLS is designed to secure applications based on the Transmission Control Protocol (TCP), DTLS is based on the User Datagram Protocol (UDP). In [14], the authors have optimized the implementation of DTLS over CoAP for the IoT. Their implementations are based on ECC and conducted on a platform named MagoNode, which features the Atmel Atmega128RFA1 with a 2.4 GHz low-power transceiver for the IEEE 802.15.4 standard.

However, both TLS and DTLS need to be updated to resist attacks using quantum computers. The work in [15] moved forward toward this goal by providing ciphersuites for TLS, in which the security of the key exchange protocol is based on the ring-LWE problem. Thus, an intriguing problem is whether the latticed-based key exchange schemes work well for DTLS over CoAP. Furthermore, in [15] only the key exchange scheme is quantum resistant. Therefore, another interesting problem is the performance of both TLS and DTLS if all the components are replaced by the aforementioned quantum-resistant cryptographic schemes.

Schemes	Key size private (kB)	Key size public (kB)	Signature size (bit)	Time sign (ms)	@32 MHz verify
BLISS	2	7	7,680	329	88
enTTS	12.7	229.5	704	66.9	962.2
Rainbow	95.4	132.7	632	257.1	288.0

Table 3. Performance and parameters of BLISS, Rainbow, and enTTS (128-bit security).

Ongoing Projects and Developments

We summarize ongoing projects and developments that will help develop the future security solutions for the IoT. The research on quantum-resistant cryptography, which is known as "post-quantum cryptography," is active, and has attracted much attention from government, industry, and academia. Two recent announcements by the U.S. National Security Agency (NSA) and NIST have indicated the increasing necessity for transitions to quantum-resistant schemes [4]. In August 2015, NSA declared its plan to turn to quantum-resistant algorithms on its website. Just recently, at PQCrypto 2016, a leading conference for post-quantum cryptography held in February 2016, NIST announced its plan for a public call for quantum-resistant schemes, leading the way to new public key standards.

The European Commission has also promoted the research on post-quantum cryptosystems. A European research group, PQCRYPTO, has been funded by the European Union Horizon 2020 project, and is conducting research on post-quantum cryptography for small devices, the Internet, and the cloud. Another project supported by Horizon 2020 is SAFEcrypto, which focuses on practical and physically secure post-quantum cryptographic solutions in protecting satellite and public safety communication systems, as well as preserving the privacy of data collected by the government.

Besides that, a research project called CryptoMathCREST, which is supported by the Japan Science and Technology Agency, aims to study the mathematical problems underlying the security of post-quantum cryptography, and implement cryptosystems based on these problems to evaluate their performance in the real world.

Conclusion

Recent advances in quantum computing have demonstrated the urgency of developing quantum-resistant algorithms for securing communication in the IoT. In this article, we have shown the impacts of large-scale quantum computers on the security of the cryptographic schemes widely used today, followed by an overview of the recommendations for cryptographic schemes that can be secure under the attacks of both classical and quantum computers. After that, the recent implementations of quantum-resistant cryp-

The research on quantum resistant cryptography, which is known as "post-quantum cryptography," is active and has attracted much attention from government, industry, and academia. Two recent announcements by the U.S. National Security Agency and NIST have indicated the increasing necessity for transitions to quantum-resistant schemes.

tographic schemes for constrained devices have been introduced. Although ongoing projects are taking steps to develop new quantum-resistant security solutions for the IoT, more work is needed to prepare the IoT system for the quantum world.

Acknowledgments

The work presented in this article was supported in part by the National Natural Science Foundation of China under Grant nos. 61301166, 61672029, 61363069, and 61662016, the Fundamental Research Funds for the Central Universities, China University of Geosciences (Wuhan) (Grant Nos. CUGL150831, CUGL150416), and the JSPS KAKENHI, Grant Nos. 26.04347 and 15F15350.

References

[1] S. Sicari *et al.*, "Security, Privacy and Trust in Internet of Things: The Road Ahead," *Computer Networks*, vol. 76, 2015, pp. 146–64.
[2] J. Granjal, E. Monteiro, and J. Silva, "Security for the Internet of Things: A Survey of Existing Protocols and Open Research Issues," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 3, 2015, pp. 1294–1312.
[3] T. Monz *et al.*, "Realization of a Scalable Shor Algorithm," *Science*, vol. 351, no. 6277, 2016, pp. 1068–70.
[4] NIST, *Report on Post-Quantum Cryptography*, NISTIR 8105 DRAFT; http://csrc.nist.gov/publications/drafts/nistir-8105/nistir_8105_draft.pdf, accessed Oct. 4, 2016.
[5] A. Daniel *et al.*, "Initial Recommendations of Long-Term Secure Post-Quantum Systems"; http://pqcrypto.eu.org/docs/initial-recommendations.pdf, accessed Oct. 4, 2016.
[6] J. Buchmann *et al.*, "Post-Quantum Cryptography: State of the Art," *The New Codebreakers*, Springer, 2016, pp. 88–108.
[7] J. Buchmann, E. Dahmen, and A. Hülsing, "XMSS-A Practical Forward Secure Signature Scheme Based on Minimal Security Assumptions," *Post-Quantum Cryptography*, Springer, 2011, pp. 117–29.
[8] D. J. Bernstein *et al.*, "SPHINCS: Practical Stateless Hash-Based Signatures," *Advances in Cryptology--EUROCRYPT 2015*, Springer, 2015, pp. 368–97.
[9] C. Peikert, "A Decade of Lattice Cryptography," *Cryptology ePrint Archive*, Rep. 2015/939, 2015, http://eprint.iacr.org/2015/939.pdf, accessed Oct. 4, 2016, 2016.
[10] C. Bormann *et al.*, "Terminology for Constrained-Node Networks," IETF RFC 7228, DOI 10.17487/RFC7228, May 2014; http://www.rfc-editor.org/info/rfc7228, accessed Oct. 4, 2016.
[11] T. Oder *et al.*, "Beyond ECDSA and RSA: Lattice-Based Digital Signatures on Constrained Devices," *51st Annual ACM Design Automation Conf. 2014*, San Francisco, CA, June 1–5, 2014.
[12] T. Pöppelmann, T. Oder, and T. Güneysu, "High-Performance Ideal Lattice-Based Cryptography on ATxmega 8-Bit Microcontrollers," *Progress in Cryptology-LATINCRYPT 2015*, Springer, 2015, pp. 346–65.
[13] P. Czypek *et al.*, "Efficient Implementations of MQPKS on Constrained Devices," *Cryptographic Hardware and Embedded Systems 2012*, Springer, 2012, pp. 374–89.
[14] A. Capossele *et al.*, "Security as a CoAP Resource: An Optimized DTLS Implementation for the IoT," *2015 IEEE ICC*, 2015, pp. 549–54.
[15] J. Bos *et al.*, "Post-Quantum Key Exchange for the TLS Protocol from the Ring Learning with Errors Problem," *2015 IEEE Symp. Security and Privacy*, 2015, pp. 553–70.

Biographies

Chi Cheng [M'15] (chengchizz@gmail.com) received his B.S. and M.S. degrees in mathematics from Hubei University in 2003 and 2006, respectively, and his Ph.D. degree in information and communication engineering from Huazhong University of Science and Technology in 2013. He is currently an associate professor in the School of Computer Science, China University of Geosciences, Wuhan, China, and a JSPS postdoctoral researcher at Kyushu University, Japan. His research interests include applied cryptography and network security.

Rongxing Lu (rlu1@unb.ca) has been an assistant professor at the Faculty of Computer Science, University of New Brunswick, Canada, since August 2016. Before that, he worked as an assistant professor at the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, from May 2013 to August 2016. His research interests include applied cryptography, privacy enhancing technologies, and IoT-big data security and privacy. He currently serves as the Secretary of IEEE ComSoc CIS-TC.

Albrecht Petzoldt (albrecht.petzoldt@gmail.com) received a Diploma in mathematics from FAU Erlangen-Nürnberg in 2008, and a Ph.D. in computer science in 2013 at the Technical University of Darmstadt (TU Darmstadt), Germany. He is currently working as a Japan Society for the Promotion of Science (JSPS) postdoctoral researcher at Kyushu University. His main research interests are multivariate cryptography and post-quantum digital signature schemes.

Tsuyoshi Takagi (takagi@imi.kyushu-u.ac.jp) received his B.Sc. and M.Sc. degrees in mathematics from Nagoya University in 1993 and 1995, respectively, and his Ph.D. from TU Darmstadt in 2001. He is currently a professor in the Institute of Mathematics for Industry at Kyushu University. His current research interests are information security and cryptography. He has received the DOCOMO Mobile Science Award in 2013, IEICE Achievement Award in 2013, and JSPS Prize in 2014, and is a Program Chair of PQCrypto 2016.

Game Theoretic Mechanisms for Resource Management in Massive Wireless IoT Systems

Prabodini Semasinghe, Setareh Maghsudi, and Ekram Hossain

ABSTRACT

As a result of rapid advancement in communication technologies, the Internet of Things (i.e., ubiquitous connectivity among a very large number of persons and physical objects) is now becoming a reality. Nonetheless, a variety of challenges remain to be addressed, one of them being the efficient resource management in IoT. On one hand, central resource allocation is infeasible for large numbers of entities, due to excessive computational cost as well as immoderate overhead required for information acquisition. On the other hand, the devices connecting to IoT are expected to act smart, making decisions and performing tasks without human intervention. These characteristics render distributed resource management an essential feature of future IoT. Traditionally, game theory is applied to effectively analyze the interactive decision making of agents with conflicting interests. Nevertheless, conventional game models are not adequate to model large-scale systems, since they suffer from many shortcomings including analytical complexity, slow convergence, and excessive overhead due to information acquisition/exchange. In this article, we explore some non-conventional game theoretic models that fit the inherent characteristics of future large-scale IoT systems. Specifically, we discuss evolutionary games, mean field games, minority games, mean field bandit games, and mean field auctions. We provide the basics of each of these game models and discuss the potential IoT-related resource management problems that can be solved by using these models. We also discuss challenges, pitfalls, and future research directions.

INTRODUCTION

With the remarkable growth of the Internet and communication technologies over the past few decades, the world is now stepping into a new era identified by strong connectivity among billions of humans and machines. The concept of providing ubiquitous connectivity for anyone and anything (covering wireless portables, sensor devices, smart meters, etc.) at any time and any place is identified as the Internet of Things (IoT).

Basically, the devices connected to the IoT are expected to be smarter than the devices interconnected via traditional internet; consequently, IoT devices not only are in charge of making decisions, but also perform their tasks without human intervention. As a result, all forms of communications, including human-to-human, human-to-machine, and machine-to-machine (M2M) might take place in IoT [1]. In other words, IoT connects billions of persons and physical objects together, so they can communicate, coordinate, and share information with each other in order to make decisions and perform their individual tasks efficiently. Applications of the IoT include remote healthcare monitoring, waste management, smart homes, smart parking, and so on, which can have a dramatic impact on many aspects of our lives. Figure 1 shows a schematic diagram of a few end-user applications of IoT.

Despite its great potential to improve the quality of our lives, facilitating and implementing wireless IoT gives rise to several challenges such as security (identity management, information privacy), compatibility (technical standardization, interoperability), and resource allocation (spectrum management, energy efficiency) [1]. Specifically, due to the massive number of devices, IoT systems could be ultra-dense, and the use of efficient resource management techniques is of vital importance. Inefficient resource management can drastically reduce the performance of IoT systems. Resource management in the context of wireless IoT may include several aspects, some of which are:

- Spectrum allocation
- Power control
- Interference management
- Backhaul resource allocation
- Allocation of storage and computing resources in the cloud

In particular, distributed resource management schemes are desirable for IoT systems due to following reasons:

- Future machines are foreseen to be able to make decisions and perform tasks independently; this capability should be exploited by system designers instead of being neglected.
- For ultra-dense systems with very large numbers of devices, centralized control yields excessive complexity and computational cost.
- Efficient centralized resource allocation necessitates information availability at the controller. Nonetheless, acquiring such information may not be feasible due to limited fronthaul/backhaul capacity.

Over the past decade, traditional learning and

The authors explore some non-conventional game theoretic models that fit the inherent characteristics of future large-scale IoT systems. Specifically, they discuss evolutionary games, mean field games, minority games, mean field bandit games, and mean field auctions. They provide the basics of each of these game models and discuss the potential IoT-related resource management problems that can be solved by using these models.

Prabodini Semasinghe is with Price Industries Limited; Setareh Maghsudi is with Yale University; Ekram Hossain is with the University of Manitoba.

Digital Object Identifier:
10.1109/MCOM.2017.1600568CM

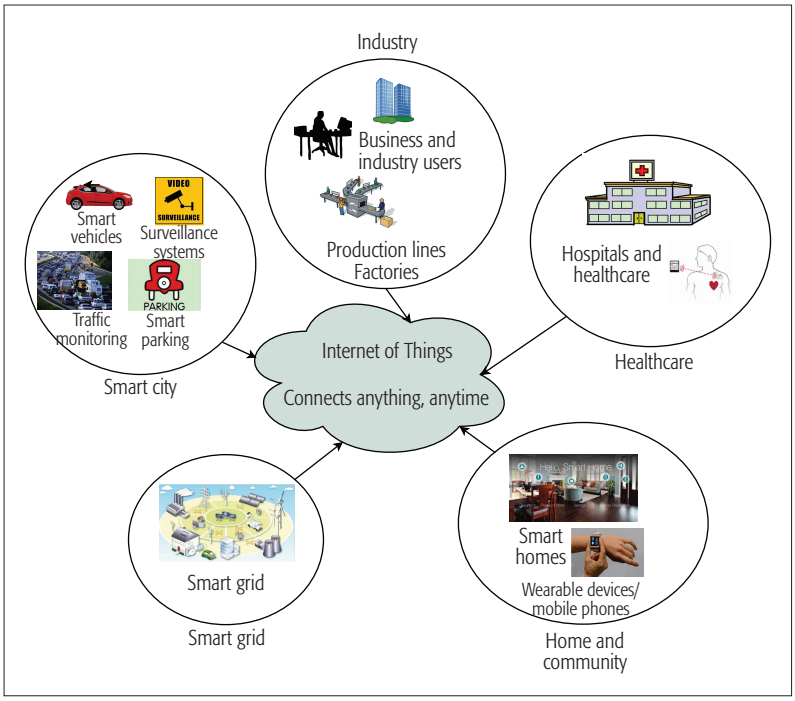


Figure 1. Future IoT: End-user applications.

game-theoretical models have been extensively applied to analyze interactive decision making of agents with conflicting interests. In essence, such models are beneficially used in a variety of contexts, among them efficient resource management for wireless heterogeneous networks, sensor networks, and M2M communications. Nevertheless, traditional learning and game-theoretical models suffer from many shortcomings, which render them inadequate to analyze the problems associated with the emerging ultra-dense network infrastructures. In particular, arguments against using traditional game models for large-scale systems include, but are not limited to:

- The immense overhead caused by information acquisition
- The slow convergence to equilibrium
- The inefficiency of equilibrium in terms of social welfare
- The excessive computational complexity
- The theoretical complexity of characterizing the equilibrium set

Therefore, it becomes imperative to rethink current analytical models of resource management, and move toward less conventional models that fit the characteristics of future wireless networks more appropriately. To this end, in this article, we provide a brief overview on some learning and game-theoretical mechanisms that can be taken advantage of when confronting distributed resource allocation problems in ultra-dense IoT systems. Specifically, we explore the following game models:

- Evolutionary games
- Mean field games
- Minority games

Furthermore, we also review two mean field approximation models for large-scale multi-agent systems:

- Mean field bandit games
- Mean field auctions with learning

The contribution and organization of this article can be summarized as follows. We first discuss some inherent characteristics, requirements, and limitations of large-scale IoT systems, which should be taken into account to enable appropriate design and implementation of resource management schemes for such systems. We thereby justify the importance of taking a step back from the traditional game models and moving forward to new models that not only fit the recently emerged requirements, but also are able to combat new challenges. As examples of such game models, we review evolutionary games, mean field games, minority games, mean field auctions, and mean field bandit games in a tutorial manner. Afterward, we provide a summary and compare different features, pros, and cons of the game models discussed. Moreover, we discuss challenges and possible pitfalls when modeling resource management problems of large-scale systems using those game models. We also outline possible future research directions.

Challenges in Distributed Resource Management for Wireless IoT

In the following we briefly explore some inherent features of massive wireless IoT systems. We thereby identify the fundamental challenges that arise when designing distributed resource management schemes for such systems.

Random Deployment

As a result of densification, deployment of IoT devices will be random and unknown to service providers. While such randomness and uncertainty make efficient management of scarce resources challenging enough, the problem becomes aggravated when device mobility is taken into account, which results in unpredictable and frequent change in device locations.

Scalability

Different types of IoT devices, used for different purposes, may connect to the network at different times. Consequently, the size of a network consisting of many (potential) devices may vary. Resource management algorithms are thus expected to be scalable. That is, the computational complexity of the algorithms should stay bounded even under an abrupt growth in the number of IoT devices connected to each other.

Limited Fronthaul/Backhaul

While providing fronthaul/backhaul connectivity is challenging even in networks of small or medium size, it is evident that in a wireless IoT system with thousands of connected devices, not all nodes would be able to use the limited backhaul/fronthaul connections for information exchange. In order to cope with this problem, it is important that the resource management algorithms rely on little or no information exchange among the wireless IoT devices and the controller device (e.g., a cellular base station).

Inhomogeneity

As described before, the IoT is a collection of a wide variety of different types of devices that are used for different purposes. Consequently, each

device has different requirements such as energy constraint and rate requirement. Resource management thus has to be done while taking the inhomogeneity and the unique characteristics of the devices into account.

Non-Guaranteed Energy Supply

Due to irregular deployment and high-speed mobility, many IoT devices may not be connected to a power grid. Thus, they are either battery operated, which necessitates frequent recharge, or powered through energy harvesting, which is inherently opportunistic and random. In order to mitigate such limitations, resource management techniques are expected to be power-efficient.

Uncertain and Incomplete Information

To execute certain resource management algorithms, IoT devices may be expected to exchange information with nearby nodes (i.e., local interactions). This information, however, can be distorted as a result of the noisy backhaul. Moreover, it is very likely to be delayed due to the time taken in processing and transmission. Channel state information (CSI) might also be distorted or temporarily unavailable, as a consequence of fading experienced by feedback channels. Furthermore, if the status of each channel is estimated by spectrum sensing, the sensing result can be imperfect. Clearly, efficient resource allocation becomes particularly demanding when the designer has to deal with such uncertain and incomplete information.

Game Models for Resource Management in Massive IoT Systems

Evolutionary Games

The evolutionary game [2], which was first developed to model the evolving populations of the biological entities, allows players with bounded rationality to learn from the environment and make individual decisions on their behavior. In particular, in such games, players move strategically by replicating more successful actions rather than the possible outcomes of every joint action profile, that is, the actions that are used more frequently among the players. Without going into mathematical details, in the following we describe basic elements of an evolutionary game.

In an evolutionary game, the set of players is called the *population*. Moreover, the collection of fractions of the population selecting different actions at a given step is referred to as the *population state* at that step. Players adapt their strategies by replicating a more successful action in terms of the occurrence frequency until the system reaches an equilibrium. The process of action adaptation is generally modeled by a set of ordinary differential equations called *replicator dynamics*. For an action a, the replicator dynamics can be written as follows:

$$\dot{x}_a(t) = x_a(t)\,(u_a(t) - \bar{u}(t)), \tag{1}$$

where $u_a(t)$ is the utility of a player who selects action a at step t, and $\bar{u}(t)$ is the average utility of the population. Moreover, $x_a(t)$ is the fraction of population selecting action a at step t, and $\dot{x}_a(t)$ denotes the rate of change of the fraction of population selecting action a at step t. The fixed point of this replicator dynamics is an *evolutionary equilibrium*. Although the evolutionary equilibrium can be obtained by simultaneously solving the replicator dynamics in Eq. 1, it is of utmost importance to develop distributed algorithms that can be executed individually at every player in an asynchronous manner. One such algorithm with linear time complexity is given in [3]. In games that assume full rationality of players, every player has to track the moves of all other players. Therefore, using conventional games to model resource allocation problems in massive IoT systems can result in very complex algorithms and excessive feedback information exchange due to the large number of interconnected devices. Conversely, in evolutionary games, instead of relying on the knowledge of decisions made by others, players simply adapt their moves only based on the system's average utility. This characteristic makes the evolutionary game model a good candidate to develop resource allocation algorithms with low complexity, which are also well appropriate for massive IoT systems with limited backhaul/fronthaul connectivity. Moreover, as evolutionary equilibrium provides identical utilities for all players in the system, resource management schemes based on this model can ensure fairness among all players. Power control, sub-carrier allocation/medium access control, and joint power-subcarrier allocation, transmission mode/network selection, and analyzing the behavior of the system under denial of service (DoS) attacks are possible applications of evolutionary games in the context of massive IoT. Evolutionary-game-based algorithms can also withstand a certain amount of delay in information exchange [3]. However, the performance of the evolutionary-game-based algorithms under information delay is highly subjective to the system and network parameters such as channel gains and number of devices in the network. Nonetheless, evolutionary game models may not be capable of modeling uncertainty and the stochastic nature of parameters such as queue dynamics and non-guaranteed energy supply. Moreover, evolutionary games assume homogeneity of players. Therefore, modeling the interconnection of different types of IoT devices may be difficult.

Mean Field Games

As discussed previously, evolutionary games assume that players are bounded rational, and thus they replicate the actions that result in higher payoffs. Due to the bounded rationality, in such games, players do not necessarily implement the best response dynamics. However, there are some scenarios in which IoT devices are in fact rational and interested in making the best decision based on other agents' actions. In this case, conventional games analyze the pair-wise interactions of players in order to achieve an efficient equilibrium. In large-scale systems such as IoT, such analysis requires heavy information exchange and results in excessive complexity. Recently, the concept of mean field games (MFGs) [4] has been developed to cope with the aforementioned issues; that is, to effectively model and analyze the interactions of a large number of rational entities. MFGs are a special form of differential games, where each player has a state, a set of actions, and a control policy.

Due to irregular deployment and high-speed mobility, many IoT devices may not be connected to a power grid. Thus, they are either battery operated, which necessitates frequent recharge, or are powered through energy harvesting, which is inherently opportunistic and random.

The problem is to find an optimal mapping of each possible winning history to an action, called a strategy. Obtaining the mapping of each possible winning history to an action can be done in a distributed manner at every player, by implementing inductive learning.

The control policy maps every state into an action over a pre-defined period of time. Instead of modeling each player's interaction with every other player, MFG models an individual's interaction with the effect of the collective behavior of all the players (i.e., the mean field). Thus, at each step of the game, the mean field is simply the fraction of players at every state. Let the game involve K players, and s denotes any arbitrary state defined by the game model. Moreover, at every step t, the state of player K is shown by $s_k(t)$. Then, at every step t, the mean field function of the game is formally defined as

$$m(t,s) = \lim_{K\to\infty} \frac{1}{K} \sum_{\forall k\in\mathcal{K}} \mathbb{I}_{\{s_k(t)=s\}}, \quad (2)$$

where $\mathbb{I}_{\{x=y\}}$ stands for the indicator function, which returns 1 if the given condition is true (i.e., if $x = y$), and 0 otherwise.

In MFGs, the interaction of every player with the mean field is modeled by a Hamilton-Jacobi-Bellman (HJB) equation, which is also called the *backward equation*. Then the movement of the mean field function according to the players' actions is defined by a Fokker-Planck-Kolmogorov (FPK) equation. This FPK equation is called the *forward equation*. In MFGs, the objective of each player is to come up with a sophisticated control that maximizes its own utility over a pre-defined period of time, considering the collective behavior of all other players.

We omit the derivation of mean field equations due to the complexity as well as the brevity of this article. A comprehensive discussion on the derivation of mean field equations can be found in [5]. The mean field equilibrium can be obtained by solving mean field equations (i.e., FPK and HJB equations) simultaneously. Obtaining the mean field equilibrium is challenging since there is no general technique to solve the mean field equations. There are, however, several approaches proposed in the literature. For instance, in [5], a finite difference technique is proposed to obtain the mean field equilibrium. This approach is applied in [6] to develop a distributed power control algorithm for dense small cell networks. As another example, [7] uses a drift plus penalty (DPP) approach in the framework of Lyapunov optimization to solve the mean field equations.

When it comes to modeling the resource allocation problems for massive IoT systems, the most significant aspect of MFGs is the ability to describe the behavior of a large system solely with two equations. In addition, mean field equations model the behavior of the system for a period of time rather than maximizing the instantaneous utility shortsightedly. Furthermore, since the MFG is a special form of differential games, when applying it to solve the resource allocation problem, the stochastic nature of the system (battery dynamics, channel dynamics, variation of the queue length, etc.) can be taken into account. Therefore, the MFG is a good candidate to develop resource allocation techniques considering the dynamic nature of the system. Energy-aware power control, resource management considering the mobility of the IoT devices, and queue-aware resource allocation are potential applications of MFGs in the context of massive IoT. Besides, with MFG models it is possible to derive offline algorithms that do not rely on any information exchange among devices while executing the algorithm. In fact, devices are supposed to gather the required information at the initialization phase prior to the execution of the algorithm. This feature makes MFGs even more appealing since it requires limited backhaul/fronthaul connectivity, and network operators do not have to worry about information delay. However, with MFG formulations, taking the incompleteness of information into account is challenging.

Minority Games

Similar to evolutionary games, minority games are a branch of game theory that model the behavior of entities with bounded rationality, in particular, in scenarios where belonging to the population's minority is more advantageous. The minority game was first introduced as a mathematical model for the El Farol Bar problem [8], stated as follows: An odd number of players take part in a repeated game, where at every iteration, each agent decides whether to go to a bar or not. On one hand, if the bar is too crowded, no one would enjoy the bar, so the better choice is not to go to the bar. On the other hand, if the bar is deserted, everyone in the bar enjoys it; hence, it is desirable to go to the bar. As a result, in this setting, being in the minority is always beneficial. The general setting of a minority game is similar: An odd number of bounded rational players have two actions from which to select. At every iteration, players selecting the minority's action are defined as winners. At the end of the each step, each player knows the outcome of the game (i.e., the winning side). This information is called the winning history of the player. Each player will make a decision on the action at the next step based on its winning history, which implies that the players do not evaluate the possible outcome of every joint action profile. Thus, the problem is to find an optimal mapping of each possible winning history to an action, called a strategy. Obtaining the mapping of each possible winning history to an action can be done in a distributed manner at every player by implementing inductive learning [8].

Different from the two previously explained game models, resource allocation algorithms that can tolerate imperfect information can be implemented using minority game models. The inductive learning algorithm is also simple, and hence the corresponding resource allocation technique is scalable. However, having a binary action set can limit the applications of minority games. Deciding whether to transmit or not in a slotted ALOHA system, transmission mode/network selection, and interference management by deciding whether to transmit or not are a few resource management problems in large-scale IoT that could be possibly modeled as minority games. Moreover, similar to evolutionary games, minority games are incapable of modeling the inhomogeneity of IoT devices. As a solution, when different categories of devices are present in the system, it may be possible to model the resource allocation problem as a hierarchical game where intra-category interactions (interaction among the devices that belong to the same category) are modeled as different levels of the game.

Mean Field Bandit Games

Multi-armed bandits (MABs) are a class of sequential optimization problems, where in successive rounds a player pulls an arm from a given a set of arms to receive some a priori unknown reward. The player observes only the reward of the played arm. As a result of information shortage, there might be a difference between the maximum achievable reward and the reward of the actual played arm, which is referred to as *regret*. The player intends to optimize some regret-based objective function over the game horizon by selecting arms according to some decision-making policy. Thus, the problem is to find a balance between gathering immediate rewards and obtaining information to achieve a large reward in the future, known as the exploration-exploitation dilemma.

While the basic bandit problem involves only one agent, the problem can be generalized to the multi-agent case, where the agents arbitrarily affect each other's rewards; thus, it is important to reach some sort of system stability or equilibrium. When only a few agents are involved in the bandit game, conventional equilibrium notions, such as correlated, Nash, or perfect Bayesian equilibria, might be practical. For a large number of agents, however, such equilibrium notions are not feasible due to excessive complexity and long convergence time. To mitigate this problem, similar to games with complete information, mean field approximation can be used to analyze large-scale bandit games. As is conventional, in such approximation every agent regards the rest of the world as being stationary, considering individual moves of agents as unimportant details. While research in this area is still in its infancy, Gummadi *et al.* provide basic elements of such models in [9]. However, there remains a wide range of open issues to be addressed in future research. In particular, the validity of mean field approximation, as well as the existence and uniqueness of mean field equilibrium for general bandit games with arbitrary reward distribution, are still open to investigate.

A mean field bandit model assumes that no prior knowledge is available to decision makers. Therefore, these models are very appropriate for IoT systems with limited backhaul/fronthaul connectivity. Specific problems include user association, transmit power and spectrum allocation for wireless IoT devices, transmission mode selection, and channel access. An example can be found in [10].

Mean Field Auctions

As is well known, an auction corresponds to a process of buying/selling goods, where an auctioneer offers the good(s), takes bids from participants (bidders), adjusts prices, and finally sells the good(s) to the participant making the highest bid. The actual price paid by the winner depends on the auction rule. For instance, in a first-price auction, the highest offered price has to be paid by the winner, whereas in a second-price auction, only the second highest price is paid. While static auctions are performed only once (similar to one-shot games), dynamic auctions are performed repeatedly. Such repetition provides the possibility of incorporating learning theory in the model, the valuation of the object being sold is not known to the bidders a priori. Nevertheless, as is natural in a dynamic setting, the agents react to each other, so the system requires to converge to some steady state or equilibrium.

Since providing a survey on auction models and their application is out of the scope of this article, it suffices to mention that various auction models have been used intensively by researchers to model a variety of wireless networking problems, with a very famous example being the spectrum auction. Nonetheless, similar to any other multi-agent setting, dynamic auctions become computationally infeasible even for a small number of agents, since the auctioneer has to solve for optimal prices given a large set of bids. Thus, here as well, mean field approximation can play an important role to make the problem easier to deal with. Large-scale dynamic auctions with learning have been studied by Iyer *et al.* in [11, 12] using a mean field model.

Similar to mean field bandits, mean field auctions are particularly appropriate when addressing the decision problems under information shortage. However, here the presence of a coordinator might be necessary to perform the auction process. Channel access, load balancing, backhaul/fronthaul channel allocation, and resource allocation for mobile cloud computing are among the potential applications in the context of massive IoT.

In Table 1, we summarize the main features of the game models discussed. Moreover, in Table 2, we clarify to what extent each game model would be able to address the wireless IoT-specific challenges described earlier. However, it is important to note that sometimes game models can be adapted to the specific problem to eliminate some of the shortcomings.

As has been mentioned before, modeling the random deployment of IoT devices is challenging. For all the aforementioned game models, random deployment of the network nodes has to be taken into account when deriving the utility functions. However, derivation of the utility functions in order to capture the randomness of the network is problem specific.

Challenges and Future Directions

The aforementioned game models provide vast opportunities to develop computationally efficient distributed solutions for resource management in wireless IoT systems. Nevertheless, they are associated with some shortcomings and challenges as well. In what follows, we discuss some common pitfalls and outline potential future research directions.

Rigorous Design of Utility Functions

In general, designing proper utility functions is a challenging task. One one hand, a utility function should be designed in such a way that it reflects the objective of the IoT application. On the other hand, it should satisfy the requirements of the game model that is being used.

For example, as discussed in evolutionary games, the utility of all players that select a certain action a should be equal. Despite guaranteeing fairness, such equality may not be physically implementable for most wireless IoT systems, since a variety of factors such as channel gain and

Similar to mean field bandits, mean field auctions are particularly appropriate when addressing the decision problems under information shortage. However, here the presence of a coordinator might be necessary to perform the auction process.

Game model	Equilibrium concept	Information requirement	Rationality	Remarks	Potential IoT applications
Evolutionary games	Evolutionary equilibrium	Average system utility	Bounded	Guaranteed fairness, linear time complexity, tolerant to information delay	Power control, spectrum/subcarrier allocation, transmission mode/network selection
Mean field games	Mean field equilibrium	Initial distribution of the players' states	Full	Stochastic awareness, performance evaluation for a time period, complicated mean field equations	Energy/queue/channel-aware resource allocation, resource management under mobility
Minority games	Public perfect equilibrium	Successful action	Bounded	Binary action set, tolerant of imperfect information	Scheduling, transmission mode/network selection, interference management
Mean field bandit games	Mean field equilibrium	None	Bounded	No information, binary reward set	User association, scheduling, channel allocation
Mean field dynamic auctions	Mean field equilibrium	Bids	Bounded	No information on utility function, an auctioneer is required	User association, scheduling, channel allocation

Table 1. Summary of challenges in massive wireless IoT that can be addressed by each game model.

Game model	Random deployment	Scalability	Limited backhaul/fronthaul	Inhomogeneity	Non-guaranteed energy supply	Uncertain and incomplete information
Evolutionary games	Yes	Yes	Yes	No	No	No
Mean field games	Yes	Yes	Yes	Yes	Yes	No
Minority games	Yes	Yes	Yes	No	No	No
Mean field bandit games	Yes	Yes	Yes	Yes	Yes	Yes
Mean field auctions	Yes	Yes	Yes	Yes	Yes	Yes

Table 2. Summary of the main features of each game model.

interference impact the utility of each IoT device. In addition, the utility of each player at a certain step of the game is affected by the population state of the system, which should be properly reflected in the utility function. Also, in MFG models, it is assumed that state-interchange among players does not affect the game's outcome. In many situations, however, such an assumption may be unrealistic. Sometimes, this assumption can be relaxed through a careful design of players' utility function so that it only depends on the player's state and the mean field. However, this is not a trivial task since the existence and uniqueness of mean field equilibrium is largely affected by the design of utility functions.

Efficiency of Solutions in Evolutionary Games

Although adapting actions according to the replicator dynamics guides the players to a fair equilibrium, this equilibrium point may not be Pareto optimal. Although this drawback is sometimes justified by the bounded rationality, there might be efficient dynamics that alleviate the inefficiency of equilibrium in evolutionary games. Design and analysis of such dynamics might contribute to realizing efficient IoT systems.

Self-Healing Properties of Evolutionary Games

Self-healing is a significant feature of resource management, which enables the system to detect, diagnose, compensate, and recover from failures and abnormal status. Despite its importance, self-healing is a rarely addressed topic in the current literature. With regard to evolutionary games, after converging to some desired operating point, a unilateral deviation of an IoT device might be triggered due to imperfect/delayed information or by an attack by an intentional mutator. This wrong decision is considered as a *mutation* in the strategy adaptation, which yields the *replicator mutator dynamics* [13]. In [14], the authors show that under certain conditions, mutation can lead the system to sustained oscillations, referred to as *limit cycles*. Therefore, analyzing the limit cycle behavior of resource management algorithms based on evolutionary games is a potential direction for future research.

Modeling the Collective Behavior of Players in Mean Field Game Models

A well developed MFG model realistically describes the impact of players' collective behavior (mean field) on the utility of each individual player. However, this is not a trivial task. In [6], a solution is proposed using stochastic geometry; nonetheless, developing new solutions is an interesting line of future research.

Solving the Mean Field Equations

As mentioned previously, solving the coupled FPK and HJB equations is challenging. Although a variety of methods have been developed so far, each method has its own pros and cons; hence, the solution approach has to be chosen based on the problem. Exploring new techniques to solve the mean field equations and adapting existing tech-

niques to different problems is an open research direction.

Modyfing Game Models

It can be seen from Table 2 that one game model is not capable of addressing all the challenges that arise when designing distributed resource management schemes for massive IoT. However, modifying game models to address those challenges is always possible. Hence, it is an important future direction.

Conclusion

Future wireless networks are expected to consist of billions of interconnected human-driven devices and machines. The emergence of such structures, also known as IoT, necessitates a search for new mathematical tools that can be beneficially used to analyze ultra-dense networks such as massive IoT. In this article, we have provided a comprehensive discussion of some unconventional analytical tools, mostly based on game theory and learning, that are particularly appropriate for handling the resource management problem in large-scale IoT systems. Specifically, we have discussed evolutionary games, mean field games, minority games, mean field bandit games, and mean field auctions. Moreover, we have discussed how each game model can provide solutions to the problems arising in resource management in massive IoT. We have also discussed potential future research directions by explaining the challenges and difficulties that arise when modeling resource management problems using the aforementioned game models in large-scale systems.

References

[1] P. Corcoran, "The Internet of Things: Why Now, and What's Next?," *IEEE Consumer Electronics Mag.*, vol. 5, no. 1, Jan. 2016, pp. 63–68.

[2] R. A. Fisher, *The Genetical Theory of Natural Selection: A Complete Variorum Edition*, Oxford Univ. Press, 1930.

[3] P. Semasinghe, E. Hossain, and K. Zhu, "An Evolutionary Game for Distributed Resource Allocation in Self-Organizing Small Cells," *IEEE Trans. Mobile Computing*, vol. 14, no. 2, Feb 2015, pp. 274–87.

[4] M. Huang, P. E. Caines, and R. P. Malhamé, "Large-Population Cost-Coupled LQG Problems with Nonuniform Agents: Individual-Mass Behavior and Decentralized ε-Nash Equilibria," *IEEE Trans. Automatic Control*, vol. 52, no. 9, 2007, pp. 1560–71.

[5] M. Burger and J. M. Schulte, "Adjoint Methods for Hamilton-Jacobl-Bellman Equations," 2010.

[6] P. Semasinghe and E. Hossain, "Downlink Power Control in Self-Organizing Dense Small Cells Underlaying Macrocells: A Mean Field Game," *IEEE Trans. Mobile Computing*, vol. 15, Feb 2016, pp. 350–63.

[7] S. Samarakoon *et al.*, "Energy-Efficient Resource Management in Ultra Dense Small Cell Networks: A Mean-Field Approach," *IEEE GLOBECOM 2015*, Dec 2015, pp. 1–6.

[8] D. Challet *et al.*, "Minority Games: Interacting Agents in Financial Markets," *OUP Catalogue*, 2013.

[9] R. Gummadi, R. Johari, and J. Y. Yu, "Mean Field Equilibria of Multi-Armed Bandit Games," *50th Annual Allerton Conf. Commun., Control and Computing*, 2012, p. 1110.

[10] S. Maghsudi and E. Hossain, "Distributed Cell Association for Energy Harvesting IoT Devices in Dense Small Cell Networks: A Mean-Field Multiarmed Bandit Approach," *CoRR*, vol. abs/1605.00057, 2016.

[11] K. Iyer, R. Johari, and M. Sundararajan, "Mean Field Equilibria of Dynamic Auctions With Learning," SSRN, Feb 2012.

[12] K. Iyer, R. Johari, and M. Sundararajan, "Mean Field Equilibria of Dynamic Auctions With Learning," *SIGecom Exch.*, vol. 10, no. 3, Dec 2011, pp. 10–14.

[13] K. M. Page and M. A. Nowak, "Unifying Evolutionary Dynamics," *J. Theoretical Biology*, vol. 219, 2002, pp. 93–98.

[14] D. Pais and N. E. Leonard, "Limit Cycles in Replicator-Mutator Network Dynamics," *IEEE 50th IEEE Conf. Decision and Control and Euro. Control Conf.*, 2011, pp. 3922–27.

Biographies

Prabodini Semasinghe (semasilp@myumanitoba.ca) received her Ph.D. in electrical and computer engineering from the University of Manitoba, Canada, in 2016, her M.Eng. degree in telecommunications from the Asian Institute of Technology, Thailand, in 2012, and her B.Sc. degree in electrical and electronic engineering from University of Peradeniya, Sri Lanka, in 2009. She is currently working as an embedded software designer at Price Industries Limited, Canada. Her major research interests are in distributed resource allocation in IoT and wireless small cell networks, applied game theory, and designing software for embedded HVAC systems.

Setareh Maghsudi (setareh.maghsudi@yale.edu) received her B.Sc. degree from Iran University of Science and Technology in 2007, her M.Sc. degree from the University of Kiel, Germany, in 2010, and her Ph.D. degree (with distinction) from the Technical University of Berlin, Germany, in 2015, all in electrical engineering. From August 2015 to August 2016 she was a postdoctoral fellow at the Department of Electrical and Computer Engineering, University of Manitoba. Currently, she is a postdoctoral fellow at Yale Institute for Network Science, Yale University. Her research interests include online learning and decision making, as well as game theory, with applications to cyber-physical systems.

Ekram Hossain [F'15] (ekram.hossain@umanitoba.ca) is a professor (since March 2010) in the Department of Electrical and Computer Engineering at the University of Manitoba. He is a member (Class of 2016) of the College of the Royal Society of Canada. He received his Ph.D. in electrical engineering from the University of Victoria, Canada, in 2001. His current research interests include design, analysis, and optimization of wireless/mobile communications networks, cognitive radio systems, and network economics. He has authored/edited several books in these areas (http://home.cc.umanitoba.ca/~hossaina). He serves as an Editor for *IEEE Wireless Communications*. Also, he is a member of the IEEE Press Editorial Board. Previously, he served as the Editor-in-Chief of *IEEE Communications Surveys & Tutorials* from 2012 to 2016 and an Area Editor for the *IEEE Transactions on Wireless Communications* in Resource Management and Multiple Access from 2009 to 2011, an Editor for *IEEE Transactions on Mobile Computing* from 2007 to 2012, and an Editor for the *IEEE Journal on Selected Areas in Communications — Cognitive Radio Series* from 2011 to 2014. He has won several research awards including the IEEE Communications Society Transmission, Access, and Optical Systems (TAOS) Technical Committee's Best Paper Award at IEEE GLOBECOM 2015, the University of Manitoba Merit Award in 2010, 2014, and 2015 (for Research and Scholarly Activities), the 2011 IEEE Communications Society Fred Ellersick Prize Paper Award, and the IEEE Wireless Communications and Networking Conference 2012 Best Paper Award. He was elevated to an IEEE Fellow "for spectrum management and resource allocation in cognitive and cellular radio networks." He was a Distinguished Lecturer of the IEEE Communications Society (2012–2015). Currently, he is a Distinguished Lecturer of the IEEE Vehicular Technology Society. He is a registered Professional Engineer in the province of Manitoba, Canada.

Future wireless networks are expected to consist of billions of interconnected human-driven devices and machines. The emergence of such structures, also known as IoT, necessitates a search for new mathematical tools that can be beneficially used to analyze ultra-dense networks such as massive IoT.

UAV-Based IoT Platform: A Crowd Surveillance Use Case

Naser Hossein Motlagh, Miloud Bagaa, and Tarik Taleb

The authors demonstrate how UAVs can be used for crowd surveillance based on face recognition. To evaluate the use case, they study the offloading of video data processing to a MEC node compared to the local processing of video data onboard UAVs. For this, they developed a testbed consisting of a local processing node and one MEC node.

ABSTRACT

Unmanned aerial vehicles are gaining a lot of popularity among an ever growing community of amateurs as well as service providers. Emerging technologies, such as LTE 4G/5G networks and mobile edge computing, will widen the use case scenarios of UAVs. In this article, we discuss the potential of UAVs, equipped with IoT devices, in delivering IoT services from great heights. A high-level view of a UAV-based integrative IoT platform for the delivery of IoT services from large height, along with the overall system orchestrator, is presented in this article. As an envisioned use case of the platform, the article demonstrates how UAVs can be used for crowd surveillance based on face recognition. To evaluate the use case, we study the offloading of video data processing to a MEC node compared to the local processing of video data onboard UAVs. For this, we developed a testbed consisting of a local processing node and one MEC node. To perform face recognition, the Local Binary Pattern Histogram method from the Open Source Computer Vision is used. The obtained results demonstrate the efficiency of the MEC-based offloading approach in saving the scarce energy of UAVs, reducing the processing time of recognition, and promptly detecting suspicious persons.

INTRODUCTION

Unmanned aerial vehicles (UAVs), also known as drones, are expected to provide diverse civilian, commercial, and governmental services. The use of UAVs has currently started in different civilian sectors. UAVs are used for environmental monitoring to monitor land pollution and industrial accidents. In agriculture, they are employed to monitor the general health of plants by showing water and nutritional stress as well as finding insect damage [1]. One of the main applications of UAVs has been in disaster relief and management. In [2], a cloud-supported UAV framework is proposed for disaster sensing applications in disconnected, intermittent, and resource-limited environments. Moreover, during the Japan East great earthquake, UAVs were used:

- To coordinate disaster relief efforts
- To capture images of the damaged reactors at the Fukushima Daiichi nuclear power plant for site assessment
- To provide real-time data of radiation levels at the nuclear power plant
- To assess the state of the cleanup and reconstruction efforts taking place in Fukushima prefecture [3].

In addition to the aforementioned applications, UAVs are used in law enforcement for border control by detecting the locations of people intending to cross national borders. In a rescue and border control use case, UAVs are daily used to rescue migrants in the Mediterranean Sea [4] with the help of video surveillance systems. Furthermore, UAVs are used for public safety, through crowd surveillance, to provide safety for crowds of people through recognizing criminals and detecting any other suspicious human activities. A potential use case of UAVs can be crowd surveillance [5]. In such a use case, cameras are mounted on UAVs; by applying face recognition methods on streamed videos, suspicious people can be detected in real time in an efficient manner.

Due to the computational overhead required by such a use case and given the limited power supply of UAVs, the processing of collected data by a UAV is a challenging issue. Nowadays, depending on the UAV type, batteries available in the market do not allow UAV flights longer than 90 minutes, and that is without doing any processing onboard UAVs [6]. Therefore, in order to ensure a flight time long enough for UAVs, the computational overhead onboard UAVs should be as lightweight as possible. The offloading process of video data processing to an edge cloud may be regarded as a solution. However, depending on the underlying radio access technology (RAT), that is, WIFI or LTE, streaming videos from UAVs to an edge cloud, that is, mobile edge computing (MEC), still requires an important amount of energy. For this reason, it is mandatory to distinguish between the applications that could be executed onboard of UAVs and those that should be offloaded to MEC. In this article, we consider the UAV-based crowd surveillance use case and investigate the benefits (or drawbacks) of the offloading process in terms of energy consumption and processing time. Indeed, along with the ongoing advances in wireless communications technologies, MEC will facilitate the offloading process from UAVs due to its expected wide deployment in the network, meaning that a UAV does not need to travel to carry out the data offload. In this article, a testbed is developed for performing face recognition using the Local Binary Pattern Histogram (LBPH) method from Open Source

Digital Object Identifier:
10.1109/MCOM.2017.1600587CM

Naser Hossein Motlagh, Miloud Bagaa, and Tarik Taleb are with Aalto University; Tarik Taleb is also with Sejong University.

Computer Vision (OpenCV), which is a precise and more accurate algorithm [7].

We used the algorithm to recognize particular faces from a database of 40 faces, each stored in a separate directory, considering that each person has 10 different facial details and expressions (e.g., open vs. closed eyes, smiling or not smiling, face with or without glasses). To do the experiment, 10 videos of different lengths are taken in real life by a camera, each of which contains a group of people. Thus, we compared the performance of the offloading process against the local processing of the face recognition onboard UAVs. To make this comparison, we performed two experiments. First, we sought to recognize the faces of five suspected people while varying the video lengths. The results of this experiment demonstrate that it is highly efficient to offload the face recognition operation to MEC rather than processing it locally onboard UAVs. Second, we looked for face recognition by setting the video duration to 1 s while the number of suspected people was varied. The results of the second experiment show that the energy required by a UAV and the processing time when the video processing is offloaded remains the same regardless of the number of profiled persons. The envisioned UAV-based crowd surveillance use case is implemented as part of the target UAV-based IoT platform described below.

The article is organized in the following fashion. We describe the envisioned UAV-based IoT platform. We discuss the potential of UAVs and their integral role in fifth generation (5G) mobile systems. We introduce the target UAV-based crowd surveillance use case and discuss the results obtained from a real-life implementation of the use case. The article then concludes.

UAV-Based IoT Platform

While UAVs are used for their original tasks (e.g., parcel delivery by Amazon, power line monitoring by SharperShape), they can be simultaneously applied for offering numerous value added services (VASs), particularly in the Internet of Things (IoT) when they are equipped with remotely controllable IoT devices. In such a way, UAVs will form an innovative UAV-based IoT platform operational in the sky [1]. This shall decrease the capital and operational expenses for creating a novel ecosystem. Through this platform, IoT data can be collected via remotely controllable IoT devices mounted on UAVs whenever triggered on and off at the right time, at the intended positions, and/or per specific events. Based on the required energy, the collected data can be processed locally onboard UAVs or offloaded to cloud servers on the ground. To build an efficient UAV-based IoT platform, there is need for a platform orchestrator (centralized or distributed) that is aware of diverse contextual information about UAVs, such as their flying routes, their IoT equipment, and their battery status. For instance, in a scenario when a police department requests a video record from a specific position, the appropriate flying UAV has to deviate from its original path to execute the task. To do this, knowledge on the current state of the UAV such as its current geographical position and its remaining energy becomes mandatory [8]. Figure 2 shows our envisioned architecture for the UAV-based IoT platform. The figure demonstrates a widespread network of flying UAVs, each assigned to a specific task: some are flying, and some are ready to fly when needed.

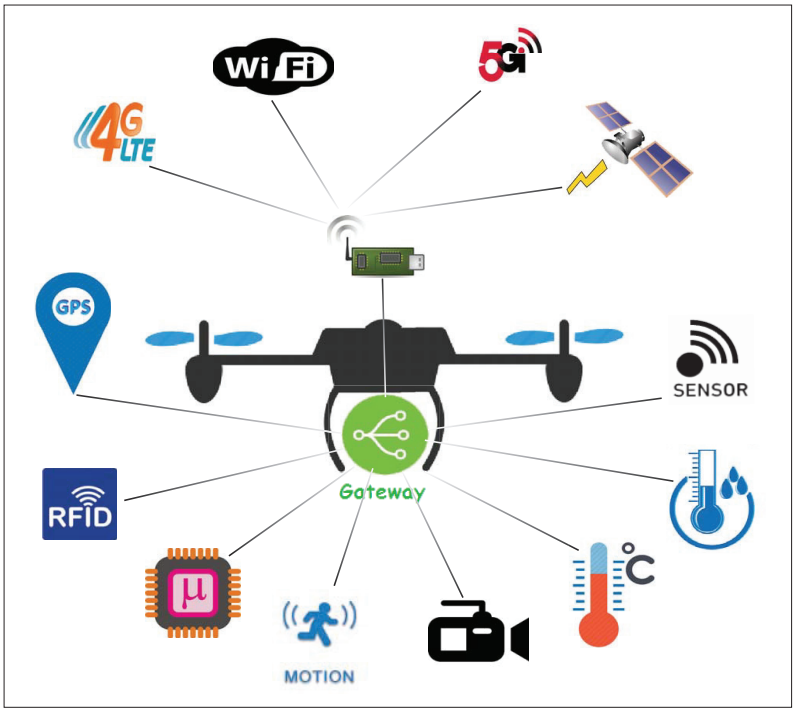


Figure 1. UAVs equipped with diverse IoT devices.

The data delivery from UAVs is performed by any wireless technology that suits the target UAV application such as WiFi and cellular networks (i.e., 4G-LTE, 5G). The choice of wireless technology may depend on diverse factors such as required security, reliability, and system responsiveness. Instead of UAV-to-ground communications, UAVs may also form clusters, in a flying ad hoc networking (FANET) manner, leveraging their short-range wireless communication technologies (e.g., Bluetooth and WiFi) to benefit from sharing their onboard IoT devices, computation resources, and data transmission links. In a cluster, a suitable UAV could be elected as the cluster head to transfer the collected IoT data on behalf of other UAVs to the ground station. Such a clustering approach may be beneficial in situations where UAVs do not have enough individual power/computation resources to accomplish a task or may need to complement each other's IoT devices to carry out an IoT task. Figure 2 depicts the system orchestrator (SO), which coordinates the operations of UAVs and their IoT devices and handles requests from users for IoT services. To satisfy a request for an IoT service, the SO first selects the most suitable UAVs based on many metrics such as UAVs' current routes, their onboard IoT equipment, their residual energy level, and the priority level of their current mission [8]. The SO also coordinates the flying paths of UAVs, ensuring collision-free travel. For secured communications between UAVs and ground stations, the SO instructs UAVs on which access technology to employ and when, and specifies where the data should be delivered (e.g., edge vs. central cloud). The SO is assumed to have all necessary intelligence to be self-capable to autonomous-

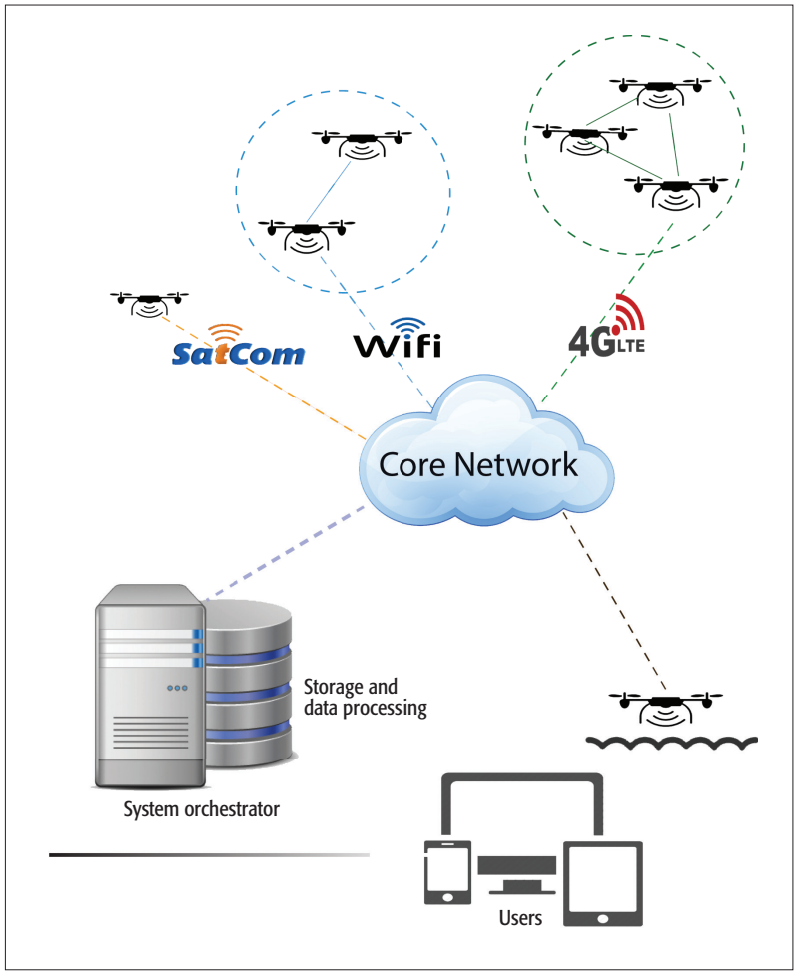


Figure 2. High-level view of the envisioned UAV-based IoT platform.

ly self-operate, self-heal, self-configure, and adequately resolve any possible conflicts from diverse policies [1].

UAV Potential in Advanced Mobile Communication Systems

UAVs exhibit outstanding characteristics compared to manned airplanes. They have unique features for being dynamic, easy to deploy, easy to reprogram in flight, able to measure anything anywhere, and able to fly in a controlled air space with a high degree of autonomy [9]. As mentioned earlier, UAVs can be used to provision diverse services, ranging from civilian to commercial and governmental. Using suitable IoT devices, cameras, and communication devices, countless use cases can be defined for UAVs. For instance, using high-resolution cameras and a suitable communication system such as LTE, UAVs can be used for crowd surveillance, the core topic of this article. This use case can obviously be considered for security reasons to monitor any suspicious activity among crowds of people. When equipped with suitable IoT devices, UAVs can be used to collect IoT data from great heights. Depending on the energy required for the computation of the IoT data and the urgency of the IoT task, the collected IoT data can be processed locally or delivered to an adequate server using a suitable RAT [10]. The data can be gathered from any sensor (e.g., temperature and humidity) or any imagery device (e.g., digital camera). The latter can be used for surveillance, inspection, mapping, or modeling. Most existing UAVs have the ability to deliver data in real time to a ground control station (GCS). Some have local data storage and processing capabilities, enabling them to carry out computational tasks onboard. Most IoT devices onboard UAVs (e.g., sensors, cameras, actuators, and RFIDs) are remotely controllable (Fig. 1).

In addition, UAVs can employ FANET principles to deliver data to a server/GCS. FANET resolves several design limitations associated with the infrastructure-based architecture approach. It solves the communication range restriction between UAVs and GCS, and provides a certain level of reliability for the communication [11]. An important issue in UAV communications pertains to the type of communication technology to be employed onboard UAVs. Due to the dynamic and mobility features of UAVs, there is a need to guarantee reliable communication among them (i.e., good coverage, stable connectivity, and sufficient throughput). The advanced communication systems (i.e., LTE 4G and 5G mobile networks) will be the communication standard to support the long distance, high altitude, and high mobility nature of UAVs. UAVs will use these communication technologies to transfer or exchange data with diverse IoT devices on the ground in a machine-to-machine (M2M) manner as well as to communicate with GCS. Indeed, current LTE 4G systems are used to increase network expandability up to hundreds of thousands of connections for low-cost, long-range, and low-power machine type communication (MTC)/IoT devices. In addition, 5G networks will be designed to offer high data speed (i.e. exceeding 10 Gb/s) and extremely low latency (i.e., 1 ms) [12]. These networks will provide ubiquitous coverage, including at high altitudes. They will support 3D connectivity; a characteristic referring to the ultra-high reliability, ultra-high availability, and ultra-low latency features of UAVs. One of the most important features of these mobile networks shall be support for extreme real-time communications such as real-time mobile video surveillance and streaming. Furthermore, they shall provide broadband access enabling high-definition video and photo sharing in a densely populated area. These mobile networks are also expected to support UAVs in avoiding physical collisions among them by supporting remote planning and alteration (when needed) of their flying routes.

Along with MEC, these advanced communication systems could lift the computing and storage resource restrictions of UAVs, enabling them to offload intensive computations to the edge cloud. Indeed, MEC aims to place generic storage and computing close to the network edge in a mobile network environment. MEC also aims to enable billions of mobile devices to operate for real-time and computation-intensive applications directly at the network edge. MEC can be applied for different use cases as video analytics, location services, IoT, augmented reality, optimized local content distribution, and data caching. The outstanding characteristics of MEC are its service mobility support, closeness to end users, and the dense

geographical deployment of the MEC servers [13]. These capabilities will contribute to wide deployment of UAVs, such as the Unmanned Aerial System (UAS) Traffic Management (UTM) system envisioned by the U.S. National Aeronautics and Space Administration (NASA) [14]. With such wide deployments of UAVs, new business models will appear whereby UAVs can be used as a backbone for the ground Internet and/or to complement the coverage of 5G. In this regard, it is worth mentioning Google's project, SkyBender, which uses UAVs to deliver Internet at speeds 40 times faster than 4G systems in the Mexico desert [15]. However, there is a delivery range restriction as the project employs high-frequency millimeter-wave technology that has shorter communication range in comparison to the traditional wireless communication technologies.

UAV-Based Crowd Surveillance

In public places such as stadiums or during parades, it is important to protect civilians from threats. Indeed, in recent years, the rate of crimes in urban areas, such as street crimes, vandalism, and terrorism, has increased. Therefore, anticipating crimes through detection and recognition of criminals among crowds of people is an important approach. In traditional patrol systems, there is a need for many security guards and a huge amount of human effort to provide necessary safety for people. In this vein, UAVs can be used to assist security guards by remotely surveilling people at places of interest. UAVs can provide immunity from any hazard and help not just to control but to track, detect, and recognize criminals adopting face recognition methods. Employing UAVs with appropriate IoT devices, such as video cameras, can offer an efficient crowd surveillance system; detect any eccentric motion and suspicious action; and recognize criminals' faces. The use of this technology provides a bird's eye view for crowd surveillance and face recognition. Therefore, crowd safety and security can be enhanced, while at the same time, the number of security guards deployed on the ground can be reduced. The process of face recognition consists of well defined steps: facial features extraction, database creation of known faces, and face detection matching videotaped faces with profiled ones. Different video analytic tools are available. Many of them can cope with the high mobility feature of UAVs and can achieve face recognition with high accuracy. Recognition of multiple faces at the same time is also possible. The processing of recorded video for face recognition can happen locally as well as at remote servers, enabling the offloading of the face recognition operation to MEC. OpenCV presents noticeable algorithms for face recognition. It employs machine learning to search for profiled faces within a video frame. Indeed, OpenCV uses LBPH with its associated libraries and databases. The approach of LBPH is to summarize the local structure in an image by comparing the pixels with its adjacent ones. LBPH results in accurate face recognition.

In the remainder of this article, we demonstrate how much impact the offloading of face recognition computation has on the energy consumption of UAVs and the overall processing time. Figure 3 depicts the envisioned experiment scenario. In this scenario, we consider a UAV equipped with a video camera and connected to the GCS through LTE cellular network. Figure 4a shows the UAV used in our experiment. The figure also shows the LTE eNodeB used (donated by Nokia). The underlying LTE network is exclusively used for research, and offers low latency and a high bit rate as well as extended coverage to support a variety of scenarios, where measurements can be carried out horizontally, vertically, at higher altitudes, with line of sight (LoS) and beyond LoS. The network includes edge computing resources co-located with the LTE base stations deployed in the Aalto University campus, thus enabling dedicated high-speed low-latency access to critical resources. This is schematically represented by MEC in Fig. 3. The used UAV is a built-in hexacopter equipped with an LTE modem, a gimbal with a high-resolution digital camera, as well as several computing and sensing resources. They include a flight controller (FC) module for stable flight, equipped with gyroscopes, accelerometers, and a barometer; and an embedded Linux system (i.e., a Raspberry Pi) interconnecting the LTE modem to the FC. To set up an LTE connection, any PC can be used as a GCS. On the PC, flight control software, such as Mission Planner, is installed. The PC is used for controlling the FC via a connected LTE modem.

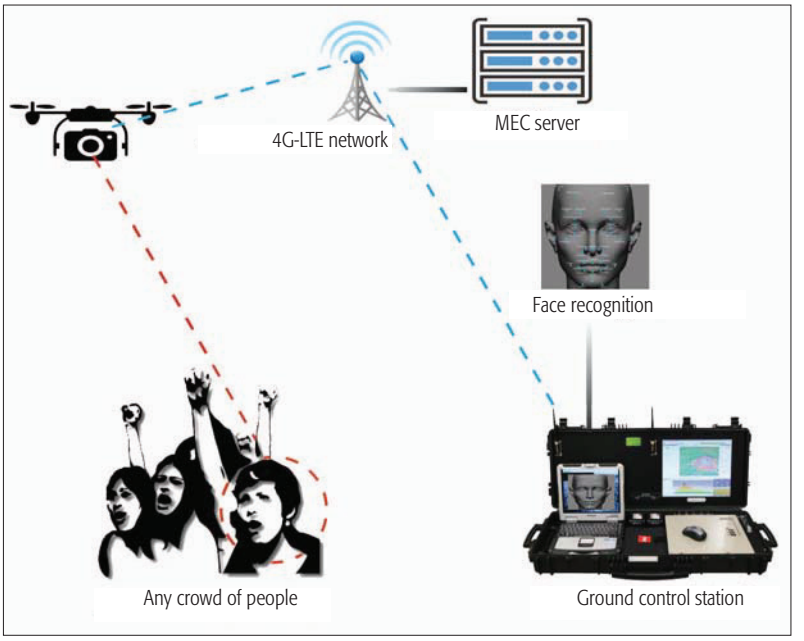


Figure 3. High-level diagram of the envisioned experiment scenario.

The hexacopter can carry 1.5 kg of payload, including laboratory equipment and metering devices. With a completely charged battery, its flight time is around 30 minutes with the full payload. It also has a safe landing scheme to cope with unlikely motor failure situations. In the envisioned scenario, security guards access the control station and continuously surveille the people. Upon noticing uncommon behavior from a particular person (or group of persons), they command the UAV to take a video of the person(s) and apply facial recognition on the captured video to identify the suspicious person(s) and verify if he/she/they have any criminal records. To investigate the benefits of computation offloading of the facial recognition operation to MEC vs. its local

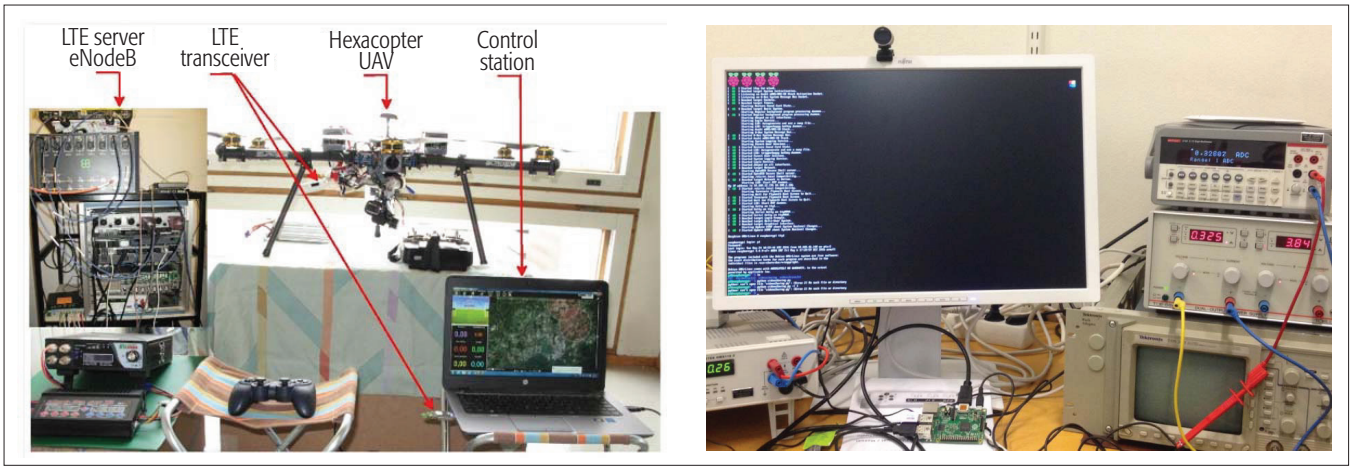


Figure 4. Experiment setup (testbed): a) used UAV along with LTE-based system for UAV control; b) testbed for energy consumption measurement.

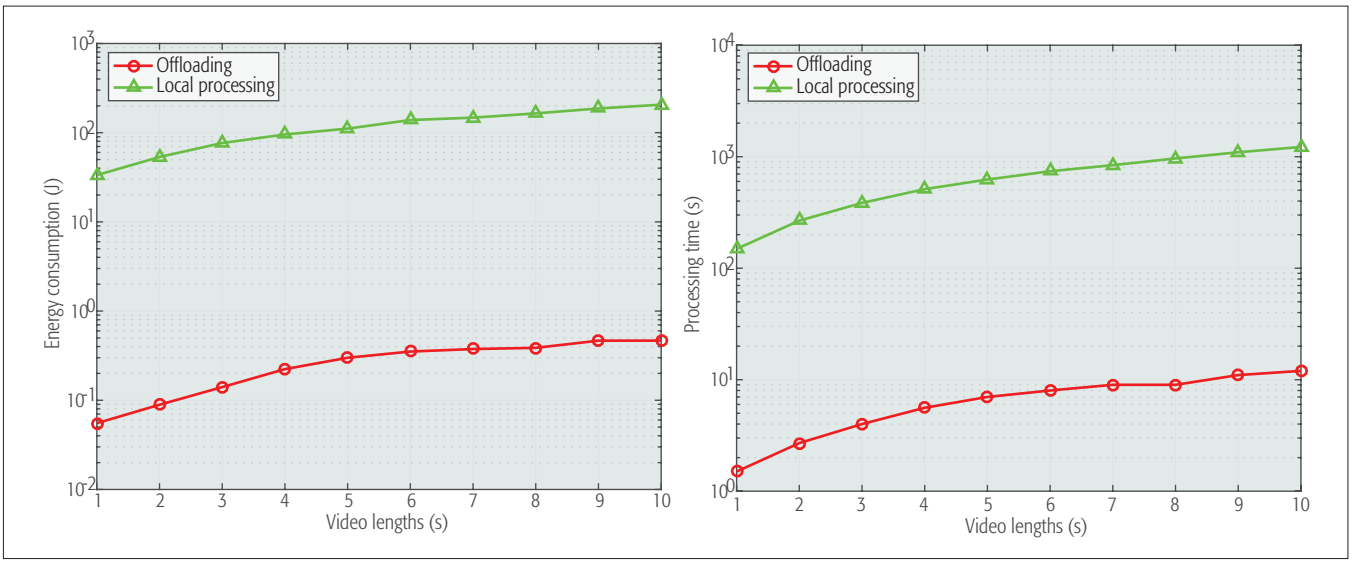


Figure 5. Performance evaluation when different videos of different lengths are processed locally onboard a UAV and when their computation is offloaded to MEC.

processing, we developed a small-scale testbed as shown in Fig. 4b. The testbed environment consists of a Raspberry Pi (RPi) and a laptop that serves as a MEC node. The RPi works as the local processing unit onboard the UAV. In addition, the laptop works as the command and control station of the UAV's gateway for turning the camera on/off, or to command it to locally process the face recognition or offload the processing to the MEC node.

In the experiment and as stated earlier, we used OpenCV's LBPH algorithm to recognize particular faces from a database of 40 faces, each stored in a separate directory. Each person has 10 different faces, varying in brightness/contrast, facial expression (open vs. closed eyes, smiling, not smiling), and facial details (e.g., with or without glasses). The code source in the testbed was developed using the Python programming language. In the experiments, we used TOE8842 dual output power supply as the DC power generator of RPi and set it to 5 V. For the energy consumption measurement, we used a 6-digit resolution digital multi-meter to measure the current (I). In the testbed, 10 videos of different lengths were taken in real life from the camera, each of which contains a group of people. The duration of the *i*th video is *i* s (i.e., duration of the 5th video is 5 s). Therefore, we evaluate the performance in terms of energy consumed and processing time when the facial recognition operation is carried out onboard the UAV and when it is offloaded to MEC. Figure 5 shows the results of our first experiment where we looked to recognize the faces of five suspected people while varying the video lengths. The figure shows that it is far more efficient to offload the facial recognition operation to MEC rather than processing it locally onboard a resource-constrained UAV. Indeed, local processing of the video data consumes a significant amount of energy and drains the UAVs scarce battery. Moreover, the offloading process drastically reduces the processing time compared to performing the facial recognition locally onboard the UAV. From this figure, we observe that the offloading process reduces the energy consumption and processing time more than 100 times compared to performing local processing of video onboard UAVs. Figure 6 shows the results of the second experiment, where the video duration is set to 1 s and the number of suspected people is

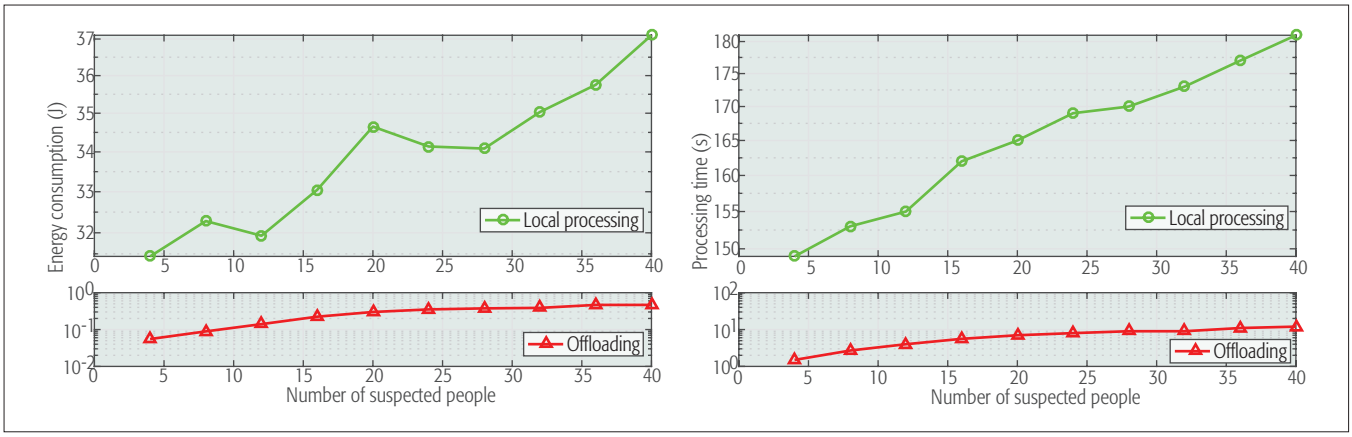


Figure 6. Performance evaluation when a 1s-long video is processed locally on board of UAV and when its computation is offloaded to MEC to recognize different numbers of profiled persons.

varied. The results show that the energy required by the UAV and the processing time if the video processing is offloaded remains the same regardless of the number of profiled persons. However, when the video is processed locally, the required energy and the processing time increase somewhat linearly along with the number of profiled people when the video is processed locally.

Conclusion and Future Work

UAVs are gaining lots of momentum. When equipped with diverse IoT devices, they can be used to form an integrative IoT platform operational in the sky. In this article, we present a high-level view of such a UAV-based IoT platform. As a specific use case of the platform, the article introduces the case of UAV-based crowd surveillance applying facial recognition tools. A testbed is developed using a built-in UAV along with a real-life LTE network. The article compares two cases: when videos are processed locally onboard UAVs and when their processing is offloaded to MEC. The obtained results demonstrate clearly the benefits of computation offloading in saving energy and significantly improving system responsiveness in quickly detecting and recognizing suspicious persons in a crowd. Improvement in the performance becomes more noticeable for longer videos and also when the number of profiled persons is high.

In the future, we will work toward performing crowd surveillance and facial recognition when a cluster of UAVs are employed. In our study, we will investigate the energy consumption by means of local processing when the UAVs share the processing tasks among themselves vs. offloading the computational tasks to MEC. We will also use more than one MEC node to study the efficiency of processing time when the tasks are performed locally by the cluster members vs. when they are offloaded to the MEC nodes. In addition, we are seeking to use more efficient algorithms for testing the crowd surveillance use case.

References

[1] N. H. Motlagh, T. Taleb, and O. Arouk, "Low-Altitude Unmanned Aerial Vehicles-Based Internet of Things Services: Comprehensive Survey and Future Perspectives," *IEEE Internet of Things J.*, vol. 3, no. 6, Dec.. 2016, pp. 899–922.

[2] C. Luo *et al.*, "A UAV-Cloud System for Disaster Sensing Applications," *IEEE VTC-Spring*, Glasgow, U.K., May 2015, pp. 1–5.

[3] S. Jacek, "Fukushima Plants Radiation Levels Monitored with an UAV," https://theaviationist.com/2014/01/29/fukushima-japan-uav/, Jan. 2014, accessed 12 June 2016.

[4] M. Elizabeth, "Human Rights at Sea: The Success of The Migrant Offshore Aid Station (MOAS)," https://www.humanrightsatsea.org/thesuccess-of-the-migrant-offshore-aid-station-moas/, Dec. 2014, accessed 12 June 2016.

[5] S. Qazi, A. S. Siddiqui, and A. I. Wagan, "UAV Based Real Time Video Surveillance Over 4G LTE," *Int'l. Conf. Open Source Systems Technologies*, Lahore, Pakistan, Dec. 2015, pp. 141–45.

[6] Microdrones, "UAV/Drone Equipment: Multispectral Camera, Thermal Imaging, GPS," https://www.microdrones.com/en/products/equipment/, 2016, accessed 08 Mar. 2016.

[7] OpenCV Development Team, "FaceRecognizer — Face Recognition with OpenCV," http://docs.opencv.org/2.4/modules/contrib/doc/facerec/, 2016, accessed 12 June 2016.

[8] N. H. Motlagh, M. Bagaa, and T. Taleb, "UAV Selection for a UAV-Based Integrative IoT Platform," *Proc. IEEE GLOBECOM 2016*, Washington, DC, Dec. 2016.

[9] S. Colin, "Why Drones Are the Future of the Internet of Things," http://droneanalyst.com/2014/12/01/drones-are-the-future-of-iot/, 2016, accessed 12 June 2016.

[10] S. Koulali *et al.*, "A Green Strategic Activity Scheduling for UAV Networks: A Sub-Modular Game Perspective," *IEEE Commun. Mag.*, vol. 54, no. 5, May 2016, pp. 58–64.

[11] H. Tareque, S. Hossain, and M. Atiquzzaman, "On the Routing in Flying Ad Hoc Networks," *2015 Federated Conf. Computer Science and Info. Systems*, vol. 5, Lodz, Poland, Sept. 2015, pp. 1–9.

[12] H. Shariatmadari et al., "Machine-Type Communications: Current Status and Future Perspectives toward 5G Systems," *IEEE Commun. Mag., Standards Supplement*, vol. 53, no. 9, Sept. 2015, pp. 10–17.

[13] T. Taleb *et al.*, "Mobile Edge Computing Potential in Making Cities Smarter," to appear, *IEEE Commun. Mag.*

[14] R. Joseph, "Unmanned Aircraft System (UAS) Traffic Management (UTM)," https://utm.arc.nasa.gov/, July 2016, accessed 10-October-2016.

[15] H. Mark, "Project Skybender: Googles Secretive 5G Internet Drone Tests Revealed," http://www.theguardian.com/technology/2016/jan/29/projectskybender-google-drone-tests-internet-spaceport-virgin-galactic/, Jan. 2016, accessed 04 June 2016.

Biographies

Naser Hossein Motlagh [S'15] (naser.hossein.motlagh@aalto.fi) received his B.Sc. degree in information technology (communications and networking) and M.Sc. degree in telecommunications engineering from Vaasa University of Applied Sciences and the University of Vaasa, Finland, in 2009 and 2012, respectively. Currently, he is pursuing his Ph.D. degree at the Department of Communications and Networking, School of Electrical Engineering, Aalto University, Finland. His research interest includes device-to-device communications, the Internet of Things, and unmanned aerial vehicles.

Miloud Bagaa (miloud.bagaa@aalto.fi) received his Engineer's, Master's, and Ph.D. degrees from the University of Science and Technology Houari Boumediene (USTHB), Algiers, Algeria, in 2005, 2008, and 2014, respectively. From 2009 to 2015, he

was a researcher with the Research Center on Scientific and Technical Information (CERIST), Algiers, where he was a member of the Wireless Sensor Networks team, DTISI Division. From 2015 to 2016, he was granted a postdoctoral fellowship from the European Research Consortium for Informatics and Mathematics, and worked at the Norwegian University of Science and Technology, Trondheim, Norway. Currently, he is a senior researcher with the Communications and Networking Department, Aalto University. His research interests include wireless sensor networks, the Internet of Things, 5G wireless communication, security, and networking modeling.

Tarik Taleb [S'04, M'05, SM'10] (tarik.taleb@aalto.fi) received his B.E. degree in information engineering, and M.Sc. and Ph.D. degrees in information sciences from Tohoku University, Sendai, Japan, in 2001, 2003, and 2005, respectively. He is currently a professor at the School of Electrical Engineering, Aalto University. He has been a senior researcher and 3GPP standardization expert with NEC Europe Ltd., Heidelberg, Germany. He previously worked as an assistant professor at the Graduate School of Information Sciences, Tohoku University, Sendai, Japan. His current research interests include architectural enhancements to mobile core networks, mobile cloud networking, mobile multimedia streaming, and social media networking. He has also been directly engaged in the development and standardization of the Evolved Packet System as a member of 3GPP's System Architecture working group. He has received many awards, including the IEEE ComSoc Asia Pacific Best Young Researcher Award in June 2009, and some of his research work has also received best paper awards at prestigious conferences.

Business Development in the Internet of Things: A Matter of Vertical Cooperation

Amirhossein Ghanbari, Andres Laya, Jesus Alonso-Zarate, and Jan Markendahl

ABSTRACT

Smart and connected devices can improve industrial processes, and generate new and better services. While this premise is well understood within the ICT industry, there is a challenge in extending this knowledge to vertical industries. The potential of the Internet of Things lies in the interaction among industries working together toward value co-creation. Firms need to look beyond their internal business models and explore cooperative perspectives to define new business opportunities. In this article, we look into the relevance of vertical cooperation in the area of IoT and highlight the need to develop new value networks that leverage this cooperation and enable the creation of new business models. To lead our discussions, we use the examples of two major building blocks of smart cities: intelligent transport systems and health and well being services based on connected devices and solutions.

INTRODUCTION

The vision of the Internet of Things (IoT) as a hyperconnected world of interconnected devices can improve industries and facilitate better services. While technology and standards for the IoT are on their way to maturity, there are also challenges and opportunities to be faced from the business perspective. The challenges lie in the fact that there are no predefined processes to follow or well established boundaries in the market. And the opportunities are that IoT has the potential to create completely new possibilities in uncontested markets.

The IoT concept first caught the attention of the information, communications, and technology (ICT) industry, where the potential of connected devices was clearer. Therefore, it has become necessary for the ICT technical community to show grounded evidence that other industries can benefit from IoT solutions. It is clear that mere technology availability is not enough to drive the creation of added-value solutions, and it is thus necessary to explore and define the business context and the opportunities to build valuable propositions.

We argue that in order to support creation of solutions based on IoT, the ICT sector needs to have closer involvement in the development of services and understand how it can be profitable for other industries.

We take the stance that IoT corresponds to a set of technologies, principles, and systems associated with Internet-connected objects. The concept of IoT encompasses data sensing and acquisition, processing, transmission, storage, analysis, actuation, and sharing; all this, with the ultimate purpose of creating value. The stages of value creation in IoT are represented in Fig. 1 [1]. The inner loop represents the smart device information loop, and it refers to products with embedded sensor and computational capabilities used to improve their own efficiency. Information is sensed and analyzed locally to generate an action. The outer loop represents the connected device information loop, which allows the device to exchange data with the outer world. Data can be communicated and aggregated by third parties to create added value. Even though these two loops only represent the ICT perspective of value creation in IoT, they already highlight the associated complexity, where different resources, players, and expertise are necessary to create valuable solutions.

The creation of viable IoT solutions needs coordinated interaction among industries. This leads to value co-creation among involved entities, which blurs the differentiation between providers and consumers. This is different from traditional mobile broadband business, where communication providers usually interact directly with end users. In the IoT context, other industry verticals are directly involved in the value creation process, and in many cases have the direct relationship with end users. Hence, the ICT industry is facing a new set of actors between themselves and the end users. These actors are then both co-creators of IoT solutions as well as consumers of ICT service enablement for IoT.

As the main purpose of any business, co-creation should also lead to a profitable outcome for involved entities. At the same time, the customer should also benefit from the created solution. However, in order to find and define business opportunities in the emerging IoT context, the burning question of *how can we make profit with the IoT?* should be preceded by *how can we create value by using the IoT?* The latter question has two angles: first, creating value for end users, and second, co-creating the value together with other vertical industries.

In this article, we elaborate on the limitations of existing value chains for defining business models for the IoT. We describe the interactions

The potential of the Internet of Things lies in the interaction among industries working together toward value co-creation. Firms need to look beyond their internal business models and explore cooperative perspectives to define new business opportunities. The authors look into the relevance of vertical cooperation in the IoT and highlight the need to develop new value networks that leverage this cooperation and enable the creation of new business models.

Andres Laya, Amirhossein Ghanbari, and Jan Markendahl are with KTH Royal Institute of Technology; Jesus Alonso-Zarate is with the Centre Tecnologic de Telecomunicacions de Catalunya (CTTC/CERCA).

Digital Object Identifier: 10.1109/MCOM.2017.1600596CM

In the highly cross-industrial ecosystems around the IoT, it is unlikely that a single firm will provide a complete IoT solution. Companies require resources and knowledge from different fields which do not necessarily belong to a single industry. Therefore, relationships must be built within and across industries.

required in future ecosystems of converging industries leveraging on the IoT. We discuss the emerging perspective of value networks as drivers of co-creation models for the IoT. We describe two examples of use cases for the smart city to exemplify the need for a new business perspective. Finally, we conclude the article with final remarks.

Limitations of Existing Value Chains and Business Models for the IoT

Business models constitute tools that can help define processes and mechanisms for value creation. They allow defining for whom the value is being created and how revenues can be obtained. However, business models are largely focused on the perspective of a single firm. This means that they typically explain how one company can deliver value to its customers and stakeholders by establishing relationships with partners and suppliers within the internal logic of a firm.

In the cross-industrial ecosystems around the IoT, it is unlikely that a single firm will provide a complete IoT solution. Companies require resources and knowledge from different fields that do not necessarily belong to a single industry. Therefore, relationships must be built within and across industries. Tight collaboration between the telecommunication industry and service providers from different industries, such as health, automotive, utility, and industrial, is needed to implement IoT-based solutions that can provide value. According to [2], business relationships can be put into four major categories:

- *Cooperation*: Firms interact with each other and combine complementary know-how or resources to achieve a common goal.
- *Competition*: Firms compete over similar goals that can generally translate into customer and market shares.
- *Coopetition*: This comprises competition and cooperation happening simultaneously; for instance, mobile network operators (MNOs) competing on customer acquisition but cooperating on network sharing.
- *Coexistence*: Firms exist and affect each other indirectly, without direct business interactions. They exist as part of the same industry, but do not cooperate or compete over similar goals.

These four business relationship categories fall within two major setups: *vertical* and *horizontal*. If we consider the traditional value creation process of *value chains*, business interactions mainly consist of vertical relationships between sellers and buyers. In contrast, horizontal relationships consist of interactions among similar firms, which usually involve a certain level of competition.

As discussed before, the IoT will necessarily bring more cooperation and coopetition, where players will need to establish new ways of creating business. A traditional example of vertical cooperation is the one established between telecommunication equipment vendors (TEVs) and MNOs. TEVs are the suppliers and MNOs are the customers, but they do work together to provide added value toward final customers. In recent years, this relationship clearly goes beyond just selling and buying telecom equipment. This vertical cooperation can also occur across industries. We can take the example of TEVs and MNOs cooperating with automotive companies to enable telemetry, infotainment, and even autonomous driving in vehicles. Hence, vertical cooperation refers to the concept of suppliers collaborating with customers in order to satisfy their needs toward the service they offer to end users.

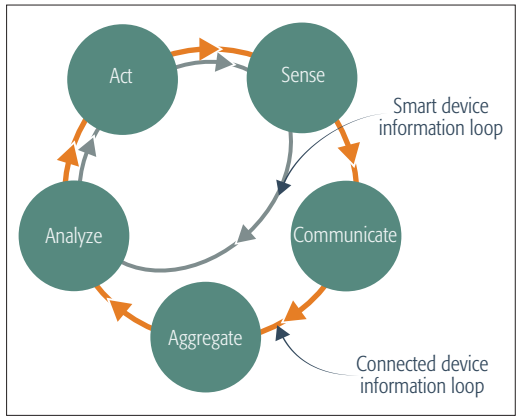


Figure 1. Smart and connected value creation as information loops in the Internet of Things. Adapted from [1].

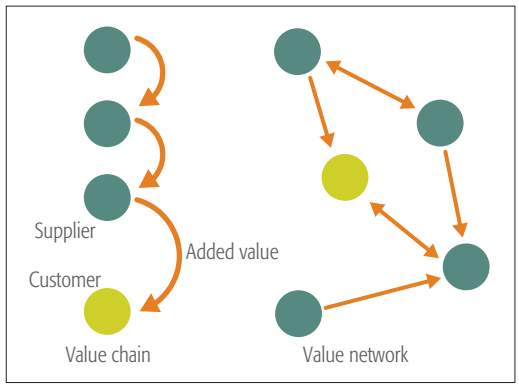


Figure 2. Value chain and value network representation.

Horizontal cooperation/coopetition can happen for various reasons such as lack of resources, external market forces, or simply cost efficiency. This could be the example of a number of MNOs sharing telecom infrastructure or licensed spectrum to reduce costs. These cooperative relationships can be articulated in the form of alliances, partnerships, joint ventures, consortia, and supply and marketing agreements [3].

Given this context, where cooperation and coopetition will become key enablers for the IoT market, it seems clear that business relationships in IoT must be discussed in more comprehensive topologies rather than value chains. Indeed, we must use the concept of *value networks* to discuss, evaluate, and propose new ways of cooperation (Fig. 2). According to [3], there are at least two reasons to establish cooperative business relationships:

- To reduce transaction costs and increase organizational learning: Close integration and interaction between suppliers and customers can improve efficiency, increase resource utilization, and ultimately reduce costs.

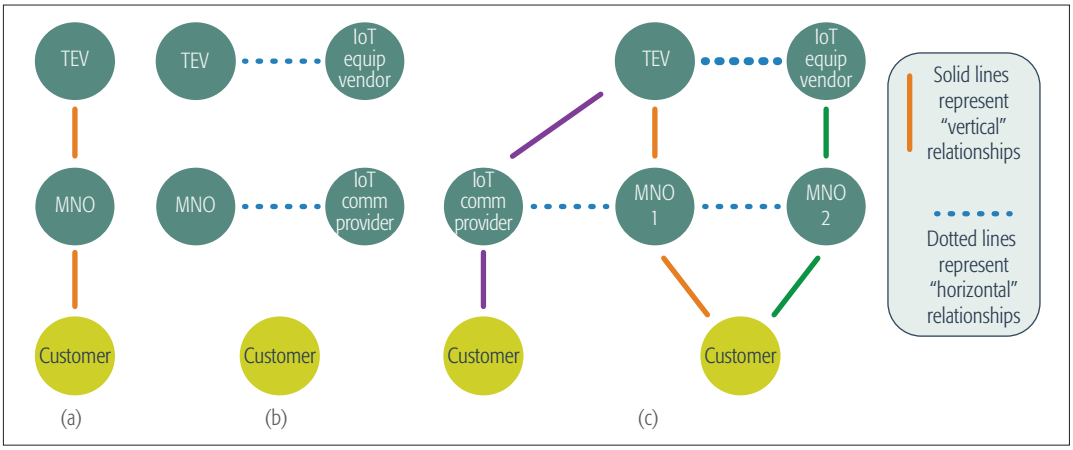


Figure 3. Vertical vs. horizontal relationships among involved actors.

- To handle resource dependency: This means that cooperative entities can share knowledge and exchange access to resources.

The second reason is particularly important in the complex products and services enabled by IoT. We can take the previous example of vertical cooperation to enable connected vehicles. In this example TEVs and MNOs cooperate to handle resource dependency while offering *connectivity* as a service to the automotive company. The automotive company, together with an MNO and a TEV, co-creates value in order to offer connected cars to end users. TEVs and MNOs are unlikely to pursue a car manufacturing venture to offer services to end users. Therefore, such vertical cooperation with the automotive industry lets them get into a new emerging market. This kind of value co-creation cannot be fully explained from a single-firm perspective.

It should be highlighted that along with the emergence of more applications of IoT, new actors are appearing to cover gaps in the market. These new actors, such as dedicated IoT communication providers, may take over the role of traditional actors such as MNOs or TEVs for enabling IoT-based solutions. Therefore, any actor who is capable of providing connectivity for the IoT solution, given that it possesses the required resources and competences, can replace traditional actors in IoT value networks as long as it is able to fulfill the needs in the vertical industries.

The creation of IoT services can benefit from this *value network* perspective to identify and define new business opportunities. This perspective is explained later. There are two concepts that are widely used in the literature and refer to this perspective: *value networks* and *business ecosystems*. In this article, we consider that both terms refer to interconnected communities of diverse actors that are interdependent on each other and interact to create and capture new value [4].

It is now clear that business models based on value chain definitions fall short of describing the complex context posed by many IoT solutions. We need to go for more complex relationship tools based on value networks instead. However, *how can we create or use business model tools when we need to embrace a value network perspective?* In order to provide an answer to this question, we look into relevant theories. To do so, we take as an example the context of smart cities, as they constitute a market where many industries come together and offer IoT services with the aid of ICT. In particular, we further expand on two of the main building blocks of any smart city:

- Intelligent transport systems (ITS); to show how ICT value chain integrates with transport value chain and creates a new value network
- Healthcare and well being (H&WB); to showcase the need to involve health, governmental, and ICT firms in service development for IoT toward the development of innovative business models

The final objective in both cases lies in highlighting the fact that different firms, based on different value chains, can cooperate and exchange their expertise and resources to co-create value.

Even though there are emerging frameworks addressing business models in value networks, there are no tools readily available in the IoT context yet that can provide a methodology to use business resources as input and obtain business opportunities as output. This might happen in the future and today constitutes a hot research area. However, even if there is no dominant design and development framework for business for IoT, this is the strategy leading to success: collaboration and service development in value networks.

Understanding Vertical and Horizontal Relationships in the IoT

In order to understand complex business relationships in the IoT value creation process, we need to differentiate between vertical and horizontal relationships (Fig. 3). Vertical relationships represent the seller-buyer, or supplier-customer, relationships (Fig. 3a), while horizontal relationships represent the case where two or more firms offer similar service (or value) in the value creation process (e.g., an MNO and an IoT communication provider, as in Fig. 3b). Since the focus of this article is on competition and cooperation as the main factors of driving value creation in IoT, in our discussions we classify the relationships into four possible categories (as presented in Table 1):

1. Horizontal competition
2. Horizontal cooperation
3. Vertical competition
4. Vertical cooperation

Even though there are emerging frameworks addressing business models in value networks, there are no tools readily available in the IoT context yet that can provide a methodology to use business resources as input and obtain business opportunities as output. This might happen in the future and today constitutes a hot research area.

As a consequence of vertical competition with cooperators, if a customer like an MNO perceives its supplier as a competitor to get a new IoT customer, it may decide to change its supplier. This can be an immediate effect of "negative feelings" as a consequence of a competitive relationship.

	Competition	Cooperation	Coopetition
Horizontal	MNO–MNO Competition over market share in mobile subscribers	MNO–MNO Roaming in mobile networks	MNO–MNO Cellular network sharing
Vertical	TEV – MNO Competition over offering IoT service platforms for third parties	TEV – MNO TEV enabling IoT platform and MNO enabling cellular connectivity	TEV – MNO Simultaneous competition over a customer and cooperation over another customer

Table 1. Examples of vertical and horizontal relationships in the IoT service provisioning.

For the discussions, even though we focus on relations among vertically positioned actors, we also benefit from similarities with horizontal relationships.

The general factors increasing level of competition within the ICT industry in the context of IoT are [5]:

1. Increased overlap between firms' competitive goals. In this case, tap into new solutions based on connected devices. We even argue that the goal is to achieve control over generated data.
2. Increased maturity of ICT industry, particularly in countries with high penetration rates of mobile technologies.
3. Increased similarity between firms in terms of offered services, due to adoption of common standards.
4. Decreased resource dependency between firms, due to availability of off-the-shelf technologies.

A clear example of vertical competition can be found when MNOs and TEVs compete for customers on IoT connectivity (Fig. 3c), where the TEV also plays the role of IoT communication provider. In this case the drivers for competition are no exception to the four factors mentioned earlier. Increased maturity can be seen in the "future telecom" market for enablers of IoT services. This maturity is leading the telecom industry toward overlapping goals between TEVs and MNOs. TEVs either possess or want to get access to resources that MNOs need to build their IoT business in order to provide connectivity for IoT. Therefore, a TEV may establish similar competitive goals with its current vertical cooperator, an MNO, thus inevitably leading to vertical coopetition.

Coopetition in the IoT for the Telecommunication Industry

In vertical coopetition, competition and cooperation are present in terms of *both/and*; it "represents a situation where it is not possible to choose between contradictory dualities" [6]; therefore, they happen simultaneously.

As a consequence of vertical competition with cooperators, if a customer like an MNO perceives its supplier as a competitor to get a new IoT customer, it may decide to change its supplier. This can be an immediate effect of "negative feelings" as a consequence of competitive relationship.

Surprisingly, observations show that in most cases firms stay in the vertical coopetitive relationship [7]}; the reasons include industrial drivers, relational drivers, and firm-specific drivers. The industrial drivers in the telecom industry relate first to maturity of the market. Even when the future telecom ecosystem includes service providers from other industries, the level of maturity is not going to decrease.

Relational drivers of vertical coopetition are straightforward. Vertical cooperation is a consequence of complementarities among actors. Relational benefits are generated when involved actors in the relationships achieve a high level of interaction toward a common goal in their business, a phenomenon present in vertical cooperation. The emergence of non-ICT service providers in the value network, the need for co-creation of value, and the transformation from value chains to value networks vouch for relational drivers.

The drivers for horizontal and vertical coopetition are not always homologous. However, while one common driver is direct economic benefits [8], these economic benefits do not share a common cause. On one hand, the high cost structure of deploying a specific product (e.g., cellular networks for IoT provisioning) gathers two competitors to cooperate in order to reduce costs. On the other hand, the high cost structure of implementing a service without the original supplier (i.e., changing supplier or developing in-house) keeps the competing cooperators in vertical coopetition. Hence, the economic benefits in vertical coopetition are generally the outcome of relational benefits.

In a future research outlook, a comprehensive tool that can accommodate the profitable perspective of all active partners of the IoT co-creation process is required, considering a new perspective on business models.

New Business Model View for the IoT Based on Value Networks

A business model can be generalized as a way to explain how firms create and capture value. Some definitions discuss business models as activity systems, including which activities are performed, how they are linked, and who performs them. Others include a value proposition, and how a firm is organized and positioned to create a profit potential, with the objective of customer-focused value creation. In general, a business model should express the logic of a firm; describing the value a company offers to customers and the network of partners creating and delivering this value to generate profit and sustainable revenues [9].

Even though existing business model tools consider the business relationships, they are typically focused on the value generated by individual firms. A proposal on *ecosystem business models* suggests that value creation in IoT needs active

involvement from all relevant actors. In fact, development of services based on connected devices requires diversity in resources and activities from a network of actors [4]. This perspective is an emerging trend to explicitly develop business models that leverage on vertical cooperation within and across industries, creating services based on the IoT.

In the IoT value networks, firms are interconnected and depend on each other, usually evolving around a specific core, which corresponds to shared and common assets [4]. Common assets could be present in the form of platforms, technologies, processes, and standards that are fundamental in their businesses. The role of business models in networks is to exploit the common assets and coordinate the value co-creation process in vertical cooperation, while ensuring profitable outcomes for the actors involved [10]. In Table 2, we compare the two perspectives.

The network-centric business model framework described in [9] is suitable for developing IoT services; it uses the business model concept to understand business planning in a value network. The framework contains three elements:

1. Business network development: Actors enter a value network with their individual business models and determine their role within the business network.
2. Business opportunity development: Initial business opportunities are identified based on firm-centric business models. They are further developed and exploited in a network-centric business model.
3. Business model development: This starts from technological possibilities and matures toward commercial exploitation that are collectively understood in the value network.

Further on, the framework considers a time dimension that begins with service development, moving to a pilot phase and ending in a market phase. For the pilot phase, entrepreneurial activities are highlighted as the expectation of having an actor steering the development and taking care of "identifying the business opportunities to be exploited and facilitating development of the networked business model" [9]. The emerging concept of business models in networks are in an early stage and do not count with dominant frameworks of development. However, as we present in the section, they implicitly occur in different IoT activities.

In future IoT research, it would be beneficial to develop solutions and services considering resources and capacities from cooperating firms, providing interfaces for technical and business resources.

Use Case: New Models for Smart Cities

ICT is transforming business models in different vertical sectors. The mobile ecosystem has also changed from a situation where business relationships were mostly bilateral to a situation of cooperative business networks of providers that combine value propositions.[1] New business models are emerging in the digital economy, radically transforming established industries. In this section we discuss how these value networks apply to specific cases of a smart city.

	Level of analysis	
Type of business model	Firm level	Network level
Firm-centric	Focus on how a firm creates value and exploits an opportunity.	Focus on business relationships and external factors that influence the business of a firm, as either suppliers, partners, or customers.
Network-centric	Focus on the position and role that one firm has inside a value network	Business model approach that considers the configuration of a value network creating a combined value proposition.

Table 2. Firm-centric and network-centric business models, based on [10].

Smart Cities

A common definition of a smart city is the use of ICT to sense, analyze, and integrate the key information of core systems in running cities in order to optimize existing services, while offering new possible services to citizens.

We consider that a smart city consists of five major building blocks [11]:

1. Economic, social, and privacy implications
2. Developing e-Government
3. ITS
4. H&WB
5. Digitally built environment

In order to make a city *smart*, two approaches exist [12]: top-down and bottom-up. The top-down approach uses an ICT system that has an overview of all urban activities as well as the tools to interact with infrastructures, gathers vast amounts of data, and adjusts parameters to optimize the city operations through technology. The bottom-up approach is about the citizens in the center of innovation; this approach, rather than working toward centralization, embraces a distributed approach.

A better solution is a proper mix of them. By means of communications, it is possible to enable the creation of smart solutions in the bottom-up approach that address identified and local needs, while top-down solutions provide general development frameworks to ensure interoperability of subsystems. These solutions enable non-ICT service providers to offer IoT services: a set of services where the value is co-created among ICT actors and non-ICT service providers. IoT in the context of the smart city covers a broad area including several industries. In order to narrow down the scope, we discuss ITS to bring examples of merging ecosystems in order to co-create value for end users and H&WB to highlight relevant motivations toward network-centric business models for the IoT.

Intelligent Transport Systems

ITS applies advanced ICT technologies to surface transportation in order to achieve enhanced safety and mobility while reducing the environmental impact of transportation. This approach requires a mix of traditional transport services together with ICT solutions in order to optimize transport services and offer over-the-top (OTT) solutions. In order to enable optimization or OTT services, the transport operators need to cooperate with ICT actors instead of implementing the solutions by themselves. A further step toward offering yet another service, enabled by this compilation of

[1] As presented by the 5G Infrastructure Association of the European 5G Public Private Partnership (PPP) in "5G Empowering Vertical Industries." The 5G-PPP is a partnership between the European Commission and most relevant telecom actors, including industry manufacturers, telecommunications operators, service providers, SMEs and researchers..

Services based on IoT and connected devices are beginning to play a key role in preventive healthcare as they can provide remote connectivity and access to relevant patient information. These services can also support healthcare workers by disseminating clinical updates, schedules, learning material, and reminders.

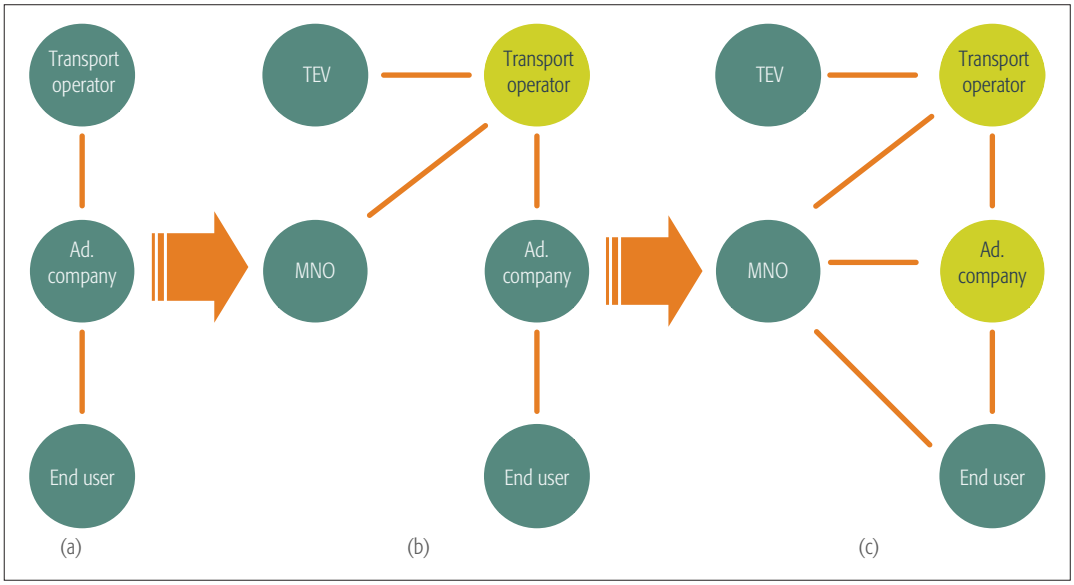


Figure 4. Gradual integration of transport and ICT value chains, which creates a new value network.

actors, is also possible. An example is utilizing bus stops as connectivity hotspots for MNO subscribers in a newly created value network.

Figure 4 illustrates the case where the transport value chain, step-by-step, integrates with the ICT value chain and creates a new value network. In the primary transport value chain, the transport operator interacts with the passengers through the advertisement company that owns the bus stations (Fig. 4a). In the next step, an MNO connects the buses and trains of the transport operator via IoT-based SIM cards, while TEV enables connectivity and manages the vehicles via an IoT connectivity platform. This process allows transport operators to offer optimized transport (as an IoT service) and multimodal route planning (as an OTT IoT service) to the passengers (Fig. 4b). Meanwhile, the passengers have an indirect relationship with the advertisement company that owns the bus stops. The presence of this advertisement company, as part of the traditional transport value chain, enables another co-created value for passengers, that is, cellular/WiFi connectivity hotspots in bus stations (Fig. 4c).

Healthcare and Well Being

Services based on IoT and connected devices are beginning to play a key role in preventive healthcare as they can provide remote connectivity and access to relevant patient information. These services can also support healthcare workers by disseminating clinical updates, schedules, learning material, and reminders [13]. Pilot projects have been useful to demonstrate the potential; however, the market implementation has been limited to date [14]. This is mainly due to lack of evidence showing the value in terms of cost and health outcomes. The lack of standardized developments also limits the assessment of the usability and value of some solutions. Moreover, isolated developments create business barriers [14]; for this reason, the development strategies should consider the integration with existing health system functions to complement objectives of current systems.

From the perspective of the telecommunication industry, MNOs and communication providers are in a position to support the needs of the healthcare industry by providing remote monitoring and access to medical data. The main challenges are related to generation of an integrated value proposition and the roles that communication providers could take. These roles can span from *just* connectivity providers to system integrators. In recent years, we have seen how incumbent MNOs are creating separate units to target the health sector. In this case, they take an active role in the development of solutions in combination with a network of service suppliers that coordinate and integrate their efforts.

The pivotal role of integrators is also crucial for home care solutions, where there is a complex environment of actors that are subject to intense collaboration. This complexity often results in unclear and unbalanced distribution of costs and benefits, since often most of the benefits are not received by the actors making the largest portion of the investment. This is a factor that slows the adoption of services, which could be overcome by developing business models from a network perspective, ensuring a profitable outcome for all involved actors.

The 5G Infrastructure Association proposes systems' integration for business models to transform healthcare delivery by ICT. The suggestion consists of having flexible strategies that allow each actor to focus on its core competences. For this, the key is to provide interfaces between business roles. The main point is that the association recognizes the importance of service development in a multi-stakeholder environment.[2]

The design of business models for healthcare services has also received attention from multidisciplinary groups that emphasize the importance of cooperation between different actors in the service design and provision. Collaboration is essential even at the stage of defining requirements, where all key actors must be involved [15]. These actors are the intermediary coordinator connecting companies, the dedicated technolo-

[2] The 5G Infrastructure Association has provided a framework for a common setup to develop and discuss roles and relationships, particularly among technology providers. The current framework is available in a white paper on the "eHealth Vertical Sector," which is available online: https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-White-Paper-on-eHealth-Vertical-Sector.pdf

gy specialists, and the healthcare specialists; they have to set business relationships to exchange value in any service [15].

Conclusions

When discussing business opportunities in the IoT, firms need to collaborate and be aware of novel network-centric business models. In this process, the key concern is how each firm can position itself within the business network in a way that guarantees its profitability as part of a larger group, not only a single firm. The challenge is that firms need to accept these new market rules where they will individually perceive less control over the final customers and the entire value proposition.

In this article, we have used the *smart city* as an example of context where ICT plays the role of enabling IoT in two representative use cases: ITS and H&WB. In both cases, the gradual integration of non-ICT industries with traditional ICT value chains are forming new emerging value networks that define how value can be created in the ecosystem. The emergence of value networks shows us the importance of co-creating value together with involved entities in the network. The value networks then change the typical structure of value creation in value chains. The change is in shifting from a unilateral relationship among suppliers and consumers toward vertical cooperation.

In the future, new tools to define these new ways of interaction among companies will need to be designed. Companies will need to define their own business models, but in connection with the value network. They will need to understand how to define a business model that is aligned with the value created by the value network in the IoT ecosystem. Understanding such complex business interactions is fundamental, and this article aims to make the ICT technical community aware of the need to pursue stronger interaction between ICT experts, vertical players, and final users, all interacting together in the IoT ecosystem.

Acknowledgments

This work has been partially funded by the research projects CellFive (TEC2014-60130-P), 2014-SGR-1551, and the Vinnova (Swedish innovation agency) IoT ecosystem Project.

References

[1] M. E. Raynor and M. J. Cotteleer, "The More Things Change: Value Creation, Value Capture, and the Internet of Things," *Deloitte Review*, no. 17, pp. 49–65, 2015, https://dupress.deloitte.com/content/dam/dup-us-en/articles/value-creation-value-capture-internet-of-things/DUP1199_DR17_TheMoreThingsChange.pdf.

[2] M. Bengtsson and S. Kock, "Cooperation and Competition in Relationships Between Competitors in Business Networks," *J. Business & Industrial Marketing*, vol. 14, no. 3, 1999, pp. 178–94.

[3] R. C. Basole, "Visualization of Interfirm Relations in A Converging Mobile Ecosystem," *J. Info. Tech.*, vol. 24, no. 2, 2009, pp. 144–59.

[4] S. Leminen, M. Rajahonka, and M. Westerlund, "Ecosystem Business Models for the Internet of Things," Jan., pp. 10–13, 2015.

[5] Y. Luo, "A Coopetition Perspective of Global Competition," *J. World Business*, vol. 42, no. 2, 2007, pp. 129–44.

[6] R. Quinn and K. Cameron, "Paradox and Transformation: Toward a Theory of Change in Organization and Management," Ballinger Series on Innovation and Organizational Change, 1988.

[7] T. Raza-Ullah, M. Bengtsson, and S. Kock, "The Coopetition Paradox and Tension in Coopetition at Multiple Levels," *Industrial Marketing Management*, Special Issue on Co-opetition Cooperation and Competition, vol. 43, no. 2, 2014, pp. 189–98.

[8] S. Lacoste, "Vertical Coopetition: The Key Account Perspective," *Industrial Marketing Management*, vol. 41, no. 4, 2012, Green Marketing and Its Impact on Supply Chain, pp. 649–58.

[9] T. Palo and J. Tähtinen, "Networked Business Model Development for Emerging Technology-Based Services," *Industrial Marketing Management*, vol. 42, no. 5, 2013, pp. 773–82.

[10] L. Bankvall, A. Dubois, and F. Lind, "Conceptualizing Business Models in Industrial Networks," *Industrial Marketing Management*, 30 Apr. 2016; http://dx.doi.org/10.1016/j.indmarman.2016.04.006.

[11] L. M. Correia and K. Wünstel, "Smart Cities Applications and Requirements," *Net!Works Technological Platform White Paper*, 2011.

[12] N. Walravens, J. Breuer, and P. Ballon, "Open Data as a Catalyst for the Smart City as a Local Innovation Platform," *Communications & Strategies*, no. 96, 2014, p. 15.

[13] K. Källander *et al.*, "Mobile Health (Mhealth) Approaches and Lessons for Increased Performance and Retention of Community Health Workers in Low- and Middle-Income Countries: A Review," *J. Medical Internet Research*, vol. 15, no. 1, 2013.

[14] A. B. Labrique *et al.*, "mHealth Innovations as Health System Strengthening Tools: 12 Common Applications and a Visual Framework," *Global Health, Science and Practice*, vol. 1, no. 2, 2013, pp. 160–71; http://www.ghspjournal.org/content/1/2/160.full.

[15] D. P. van Meeuwen, Q. J. van Walt Meijer, and L. W. Simonse, "Care Models of eHealth Services: A Case Study on the Design of a Business Model for an Online Precare Service," *JMIR Research Protocols*, vol. 4, no. 1, 2015, p. e32.

Biographies

Amirhossein Ghanbari is a strategy management consultant in the telecom industry at Northstream AB, Stockholm, Sweden. He received his M.Sc. degree in electrical engineering and Licentiate degree in ICT techno-economics from KTH Royal Institute of Technology, Stockholm. While at KTH, he was involved in projects with Huawei, Ericsson, Nokia, Samsung, Deutsche Telekom, Orange, and Siemens in the area of the Internet of Things. His research interests are the economics and business relationships in wireless communications, and IoT in specific.

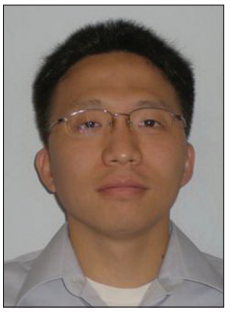
Andres Laya [S'12] is a Ph.D. student at the Communication System Department of KTH. He holds an M.Sc. degree in ICT from BarcelonaTECH (Universitat Politecnica de Catalunya, UPC), Spain. He has been involved in projects with Ericsson, Nokia, Orange, Telecom Italia, Sony Mobile, and Aalto University in the area of machine type communications. His research interests are in the area of the Internet of Things, and the business implications of connected devices in different industries.

Jesus Alonso-Zarate [SM'13] received his Ph.D. (2009) and M.B.A. (2016) from UPC). He is a senior researcher, head of the M2M Communications Department, and manager of the Communications Technologies Division at CTTC, Barcelona. Since 2010, he has published more than 150 peer-reviewed scientific papers in the area of M2M communications. He is a recipient of various best paper awards and is very active in internationally collaborative R&D projects (with industry, ESA, and H2020).

Jan Markendahl received his Ph.D. in techno-economics from KTH in 2011, where he was appointed associate professor in wireless infrastructure deployment and economics in 2012. He worked more than 20 years in R&D, business development, management, and marketing at Ericsson, Nokia Networks, Telia, and consultancy companies. His research interests include mobile networks and services, the Internet of Things, business modeling, and cost structure modeling and analysis. He is currently responsible for the VINNOVA project "IoT Ecosystems."

Understanding such complex business interactions is fundamental, and this paper aims at making the ICT technical community aware of such need of pursuing a stronger interaction between ICT experts, vertical players, and final users; all interacting together in the IoT ecosystem.

ADVANCES IN OPTICAL COMMUNICATIONS TECHNOLOGIES

Xiang Liu Zuqing Zhu

Optical communications and networking technologies continue to play an important role in realizing cost-effective interconnection of a wide variety of resources over highly distributed network environments for massive data exchange and processing. In 2016, with momentum gained from physical network elements virtualization, software-defined networking (SDN), and network functions virtualization (NFV), optical communication networks are becoming more flexible, programmable, and application-aware to enable service providers to deliver shortened time to market, and elastic and cost-effective services and solutions. Following this trend, we expect to witness continued global expansion of optical and data center networks, and faster convergence of optical and IT infrastructures in 2017.

In this first Optical Communications Series (OCS) issue of 2017, we have selected three contributions that address laser diode-based visible light communications (LD-VLC) systems, network hardware virtualization in core networks, and reconfigurable add-drop multiplexer (ROADM) contention in optical networks.

In the first contribution entitled "Laser Diode-Based Visible Light Communication: Toward Gigabit Class Communication," F. Zafar, M. Bakaul, and R. Parthiban present an overview of LD-VLC systems. VLC systems based on light-emitting diodes (LEDs) have been proposed to address the current limitations in capacity and RF spectrum availability in wireless communications. Although LEDs have been considered for their superior switching capabilities compared to other light sources, their modulation bandwidth, typically in the range of 20 MHz, is insufficient to support data rates over 1 Gb/s. In this contribution, laser diodes (LDs) that can support much higher direct modulation rates and provide higher electrical-to-optical conversion efficiencies are proposed as better-performing front-end transmitters to realize high data rate VLC systems. However, despite their superior performance, the usage of LDs can be restricted by their cost and potential health hazards. The authors provide a detailed overview of the advantages of LDs over LEDs and discuss various configurations of LD-VLC systems. They summarize the data rates achieved by the LD-VLC systems and the illumination capabilities for commercial implementations. The market opportunities and potential applications of LD-VLC systems are also discussed along with such challenges as the presence of speckles, power limitations, thermal management, and cost.

In the second contribution, "Network Hardware Virtualization for Application Provisioning in Core Networks" A. Gumaste, T. Das, K. Khandwala, and I. Monga present their summary of the trends and developments in service providers' transition to virtualized core networks. Service providers and their suppliers have been seeking approaches to realize faster and more efficient service provisioning to meet the needs and requirements of their customers in a timely and cost-effective manner. There have been consensus and progress toward achieving faster service provisioning by replacing physical network elements with programmable hardware that has non-proprietary interfaces. Such a programmable network would have lent itself to satisfy user requirements precisely and permit much faster provisioning of next generation services to meet users' evolving business needs. In this contribution, the authors outline the impact of virtualization in core networks for growing service provider businesses. The article provides an outline of a decision map for the applications with network virtualization technologies. It also presents an approach to integrate the services of over-the-top (OTT) content providers with those of traditional service providers using network virtualization in hardware. The authors describe the concept of virtualized network equipment partitioning and propose the policies for each service need/requirement.

In the third contribution, "A Closer Look at ROADM Contention," J. M. Simmons presents an overview of the contention issue when using reconfigurable optical add/drop multiplexers (ROADMs) in optical transport networks. ROADMs provide wavelength switching capability in the optical transmission layer of transport networks. As the needs for rapid service provisioning grow, service providers use ROADMs that enable greater flexibility to configure/reconfigure their networks. While the global industry has been pursuing the development of ROADMs for colorless, directionless, gridless, and contentionless operation, it has been proven to be challenging to ensure contentionless operation since ROADM contention appears in various forms. Thus, this contribution is aimed at examining the ROADM contention issue. It focuses on wavelength contention, which is defined as the unavailability of a wavelength on a network fiber for use by a ROADM port. As described in the article, wavelength contention can occur due to a limited number of add/drop ports, limited edge configurability, and/or limited pre-deployment equipment. Along with the conditions under which wavelength contention can occur, network architectures to avoid contention and to ensure contentionless operation are also proposed.

This OCS issue marks the beginning of our service term as the new Series Editor Team. We thank the former Series Editor, Dr. Admela Jukan, for her service. With the continuing support from our authors and reviewers, and valuable feedback from our readers, the Optical Communications Series is expected to continue contributing to future advancements in the field of optical communications, networks, and applications.

BIOGRAPHIES

XIANG LIU [F'17] (xiang.liu@huawei.com) received his Ph.D. degree in applied physics from Cornell University in 2000. He is currently the senior director of Optical Access Networks Research at the U.S. R&D Center of Huawei Technologies, focusing on next-generation optical access technologies. He spent the early part of his career at Bell Laboratories in New Jersey, working on high-speed optical fiber transport technologies. He is a Fellow of the OSA and a Deputy Editor of *Optics Express*.

ZUQING ZHU [SM'12] (zqzhu@ieee.org) received his Ph.D. degree from the University of California, Davis, in 2007. He is currently a full professor at the University of Science and Technology of China. Prior to that, he worked in the Service Provider Technology Group of Cisco Systems, San Jose, California. His research focuses on optical networks, and he has received Best Paper Awards from IEEE ICC 2013, IEEE GLOBECOM 2013, IEEE ICNC 2014, and IEEE ICC 2015.

CALL FOR PAPERS

IEEE COMMUNICATIONS MAGAZINE

HUMAN-DRIVEN EDGE COMPUTING AND COMMUNICATION

BACKGROUND

The vision of edge Computing considers that tasks are not exclusively allocated on centralized Cloud platforms, but are distributed towards the edge of the network (as in the Internet-of-Things and Fog Computing paradigms), and transferred closer to the business operations via the Content Delivery Networks. The traditional gateway becomes a set-top-box machine, with additional computation and storage capabilities, where micro tasks can be offloaded first, instead of directly to the Cloud. Mobile Edge Computing can also be a more suitable approach to extract knowledge also from privacy sensitive data, which are not to be transferred to third party entities (global cloud operators) for processing. The proliferation of the networking connectivity and the progressive miniaturization of the computing devices have paved the way to the sensor networks and their success in the automation of the several monitoring & control applications. Such networks are built in an ad hoc manner and deployed in an unsupervised manner, without an a-priori design. The consequent availability of long-range communication means at certain nodes of those networks has enabled the possibility of the Internet connection of the sensor network, to make use of cloud-based services.

The new challenge addressed by this Feature Topic (FT) is how to put users in the loop so that they can retake control of their information. The massive proliferation of personal computing devices is opening new human-centered designs that blur the boundaries between man and machine.

In addition, Edge services are also used to exchange the data collected and processed within the context of the IoT towards external services and/or to visualize them through traditional browser by the users. Now, the frontier for the research on the data management is related to the so-called Edge Computation and Communication, consisting of an architecture of one or more collaborative multitude of computing nodes that are placed between the sensor networks and the cloud-based services. Such a mediating level is responsible for carrying out a substantial amount of data storage and processing to reduce the retrieval time and have more control over the data with respect to the Cloud-based services and to consume less resources and energy to reduce the workload. The interdependencies among those three different levels of storage and computing within an IoT solution are complex and determining at which data should be collocated and elaborated is demanding but not simple to handle. Such a complex situation is further exacerbated if we consider to achieve Quality-of-Service goals such as reliability, availability, security, mobility and energy efficiency, without compromising the correct behavior of the system and the service duration of the devices batteries. Moreover, the interconnection between the sensor networks and the upper level is not simple to be supported, in fact, falls within those situations where traditional Internet architectures fail to provide it effectively. This is because the sensor networks are deployed on hostile and challenging environments implying intermittent connectivity, a heterogeneous mix of nodes, frequent nodal churn, and widely varying network conditions.

The analysis of human activity and their interactions with physical and digital artefacts will also be extremely useful for closing the control loop of adaptive distributed systems. This may open a new research playground for distributed systems that adapt to user behaviours in different contexts, moving more and more to the network edge through devices such as the 5th Generation mobile networks or 5th Generation wireless systems. The second aspect of the frontier of the current research is therefore related to the application of challenging networking solutions to support the Fog Communication and Computation in the Internet of Things.

The aim of this FT is to solicit novel contributions to the current debate on realizing the Edge Computing perspective to the Cloud platforms and Internet of Things by focusing on the human-driven resource management, challenging networking aspects and communication issues, by also seeking practical experiences in using these intelligent solutions in concrete use cases.

Topics of interest include, but are not limited to:

- Novel models and architectures of Edge-centric computing
- Fog-to-Cloud integration and protocols
- Communication protocols and issues
- Crowdsensing and crowdsourcing information
- Human-driven design and implementation of edge computing
- Novel socially-informed architectures
- Delay-tolerant networks, opportunistic communication and computing
- Reliability and availability, mobility and connectivity in edge-centric computing
- User-guided management of Fog systems and services
- Resource management and provision
- Data harvesting and analytics in challenged networking
- Information centric and content-centric networking
- Distributed storage services
- Heterogeneity of edge systems
- Energy-efficient communication and computation
- Security and privacy, attacks and resiliency
- Secure and sensitivity-aware applications
- Novel safe methods for including humans in the data-analysis loop
- QoS-aware communication protocols
- Daily use applications and programming models
- Test and simulation tools for evaluating challenged systems
- Modelling and simulations

SUBMISSIONS

Articles should be tutorial in nature, with the intended audience being all members of the communications technology community. They should be written in a style comprehensible to readers outside the specialty of the article. Mathematical equations should not be used (in justified cases up to three simple equations are allowed). Articles should not exceed 4500 words (from introduction through conclusions). Figures and tables should be limited to a combined total of six. The number of references is recommended not to exceed 15. In some rare cases, more mathematical equations, figures, and tables may be allowed if well-justified. In general, however, mathematics should be avoided; instead, references to papers containing the relevant mathematics should be provided. Complete guidelines for preparation of the manuscripts are posted at http://www.comsoc.org/commag/paper-submission-guidelines. Please submit a PDF (preferred) or MS WORD-formatted manuscript via Manuscript Central (http://mc.manuscriptcentral.com/commag-ieee). Register or log in, and go to the Author Center. Follow the instructions there. Select "November 2017/ Human-Driven Edge Computing and Communication" as the Feature Topic category for your submission.

IMPORTANT DATES

- Manuscript Submissions Due: April 1, 2017
- Decision Notification: July 1, 2017
- Final Manuscripts Due: August 15, 2017
- Publication: November 2017

GUEST EDITORS

Florin Pop
University Politehnica of Bucharest, Romania
florin.pop@cs.pub.ro

Aniello Castiglione
University of Salerno, Italy
castiglione@ieee.org

Jiannong Cao
The Hong Kong Polytechnic University, Hong Kong
csjcao@comp.polyu.edu.hk

Giovanni Motta
Google Inc., USA
giovannimotta@google.com

Yang Yanjiang
Huawei Singapore Research Centre, Singapore
yang.yanjiang@huawei.com

Wanlei Zhou
Deakin University, Australia
wanlei.zhou@deakin.edu.au

Laser-Diode-Based Visible Light Communication: Toward Gigabit Class Communication

Fahad Zafar, Masuduzzaman Bakaul, and Rajendran Parthiban

The authors discuss the principles of LD-based VLC systems, highlighting the operational characteristics and link configurations. The unique features of LDs compared to LEDs are described alongside the communication and illumination aspects of different classifications of LD-VLC. The challenges in terms of practical implementation and the potential applications that might arise from this communication technology are also discussed.

ABSTRACT

VLC is an emerging optical wireless communication technology that can be added as a complementary feature into existing lighting infrastructures for alleviating pressure on the rapidly dwindling radio frequency spectrum. Although LEDs have been traditionally used as transmitters in VLC, the growing urgency for higher data rates in the gigabit class range has deviated focus toward the consideration of LDs as potential sources for VLC due to their unique features of high modulation bandwidth, efficiency, and beam convergence. This article focuses on the principles of LD-based VLC systems, highlighting the operational characteristics and link configurations. The unique features of LDs compared to LEDs are discussed alongside the communication and illumination aspects of different classifications of LD-VLC. The challenges in terms of practical implementation and the potential applications that might arise from this communication technology are also discussed.

INTRODUCTION

In the last two decades, the unprecedented growth of wireless communication systems along with their increasing demands on wireless data traffic is restraining the availability of the limited and expensive RF spectrum, which is driving the need for complementary wireless transmission techniques. This gave rise to visible light communication (VLC), an emerging technology in the field of wireless communication that concurrently provides communication alongside illumination and can be incorporated into existing lighting infrastructures as a complementary functionality. VLC utilizes a light source as its transmitter where information is modulated into the intensity of the emitted light. The frequency of the modulating signal is kept high enough (typically above 300 Hz) so that occupants assume the source to be normally lit. A block diagram depicting the operating principle of a general VLC system is presented in Fig. 1.

Because of the growing trend toward replacement of traditional lighting infrastructures with solid state lighting (SSL), most VLC systems to date have considered light emitting diodes (LEDs) as transmitters because of their superior switching capabilities compared to traditional incandescent and fluorescent sources. VLC systems utilizing LEDs generally use a single-chip or multi-chip approach, which assign phosphor-coated LEDs (pc-LEDs) or red-green-blue (RGB) LEDs as sources, respectively. However, for communication at much higher speeds (gigabit class range), the modulation bandwidth of these LEDs ranging around 20 MHz is still considered quite low. Smaller area LEDs can be modulated at higher frequencies, which is attributed to their ability to be driven at higher current densities. Techniques including the use of micro LED (μ-LED) demonstrated significant rise in modulation bandwidth allowing data rates of $\sim$ 3 Gb/s. However, at high current densities, LEDs suffer from *efficiency droop*, which occurs due to electron overflow, and hence the maximum radiant flux specified for an LED may not be at the optimal point of efficiency. Efficiency droop limits drive currents, leading to a higher initial cost per lumen of LEDs, which undermines one of the fundamental features of economical solid-state lighting.

In contrast to LEDs, laser diodes (LDs) have much higher direct modulation speed and have the highest electrical-to-optical conversion efficiency without the droop issue, making them a better candidate for front-end transmitters when communication is required at much higher speeds. In the case of point-to-point transmission associated with static VLC applications, LD is expected to provide better performance than LED because of the characteristics of high optical power and light beam convergence. Despite numerous advantages, the use of LDs for VLC is still uncertain due to high cost, health hazard issues, color mixing complexity and efficient homogeneous illumination.

This article presents the latest concepts and methodologies involving the newly emerging LD-based VLC systems. It gives a detailed overview of the advantages of LDs compared to LEDs and different configurations of LD-VLC systems. It summarizes the data rates achieved by these systems alongside the illumination aspects requiring attention for commercial implementation. The market prospects of LD-VLC and the promising applications that might arise from the use of this technology are also discussed.

Digital Object Identifier:
10.1109/MCOM.2017.1500672CM

The authors are with Monash University Malaysia.

Laser Diode Overview

An LD is an electrically pumped semiconductor device that produces coherent radiation in the visible or infrared (IR) spectrum when current passes through it. In LDs, an effective laser resonance is stimulated due to the presence of coated or uncoated end facets that behave like mirrors with different reflectivities, resulting in an eventual gain in stimulated emission of highly directional photons. This allows precise control over the emission profile of the transmitter front-end, which cannot be matched by any alternative technology for wireless connectivity such as RF waves, millimeter-waves, or terahertz waves, allowing improvement in link directionality, coverage uniformity, interference management, and high data rate densities by beamforming. LD-VLC can operate over a huge amount of unlicensed spectrum ($\sim$ 400 THz) compared to RF waves traditionally used for communication ($\sim$ 300 GHz), although there is uncertainty regarding the extent to which the available spectrum can be utilized, and expensive components are required for exploiting the carrier frequency of light. However, compared to RF communication, LD-VLC provides much higher security and does not induce electromagnetic interference (EMI), making them ideal for low-coverage indoor wireless communication systems.

In comparison, an LED operates with spontaneous incoherent emission of radiation where the emitted photons are diffused over a larger area by lenses. Thus, laser beam steering and focusing is much easier than focusing an LED output [1]. The emission profile of LD is on the order of 2–3 nm, which favors efficient utilization of the available visible spectrum by wavelength-division multiplexing (WDM) and allows the use of narrow optical filter for greater rejection of ambient noise [2, 3].

The modulation speed in the case of LDs is much higher because it is controlled by the photon lifetime (on the order of picoseconds), which is much shorter than material carrier lifetime in the case of LEDs. To generate enough lumens, a commercial LED bulb must either contain a small number of LEDs operated at high power density (increasing the cost of additional heat sinking due to less efficient chips) or many LEDs operated at low power density (increasing the cost of additional processing and packaging of semiconductor material). One option for saving the cost per lumen is utilizing LDs (v. 2). Although the peak wall plug efficiency for state-of-the-art blue LDs is lower (38 percent) than for blue LEDs (> 70 percent), the LD peak efficiency occurs at an input power density ($\sim$ 25 kW/cm^2) that is over three orders of magnitude higher than where the LED peak efficiency occurs ($\sim$ 3 W/cm^2). Thus, the brightness per epitaxial area is tremendously enhanced for LDs [4]. Commercial IR lasers are already considered to be the most efficient converters of electrical-to-optical energy. If similarly efficient lasers could be developed at visible wavelengths, economical solid-state lighting could enter the "ultra-efficient" domain, indicating efficiency on the order of 70 percent relative to that of an optimal multi-component white light source [5].

Traditionally in telecommunications, LDs have been used in short-haul optical fiber communication as data transmitters. LD-VLC systems with plastic optical fiber (POF) backbone have been demonstrated in recent works that have shown data rates of up to 1.25 Gb/s [6, 7]. Usually, laser beam coupling into an optical fiber is much more efficient than LED beam coupling, with the latter requiring thicker fibers. This opens up scopes for implementing VLC with fiber backbone networks where LDs can be used as sources for both data transmission and illumination.

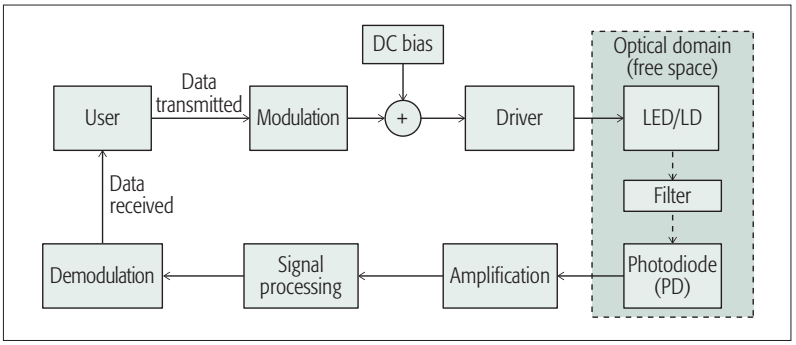


Figure 1. A block diagram of a general IM/DD-based VLC system. The data that needs to be transmitted by the user is modulated, and a DC bias is added to make sure that the signal is in the operating region of the source, which can be either an LED or an LD. The source converts the electrical signal into an intensity modulated optical signal, which then travels across a free space optical channel, and is passed through an optical filter and detected by the photodiode. The photodiode generates an equivalent electrical signal, which is amplified and demodulated to retrieve the original transmitted data.

Illumination Aspects of LD-VLC

The mechanism of white light generation in the case of LDs is similar to LEDs, which is demonstrated in Fig. 3. The first approach involves the combination of a blue LD with a yellow phosphor. Laser-based VLC will suffer less from the phosphor response than LED systems due to its higher power operation at high current density and larger inherent bandwidth than LEDs. Denault *et al.* demonstrated 442 nm laser-based white lighting using YAG:Ce phosphor with a luminous efficacy of 76 lm/W [8]. The LD-VLC system demonstrated by Chun *et al.*, utilizing remote phosphor technique, had a correlated color temperature (CCT) of around 7000 K, which is still considered to be much "cooler" for general lighting [9]. In addition to high efficiency and low cost per lumen of light produced, blue laser pumped phosphor would have advantages in terms of facile spatial controllability of the blue light accompanied by more options for placement of remote phosphors [5]. However, white light generation through remote phosphor and diffuser has a negative impact on the communication capacity of an LD-VLC system. A recent work utilizing phosphorous diffuser diverged blue LD demonstrated that the throughput response is degraded by 24 dB due to high reflection, scattering, and absorption of the laser beam when passing through the phosphorous diffuser film [10].

The second approach involves the mixing of LDs operating at different wavelengths to generate white light. The use of color mixed laser light would enable in situ tailoring to human or economic preferences of chromaticity, color rendering quality, or health. After being emitted from a source, light reflects off the surface of objects.

Proper design of phosphor film and diffuser optics to mitigate the loss in throughput is a key area requiring attention. The challenges associated with increased color temperature also need to be addressed to enable the widespread acceptance of LD-based standard lighting systems.

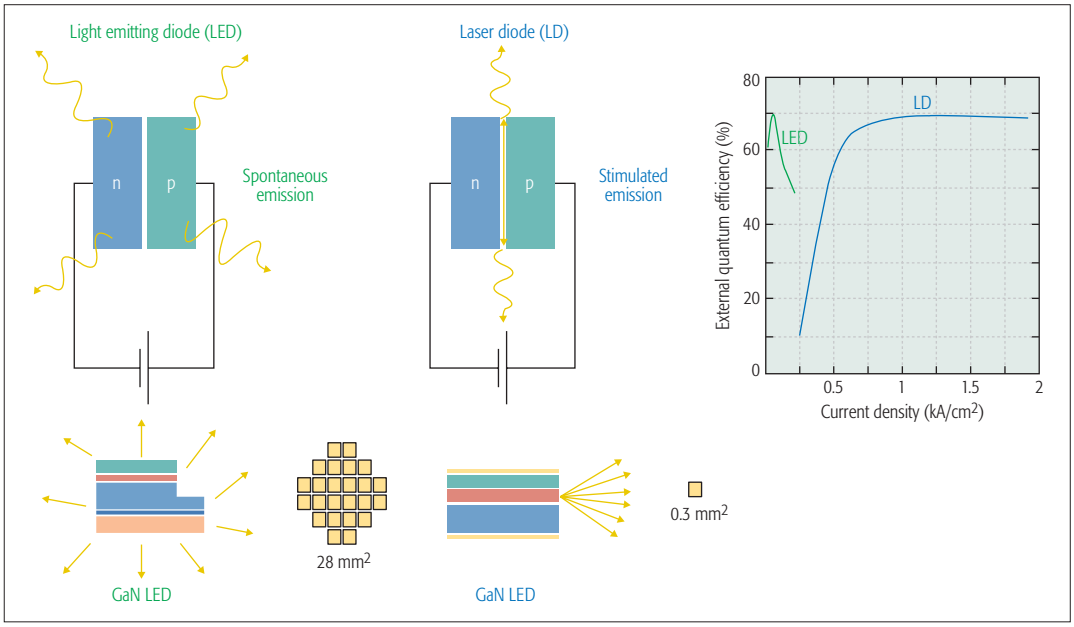


Figure 2. The structural difference between GaN-based LEDs and LDs is presented. LEDs typically emit light by spontaneous emission generated by radiative recombination of electrons and holes in the p-n junction, which emits photons that are out of phase (incoherent). LDs, on the other hand, emit light by stimulated emission of photons that are in phase (coherent). The typical areas of illumination of commercial LEDs and LDs are also illustrated. The relative distribution of external quantum efficiency of LEDs and LDs as a function of current density clearly demonstrates the efficiency droop that the LEDs suffer at higher current density.

If objects reflected only narrow spectral bands, sources with narrow emission profiles would risk not emitting any light that could be reflected. However, the reflectance spectra of virtually all objects are broad, smooth, and continuous. Therefore, some light will be differentially reflected off these object surfaces provided that the narrowband light sources are not too widely spaced in wavelength. The reflected light then travels to and is absorbed by the three cone photopigments in the human eye. The culminated excitation of all three cones from this reflected light causes perception of light color. It does not matter whether the light is evenly distributed over all, or concentrated in a narrow set of wavelengths within a cone's photopigment range. Thus, spectrally discontinuous sources (LDs) can be analogous to spectrally continuous light sources in terms of their effect on the human visual system. A study by Neumann *et al*. demonstrated that diffused laser light does not compromise the user experience compared to conventional light luminaires, and pioneered the work toward LD-VLC [5]. Tsonev *et al*. recently demonstrated an RGB LD-VLC system that achieved CCT ~ 8000 K; it was suggested that warmer light could be achieved by increasing the red and green components. Another work on RGB LDs by Janjua *et al*. highlighted that blue LD bias condition needed to be compromised in order to maintain a good color temperature (CCT ~ 5800 K) compared to encoding with blue LD, which causes much cooler diffused light (CCT ~ 12,000 K) [11].

A low divergent laser beam is remarkably easier to control (focus, steer, and mix) than Lambertian light from LEDs, which would facilitate new system-level lighting architectures. Glare minimization from reflective surfaces can also be achieved by implementing polarization control, which is an exclusive feature for LD-based light sources. Proper design of phosphor film and diffuser optics to mitigate the loss in throughput is a key area requiring attention. The challenges associated with increased color temperature also need to be addressed for enabling the widespread acceptance of LD-based standard lighting systems.

Communication Aspects of LD-VLC

VLC links utilizing LDs generally use intensity modulation (IM) to encode data at the transmitter end and direct detection (DD) for conversion of an optically modulated signal into an electrical signal at the receiver end with single- or multi-source link configurations. The IM/DD channel can be modeled as a baseband linear system. The channel model for LD-VLC is similar to the IR channel, although the reflectivity of the surfaces enclosing the system differs causing changes in delay spread,[1] which is a key parameter for diffuse link configurations. However, it is worth noting that the narrow beam direction yields a less dispersive channel in time due to the absence of multipath. Different single- and multi-carrier modulation schemes are generally used to encode information that needs to be transmitted. In multi-carrier systems, the delay spread is significantly shorter than the symbol duration, which simplifies the equalization process to single-tap equalization. Inter-symbol interference (ISI) and inter-carrier interference (ICI) are completely eliminated by using a larger number of subcarriers and a cyclic prefix (CP) of length equal to or longer than the maximum excess delay of the channel. The CP is an additional overhead, which generally results in some loss of spectral efficiency. However, since the delay spread in VLC is much less than in RF

[1] Delay spread is a measure of the multipath richness of a communication channel and is a critical performance criterion for determining the upper bound of transmission data rate.

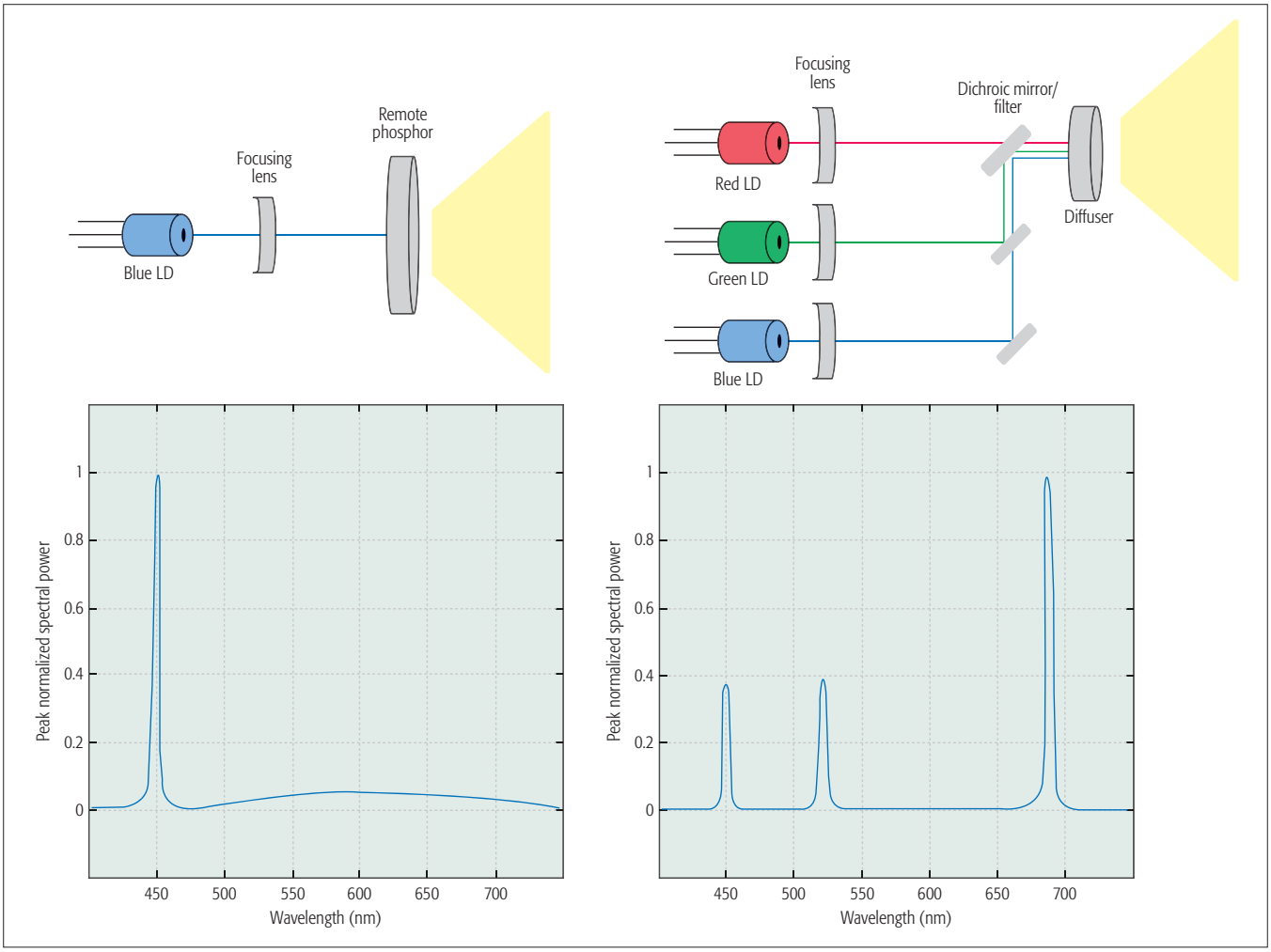


Figure 3. White light generation by utilizing the single source approach (left) and the multi-source approach (right). The relative normalized power spectral distribution for each configuration is also presented.

systems, the expected spectral efficiency loss is insignificant.

The work done on single-source-based LD-VLC systems mainly focuses on GaN-based blue LDs. Watson *et al.* demonstrated error-free data transmission at 2.5 Gb/s using blue LD (3 dB modulation bandwidth of 1.4 GHz) and non-return-to-zero on-off keying (NRZ-OOK) modulation scheme, which was further extended to 4 Gb/s by utilizing an ultra-violet (UV) extended high-speed photodiode (PD) [6, 12].

Higher data rates can be achieved by utilizing more spectrally efficient modulation schemes like orthogonal frequency-division multiplexing (OFDM). Since an LD-VLC system uses IM/DD, the information signal needs to be both real and non-negative. Traditionally, the OFDM signal is complex and bipolar, and thus needs to be modified for use in VLC. A bipolar OFDM signal can be used to modulate an LD if a suitable positive operating point is set for the LD around which the bipolar signal can be realized; this scheme is referred to as DC-biased optical OFDM (DCO-OFDM). By implementing DCO-OFDM, Chi *et al.* enhanced the modulation bandwidth of LDs from 900 MHz to 1.5 GHz, enabling transmission at 9 Gb/s with a CP duration of $\sim$ 0.7 ns. However, the dual functionality of illumination was overlooked in this work but was recently taken into account by Chun *et al.* for a 6.52 Gb/s LD-VLC system operating at 1000 lux [9].

Combining different LDs operating at distinct wavelengths opens the scope for utilizing WDM and color controllability while increasing cost and complexity. Janjua *et al.* demonstrated an RGB LD-VLC system where red LD was used for communication, and blue and green LDs for illumination. By implementing 16-quadrature amplitude modulation (QAM) OFDM with a CP duration of $\sim$ 0.7 ns, a data rate of 4 Gb/s was achieved [11]. Tsonev *et al.* demonstrated a 4 Gb/s VLC system at a distance of 2.88 m utilizing RGB LDs and OFDM modulation [2]. It was suggested that a data rate of 100 Gb/s is achievable by using WDM with 36 parallel channels. Based on simulations, Hussein *et al.* demonstrated OOK-based LD-VLC systems capable of achieving 5 Gb/s data rate by utilizing imaging receivers. By adding relay assistance to the receivers proposed, the data rate was further increased to 10 Gb/s [7].

The choice of coherent and incoherent modulation schemes like QAM and OOK, respectively, depends on the specification of the VLC system in terms of available resources. The OOK scheme has an inherent transmission capacity limitation due to lower spectral efficiency compared to OFDM. However, QAM-based OFDM increases system complexity and cost. If the channel band-

Data rate (Gb/s)	Distance (m)	Transmitter	Modulation	CCT (K)	Ref.
2.5	–	GaN blue LD	NRZ-OOK	–	[6]
3.43	2.88	RGB off-the-shelf LDs	DCO-OFDM	8000	[2]
4.4	0.2	Red LD (blue and green for illumination)	DCO-OFDM	5835	[11]
4	0.15	GaN blue LD	NRZ-OOK	–	[12]
6.52	1	Blue LD with phosphor	DCO-OFDM	7092	[9]
9	5	GaN blue LD	DCO-OFDM	–	[13]

Table 1. Data rates experimentally demonstrated for LD-VLC systems.

width can be increased, simple schemes are sufficient to yield high data rates as demonstrated in [7]. In case of high-end systems where complexity is not the primary concern, multi-carrier schemes can be utilized.

Despite the very high bandwidth of LDs, the maximum achievable data rate of LD-VLC systems is generally limited due to the absence of commercially available high-speed PDs. In the case of LED-VLC, this was not an issue since the low bandwidth of LEDs was easily covered by available silicon PDs. As mentioned earlier, a 422 nm blue laser was reported to achieve error-free communication at 2.5 Gb/s with a system bandwidth of 1.4 GHz limited by the response of the PD [6]. The modulation bandwidth was further extended to 2.6 GHz to enable transmission at 4 Gb/s by utilizing a high-speed UV extended PD (bandwidth ~ 7 GHz) where the system bandwidth was no longer limited by the receiver but by the source itself [12]. Thus, to fully capitalize on the high modulation bandwidth of LDs, high-speed PDs need to be developed on a commercial basis.

The data rates obtained by different practical LD-VLC systems are summarized in Table 1. It is clearly evident that the communication aspect of VLC is quite promising. Most of the OFDM-based schemes have a typical CP duration of roughly 1 ns, which in turn represents the delay spread of the channel. However, a WDM-based VLC system utilizing higher numbers of LDs operating at different wavelengths needs to be demonstrated practically, which would reveal new challenges in beam alignment, spatial multiplexing, and packaging of transmitters.

Challenges

Health and Illumination

The major issue that might hamper the rapid growth of LD-VLC technology is safety concerns involving the use of LDs. In terms of wavelength, the IEC 60825 rulings permit longer-wavelength devices to output much more power than shorter ones. The safety standards for point source emitters based on transmission power were classified into four classes: Class 1 (up to 0.2 mW), Class 2 (0.2 to 1 mW), Class 3A (1 to 5 mW), and Class 3B (5 to 500 mW). Indoor systems are recommended to be Class 1 eye safe under all conditions. LEDs, being large area devices, can utilize much higher launch powers (250–750 mW) depending on the wavelength of the source. In contrast, the power is limited to only 0.2 mW for LDs to be Class 1 eye safe. Lasers operating in Class 3B can be rendered as Class 1 eye safe by passing the beams through diffusers, allowing higher launch power (~ 100 mW). While such diffusers can achieve efficiencies of about 70 percent, computer aided holograms offer a means to generate custom tailored radiation patterns with efficiencies approaching 100 percent. The maximum achievable data rate of a system at an acceptable bit error rate (BER) is dependent on the signal-to-noise ratio (SNR), energy per bit, and bandwidth. Although the power constraint due to eye safety yields lower SNR for LDs compared to LEDs, the drawback can be overcome by the higher bandwidth of these sources.

Information regarding the performance parameters for an RGB LD-VLC system is compared to an RGB LED-VLC system in Table 2, where both systems utilize a QAM-OFDM modulation scheme. It can be seen that the LD-based VLC system is operating outside its safety limits, but since a holographic diffuser was used, the illumination was demonstrated to be safe. Although the illuminance is quite low for a single LD triplet source, for higher levels of illuminance (~ 840 lx), the data rate was demonstrated to be much higher (~ 14 Gb/s) provided the number of sources were increased.

Another important aspect to be considered is the presence of speckle patterns due to high coherence of lasers. Laser speckles in an illuminated area can be visually disturbing, which might induce health complications. Speckle contrast of a laser pumped phosphor-based system was reduced to 1.7 percent at 5000 lm by proper engineering design, which is similar to that of a blue LED [1]. Other solutions involve the use of glass diffusers, hadamard matrices, vibrating reflectors, and speckle reduction actuators, but their usage comes at the cost of system efficiency. Speckle reduction below the visible level (≤ 10 percent) has been reported by Lemoptix. The techniques traditionally applied to reduce speckle pattern is further depicted in Fig. 4. The optimal selection of the speckle reduction technique for any laser-based lighting system will significantly depend on the LD output power, beam size, and the power consumption of the actuator itself.

Practical Communication Systems

When simultaneous wireless connectivity and illumination are realized with LDs, the number of devices reaches staggering values. Based on the analysis by Tsonev *et al.* in [2], for a data rate of 100 Gb/s at a link distance of 93 cm with spot radius of 19 cm, 324 separate LDs are required. The analysis by Hussein *et al.*, which considered illumination for a standard room (8 m × 4 m × 3 m), needed 72 individual RGB LDs. LDs are currently expensive, and such high numbers will greatly increase system cost.

Cost is an extremely important factor in determining the viability of any lighting product in the industry. At present, LDs are more expensive than LEDs with comparable output power. However, LDs can be packed much more densely on a chip than LEDs, which will give way to brighter

System	Channel bandwidth	Bias level (mA)	Radiant power (mW)	Data rate	Illuminance (lx)	Distance (m)
RGB LED system [14]	25 MHz	350 (R,G,B)	437 (R) 960 (G) 115 (B)	575 Mb/s	350	0.66
RGB LD system [2]	198–245 MHz	85 (R) 36 (B) 80 (G)	48.7 (R) 12.79 (G) 15.89 (B)	3.43 Gb/s	21.25	2.88

Table 2. Comparison of rate capabilities of RGB LED and LD based VLC system.

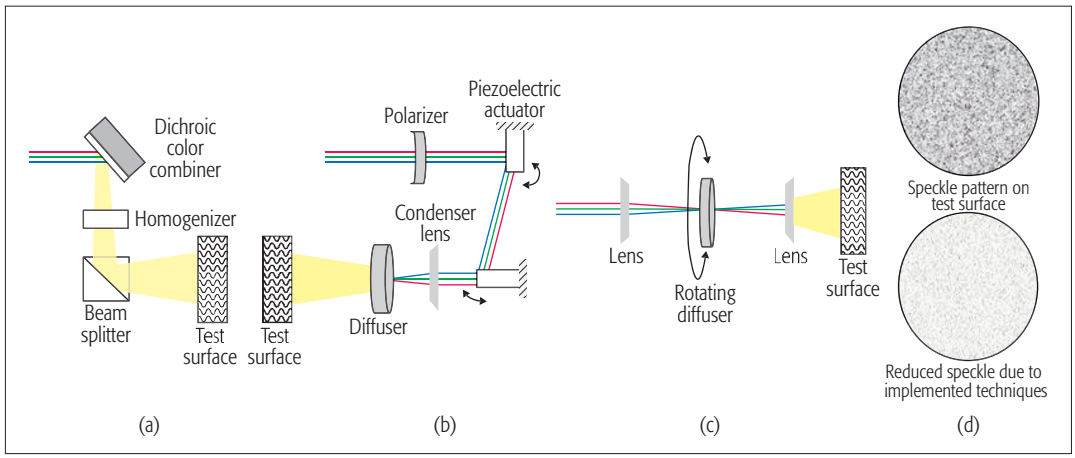


Figure 4. Illustration of three different techniques generally implemented to reduce speckle contrast for laser-based lighting sources: a) a homoginizer and beam splitter to blend in the intensity across the entire illuminated area; b) the beam is steered by two piezoelectric actuators, which causes the angular diversity of the laser beam to suppress speckle; c) a rotating diffuser is used where speckle is mitigated due to angular dispersion of the incident beam; d) general speckle pattern on the test surface before and after the implementation of such techniques.

sources with higher energy efficiency on a dollars per-lumen basis. The current state-of-the-art power conversion efficiency (PCE) is 70 percent for LEDs and 30 percent for LDs which occurs at input power densities of 10 W/cm^2 for LEDs and 25 kW/cm^2, respectively. Areal chip cost necessary for economical lighting depends on input power density and achieving low enough chip cost for LEDs to be operated at input power densities at which their PCEs peak is much more challenging compared to LDs. Because of higher bandwidth, the cost per bit for LD-VLC systems will be much lower. LD manufacturing cost can also be decreased by the elimination or reconfiguration of some of the processes. By keeping much of the high-power diode fabrication at the wafer scale, the need for cleaving and facet coating can be eliminated. The laser-based semiconductor lighting industry is a maturing industry with chip development on a steep upward curve and costs falling rapidly along with some indications that the overall system cost may be lower than LED-based approaches in the future [9]. The high demand for LEDs plays an important role in their lower cost since they are produced in massive numbers. Highly stringent applications requiring LDs are foreseen to enable a rise in their demand, which would further decrease their prices.

Packing LDs densely will certainly give rise to issues regarding thermal management. Proper controller circuitry required for such multi-LD designs will be a major challenge as well. Compared to LEDs, since LDs reach their peak internal quantum efficiencies at much lower current densities, fewer LD chips are required to generate enough lumens, which can be managed by standard heat sinking techniques. On the other hand, compared to traditional pc-LEDs, remote phosphor white light generation utilizing LD allows optical design flexibility and thermal management due to the physical separation of the optical pump source and phosphor module. Another solution is to exploit the beam guiding capabilities of LDs and use large area central cooling systems for hidden sources, which would help avoid expenses for sophisticated small-scale cooling mechanisms for distributed sources as demonstrated in Fig. 5.

LDs are highly directional radiation sources capable of providing high power within a small area, but for practical systems utilizing diffuse lighting, deterioration in link performance needs to be considered. In the case of diffuse links, due to limited power of a single source, the links are only suitable for small distances (typically tens of meters). On the other hand, on comparing scattered high-power laser light from different surfaces in a room to establish a VLC link to direct line-of-sight (LOS) paths between transmitter and receiver, it was found out that the directional approach allows higher spatial reuse, thus providing better data rates [15].

Although most of the LD-VLC systems demonstrated so far strictly use an LOS topology; in practical mobile systems, beam steering methods

The high demand for LEDs plays an important role for their lower cost since they are produced in massive numbers. Highly stringent applications requiring LDs is foreseen to enable a rise in their demand which would further decrease their prices.

The unique features of LDs provide a new level of flexibility for applications in different environments including flying aircraft, hospitals, indoor wireless networks, underwater communication and so on. In the case of underwater communication, LD-VLC can piggyback on the progress achieved by LED-VLC and offer communication at much higher speeds.

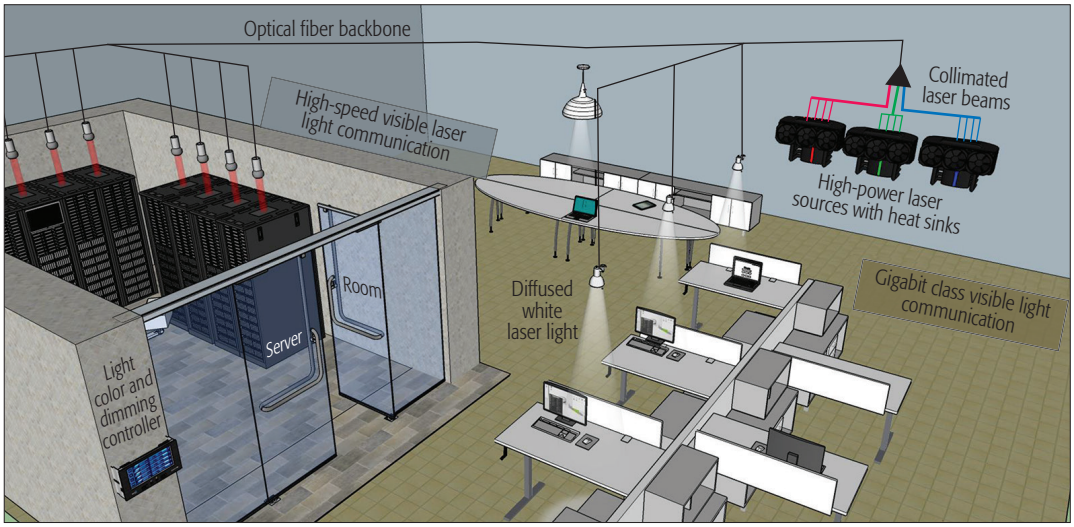


Figure 5. The concept of integration of different LD-based communication technologies in an open office. The fiber backbone of the office space can use plastic optical fiber or multimode fibers (for higher speed), which would carry the data and coupled light to the laser sources. The laser sources generating high-power beams can be installed at a hidden location where the thermal management can be taken care of by using heat sinks and cooling fans. The server room, which generally holds cloud-based data centers and requires data rates in the area of 100 Gb/s, would utilize direct laser-based communication (VLLC), where illumination aspects are not crucial. The open office space, on the other hand, would require illumination, which would be generated by diffused laser light along sideproviding gigabit class VLC. The light color and dimming controller would further allow occupants to control and tune the light based on preferences and activities.

primarily used in free space optics (FSO) systems can be used. An adaptive framework also needs to be developed considering feedback regarding link stability and instantaneous data rates. So far, channel adaptive bit and power loading have been demonstrated for an LD-VLC system where the bits are assigned and power is loaded on each subcarrier based on channel condition [9]. A delay adaptation technique was also demonstrated to achieve a higher degree of freedom in link design, thereby providing higher data rates [7].

Future Prospects and Potential Applications

In the current context, LDs can steadily make its way toward incorporation into value added lighting systems for applications that are demanding enough to exploit the unique features of lasers including high brightness, collimated beam propagation, compact size, and so on. Value added lighting systems merge different industrial sectors including electric utilities, Internet providers, facility management, and luminaire designers into the lighting market, creating huge scopes of profit and modernization. Laser-based lighting innovations can improve our quality of life in terms of our ability to access rapid data downloads in highly secure light networks, to tune decorative and room lighting to different needs, and to improve our health through lighting-based therapies. It may be possible to offset the cost of integration of expensive laser-based systems by providing added value in the overall system, compensating higher end pricing [4]. A concept of such a system incorporating various applications of LDs is depicted in Fig. 5.

LDs are extensively becoming popular for use in automotive lighting, where the system was reported to be much brighter and more efficient than their LED-based counterparts. They offer better design flexibility and more space for mechanical parts. LDs coupled with diffractive optics elements can generate specific signs (e.g., arrows) or patterns, making them favorable for use in traffic signals and advertisement boards. This creates opportunities for utilizing them in intelligent transport systems where the use of VLC is already being considered. The unique features of LDs provide a new level of flexibility for applications in different environments including aircraft, hospitals, indoor wireless networks, underwater communication, and so on. In the case of underwater communication, LD-VLC can piggyback on the progress achieved by LED-VLC and offer communication at much higher speeds because of its low absorption and scattering coefficients.

Conclusion

The increasing congestion in the available RF spectrum due to the growing number of wireless communication devices is driving the need for an alternative technology, which enabled VLC to be an emerging research area. Although LEDs were mostly considered as transmitters for VLC systems, LDs, because of their superior modulation bandwidth and high efficiency, have recently gained attention for gigabit class communication. Compared to traditional light luminaires, it was demonstrated that user experience is not compromised by diffused laser light, which creates a scope for considering LD-based lighting systems to concurrently provide illumination alongside communication. However, there are various challenges in terms of the presence of speckles, power limitations, thermal management, and cost that need to be addressed for facilitating expeditious progress of this technology. VLC itself is in a

developing stage, and incorporating laser diodes would help to effectively resolve some of its issues besides giving rise to numerous beneficial applications that are foreseen to have significant contribution in the field of next generation wireless communication.

References

[1] C. Basu, M. Meinhardt-Wollweber, and B. Roth, "Lighting with Laser Diodes," *Advanced Optical Technologies*, vol. 2, Aug. 2013, pp. 313–21.

[2] D. Tsonev, S. Videv, and H. Haas, "Towards a 100 gb/s Visible Light Wireless Access Network," *Opt. Express*, vol. 23, no. 2, Jan 2015, pp. 1627–37.

[3] T. Borogovac and T. Little, "Laser Visible Light Communications," *Proc. IEEE Photonics Society Summer Topical Meeting Series*, July 2012, pp. 117–18.

[4] L. Y. Kuritzky and J. S. Speck, "Lighting for the 21st Century with Laser Diodes Based on Non-Basal Plane Orientations of Gan," *MRS Communications*, vol. 5, 2015, pp. 463–73.

[5] A. Neumann *et al.*, "Four-Color Laser White Illuminant Demonstrating High Color-Rendering Quality," *Opt. Express*, vol. 19, no. S4, Jul 2011, pp. A982–90.

[6] S.Watson *et al.*, "Visible Light Communications Using A Directly Modulated 422 Nm Gan Laser Diode," *Opt. Lett.*, vol. 38, no. 19, 2013, pp. 3792–94.

[7] A. Hussein and J. Elmirghani, "10 gbps Mobile Visible Light Communication System Employing Angle Diversity, Imaging Receivers, and Relay Nodes," *IEEE/OSA J. Opt. Commun. Networking*, vol. 7, no. 8, Aug. 2015, pp. 718–35.

[8] K. A. Denault *et al.*, "Effcient and Stable Laser-Driven White Lighting," *AIP Advances*, vol. 3, no. 7, 2013.

[9] H. Chun *et al.*, "Visible Light Communication Using Laser Diode Based Remote Phosphor Technique," *Proc. IEEE Int'l. Conf. Commun. Wksp.*, June 2015, pp. 1392–97.

[10] Y.-C. Chi *et al.*, "Phosphorous Diffuser Diverged Blue Laser Diode for Indoor Lighting and Communication," *Scientific Reports*, vol. 5, 2015, pp. 18,690–99.

[11] B. Janjua *et al.*, "Going Beyond 4 Gbps Data Rate by Employing RGB Laser Diodes for Visible Light Communication," *Opt. Express*, vol. 23, no. 14, Jul 2015, pp. 18,746–53.

[12] C. Lee *et al.*, "4 Gbps Direct Modulation of 450 nm Gan Laser for High-Speed Visible Light Communication," *Opt. Express*, vol. 23, no. 12, 2015, pp. 16 232–37.

[13] Y.-C. Chi *et al.*, "450-nm Gan Laser Diode Enables High-Speed Visible Light Communication with 9-Gbps QAM-OFDM," *Opt. Express*, vol. 23, no. 10, May 2015, pp. 13,051–59.

[14] Y. Wang *et al.*, "Demonstration of 575- mb/s Downlink and 225-mb/s Uplink Bi-Directional SCM-WDM Visible Light Communication using RGB LED and Phosphor-Based LED," *Optics Express*, vol. 21, no. 1, 2013, pp. 1203–08.

[15] A. Sevincer et al., "Lightnets: Smart Lighting and Mobile Optical Wireless Networks — A Survey," *IEEE Commun. Surveys & Tutorials*, vol. 15, no. 4, 2013, pp. 1620–41.

Biographies

FAHAD ZAFAR received his B.E. degree (first class honors) in electrical and computer systems engineering from Monash University Malaysia in 2013. He is currently pursuing his Ph.D. degree at the same institution. His areas of research include modulation and dimming schemes for optical wireless communication, visible light communication, solid state lighting, and intelligent lighting systems.

MASUDUZZAMAN BAKAUL [S'02, M'06] received his M.S. and Ph.D. degrees from the University of Melbourne, Australia, in 2003 and 2006, respectively. He is currently a senior lecturer of electrical and computer systems engineering at Monash University Malaysia, which he joined in 2014. Prior to Monash, he worked for eight years for NICTA and the University of Melbourne through several collaborative appointments. His research areas include optical communications, millimeter-wave photonics and silicon photonics, in which he has contributed 90 technical articles.

RAJENDRAN PARTHIBAN [M'05] is an associate professor at the School of Engineering, Monash University Malaysia. He completed his B.E. (Hons) in 1997 and Ph.D. in 2004, both from the University of Melbourne. His research interests are in cost and energy comparisons of optical network architectures, visible light communication and positioning, and grid/cloud computing.

VLC itself is in a developing stage, and incorporating laser diodes would help to effectively resolve some of its issues besides giving rise to numerous beneficial applications that are foreseen to have significant contribution in the field of next generation wireless communication.

Network Hardware Virtualization for Application Provisioning in Core Networks

Ashwin Gumaste, Tamal Das, Kandarp Khandwala, and Inder Monga

The authors articulate the impact of NV on networks that provide customized services and how a provider's business can grow with NV. They outline a decision map that allows mapping of applications with technology that is supported in NV-oriented equipment. Analogies to the world of virtual machines and generic virtualization show that hardware supporting NV will facilitate new customer needs while optimizing the provider network from the cost and performance perspectives.

ABSTRACT

Service providers and vendors are moving toward a network virtualized core, whereby multiple applications would be treated on their own merit in programmable hardware. Such a network would have the advantage of being customized for user requirements and allow provisioning of next generation services that are built specifically to meet user needs. In this article, we articulate the impact of network virtualization on networks that provide customized services and how a provider's business can grow with network virtualization. We outline a decision map that allows mapping of applications with technology that is supported in network-virtualization–oriented equipment. Analogies to the world of virtual machines and generic virtualization show that hardware supporting network virtualization will facilitate new customer needs while optimizing the provider network from the cost and performance perspectives. A key conclusion of the article is that growth would yield sizable revenue when providers plan ahead in terms of supporting network-virtualization-oriented technology in their networks. To be precise, providers have to incorporate into their growth plans network elements capable of new service deployments while protecting network neutrality. A simulation study validates our NV-induced model.

INTRODUCTION

Provider revenues are growing primarily based on provisioning next generation services such as video, cloud, mobile backhaul, and data centers. Applications that dominate provider revenues are becoming aggressive in their network requirements [1]. If service providers do not reinvent themselves to meet application requirements, their revenue will decrease due to *over-the-top* (OTT) vendors capturing much of the newfound e-commerce revenue. For example, video distribution OTT vendors like Netflix, Amazon, Dropbox, and Salesforce are cashing in on raw bandwidth pipes provided by network operators, creating a constant feud between network providers and application providers. In the worst case scenario, a network provider could impede good quality service to application providers as they do not share revenues, given that the network is merely seen as a basic bandwidth pipe. This feud must be resolved for the larger sake of the ecosystem.

Another aspect of this feud is the drive to protect network neutrality (NN). Shown in [2] are multiple aspects of NN. While not throttling someone's service is a given, a more important aspect is how to create a new service that better facilitates the OTT operator while protecting NN. It is not a question of how long it would take for service providers (SPs) to support OTT services, but rather a question of how to support such a service. As elaborated in [1], it is about routing money, not packets.

This article studies the interaction between network providers and application providers through the use of network hardware virtualization. Network virtualization (NV) manifests itself as an excellent way to resolve this feud by facilitating the partitioning of the network hardware into qualitative domains that are responsible for providing specific service to the application provider. We see NV as an intermediate enabler for network functions virtualization (NFV), and in direct conjunction with SDN white-boxes.

In this article, we propose NV as an enabler toward solving the paradox between network operators and OTT application providers. Network operators reason that they have to invest in the network infrastructure, and licensing and maintaining the network, while application providers use the network and earn revenue from consumers. Customers of the network provider at times overuse the liberties provided by the network provider. Application providers, on the other hand, treat the network as a bunch of bandwidth pipes that pre-exist and do not see a reason to share their revenue. There are merits in both arguments from the perspectives of network and application providers. The deadlock needs to be resolved for both parties to maximize profit as well as serve the end user better.

This deadlock can be technically resolved by implementing NV. The idea is that by using NV in the network, a service provider c*an now customize services that suit the OTT application*. An OTT application provider now has the incentive to share revenue or buy a specific related service that better drives their application to their end user (the consumer).

The next obvious question is *how to implement NV in a network operator*. We begin by understanding application requirements at a broad level and mapping them to possible capabilities of networks to offer customized services. Webb

Digital Object Identifier:
10.1109/MCOM.2017.1500488CM

Ashwin Gumaste and Tamal Das are with the Indian Institute of Technology Bombay; Kandarp Khandwala is with the University of California San Diego; Inder Monga is with Lawrence Berkeley National Laboratory.

et al. [3] described ways in which an application can communicate to the network in terms of customization required for a particular application. However, rather than real-time application-level changes, most OTTs have specific and well-known requirements from the network [4]. *So can we model a network based on such requirements, mapping these requirements to NV partitions?*

To do so, we first explore if it indeed is feasible to model OTT requirements over an SP network, by isolating key services that would have strong business cases for implementing NV, and have key requirements that a provider can fulfill. To this end, the next section presents a table that manifests OTT requirements from the network including network technology choices [5]. For the sake of brevity, we focus only on the metro and core network, assuming that network pipes are essentially static entities with little scope for technology enhancement due to voluminous users, although with NFV even the access gear could be virtualized. Then we present a method for the application service provider (ASP) to interact with the SP and show how this can be implemented in four different technology classes, each using a software defined control plane. Following that we show how software defined networking (SDN) can be made to function in such a scenario, and the relationship between SDN and NV pertaining to the technology solutions. Finally, we capture results from a simulation model that validates our hypothesis.

Disparate Networking Requirements

In this section, we discuss application-level requirements of various domains and how these can be mapped to network equipment through NV. Shown in Table 1 is a list of key revenue-generating OTT services. For each service, Table 1 lists network-centric specifics desired by an ASP and plausible technology options to provision the service. Table 1 considers ASP businesses currently valued at US$1 billion plus [6]. The key driver toward ASP traffic is video. Since we ignore the access network, it is safe to say that the traffic is largely business-to-business (B2B) in nature, but can without loss of generality be extended to a business-to-consumer (B2C) model. For many of the applications, there are multiple technology solutions possible, and the ones that are commercially viable in a tier 1 provider network are listed in column 3.

The key question that Table 1 highlights is: *how can an SP provision a particular service requirement in the network?* To this end, a system must be designed that orchestrates interaction between the provider and the OTT ASP, adhering to tenets of NN. This interaction must be mapped onto network hardware so that service provisioning is possible. Our proposal is to create an SDN controller that would facilitate interaction between incoming traffic requests from ASPs mapping these onto provider hardware that adheres to NV principles. The key challenge in this approach is to *map the incoming demand into network-specific parameters that can be used for traffic engineering, bandwidth brokering, provisioning, and service support*; and *enable the network hardware to be able to provision new services with specific OTT needs.* The challenge in the latter is to be able to create services and differentiate them at the hardware layer.

Domain/OTT service	Requirement from network	Technology
Video services	Guaranteed bandwidth low jitter	IP/MPLS/WDM/CE
Mobile VAS	Unconstrained bandwidth low packet drop	MPLS/OTN/CE
Video advertising and merchandise delivery	Bandwidth on demand low jitter	MPLS/CE
Real-time events and entertainment delivery	Extreme multicast bandwidth on demand	MPLS/CE/WDM
Healthcare and telemedicine	Low downtime, high bandwidth, security, low latency	MPLS/CE
Defense networks	Minimal downtime, low latency, security, virtualization, multicast, bandwidth	MPLS/OTN/CE/WDM
Finance and banking	Virtualization, minimal latency, security	IP/MPLS/CE/OTN
Educational networks	Multicast, high bandwidth	WDM
IT virtualization	Fast switching, resiliency	MPLS
Gaming services	Extreme interaction, multicast, low latency	MPLS/CE/IP

Table 1. Service-technology matrix.

The next section describes a solution using NV principles to partition SP hardware to meet ASP service goals. The advantage of NV is that it enables an SDN controller to realize the full potential of an SDN-centric network.

Building a Solution with Virtual Network Equipment Partition

In this section, we describe a method to implement NV to meet specific ASP requirements. We assume that a request for a service arrives into a SP domain and a *network management system* (NMS) communicates to an SDN controller that would provision services. The NMS can abstract specific requests into network-centric parameters with the goal of provisioning services. The NMS maps a service request onto an abstracted network topology by considering specific service parameters. These parameters are then mapped onto all the network elements (NEs) in the path to check service provisioning feasibility. To check feasibility, there must be a parameterized relationship between incoming service requests and the equipment deployed. The SDN controller maps an incoming request to a network-virtualized hardware. The idea is that every piece of hardware is further divided into service supporting modules that are parameter-driven and have a direct relation with an SDN controller populated flow table. Virtualization happens by the creation of multiple (virtualized) instances of the data-plane at each NE. Each such instance of the data plane enables OTT-service-specific feature implementation.

Method to Implement NV in SP-ASP (OTT) Interaction

We now describe how to implement NV in a provider network. A request that enters the network is provisioned through a network interface supported by the NMS. For each new incoming

In our approach, we partition a switch/router/optical-cross-connect into VNEPs that can individually provision services. The idea is to dynamically create a VNEP that will adhere to all the system-wide parameters for a particular service, with the constraint that the sum of all the VNEPs in a NE is less than the total capacity of the switch. The union of VNEPs is not linear.

request, the NMS computes the optimal network resources to be allocated. To this end, the following steps are envisaged at the centralized NMS:

- A route is computed based on service requirements. Actual bandwidth allocation is computed along the route depending on the specified request and other requests at that instance.
- Each element along the computed route is examined from a service support perspective, whether it can satisfy *specific* requirements of the service.
- To compute the specific requirements of the service request, we propose the concept of or *virtualized network equipment partitions* (VNEPs) that enables a network equipment (e.g., a switch or router) to be partitioned to satisfy specific service parameters. An example of a VNEP is provided in the next subsection.
- If VNEPs are possible along the path to provision the request, all the network equipments are provisioned to meet the new request by the NMS through the SDN controller. Otherwise, an alternate path that maximally conforms to the VNE requirement (partially, if not fully) is provisioned.
- A VNEP created at a node may be moved to another node depending on resource availability over a period of time.

VNEP computation is now described in detail.

VNEP Computation

A VNEP is represented by the virtual partitioning of hardware such that each of the partitioned elements corresponds to fully functional entities, capable of performing all the functions as the larger hardware, but specific for a service request. The key to VNEP creation is to note that the overlaid software creates partitions by allocating hardware resources within a larger NE. Partitions could be created in switching elements, network processors, buffers, and packet classifiers. Partitions correspond to hardware resources as defined by the software and are made available strictly for a particular service or function.

Our conjecture (based on an analysis of existing network gear) is that a networking element can be divided into partitions, such that a partition can act as a completely independent networking element. We argue that the sum of parts — that is, the union of all *partitions* — does not necessarily add up to the *original element* for that *particular parameter*. Throughput, average latency, and packet loss rate are examples of parameters.

Let us consider an example: Assume a 60 Gb/s switch fabric with virtual-output-queued (VOQ) buffers, with 6-input lines and 6-output lines all at 10 Gb/s. Assume one of the lines is sending data at 2 Gb/s, the average packet size is 250 bytes, and the VOQ memory to store packets for contention resolution is 3 Mb. The maximum ingress-to-egress latency is observed as 300 s. However, we aim to estimate average latency, which is a function of the provisioned services at the other 5-ingress ports, the nature of the traffic, and type of switch fabric (cut-through, store-and-forward, shared memory, etc.).

Since the latency of a flow through a switch also depends on other flows, one way to control it is to bound the number of flows through the switch. A simple 4 × 4 cross-bar with VOQ (essentially a 12 × 4) switch (each port at 1 Gb/s) can take 4 flows each with 250-byte average packet size at full line rate (wirespeed operation), resulting in 1.2 μs switching, while the same switch will result in 2.4 μs port-to-port latency if the average packet size is 128 bytes [7]. Similarly, the switch will result in a latency of 3 μs if the packet size is 64 bytes [7]. The switch behavior becomes more erratic when the standard deviation between flows across multiple ports increases [8]. For example, the switch results in a port-to-port latency of 12 s for multicast traffic if the packet size is 64 bytes, and remaining ports have provisioned flows with packet size of 1500 bytes. The above discussion highlights the complex relationship between packet sizes, port counts, traffic distribution (random/unicast/multicast), and so on, implying that for carrier-class services, that is, with desired deterministic parameters of delay and jitter, predicting switch behavior is important but difficult. Even intricate queuing models (i.e., those deploying G/G/1 queues) tend not to converge in real time.

So our approach is to provision services without getting involved in the intricacies of computing switch-specific parameters in real time. *Our approach is technology-specific, given the enormous amount of technology deployments.*

In our approach, we partition a switch/router/optical-cross-connect into VNEPs that can individually provision services. The idea is to dynamically create a VNEP that will adhere to all the system-wide parameters for a particular service, with the constraint that the sum of all the VNEPs in an NE is less than the total capacity of the switch. The union of VNEPs is not linear. This implies that the system leads to overprovisioning, which, though undesired, is necessary to maintain many of the carrier-class attributes desired for OTT services.

VNEP creation and sizing involves the following steps:

1. An NE is viewed as the number of instances q_i of a particular parameter i such that $f(qi)$ denotes the performance criteria (e.g., bounded latency) for parameter i.
2. The value $f(qi)$ also takes into consideration another parameter whose performance criteria is $f(qi)$ is the number of instances of supporting j and which impacts $f(qi)$.
3. Note that it is mathematically nontrivial to compute $f(qi)$, and hence worst case provisioning metrics are used as acceptable practices.

The second point is supported by an example. Let i denote the service parameter for port-to-port latency. Assume a 60 Gb/s switch fabric supports 6 × 10 Gb/s connections with 250 bytes average packet size and $f(qi)$ = 3 μs. The same fabric will have an $f(qi)$ = 12 μs latency for the same number of flows if the average packet size reduces to 64 bytes. The delay increases sizably ($f(qi)$ = 50) if the number of flows increases to 60 × 1 Gb/s flows. So, now if we have to provision a service of 1 Gb/s with a latency within 3 μs, and another service of 5 Gb/s with a latency also within 3 μs, how do we do so given that the packet size of the first service is, say,

128 bytes and the second one is, say, 64 bytes? Obviously, the second service will require more overprovisioning compared to the first one, that is, that although the second service is 5× the first service, in order to achieve similar parameters, the second service may have to be provisioned through the switch with 12× resources (buffers primarily) so that the switch can meet provisioning requirements. Now how do we arrive at the number 12×? This number is a function of both volume and quality: volume, as in how much more would the service take in every parameter's domain, and quality, as in what would be the impact of the service provisioning in other parameter domains.

Shown in Fig. 1a is the actual process for creating and allocating VNEPs. From an incoming request (*Req*(*i*)), we compute the corresponding partition's impact on other partitions. The SDN controller computes VNEPs for each service at each NE. The controller then sends specific information to each node to partition itself according to its VNEP computation based on the four use cases discussed at the end of this section.

Given that there are a large number of protocols deployed leading to a variety of equipment such as IP/multiprotocol label switching (MPLS) routers, optical transport network (OTN) cross-connects, carrier Ethernet (CE) switches, and wavelength-division multiplexing (WDM) gear, a key question is how to implement partitioning. It is publicly known that many vendors are in the process of SDNizing their current gear. The question we want to answer is: *how can equipment vendors achieve network virtualization at the data plane?*

To this end, we have identified network equipment from 10 vendors who are known to be committed to SDNizing their product portfolio. These 10 vendors combined have products across the aforementioned technologies (IP/MPLS etc.). Seven of these vendors have products in the layer 2/3 (L2/L3) space, while three products are from the optical space.

On studying the equipment of these seven chipsets as well as corresponding patents, it appears the architecture follows one or a combination of the following three strategies:

- A field programmable gate array (FPGA)-based switching core or an FPGA as a processing element
- An application-specific integrated circuit (ASIC) or merchant silicon-based switching core with an FPGA or a processor guiding the ASIC
- A network processor (NP)-based switching core

In Table 2, we captured the key chipsets that are used for creation of the products for the various equipment vendors. The table also shows how the data path can be partitioned.

Shown in Table 2 are seven implementations of a switching plane used for L2/L3 equipment. Additionally, we have also considered three reconfigurable optical add/drop multiplexer (ROADM) implementations using liquid crystal on silicon (LCOS)-based wavelength selective switches (WSSs) of 1 × *M* and *M* × *N* configurations and another WSS based on digital light-

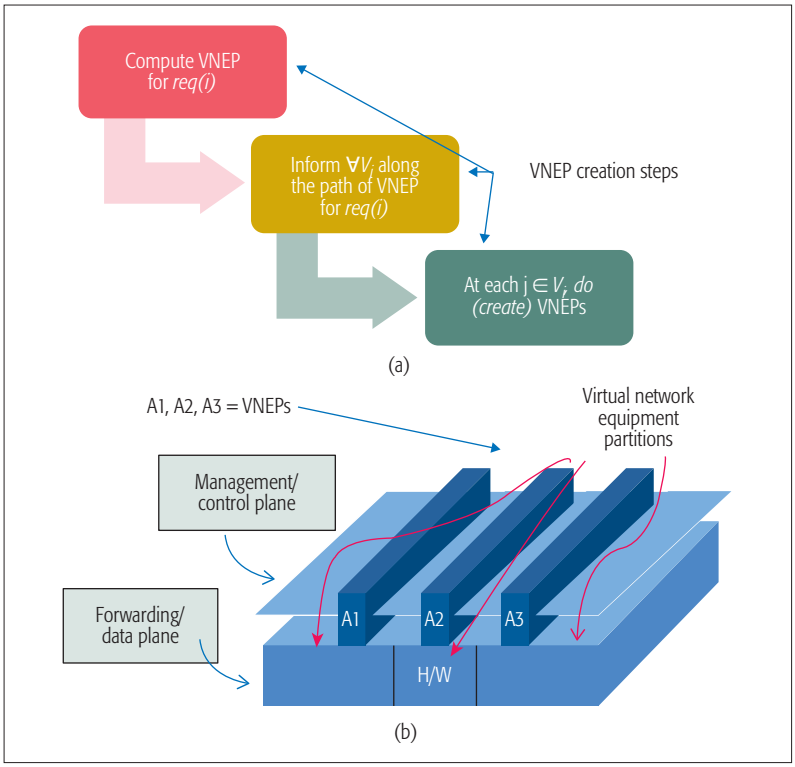


Figure 1. VNEP.

wave processing technology. The initial seven L2/L3 cases are shown in the table. Two types of FPGAs (FPGA 1 and 2), two types of network processors (NP 1 and 2) and three types of ASICs with FPGAs and NPs (ASICs 1–3) are compared.

The key takeaway from Table 2 is to show that irrespective of the technology deployed, it is indeed possible to create VNEPs. To this end, Table 2 showcases the sliceability parameter at what granularities can we slice a fabric. The impact of slicing is on the throughput (speed of the device) and latency. The memory capacity also has a direct impact on the throughput: the more slicing, the more memory is required; hence, latency suffers. Larger numbers of flows require either more interconnected fabrics (multicard designs) or use of large ASICs (columns 7 and 8). The latency is impacted by sliceability as well as protocol (quality of service [QoS], more processing, etc.).

VNEP Partitioning Analogous to Virtual Machine Creation and Migration: VNEP creation and NV using VNEPs is analogous to virtual machine (VM) creation and migration in hardware. Shown in Fig. 1b is the analogy of the forwarding plane in an NE with a VM hypervisor. VMs can be dynamically created in a processing environment. The same analogy is used for VNEP creation, whereby VNEPs are, like VMs, created on the fly and use the switch fabric resources independently. As in Fig. 1b, VNEPs are created by the control plane (SDN-based) and implemented within the NE through NV.

From the perspective of Table 2, we can create VNEPs as slices in different implementations of L2/L3 equipment or as independent optical switches virtually superimposed on a ROADM, as shown next.

	FPGA 1	FPGA 2	NP1	NP2	ASIC 1	ASIC 2	ASIC 3
Sliceable or not	Yes (425K logic blocks)	Yes (693K logic blocks)	Yes	Yes	Yes	Yes	Yes
Min. SW granuality IO (Mb/s)	1	≤ 1	1	0.064	0.128	1	0.064
SW granularity every component (Gb/s)	240	800	120	640	40	1280	50
How many parallel lines (10 Gb/s)	24 (16 standard 8FX)	80 (GTX)	12 × 10G OR 3 × 40G	64 × 10G and 16 40G	4 × 10G and 24 × 1G and 12 × 2.5G	12 8 × 10G	2 × 25G
Switch capacity	360 Mp/s	1600 Gb/s	≤ 2 GHz	1.25 GHz	60 Mp/s	1440 Mp/s	30 Mp/s
Average latency	400 ns	500 ns	NA	NA	NA	150–650ns	120 ~750ns
Protocol	L2/3/4	L2/3/4	L2/L3	L2/L3/L4	L2/L3	L2/L3	L2
Memory capacity	50 Mb	52 Mb	4 MB	7.5 MB	8.5 MB	12 Mb	–
Number of switching blocks	Variable	Variable	120,000	240,000	40,000	1,280,000	50,000

Table 2. VNEPs in pragmatic network elements.

Use Cases

Use Case 1: IP/MPLS-over-WDM: For IP/MPLS overlay and WDM ROADM underlay, IP/MPLS label switched routers (LSRs) are partitioned based on supported flows, and WDM ROADMs are partitioned to support non-blocking connections. VNEPs in the ROADM require support of colorless, directionless, and contentionless (CDC) as well as gridless properties. A VNEP in an LSR is an MPLS tunnel.

Use Case 2: MPLS-over-OTN with WDM: In the case of MPLS-over-OTN with WDM underlay, VNEP partitions take into consideration OTN pipes at MPLS-LSR interfaces that further feed into a WDM network. We assume services are sub-wavelength granular, implying wavelength assignment as a multi-service aggregation and provisioning problem. Partitioning happens at the LSR forwarding plane and OTN-based ODU (optical data unit) switch fabric.

Use Case 3: CE+OTN-over-WDM: In this case, we partition the CE switch fabric into discrete switching chunks so that an Ethernet switched path (ESP) is mapped onto an OTN ODU port. The VNEPs are portions of the CE switch fabric implemented.

Use Case 4: IP-over-CE+OTN-over-WDM: In this case, IP routers are at select locations as an overlay with a CE underlay, all over a ROADM-based WDM network. Whenever a service has granularity that is near a wavelength's full capacity (10/100Gb/s), it is routed all-optically by the ROADM. Whenever a service can be routed at layer 2 through the use of an ESP, it is done so using the CE network used for aggregation and switching. However, when layer 2/1 provisioning is not possible, the service is handled exclusively through the IP layer. VNEP information created by the centralized controller is used to partition switching resources at any or all of the CE/IP layers that use FPGAs/ASICs/NPs.

Interaction between SDN and NV

Figure 2 shows a switch architecture to implement SDN with NV with a controller connected to the switch's northbound interface. The switch could support L2/L3 protocols, and the interfaces would be mapped onto wavelengths. Incoming flows are segregated at the input buffers (which are further segregated to support VOQs). Flow headers are worked on by a control state machine (CSM) that also populates SDN tables. All protocol functioning happens at the controller. To support scalability, we assume that the controller runs on a VM.

The architecture in Fig. 2 can have multiple manifestations including the use of FPGAs/ASICs/NPs. In one embodiment, we assume an IP/MPLS LSR in which the CSM and SDN flow tables (SDNFTs) are implemented in an FPGA, while other modules are implemented in an ASIC. In another CE design, the SDNFT, CSM, and VOQs are implemented in an NP, while the switch fabric and memory are implemented in an ASIC. Yet another design includes a smaller CE device that has the entire design except the SDNFT in an FPGA, with the flow tables in a TCAM ASIC.

We propose the following three policies for VNEP partitioning:

Policy 1: Throughput Maximization: In this policy, VNEP computation maximizes the throughput at every NE. This is a non-carrier-class policy implying that the port-to-port latency per NE is non-deterministic. This implies an additive increase of throughput, and hence, whenever a new request arrives at the SDN control plane, a VNEP is created with a view to maximize network-wide throughput. The CSM partitions the hardware as per the specifications of Table 2.

Policy 2: Latency-Bounded Partitioning: In this policy, a VNEP is created such that the corresponding service is guaranteed to meet end-to-end latency requirement through every NE by bounding latency. This policy requires double optimization: route selection and associated appropriate amount of partitioning at a node.

Policy 3: Latency-Sensitive Service Maximi-

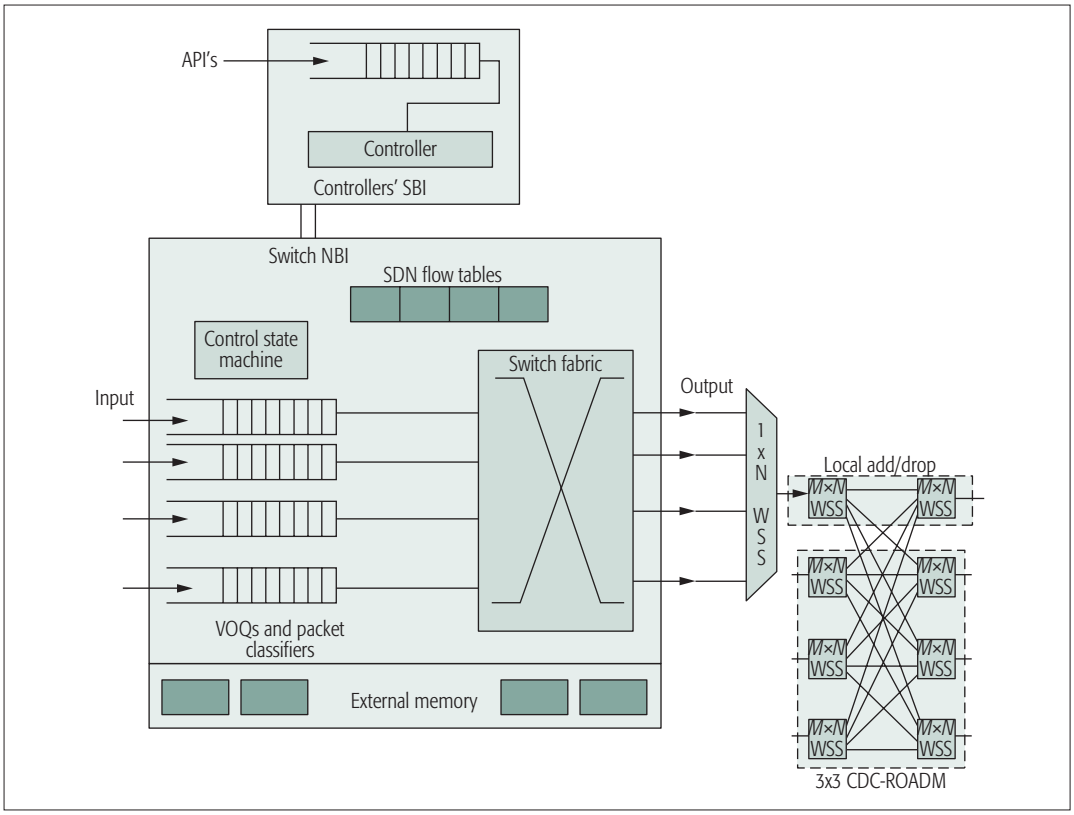


Figure 2. Switch Architecture to Implement SDN.computation (top) and VM migration analogy (bottom).

zation (LSSM): In this globally active policy, the approach is to maximize the number of services through an NE. The controller creates VNEPs such that they balance each other in terms of parameterized requirements. For example, services with similar delay and bandwidth requirements are load balanced. The controller also provides for equal cost multiple paths (ECMPs) to load balance the service.

In our simulation model, we rationalize service requirements based on their utility to the network (revenue for the provider) and normalize the utility over delay constraints. We then provision services such that the delay constraints are met, while bundling as many services together as possible. The LSSM policy is a greedy heuristic, and its complexity is fourth-order polynomial in terms of number of links in the network; hence, its functioning depends on graph size.

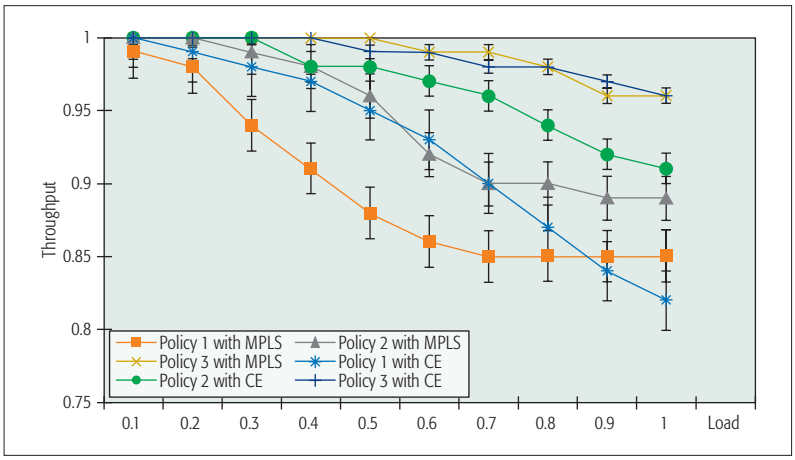


Figure 3. Throughput as a function of load for different policies.

Simulation Model and Hypothesis Verification

A simulation model was built to test our VNEP hypothesis as a method to facilitate interaction between SPs and ASPs. We model a provider network with two autonomous systems (ASs) and five metropolitan regions, with each region divided randomly into 20, 40, 60, 80, and 100 access regions. The backbone and metro networks use fiber, while the access networks could be wireless/fiber/coaxial cable-based. Our goal is to evaluate the impact of NV over different technologies by provisioning OTT services. To this end, the simulation model implements each technology solution using proposed VNEP creation policies.

Each access region has between 10,000 and 100,000 subscribers, and is connected to a metro network with multiple metros backhauled to a core network (wholly viewed as a single AS). The point of presence (POP) connecting the access to the metro supports ROADMs. The overlay depends on the technology being simulated; we study IP/MPLS, MPLS, OTN, and CE technologies, and the seven cases of Table 2 are deployed randomly. The control plane is implemented as an SDN overlay that consists of controllers, one for an AS of 10,000 users and hierarchically arranged thereafter.

The simulation model works as follows: Randomly generated service requests have specific QoS parameters. Services are organized into two levels — services and sessions. Services are exponentially distributed with a mean holding time of

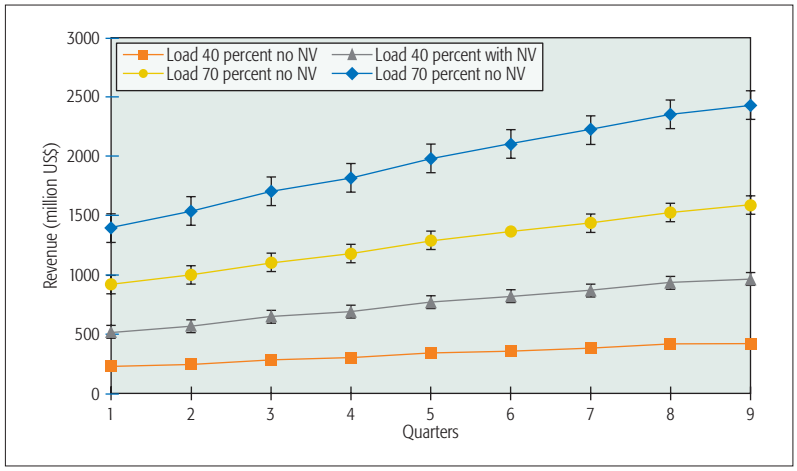


Figure 4. SP revenue with and without ASP revenue-sharing through NV.

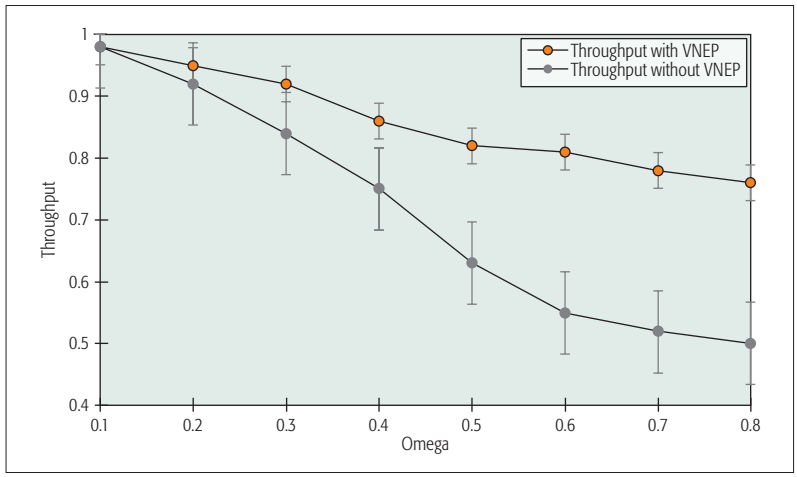


Figure 5. Throughput versus ratio of MP2MP/P2P traffic.

6 months, while session holding time is exponentially distributed with a mean time equivalent to a 100 MB video file download session. The service is guided to the appropriate controller, which uses one of the three VNEP creation policies and evaluates whether provisioning is possible. Services are lumped through pre-assigned aggregation policies. Once a service is provisioned, we compute service and switch statistics.

Load is computed as average occupancy of all the services to the maximum allowable input rate across all the ingress ports. MPLS LSRs have 1 Gb/s and 10 Gb/s interfaces and a net switching capacity of 640 Gb/s [9]; CE switches have 1 Gb/s and 10 Gb/s interfaces and 80 Gb/s fabric that is stacked to create a 640Gb/s node [7]. ODU switching is assumed at ODU0/1/2e [10]. Transport wavelengths can be generated by an MPLS/CE/IP forwarding plane and support 10 Gb/s, 40 Gb/s, and 100 Gb/s. Cost is computed as in [11] for both capital expenditure (CAPEX) and operational expenditure (OPEX), while we assume that for provisioning OTT services, the OTT ASP shares 20 percent of its revenue with the SP.

Figure 3 compares all three policies used for VNEP computation using MPLS and CE technologies using the FPGA, FPGAs+ASIC, and NP+ASIC approaches. We show throughput vs. load with error bars indicating stability of results. MPLS and CE were chosen as likely candidates for cost considerations. A peculiar behavior is that policy 3 has the best throughput, while being able to take service latency into consideration.

Figure 4 highlights the effect of ASP revenue sharing through NV on the SP revenue. It shows that there is sizable incentive for ASPs to share their revenue as the providers would be able to grow the network, thereby facilitating larger and qualitatively superior reach for the ASPs. Figure 4 is generated as follows: We first measure ASP revenue without NV and no revenue sharing. NV is implemented using the most popular approach, FPGA+ASIC, and uses the LSSM policy. We compute the revenue by pegging each service at 30–40 percent higher price than before. For example, a 12 Mb/s HD video pipe was priced at US$20 per month with no revenue sharing and hence no NV support. The same pipe with guaranteed bandwidth (no packet loss) is priced at US$26, while it is priced at US$30 with bounded latency and 50 ms restoration of service in case of fiber cut/equipment failure.

Figure 5 studies the impact of VNEP on throughput in the network as a function of the ratio (defined as Omega) of multipoint-to-multipoint (MP2MP) traffic to point-to-point traffic. The graph is generated for the case of MPLS. As Omega increases, the throughput without VNEP decreases rapidly, while that with VNEP decreases gradually. This is a critical result showing the maximum benefit of the use of VNEP, which is modeled as implemented in FPGA+ASICs. The graph shows how VNEPs can impact new service support such as multicast services that are poorly handled at higher loads.

Conclusion

We present an approach to integrate OTT application providers with service providers using network virtualization in hardware. Our work is inspired by [9, 12]. We propose the concept of virtual network equipment partitions that enable an NE to be partitioned as per service requirement, thereby benefiting from programmability of the control plane. Policies to partition an NE are discussed. Results from a simulation study show the benefit for ASPs in a provider network using NV-compliant hardware.

References

[1] V. Mishra, "Routing Money, Not Packets" *Commun. ACM*, vol. 58 no. 6, pp. 24–27.
[2] FCC Report: Protecting and Promoting the Open Internet, FCC 15-24, GN Docket No 14-28.
[3] K. C. Webb, A. C. Snoeren, and K. Yocum, "Topology Switching for Data Center Networks," *Proc. Hot-ICE*, 2011.
[4] J. Mogul and L. Popa, "What We Talk about When We Talk about Cloud Network Performance," *ACM Proc. Sigcomm 2013*, Chicago, IL.
[5] A. Gumaste and S. Akhtar, "Evolution of Packet-Optical Integration in Backbone and Metropolitan High-Speed Networks: A Standards Perspective" *IEEE Commun. Mag.*, vol. 51, no. 11, Nov. 2013, pp. 105–11.
[6] Infonetics Research, "Data Center and Enterprise SDN Hardware and Software Report," 2013.
[7] S. Bidkar *et al.*, "On the Design, Implementation, Analysis, and Prototyping of a 1-ms, Energy-Efficient, Carrier-Class Optical-Ethernet Switch Router" *IEEE/OSA J. Lightwave Tech.*, vol. 32, no 17, pp 3043–60.
[8] S. Das, G. Parulkar, and N. McKeown, "Rethinking IP Core Networks," *IEEE J. Optical Commun. Networking*, Dec. 2013.
[9] P. Bossharty, "Forwarding Metamorphosis: Fast Programmable Match-Action Processing in Hardware for SDN," *ACM Proc. Sigcomm 2013*, Hong Kong, China.

[10] Per Harald Knudsen-Baas, *OTN Switching*, Master's thesis, Dept. of Telematics, Norwegian Univ. of Science and Technology, June 2011.
[11] A. Mathew *et al.*, "Multi-Layer High-Speed Network Design in Mobile Backhaul Using Robust Optimization" *IEEE/OSA J. Opt. Commun. and Networking*, vol. 7. no. 4, Apr. 2015. pp 352–67.
[12] E. Haleplidis *et al.*, "Network Programmability with ForCES," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 3, 2015.

Biographies

Ashwin Gumaste is currently an associate professor in the Department of Computer Science and Engineering at the Indian Institute of Technology (ITT) Bombay, Mumbai. He was the Institute Chair Associate Professor (2012–2015) and the JR Isaac Chair (2008–2011), and from 2008 to 2010 he was a visiting scientist at the Massachusetts Institute of Technology, Cambridge. He has held positions with Fujitsu Laboratories (USA) Inc and Cisco Systems, and has been a consultant to Nokia Siemens Networks. He has also held short-term positions at Comcast, Lawrence Berkeley National Labs, and Iowa State University. His work on light-trails has been widely referred to, deployed, and recognized by both industry and academia. His recent work on omnipresent Ethernet has been adopted by tier 1 service providers and also resulted in the largest ever acquisition between any IIT and the industry. This has led to a family of transport products under the premise of carrier Ethernet switch routers. He has 23 granted U.S. patents and has published about 150 papers in referred conferences and journals. He has also authored three books on broadband networks. For his contributions he was awarded the Government of India's DST Swaranajayanti Fellowship in 2013, the DAE-SRC Outstanding Research Investigator Award in 2010, the Vikram Sarabhai research award in 2012, the IBM Faculty award in 2012, the NASI-Reliance Industries Platinum Jubilee award in 2016, as well as the Indian National Academy of Engineeringís (INAE) Young Engineer Award in 2010.

Tamal Das is a research scientist at IIT Bombay. Prior to this, he was a postdoctoral researcher at the Technical University of Braunschweig, Germany. He received his Ph.D. from IIT Bombay, and B.Tech.+M.Tech. from IIT Delhi. His research interests are in stochastic analysis, telecommunication networks, and network algorithms. He has authored over 25 high-quality scientific publications, and was a recipient of the IEEE ANTS 2010 Best Paper Award.

Kandarp Khandwala is currently a Master's student at the University of California San Diego, specializing in human-computer interaction. He is the co-inventor of two U.S. patents applied for in the areas of online shopping cart abandonment and (predicting) permission email unsubscription. He received his Bachelor's degree in computer science and engineering from IIT Bombay in 2015, where he conducted research under Prof. Ashwin Gumaste. He was also a finalist for the Aditya Birla Scholarships Programme, 2011.

Inder Monga serves as the division director of the Scientific Networking Division, Lawrence Berkeley National Lab, and executive director of the Energy Sciences Network, a high-performance network interconnecting the National Laboratory System in the United States. In addition to managing the organization, his research interests include developing and deploying advanced networking services for collaborative and distributed big data science. Recently, his focus is on the broad adoption of SDN in the wide area network including recent work on operationalizing SDN, multi-layer SDN, and software-defined exchanges. He actively contributes to the Open Networking Foundation (ONF), the standards organization for SDN, as Chair of the ONF Research Associates and a member of the Open Source Software Leadership Council. He currently holds 20 patents and has over 20 years of industry and research experience in telecommunications and data networking.

A Closer Look at ROADM Contention

Jane M. Simmons

ROADMs provide the wavelength-switching capability in the optical layer of most transport networks. As the need for greater network configurability grows, ROADMs continue to evolve to provide greater flexibility. More specifically, ROADMs that provide colorless, directionless, gridless, and contentionless operation are on the roadmap of many equipment vendors and service providers. We examine ROADM contention in more detail in order to provide greater clarity regarding contentionless ROADMs.

ROADMs provide the wavelength-switching capability in the optical layer of most transport networks. As the need for greater network configurability grows, ROADMs continue to evolve to provide greater flexibility. More specifically, ROADMs that provide colorless, directionless, gridless, and contentionless operation are on the roadmap of many equipment vendors and service providers. We examine ROADM contention in more detail in order to provide greater clarity regarding contentionless ROADMs.

Introduction

Reconfigurable optical add/drop multiplexers (ROADMs) provide wavelength-granularity switching and the ability to add/drop individual wavelengths to/from a wavelength-division multiplexed (WDM) signal. Their ability to perform these operations in the optical domain has led to a dramatic reduction in the amount of electronics required at network nodes, thereby providing benefits in cost, power, space, and reliability. First deployed in carrier networks in the 2000 timeframe, ROADMs have undergone numerous architectural improvements over the past 15 years, as described in [1–4]. Most recently, the focus has been on adding operational flexibility to ROADMs, to better enable configurable (as opposed to quasi-static) networks.

Four properties have received the most attention with respect to flexible ROADMs: colorless, directionless, gridless, and contentionless. The first three of these properties are relatively simple to describe. *Colorless* indicates that any slot of the ROADM can accommodate a transponder of any wavelength. (A transponder is a transmit/receive card. A slot is the physical position on a ROADM shelf where the transponder card is inserted.) This is especially beneficial when using tunable transponders as it allows a transponder to be tuned to a different wavelength without having to manually move the transponder to a different ROADM slot. It also eliminates operational errors where a transponder is inserted in the "wrong" slot. *Directionless* refers to the ability of a transponder to access any of the network fibers that enter/exit a ROADM. Thus, a transponder that initially launches a connection on the eastbound fiber may later launch a connection on the westbound fiber, without requiring any manual intervention. This is useful for optical-layer restoration and for dynamic networking in general. *Gridless* indicates that the ROADM filters can accommodate a range of wavelength spacings and modulation formats. This is needed to support, for example, flexible transponders and elastic optical networks [5]. Gridless ROADMs are also referred to as bandwidth-variable ROADMs.

ROADM contention can take on many forms, leading to some confusion as to what constitutes a contentionless ROADM. We define the contentionless property as the ability of the ROADM to support any connection that can be supported by the current network configuration. Here we focus on wavelength contention (contention can also arise, for example, due to limits on the amount of add/drop at the ROADM [6]). In that context, contentionless operation implies that if a wavelength is available to be used on a network fiber, it will not be blocked from being used by the ROADM. Wavelength contention in a ROADM can largely be avoided through the use of effective algorithms (e.g., by treating the ROADM add/drop ports as "network links" in standard routing and wavelength-assignment (RWA) algorithms). Simulation studies that demonstrate the benefits of intelligent algorithms with regard to ROADM wavelength contention can be found in [7]. Nevertheless, ROADM wavelength contention still occurs when the network is heavily loaded and/or highly dynamic. There are also restoration schemes that have a tendency to engender wavelength contention, as discussed below. Thus, there is continuing effort to develop cost-effective contentionless ROADMs.

There are three canonical ROADM architectures (see [1–3] for more details of the architectures). For illustrative purposes regarding wavelength contention, we use the broadcast-and-select (B&S) architecture with optical splitters on the input ports and small wavelength-selective switches (WSSs) on the output ports. A directionless configuration is assumed, with the add/drop ports treated similarly to the network-fiber ports. (The add/drop ports are the interface to the "clients" in the electronic layer; the network-fiber ports are the interface to the fiber pairs that run between neighboring nodes.) Most of the contention scenarios enumerated below apply to the route-and-select (R&S) ROADM architecture as well, where WSSs are utilized on both the input and output ports. The third ROADM architecture, the wavelength-selective architecture, is based on a large centralized all-optical switch [8]. It is inherently contentionless and is not discussed further here. It has found

Digital Object Identifier:
10.1109/MCOM.2017.1600829CM

J. M. Simmons is with Monarch Network Architects.

limited deployment due to concerns about its reliability and scalability (although [8] presents a design that addresses this latter concern).

In the next section, we illustrate the ROADM contention that may occur when the number of add/drop ports is less than the number of network fibers. One of the more common misconceptions is that if the number of add/drop ports equals the number of network fibers, wavelength contention will not occur within the ROADM. We provide several examples of where this is not true. Another assumption that has appeared regarding ROADMs is that the combination of the colorless and directionless properties implies contentionless operation. This is shown to be an erroneous assumption as well.

It is important to understand the circumstances under which contention occurs in order to design contentionless ROADMs. There have been claims of ROADM architectures being contentionless when in fact they have been shown to exhibit contention under particular scenarios (an example of this is explored in [9]). As we illustrate the various forms of ROADM contention, we also include a discussion of how the contention can be avoided. The various architectures for avoiding contention have ramifications for cost and ancillary supported functionality. Finally, we address how contentionless operation is currently being designed into ROADMs.

Wavelength Contention Due to a Limited Number of Add/Drop Ports

The first example of wavelength contention is shown using the B&S ROADM of Fig. 1. The inputs to the ROADM are shown at the left of the figure; the outputs are shown on the right. The ROADM is assumed to be deployed at a degree-three node (the nodal degree indicates the number of input/output network fiber-pairs at the node).

Note that the add/drop ports carry WDM signals. While the number of add/drop ports typically equals the number of network fibers, in this particular example there are only two add/drop ports rather than three. This design decision may be made to reduce cost: only five WSSs are needed rather than six, and the WSSs are of size 5 × 1 as opposed to 6 × 1.

Assume that three connections arrive to this ROADM, where all three connections need to be dropped to the electronic layer (e.g., an IP router at the node). In addition, assume that all three connections have been routed using λ1 (clearly they must be routed on three different network fibers). Multiple connections on the same wavelength cannot be carried by a given add or drop port due to the resulting contention in the WDM signal. Thus, at most two of the λ1 connections can be accommodated at this node; the remaining connection will be blocked because of ROADM contention. The origin of the contention (i.e., too few add/drop ports) is readily seen in this simple example.

It should also be noted that the B&S ROADM in Fig. 1 is directionless. A transmitter/receiver on any add/drop port can access any of the network fibers. If the mux/demux operation on the add/drop ports (as represented by the trapezoids in Fig. 1) is provided by, for example, WSSs, the ROADM is also colorless. Thus, as illustrated by this example, and as stated earlier, the combination of the directionless and colorless properties does not imply the contentionless property. (In fact, in the earliest B&S ROADMs around 2000, the add/drop ports were simple taps off of each of the network fibers, resulting in a *non-directionless* architecture [1, 10]. This early architecture could be considered contentionless, though limited in flexibility, because an add/drop port carried the same WDM signal as its associated network fiber. Thus, if a wavelength was free to be used on a network fiber, it was free to be used on the add/drop port.)

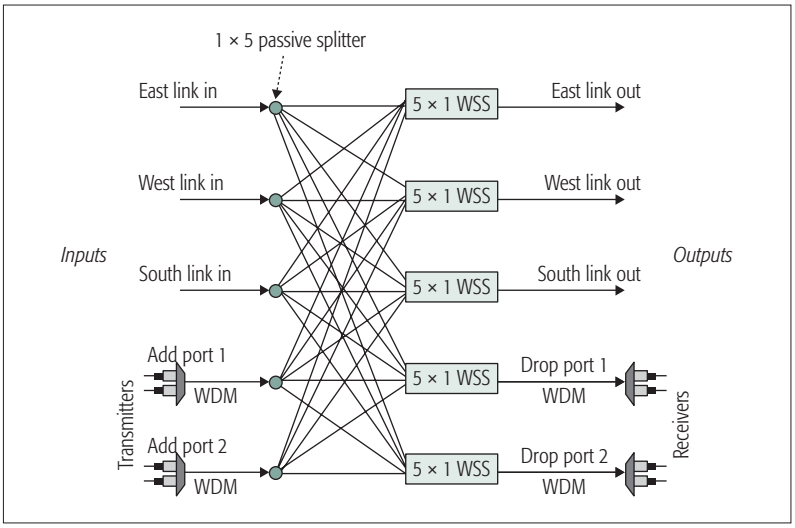


Figure 1. A B&S ROADM with three network input/output fiber pairs and two add/drop ports. The ROADM inputs are on the left; the outputs are on the right. If it is desired that three connections, all routed on λ1, be sourced/terminated at this node, one of these connections will be blocked due to wavelength contention on the add/drop ports.

To address the contention described above with respect to Fig. 1, one can add a third add/drop port to the design such that the number of add/drop ports equals the number of network fiber ports. This would allow three λ1 connections to be added/dropped. However, note that the optical splitter on each ROADM input enables multicast in the B&S architecture. For example, a signal that enters the ROADM from the East network fiber could be multicast to drop ports 1 and 2. If we modify the above scenario such that one of the three λ1 connections is required to be multicast to two different drop ports, having three drop ports is insufficient to prevent contention. Again, one of the desired λ1 connections will be blocked. More generally, if one considers multicast drops, even with a large number of drop ports, wavelength contention is not theoretically eliminated; however, if one has enough ports, such contention would be unlikely to occur in practice.

Wavelength Contention Due to Limited Edge Configurability

Client Access

An example of wavelength contention that cannot be ameliorated by an increase in the number of add/drop ports is shown in Fig. 2. For simplicity, a single network client is shown; that is, an

IP router that is connected to add/drop port 1. Assume that the IP router wants to establish two connections, one on the East link and one on the West link. Assume that λ1 is the only available wavelength on these two links. Then one of the connections will be blocked because two λ1 connections cannot be carried on add/drop port 1.

In this scenario, the wavelength contention arises from the lack of edge flexibility. The IP router interfaces to just a single add/drop port. In order to eliminate this type of contention, one solution is to have the router directly interface to transponders on all of the add/drop ports. Alternatively, an edge switch can be added to allow the IP router to access all of the add/drop ports. Adding an edge switch can also support other desirable functions, such as path restoration, dynamic connection establishment, and 1:*N* protection of the transponders.

There are two possible locations for inserting the edge switch, as shown in Fig. 3 (only the add side is shown in this figure). One option is to place the switch between the clients and the transponders, as shown in Fig. 3a. Alternatively, it could be placed between the transponders and the multiplexers/demultiplexers, as in Fig. 3b. Either option provides the edge configurability necessary to avoid the wavelength contention described above. However, there are subtle implications in the option that is chosen. First, there are likely to be differences as to how many transponders are needed to support a particular traffic pattern. The option that results in fewer required transponders depends on the traffic, as examined in more detail in [1]. Second, in Fig. 3a, the edge switch operates on a standard 1310-nm signal, thus allowing this switch to be electronic. In Fig. 3b, the edge switch operates on the WDM-compatible outputs of the transponders and must be an all-optical (i.e., photonic) switch. Because of concern regarding optical-amplifier transients, an all-optical switch will likely need to be switched more slowly than its electronic counterpart. Thus, from the perspective of restoration and rapid dynamic networking, the architecture of Fig. 3a may be preferred.

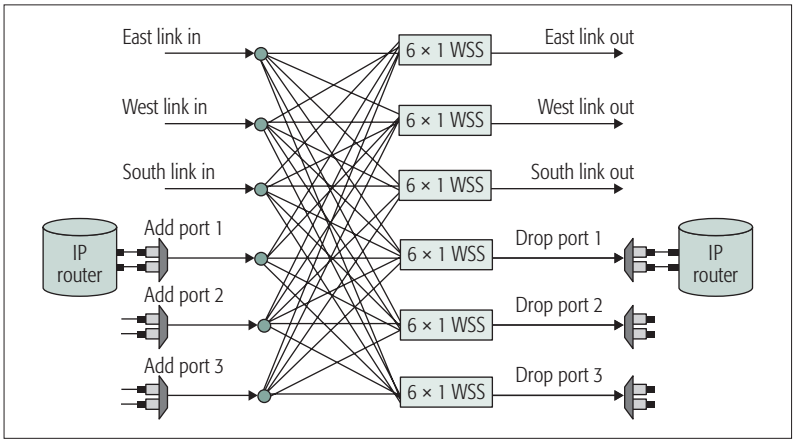


Figure 2. An IP router, which interfaces to add/drop port 1, wants to establish a connection on the East and West links. If λ1 is the only available wavelength on both of these links, one of the connections will be blocked due to wavelength contention on the add/drop port.

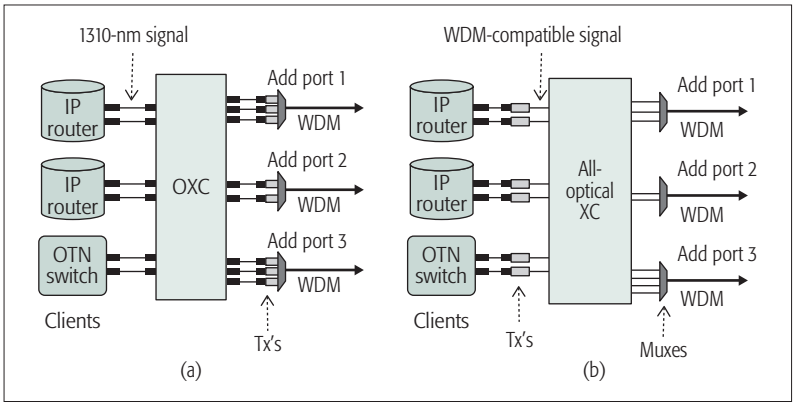


Figure 3. Adding an edge cross-connect to the ROADM. Only the add-side of the architecture, with the multiplexers (muxes) and the transmitter (Tx) portion of the transponders, is shown: a) the edge optical cross-connect (OXC) is placed between the clients and the transmitters, allowing the clients to direct traffic to any transmitter. This OXC can be electronic, as shown, or all-optical; b) the edge cross-connect (XC) is placed between the WDM-compatible output of the transmitters and the WDM multiplexers. The transmitters can access any add port. In this architecture, the XC must be all-optical, as shown.

Adding an edge switch is shown to be effective in reducing ROADM wavelength contention in the studies of [11, 12], where it is assumed that the edge switch is positioned as in Fig. 3a. Furthermore, these studies demonstrate that deploying multiple smaller edge switches is almost as effective at reducing blocking as deploying a single large edge switch. The modular architecture is advantageous in that it provides a "pay-as-you-grow" cost structure and avoids having one large cross-connect as a single point of failure.

Regeneration

Another example of wavelength contention that can be attributed to a lack of edge configurability arises when regenerator cards are used for regeneration. Regenerator cards are functionally similar to two back-to-back transponders, with the short-reach interfaces of the transponders eliminated to save cost. The signal to be regenerated enters the node on network fiber *x* using wavelength λ*i* and is dropped to the regenerator card. The signal is then added by the regenerator card using wavelength λ*j* and exits the node on network fiber *y*. Typically, a regenerator card permits λ*i* to be different from λ*j*. We assume that a regenerator card is inserted into a shelf slot associated with a single add/drop port. With this assumption, contention in the ROADM will occur when the desired λ*i* and λ*j* from the network-fiber perspective are not both available on that add/drop port.

There are a few solutions that can be considered to address this type of contention. First, it can be avoided (or at least minimized) if the ROADM architecture allows the two "halves" of the regenerator card to access different add/drop ports (e.g., via an optical backplane). This is a form of edge configurability, where the regenerator card would be connected to the add/drop port(s) with λ*i* and λ*j* available. Second, a separate edge cross-connect can be added as discussed above in relation to Fig. 3. However, note that the configuration of Fig. 3a is not compatible with regenerator cards because there is no short-reach interface (i.e., no 1310-nm signal) to feed into the edge switch. In contrast, the configuration of Fig. 3b can be utilized with regenerator cards, as

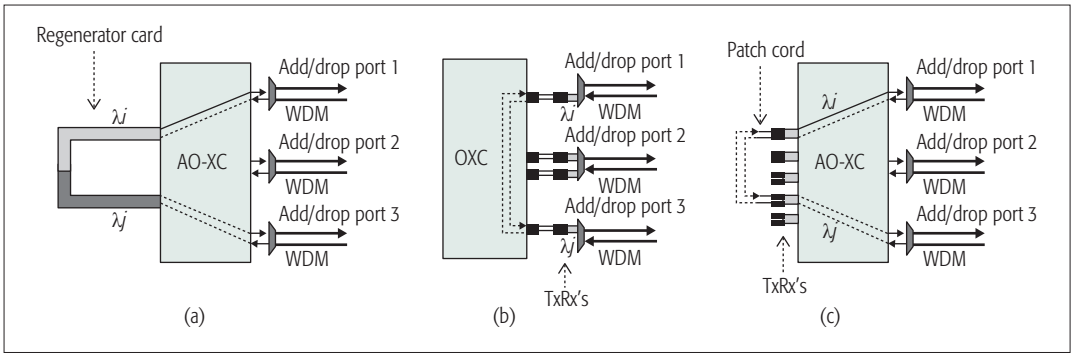


Figure 4. Regeneration at a ROADM equipped with an edge switch. Both add and drop ports are shown: a) a regenerator card in combination with an all-optical cross-connect (AO-XC). Four input/output ports on the AO-XC are utilized per regeneration; b) regeneration is achieved using two back-to-back transponders (TxRx's) interconnected by the OXC. Two input/output ports on the OXC are utilized per regeneration. (c) Two back-to-back transponders in combination with an AO-XC. Four input/output ports on the AO-XC are utilized per regeneration.

shown in Fig. 4a. This architecture utilizes four input/output ports on the edge cross-connect per regenerator card.

Alternatively, two discrete back-to-back transponders can be used for regeneration as opposed to the single regenerator card (thereby losing the cost benefit of the regenerator card). This architecture can be used in combination with the edge switch configuration of Fig. 3a, as shown in Fig. 4b. This provides flexibility with regard to utilizing the desired add/drop ports to avoid contention, as well as flexibility with regard to interconnecting transponders. Two input/output ports on the edge switch are utilized per regeneration with this architecture. There must be available transponders deployed on each add/drop port to ensure that the two interconnected transponders can support λi and λj without encountering wavelength contention.

If the edge switch utilizes the configuration of Fig. 3b, the regenerator architecture with back-to-back transponders is as shown in Fig. 4c. Note the similarity of Fig. 4c to Fig. 4a. The switch only provides flexibility with respect to the add/drop ports that are utilized. The transponders would need to be interconnected via patch cords, which is problematic if the transponders are not tunable (many wavelength combinations could be required). Furthermore, four input/output ports on the edge switch are utilized per regeneration with this configuration. Thus, this configuration addresses contention, but with less efficiency and somewhat less flexibility than the configuration of Fig. 4b.

Shared Restoration

As mentioned above, optical-amplifier transients can be problematic in an all-optical network when switching wavelengths in the optical domain, or when turning up or bringing down wavelengths on a fiber. To avoid transients when restoring traffic after a failure, various protection schemes based on pre-lit wavelengths have been proposed, where a set of pre-lit segments are stitched together at the time of failure in order to form a restoration path [13]. Because the segments are pre-lit, no change in power is experienced on the fiber links.

Using an edge switch to stitch together the pre-lit segments is straightforward, as described in [13]. However, if an edge switch is not deployed, the segments can be concatenated by reconfiguring the WSSs within the ROADM. Consider the example shown in Fig. 5. It is assumed that there are two pre-lit segments in this example, one carried on λ1 and one on λ2. The node shown in Fig. 5 serves as an endpoint of both pre-lit segments. Figure 5a depicts the baseline configuration, prior to any failure, with λ1 dropping to port 1 and λ2 dropping to port 3. Note that for each pre-lit segment, the receiver (on the right) is tied to its corresponding transmitter (on the left).

Assume that a failure occurs such that the two pre-lit segments need to be concatenated to form part of the restoration path. This can be accomplished by reconfiguring the WSSs on drop ports 1 and 3, with the resulting desired configuration shown in Fig. 5b. The wavelengths on the East and West network fiber pairs remain unchanged, thereby avoiding issues with optical transients. However, this concatenation can only be accomplished if λ1 is available on drop port 3 and λ2 is available on drop port 1. Thus, wavelength contention in the ROADM can potentially block the setup of the desired restoration path.

If an edge switch is used to perform the concatenation, the wavelengths on the network fibers and the add/drop ports are unchanged by the operation, so wavelength contention in the ROADM is not an issue. This is true for either configuration of Fig. 3. (With the configuration of Fig. 3b, the pre-lit-segment receivers need to be tied to the corresponding transmitters via a patch cord, similar to what is shown in Fig. 5. Additionally, the receivers utilized for the pre-lit segments need to be tunable.)

Sliceable Transponders

ROADM wavelength contention potentially has implications for the deployment of sliceable transponders, a technology that has been proposed to enable cost-effective elastic optical networks (EONs) [14, 15]. A single physical sliceable transponder supports multiple "virtual transponders," each of which may be assigned a flexible amount of spectrum (e.g., by adjusting the assignment of subcarriers). The "optical flows" represented by each of the virtual transponders can be routed independently in the network.

To avoid transients when restoring traffic after a failure, various protection schemes based on pre-lit wavelengths have been proposed, where a set of pre-lit segments are stitched together at the time of failure in order to form a restoration path. Because the segments are pre-lit, no change in power is experienced on the fiber links.

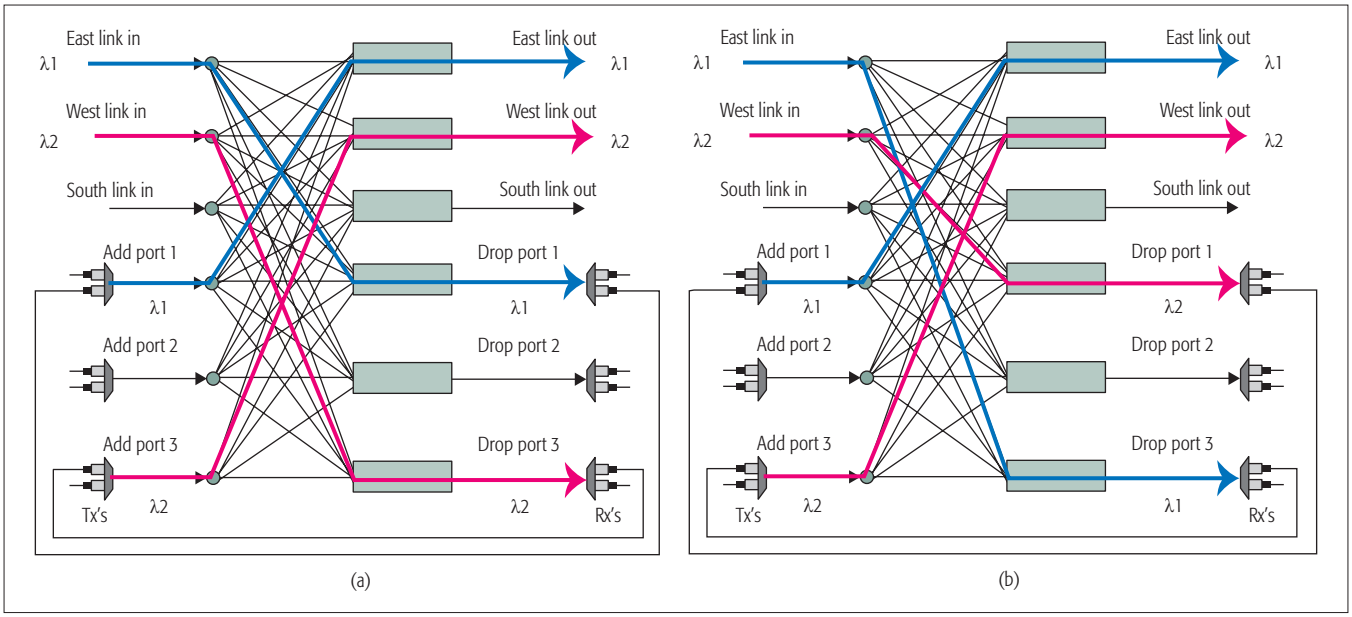


Figure 5. Shared restoration where pre-lit segments are concatenated to form the restoration path. Two pre-lit segments are shown, one on λ1 and one on λ2: a) the configuration prior to failure; b) the desired configuration after a failure occurs, where the λ1 segment must be concatenated to the λ2 segment. To support this operation, λ2 must be available on drop port 1 and λ1 must be available on drop port 3.

If sliceable transponders are used in an architecture where all optical flows that correspond to a given sliceable transponder are multiplexed together, "spectral contention" limits the flexibility of assigning spectrum to the flows. (Spectral contention in EONs is analogous to wavelength contention in conventional grid-based networks.) More specifically, the spectrum that is assigned to the set of virtual transponders associated with a given sliceable transponder cannot overlap. This is illustrated in Fig. 6a, where the edge cross-connect is located before the transponders on the add side of the ROADM (this corresponds to the architecture of Fig. 3a).

In contrast, the architecture shown in Fig. 6b potentially permits more flexibility in assigning spectrum to the flows (this corresponds to the architecture of Fig. 3b). Assuming that there are multiple interfaces between the sliceable transponder and the all-optical switch (as is shown), then the optical flows of a given sliceable transponder can be directed to different ports on the all-optical switch and ultimately to different add ports on the ROADM. This allows the optical flows of one sliceable transponder to be assigned overlapping portions of the spectrum (assuming the technology of the transponder supports this). Reference [15] includes simulations that probe the reduction in blocking probability that can be achieved with this additional flexibility in spectrum assignment.

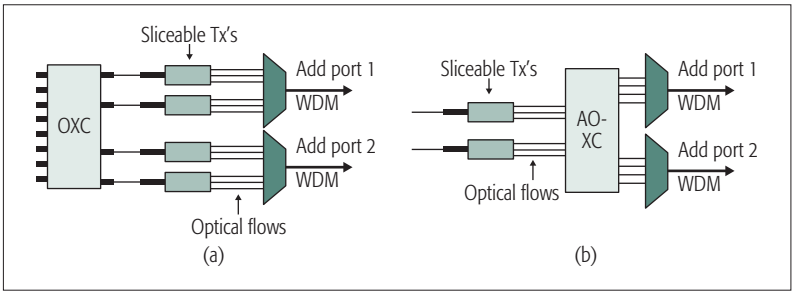


Figure 6. The impact of the edge cross-connect location on the spectrum assignment process when used with sliceable transponders. Only the add-side, with the transmitter (Tx) portion of the sliceable transponder, is shown: a) with the OXC located before the transmitters, the optical flows emanating from one sliceable transponder are muxed together onto one add port of the ROADM. The spectrum assigned to the flows must be non-overlapping; otherwise, spectral contention occurs on the add port; b) with the AO-XC located after the transmitters, the optical flows of one sliceable transponder may be directed to different add ports of the ROADM. This potentially allows the spectrum assigned to the optical flows of one sliceable transponder to overlap.

Wavelength Contention due to Limited Pre-Deployed Equipment

Another example of wavelength contention in a ROADM is portrayed in Fig. 7. In this scenario, we assume that dynamic services are supported by the network such that there must be equipment pre-deployed in the network; that is, connections must be established rapidly with no time for a "truck-roll" to install the required equipment. We assume that the pre-deployed transponders are as shown in the figure. Some of the transponders are assumed to be utilized for existing connections, as indicated. There are a total of two available transponders, one each on add/drop ports 1 and 2.

Assume that a dynamic connection needs to be established on the South link, and assume that the only free wavelength on this fiber is λ1. However, it is assumed that existing connections are already utilizing λ1 on add/drop ports 1 and 2. λ1 is available on add/drop port 3; however, there is no available transponder on this add/drop port. Thus, the new desired connection would be blocked. Note that adding an edge switch using the architecture of Fig. 3a does not address this type of wavelength contention, whereas the architecture of Fig. 3b does. Alternatively, pre-de-

ploying more transponders can minimize the frequency of such a scenario occurring, where the service provider must determine the proper tradeoff between the cost of pre-deploying extra transponders and the probability of blocking.

M × *N* WSS-Based Contentionless ROADM

As has been pointed out in the previous sections, adding edge configurability can avoid all, or at least most, instances of wavelength contention. The contentionless ROADMs that have been commercially developed thus far combine this edge configurability with the add/drop structure, where a single add/drop port equipped with an M × N WSS [16] replaces the individual add/drop ports shown in previous figures. An *M* × *N* WSS allows any of the *M* inputs to be directed to any of the *N* outputs of the WSS. *M* represents the number of transponders at the node, and *N* represents the number of network fiber pairs. See Fig. 8, where *N* equals 3.

On the input side, there is a one-to-one correspondence between the *N* input network fibers and the *N* inputs of the drop-side *N* × *M* WSS. Thus, if a wavelength is available on an input network fiber, it is available on the corresponding drop line as well. On the output side, there *potentially* could be pre-lit wavelengths on the outputs of the add-side *M* × *N* WSS that are not passed through to any of the output network fibers (perhaps in support of some type of dynamic networking scheme where pre-lit wavelengths are kept in a stand-by mode until needed to rapidly establish a new connection). Thus, *conceivably* there could be wavelength contention on the add lines. However, under normal network operation, this would not occur. Thus, this architecture is typically considered to be contentionless.

Of course, the *M* × *N* WSS itself must be contentionless in order for the ROADM to be contentionless. For example, constructing an *M* × *N* WSS from the combination of an *M* × 1 WSS feeding into a 1 × *N* WSS would lead to wavelength contention due to the single line connecting the constituent WSSs.

The *M* × *N* WSS architecture is similar, though not identical, to the edge-switch architecture of Fig. 3b. The *M* × *N* WSS essentially subsumes the AO-XC and mux/demux functionality. One aspect where the architectures potentially differ is with respect to multicast. If the *M* × *N* WSS is not capable of multicast, then multicast drop, where an incoming signal is dropped to multiple clients, is not supported in the architecture of Fig. 8. Multicast add, where an added signal is sent to multiple output network fibers, would not be supported either. In contrast, the architecture of Fig. 3b in combination with a B&S ROADM supports both of these multicast operations, though its support of multicast drop may be limited to the scenario where the clients are located on different drop ports.

The ROADM of Fig. 8 is also colorless and directionless. Such a ROADM is often denoted as CDC to indicate that it is colorless, directionless, and contentionless. The gridless property depends on the flexibility of the filter technology used in the WSSs.

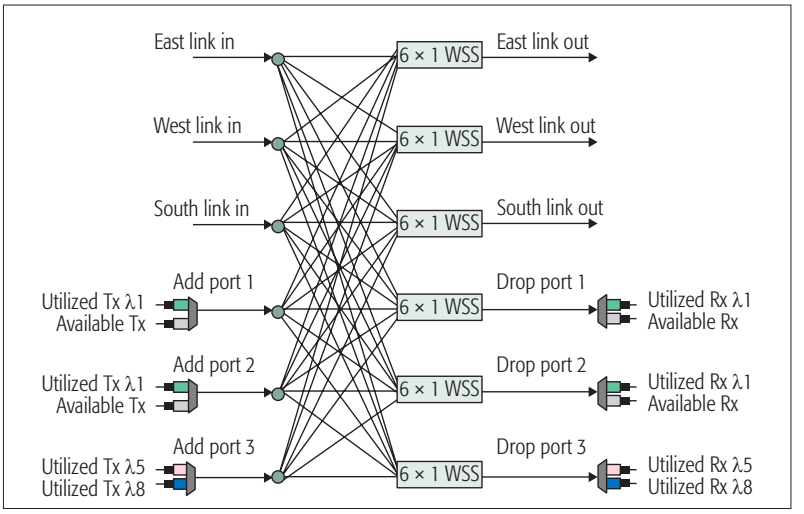


Figure 7. Assume that a new connection is desired on λ1. The only available transponders (TxRx's) are located on add/drop ports 1 and 2. However, the only add/drop port with λ1 available is port 3. Thus, the new connection would be blocked.

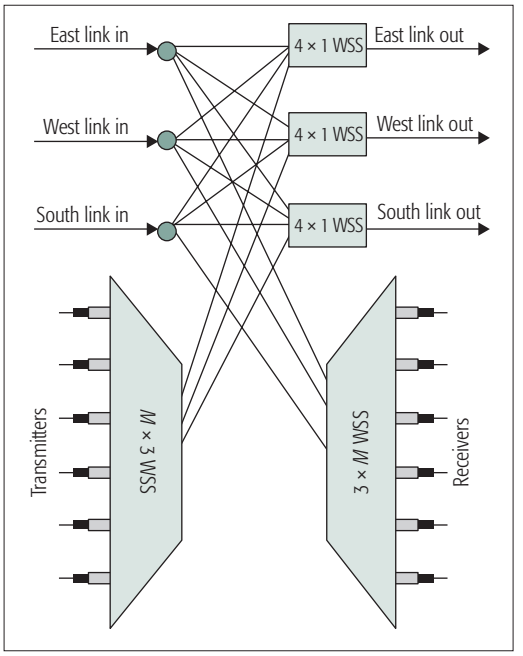


Figure 8. An *M* × *N* WSS is deployed to provide contentionless operation. Alternatively, an *M* × *N* MCS can be used.

An alternative architecture that is likely more practical replaces the *M* × *N* WSS with an *M* × *N* multicast switch (MCS) [17]. An MCS provides benefits in cost and space as compared to the WSS; it also supports multicast drop. However, it requires the receivers to be capable of selecting the desired wavelength from a WDM signal, and it typically incurs greater loss.

Conclusion

We have described a variety of situations where wavelength contention can occur within a ROADM, along with architectures to avoid the contention. No single architecture is perfect in addressing all of the potential scenarios that may arise. Furthermore, some of the solutions may add substantial cost to the ROADM. As noted in

The contentionless property is thus somewhat less crucial as compared to the colorless, directionless, and gridless properties. Nevertheless, simplifying network operation is desirable if it can be accomplished at reasonable cost; hence, further research into architecting cost-effective contentionless ROADMs is warranted.

the introduction, many of the contention scenarios can be minimized through the use of effective algorithms. The contentionless property is thus somewhat less crucial as compared to the colorless, directionless, and gridless properties. Nevertheless, simplifying network operation is desirable if it can be accomplished at reasonable cost; hence, further research into architecting cost-effective contentionless ROADMs is warranted.

References

[1] J. M. Simmons, *Optical Network Design and Planning*, 2nd ed., Springer, 2014, Chapter 2.

[2] S. Gringeri *et al.*, "Flexible Architectures for Optical Transport Nodes and Networks," *IEEE Commun. Mag.*, vol. 48, no. 7, July 2010, pp. 40–50.

[3] B. C. Collings, "Advanced ROADM Technologies and Architectures," *Proc. OFC*, Los Angeles, CA, 2015, Paper Tu3D.3.

[4] S. Perrin, "Next-Generation ROADM Architectures and Benefits," Heavy Reading White Paper, Mar. 2015; www.fujitsu.com/us/Images/Fujitsu-NG-ROADM.pdf, accessed Dec. 8, 2016.

[5] S. Poole *et al.*, "Bandwidth-Flexible ROADMs as Network Elements," *Proc. OFC/NFOEC*, Los Angeles, CA, 2011, Paper OTuE1.

[6] F. Naruse *et al.*, "Evaluations of OXC Hardware Scale and Network Resource Requirements of Different Optical Path Add/Drop Ratio Restriction Schemes," *J. Opt. Commun. Net.*, vol. 4, no. 11, Nov. 2012, pp. B26–34.

[7] P. Palacharla *et al.*, "Blocking Performance in Dynamic Optical Networks Based on Colorless, Non-Directional ROADMs," *Proc. OFC/NFOEC*, Los Angeles, CA, 2011, Paper JWA8.

[8] J. M. Simmons and A. A. M. Saleh, "Wavelength-Selective CDC-ROADM Designs Using Reduced-Sized Optical Cross-Connects," *Photon. Tech. Lett.*, vol. 27, no. 20, Oct. 15, 2015, pp. 2174–77.

[9] T. Zami, P. Jenneve, and H. Bissessur, "Fair Comparison of the Contentionless Property in OXC," *Proc. Asia Commun. and Photonics Conf.*, Hong Kong, 2015, Paper AM3G.3.

[10] A. Tzanakaki, I. Zacharopoulos, and I. Tomkos, "Optical Add/Drop Multiplexers and Optical Cross-Connects for Wavelength Routed Networks," *Proc. Int'l. Conf. Transparent Optical Networks*, Warsaw, Poland, 2003, Paper Mo.B2.4, Fig. 4b.

[11] T. Zami, "Contention Simulation within Dynamic, Colorless and Unidirectional/Multidirectional Optical Cross-Connects," *Proc. Euro. Conf. Expo Optical Commun.*, Geneva, Switzerland, 2011, Paper We.8.K.4.

[12] I. Kim *et al.*, "Performance of Colorless, Non-Directional ROADMs with Modular Client-Side Fiber Cross-Connects," *Proc. OFC/NFOEC*, Los Angeles, CA, 2012, Paper NM3F.7.

[13] J. M. Simmons, "Cost vs. Capacity Tradeoff with Shared Mesh Protection in Optical-Bypass-Enabled Backbone Networks," *Proc. OFC/NFOEC*, Anaheim, CA, 2007, Paper NThC2.

[14] O. Gerstel, "Flexible Use of Spectrum and Photonic Grooming," *Proc. Int'l. Conf. Photonics in Switching*, Monterey, CA, 2010, Paper PMD3.

[15] M. Dallaglio *et al.*, "Add and Drop Architectures for Multicarrier Transponders in EONs," *J. Opt. Commun. Net.*, vol. 8, no. 7, July 2016, pp. A12–22.

[16] B. C. Collings, "Wavelength Selectable Switches and Future Photonic Network Applications," *Proc. Int'l. Conf. Photonics in Switching*, Pisa, Italy, 2009, pp. 52–55.

[17] W. I. Way, "Optimum Architecture for $M \times N$ Multicast Switch-Based Colorless, Directionless, Contentionless, and Flexible-Grid ROADM," *Proc. OFC/NFOEC*, Los Angeles, CA, 2012, Paper NW3F.5.

Biography

Jane M. Simmons [F'11] has been involved in the research and development of optical networks for 20 years. She currently is the founder of Monarch Network Architects, which provides optical network architectural services and design tools. She served as the subject matter expert on optical networking for DARPA for six years, and made significant technical contributions to DARPA's CORONET program on dynamic, highly reliable networks. From 1999 to 2002, she was the executive engineer of network architecture and later the chief network architect of Corvis Corp. While at Corvis, she performed the network design and link engineering for the first commercially deployed all-optical backbone network. Through pioneering algorithmic and architectural optimizations, she played a significant role in the adoption of all-optical networking in telecommunications networks. Prior to Corvis, she worked at Bell Labs/AT&T Labs Research, where she conducted research on backbone, regional, and broadband access networks. She received a B.S., Summa Cum Laude, from Princeton University, and S.M. and Ph.D. degrees from MIT, all in electrical engineering. She is a member of Phi Beta Kappa, Tau Beta Pi, and Sigma Xi honor societies. She has served on the Technical Program Committees of several conferences, including OFC and ICC. She was the OFC Networks Subcommittee Chair in 2003, and was a member of the OFC Steering Committee for six years. From 2004 to 2009, she was an Associate Editor of the *IEEE JSAC* Optical Communications and Networking Series. From 2009 to 2012, she was an Associate Editor for the *Journal of Optical Communications and Networking*, and is currently on its Steering Committee. She teaches a course on optical network design at OFC, and is the author of the textbook *Optical Network Design and Planning*, now in its second edition.

CALL FOR PAPERS

IEEE COMMUNICATIONS MAGAZINE

INTERNET OF THINGS AND INFORMATION PROCESSING IN SMART ENERGY APPLICATIONS

BACKGROUND

There are four major challenges for the current electricity grid – increasing electricity demand, ageing grid infrastructure, ever-increasing penetration of renewables, and significant uptake of electric vehicles and energy storage with behind-the-meter applications for residential and commercial buildings. To address these challenges, there must be strong and low-cost communications infrastructures that can support rapid and secure information exchange as well as consistent and efficient design of communication protocols and architectures to enable automation and effective use of smart energy resources. Internet of Things (IoT) could accelerate establishment of such infrastructures. With IoT technologies, a lot more devices could be controlled and managed through the Internet, and data pertaining to the grid, commercial buildings, and residential premises can be readily collected and utilized. To derive valuable information from the data, further information and data processing becomes essential. There are, however, a number of challenges to be addressed.

This Feature Topic (FT) aims to disseminate general ideas extracted from cutting-edge research results spanning multiple disciplines. Potential authors will be able to share various viewpoints and the latest findings from research and ongoing projects relevant to smart energy applications from the perspectives of IoT and advanced information processing and communications technologies. Topics of interest include, but are not limited to:

- Machine-to-Machine (M2M) or Vehicle-to-Grid (V2G)/home communications in smart grid
- Smart metering including Behind-The-Meter (BTM) applications and autonomous demand response
- Advanced Metering Infrastructure (AMI) and information processing in smart grid
- Intelligent Energy Management System (EMS) for utility grid and commercial/residential buildings
- IoT-enabled smart energy applications, security mechanisms, and architectures
- Big data analysis and mining for IoT-based smart grid, city, and home

SUBMISSIONS

Articles should be tutorial in nature, with the intended audience being all members of the communications technology community. They should be written in a style comprehensible to readers outside the specialty of the article. Mathematical equations should not be used (in justified cases up to three simple equations are allowed). Articles should not exceed 4500 words (from introduction through conclusions, excluding figures, tables and captions). Figures and tables should be limited to a combined total of six. The number of archival references is recommended not to exceed fifteen (15). In general, however, mathematics should be avoided; instead, references to papers containing the relevant mathematics should be provided. Complete guidelines for preparation of the manuscripts are posted at http://www.comsoc.org/commag/paper-submission-guidelines. Please submit a pdf (preferred) or MSWORD formatted manuscript via Manuscript Central (http://mc.manuscriptcentral.com/commag-ieee). Register or log in, and go to Author Center. Follow the instructions there. Select "October 2017/IoT and Information Processing in Smart Energy Applications " as the Feature Topic category for your submission.

IMPORTANT DATES

- Manuscript Submission Deadline: March 1, 2017
- Decision Notification: June 15, 2017
- Final Manuscripts Due: July 15, 2017
- Publication Date: October 2017

GUEST EDITORS

Wei-Yu Chiu
Yuan Ze University, Taiwan
wychiu@saturn.yzu.edu.tw

Hongjian Sun
Durham University, UK
hongjian.sun@durham.ac.uk

Shunqing Zhang
Intel Corporation, China
shunqing.zhang@intel.com

John Thompson
Edinburgh University, UK
john.thompson@ed.ac.uk

Kiyoshi Nakayama
NEC Laboratories America, USA
knakayama@nec-labs.com

ACCEPTED FROM OPEN CALL

Wide-Area Wireless Communication Challenges for the Internet of Things

Harpreet S. Dhillon, Howard Huang, and Harish Viswanathan

The authors discuss the need for wide-area M2M wireless networks, especially for short data packet communication to support a very large number of IoT devices. They present a brief overview of current and emerging technologies for supporting wide area M2M, and then using communication theory principles, discuss the fundamental challenges and potential solutions for these networks, highlighting tradeoffs and strategies for random and scheduled access.

ABSTRACT

The deployment of Internet of Things (IoT) devices and services is accelerating, aided by ubiquitous wireless connectivity, declining communication costs, and the emergence of cloud platforms. Most major mobile network operators view machine-to-machine (M2M) communication networks for supporting IoT as a significant source of new revenue. In this article, we discuss the need for wide-area M2M wireless networks, especially for short data packet communication to support a very large number of IoT devices. We first present a brief overview of current and emerging technologies for supporting wide area M2M, and then using communication theory principles, discuss the fundamental challenges and potential solutions for these networks, highlighting tradeoffs and strategies for random and scheduled access. We conclude with recommendations for how future 5G networks should be designed for efficient wide-area M2M communications.

INTRODUCTION

The Internet of Things (IoT) consists of a network of physical devices connected with remote computational capabilities. By combining physical sensing with data analysis to create meaningful information, IoT platforms enable solutions in the realms of smart cities, smart grids, smart homes, and connected vehicles that could provide a significant qualitative improvement in people's lives [1]. IoT has also been touted as an economic engine for growth as it increases productivity, reduces cost, and improves lives. The growth in the number of IoT devices deployed is accelerating as the concept gathers broader industry momentum. General Electric has estimated the impact of the "Industrial Internet" on the world economy to be about $15 trillion [2], and analysts project that IoT related addressable revenue for mobile network operators worldwide could be $255 billion by 2020.

The notion of IoT is broad and encompasses many technologies including near field, short range and wide-area communication networks; device to device communication; device technologies for sensing, actuation, and energy harvesting; device and application software platforms for big data, security, streaming analytics, and cloud processing. The three main components of most IoT-enabled applications are the devices, the network, and the application servers. The devices sense a physical characteristic of the environment (e.g., temperature or presence of an object) and send the information through a communication network. The data is aggregated and processed by servers to provide meaningful information or an actionable output. This output could be sent back through the network to trigger a set of actuator devices (e.g., a switch for triggering a motor or alarm). This basic framework could enable new classes of applications and servers such as the management of autonomous vehicle fleets, enabling more energy efficient cities and homes, and the ubiquitous tracking of assets.

The communication network is often known as a machine-to-machine (M2M) network to distinguish it from networks that relay traffic generated or consumed by humans. While the network in general consists of wired and wireless devices, the trend is for devices to be wirelessly connected to the network edge to enable lower-cost installation, easier physical reconfiguration, and mobile applications.

IoT applications using wireless communications are highly varied and differ in their requirements. From a networking perspective, classical IoT applications can be categorized along two dimensions of range and mobility. Range refers to the geographic spread of the devices. It describes whether the devices are deployed in a small area, say within a couple of hundred feet of each other, or are dispersed over a wider area. Mobility refers to whether the devices move and if so, whether they need to communicate while on the move. Table 1 shows the five categories of applications spanning several orders of magnitude differences in range. For each category, it shows the basic device characteristics, services, and suitable networks.

For localized IoT applications, a short-range network is the most appropriate, allowing the use of unlicensed spectrum and maximizing battery life while meeting networking needs. For example, many smart home applications for environment control and monitoring would be well served using a wireless sensor network based on 802.11 or Zigbee. Shorter-range applications can be enabled using Bluetooth or NFC. The smartphone can be used as a hub to enable personal IoT applications such as health monitoring and local object tracking. Bluetooth is often used to connect to IoT devices, and an 802.11 or cellular connection provides network access.

For wide-area IoT applications such as the con-

Digital Object Identifier:
10.1109/MCOM.2017.1500269CM

Harpreet S. Dhillon is with Virginia Tech; Howard Huang and Harish Viswanathan are with Nokia Bell Labs.

Application	Range	Mobility	Device characteristics	Service characteristics	Suitable networks
• Connected car • Fleet management • Remote health monitoring	~1000m	Yes	Rechargeable battery	Managed service, highly secure	• Cellular • Satellite
• Smart metering • Parking meter	~1000m	No	Low rate, low power, low cost	Managed service	• Cellular • Dedicated network
• Hospital asset tracking • Warehouse logistics	~100m	Yes	Low rate, low power, low cost	Enterprise-deployed	• WiFi • RFID
• Industrial automation • Home automation	~10m	No	Low rate, low power, low cost	Subscription-free	• Zwave • Zigbee • Wifi • Powerline
• Personal activity • Local object tracking • Point of sale	~1m	No	Low rate, low power, low cost	Subscription-free	• Bluetooth • NFC

Table 1. M2M application categories. In this article, we focus on applications in the top two rows which have a required range of about 1000m for wide-area coverage. Applications in other rows have more established ecosystems.

The number of IoT devices per cell could be significantly larger than the number of mobile devices per cell if multiple devices are associated with each person, car and building and if additional devices are deployed throughout the environment [5]. For a given set of radio resources, more devices require improved efficiency of both the control and data planes.

nected car or fleet tracking, a mobile broadband network is more suitable because devices move over a wide area. For applications such as metering where the devices are widespread but there is little need for mobility, a wide area network is required but does not have to support seamless mobility. Although the mobile network meets the requirements for this category of applications, a dedicated network that is designed for low data rates without complex mobility management procedures can be significantly cheaper, have greater reach into buildings, and provide a substantially longer battery life for devices.

For the remainder of this article, we focus on wide-area wireless M2M communication at the physical and access layers. We describe the challenges in more detail. We briefly describe M2M solutions in current cellular standards and dedicated M2M networks. We consider the fundamental design strategies from information theory and communication theory perspective for a possible clean slate design. We conclude with thoughts on how future 5G cellular networks could be designed to accommodate M2M communication more efficiently.

Wide-area M2M Communication Challenges

In order to meet the demands of high data rate applications such as video streaming, conventional mobile broadband cellular networks are designed for high system capacity, measured in terms of data rate per unit area (bps/km^2). This capacity can be increased by using more spectrum, increasing the density of base stations, or increasing the spectral efficiency (bps/Hz) of each base station. A typical LTE base station using 20 MHz bandwidth would serve a few dozen active handset devices, each operating at a rate of up to several Mbps as opposed to a base station in a M2M network that needs to serve a very large number of devices at low rates. In contrast to low cost IoT devices, handsets are capable of performing sophisticated signal processing, and the handset battery can be charged frequently, even on a daily basis if needed.

The challenges of designing a wide-area M2M communication network are different from those of a conventional broadband network because of the characteristics of IoT applications and the constraints imposed by low-cost and low-complexity IoT devices. These differences affect the assumptions and performance metrics of the system design, and potentially motivate novel designs at the PHY and MAC layer. We highlight some key attributes of M2M networks and describe their impact on the system design.

Small payloads. In conventional broadband streaming or high data rate applications, it makes sense to invest in control overhead to establish bearers for scheduled transmission, as is done in LTE networks. In many IoT applications such as meter reading or actuation, the payload could be relatively small (~1000 bits), consisting of an encrypted device ID and a measurement or actuation command [3]. For small payloads, the control overhead for scheduled transmission may not be justified, and thus the traditional connection oriented approach of establishing radio bearers prior to data transmission will be inefficient for M2M [4]. Different IoT applications could have different latency and reliability requirements, which will impact the optimal design. For example, a meter reading for water consumption would have a longer latency requirement than a sensor for detecting a basement flood condition.

Large number of devices. The number of IoT devices per cell could be significantly larger than the number of mobile devices per cell if multiple devices are associated with each person, car and building, and if additional devices are deployed throughout the environment [5]. For a given set of radio resources, more devices require improved efficiency of both the control plane and the data plane.

Bursty demand. Certain IoT applications could exhibit highly bursty and correlated service

Recently 3GPP has been considering a number of alternatives for M2M that are based on introducing modifications to the GSM and LTE standards to meet the M2M requirements of low complexity, larger coverage, and lower device cost.

requests. For example, a sudden severe storm could activate a large number of flood sensors, which would otherwise send only infrequent heartbeat signals.

Extended range. Some sensor and actuation applications would require coverage in areas beyond a conventional cellular network, such as in basements. The system could be designed to account for an extended link budget of, say, 20 dB. Alternatively, if the extended coverage is not required, the improved link budget could be used to reduce the M2M infrastructure required to serve a given cellular coverage area and to reduce the device cost by reducing the maximum device transmit power.

Enhanced device energy efficiency. In contrast to conventional cellular devices, which can be charged on a daily basis, devices for many IoT applications may not be amenable to frequent charging [4]. Improving the device energy efficiency would reduce the operational expense of recharging or replacing batteries. One could also seek alternative energy sources such as energy harvesting from vibrational sources or light. These alternatives are often intermittent or unreliable, requiring novel approaches for resource allocation.

Reduced device cost. Compared to conventional cellular devices, IoT devices would have limited functionality and should cost less. The devices would probably not use multiple antennas, and lower-quality RF components could result in degraded link budgets. Reduced baseband processing may not allow for sophisticated decoding or encoding and could also limit encryption techniques. With reduced-complexity RF components, the devices could be restricted in the number of bands they operate on, possibly limiting global roaming capabilities.

With these insights, we now briefly describe M2M solutions in cellular standards and dedicated M2M networks.

Wide-Area M2M Technologies

Mobile communication systems standardized in 3GPP have primarily targeted feature phones, smart phones, and tablets. Over the last three decades, multiple generations (2G/3G/4G) of technologies have been standardized and deployed for voice and data communications. The capability for data communication has been exploited to use these technologies for M2M communications. Among the 3GPP technologies, GSM is the most widely used for M2M. Although GSM supports only low data rates, it is sufficient for many M2M applications. Since GSM has been operational for over two decades, device costs are substantially lower than that for 3G and 4G. Furthermore, GSM typically has better coverage than 3G and 4G in most parts of the world.

Recently 3GPP has been considering a number of alternatives for M2M that are based on introducing modifications to the GSM [6] and LTE standards to meet the M2M requirements of low complexity, larger coverage, and lower device cost.

GSM enhancements for M2M. Motivated by the low cost of GSM devices, some companies are in favor of making GSM/GPRS the de facto M2M technology. This approach, known as extended coverage (EC) GSM, seeks to make GSM more efficient for M2M by increasing uplink capacity, extending coverage of downlink for both the control channel and the data channel by 20 dB, reducing power consumption of M2M devices compared with legacy GSM devices, and reducing the complexity of the devices compared to that of legacy devices. In order to allow more devices to transmit at the same time in the same frequency in uplink, multiplexing based on overlaid code division multiple access technique is proposed where orthogonal codes are used to separate the devices simultaneously transmitting in the same time slot. Coverage is enhanced for control, synchronization, system information, and data channels, and is essentially achieved through repetition, with different repetition levels based on the coverage class the device belongs to. Other enhancements include definition of new control messages with smaller payload sizes and introduction of a new lower-power class.

Novel narrow band air-interface. Some proponents are making a case for a new narrow band air-interface that is compatible with GSM channelization of 200 KHz. In this clean slate proposal a new system is defined that is optimized for IoT. Narrow bandwidth channels based on frequency division multiplexing are defined both on the uplink and downlink within the 200 KHz bandwidth allowing frequency reuse with only 200 KHz of spectrum. Uplink channel bandwidth is 3.75 KHz with 5 KHz channel spacing and channel bonding of two or four channels is allowed for higher data rates. On the downlink channel, bandwidth of 15 KHz is proposed. FDMA access is preferred over CDMA since synchronization and closed loop power control can be avoided. Block repetition and symbol spreading CDMA are also used for extended coverage. The base station operates in RF full duplex mode in order to maximize network capacity while the devices operate in half duplex mode to reduce the RF cost. Devices could either support GMSK modulation or linear modulation such as BPSK, QPSK, and 8-PSK.

Enhancements to LTE. 3GPP standards have also incorporated a number of enhancements to the LTE for M2M since Release 11. These enhancements include creating a low priority access indicator for MTC devices facilitating access barring of only MTC devices under overload; introduction of a power saving state to increase battery life of devices that only communicate sporadically; signaling load minimization for static devices by minimizing frequency of mobility signaling such as tracking area updates; and introduction of the MTC trigger function to wake up devices [7]. Additional enhancements are being studied in Release 13, with the objective of enhancing coverage and reducing device cost compared to existing LTE networks. In the proposed approach known as NB-LTE, a 1.4 MHz narrowband version of LTE will be used by MTC devices in downlink and uplink within any system bandwidth [8]. In order to reduce cost, the devices will support a reduced set of transmission modes, for example, excluding MIMO modes and high data rates, and will have reduced maximum transmit power of 20 dBm.

A number of dedicated network technologies optimized for wide-area M2M commu-

nications have also been recently developed for dispersed, static/portable devices with low throughput requirements. Examples of these technologies include Ingenu random polarization multiple access (RPMA), Sigfox, Weightless-N, and LoRa. Each of these target an extended range compared to conventional bases stations for broadband access, with Ingenu and LoRa achieving it through spread spectrum, and with Sigfox and Weightless-N using ultra narrow bandwidth channelization (about 200Hz). Because of the extended range, a metropolitan area can be covered using fewer bases. In addition, the dedicated M2M devices are often lower power, operate at lower clock rates, and have lower cost compared to 3GPP devices. One difference between Sigfox and Weightless-N is that the latter supports only uplink communication. The LoRa technology differs from both by using a wider bandwidth (125 KHz) chirp spread spectrum signal, enabling tradeoffs between extending range by lowering data rate. While there are many vendors offering such specialized networks for M2M, there is no single global standard. All four dedicated M2M networks operate in unlicensed spectrum (2.4 GHz for Ingenu RPMA and sub-1 GHz), which could make service quality requirements difficult to guarantee. Table 2 shows a comparison of the cellular and dedicated M2M technologies [9–12].

Carrier frequency		Technology	Channel bandwidth	Representative data rate	Link budget target or max. range
Licensed cellular		LTE Cat. 0	20 MHz	DL: 1 Mb/s UL: 1 Mb/s	140 dB
		LTE Cat. M	1.4 MHz	DL: 1 Mb/s UL: 1 Mb/s	155 dB
		NB-IoT	200 kHz	DL: 128 kb/s UL: 64 kb/s	164 dB
		EC-GSM	200 kHz	DL: 74 kb/s UL: 74 kb/s	164 dB
Unlicensed	2.4 GHz	Ingenu RPMA	1 MHz	UL: 624 kb/s DL: 156 kb/s	500 km line of sight
	Sub-1 GHz	LoRa chirp spread spectrum	125 kHz	UL: 100 kb/s DL: 100 kb/s	15 km rural 5 km urban
	Sub-1 GHz	Weightless-N	200 Hz	UL: 100 b/s	3 km urban
	Sub-1 GHz	Sigfox	160 Hz	UL: 100 b/s	50 km rural 10 km urban

Table 2. Comparison of wide-area M2M technologies.

Theoretical Foundations of Wide-area M2M Communications

Massive machine communication is expected to be one of the main requirements for 5G. Since 5G will not be constrained by any backward compatibility, it is valuable to step back and study the problem from a fundamental communication theory perspective, which is done next.

Problem Statement

As discussed in the previous sections, many applications for M2M do not require the low latency or high data rates of conventional broadband. On the contrary, the central goal for M2M system design is *massive access management*, which is the main focus of this section. This shifts the design focus from downlink (current networks) to uplink in the cellular-based M2M. For the system model, we assume a single cell with a base station at the center and M2M devices uniformly distributed in the cell. We assume that for a given M2M application, each device attempts to communicate data in a random, bursty manner that can be modeled as a Poisson process with a common mean arrival rate. This assumption is relevant for a broad class of sensing applications where the underlying events occur randomly. If we further assume the devices transmit independently, then the aggregate transmissions of devices in a cell can likewise be modeled as a Poisson process with an aggregate mean arrival rate λ given by the sum of the individual mean arrival rates. For simplicity, we assume each transmitting device has the same payload of L bits that needs to be communicated to the base station in time T using the system bandwidth of W Hz. The main goal of this section is to determine the transmit power per device needed to support a given aggregate arrival rate λ at the base station as a function of W, T and L under various transmission strategies. This can also be visualized as the maximum arrival rate that can be supported for a given power constraint. While similar problems can be easily posed in terms of other performance metrics of interest, such as energy efficiency, we limit our discussion to the above setup due to space constraints. Interested readers are referred to [13] for a discussion of improving the energy efficiency of uplink transmissions in an LTE network.

Transmission Approaches

Transmission approaches can be categorized into two main classes depending upon whether the devices are transmitting over dedicated resources or over a *shared* random access channel (RACH). The former is termed *scheduled transmission* and requires that the base station already has information about the number of devices (say K) requesting to transmit in the current resource slice along with their channel gains. The latter is unscheduled (or simply RACH) transmission in which the base station does not have any information about the transmitting devices. In this work, we assume that while the base station does not know the exact number of devices in RACH, it can estimate the *average arrival rate* λ from previous time slots, and that this estimate is accurate. In the classical systems, RACH is typically used to initiate a connection with the base station by transmitting control information, such as the device identity. However, in M2M communications, since the payloads are small, it may be "optimal" to send these payloads along with control information on RACH itself. As a result, we first look at the RACH and scheduled transmissions separately and then consider a two-stage design in which the control information is first sent over RACH (stage 1) and the data is then sent through scheduled transmission (stage 2). More details about the two-stage design will be provided later.

Random Access Channel (RACH)

Optimal RACH strategy. Given K devices that have data to transmit in the current resource slice, each transmits with probability θ. Each device encodes its data with one of Q randomly chosen codebooks, each consisting of $2^{L/(WT)}$ codewords. The message is prepended with a short preamble that identifies the codebook. Each device transmits over full bandwidth W Hz for the slice duration of T sec. The receiver then jointly decodes the largest set of devices whose rate vector is in the corresponding capacity region, while treating other devices as interference. Optimizing this strategy over θ results in throughput-optimal uncoordinated performance. For other considerations, such as retransmissions, and more formal details, please refer to [14] where this strategy was recently proposed by the authors.

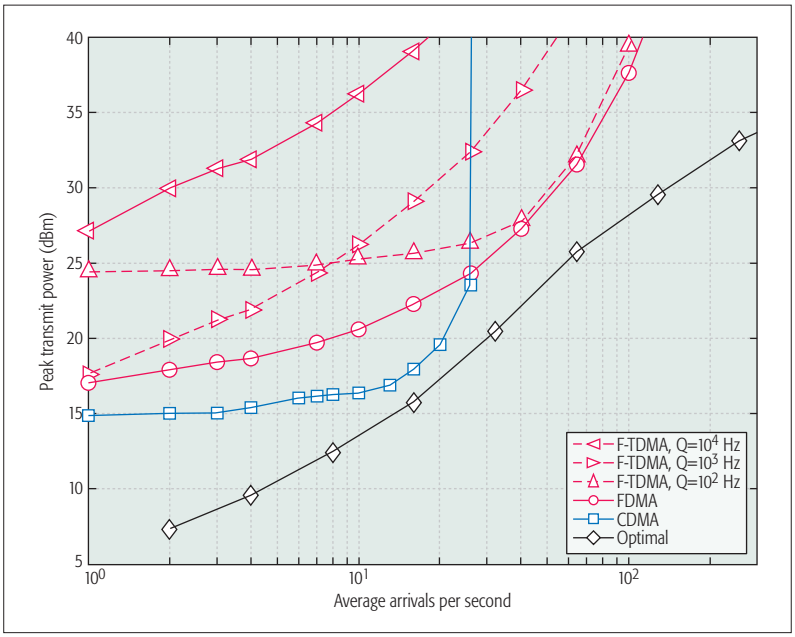


Figure 1. Peak transmit power versus arrival rate performance for RACH transmission strategies, 100KHz bandwidth, 1 second latency, 500 bit payload per device.

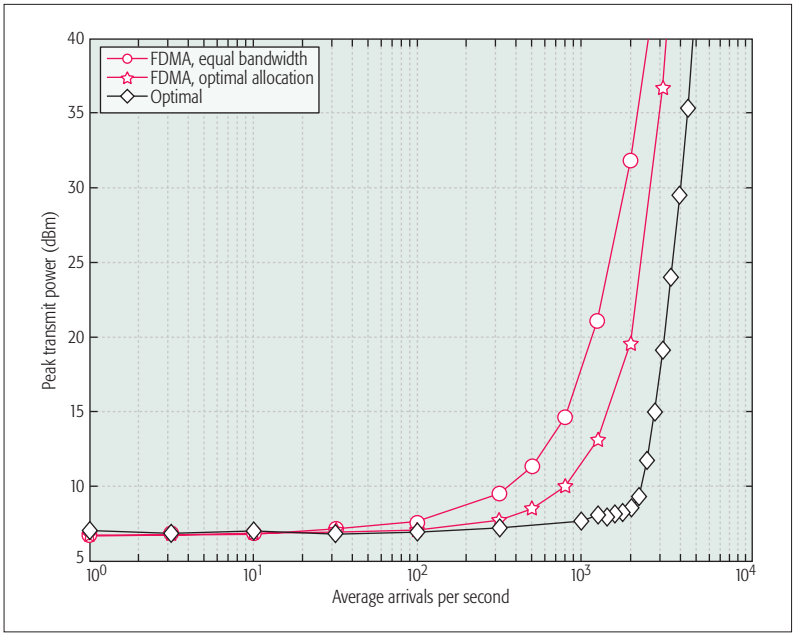


Figure 2. Peak transmit power versus arrival rate performance for scheduled transmission strategies, 100KHz bandwidth, 1 second latency, 500 bit payload per device.

Suboptimal RACH strategies. The optimal strategy discussed above is computationally intensive, which renders it unfit for practical implementations. Therefore, we consider a few suboptimal but easily implementable strategies. In particular, we consider slotted CDMA, FDMA, or a hybrid FDMA-TDMA where the bandwidth is partitioned into bins of width Q Hz and the time slot size is optimized for each arrival rate to minimize transmit power. While open loop power control can be employed to compensate for path loss and shadowing, the transmit power requires an additional fade margin to ensure reliable transmission in the presence of fading.

It should be noted that random access is an active area of research and many enhancements to the basic ALOHA protocol have been proposed. However, we focus on the above strategies that essentially correspond to optimal random access and the classic ALOHA protocol as the two extremes.

Figure 1 shows the peak (95th percentile) power of the three suboptimal strategies for transmitting a device payload of L = 500 bits with a total W = 100 KHz bandwidth in time T = 1 second and with a 0.1 probability of failure [4]. (The optimal strategy performance [14] is also shown for reference.) Devices are dropped uniformly in a cell of radius 2km, and the pathloss exponent is 3.7. If peak power is not constrained, then FDMA supports higher arrival rates. However, the bandwidth per FDMA bin may be impractically narrow at very high arrival rates. If the bin width is constrained to Q = 1000Hz, a few dB of additional power is required by F-TDMA for moderate arrival rates of λ = 100 per second. The power penalty is more significant with a Q = 10 KHz bin, which is similar to what could be achieved with LTE's 15 KHz bins. Overall, CDMA has better power performance for arrival rates below its pole capacity. The larger the bandwidth slice allocated for a CDMA channel, the lower the transmit power. However, allocating the entire bandwidth results in reduced flexibility to adapt the bandwidth allocated to M2M. We thus recommend dividing the available bandwidth into channels of smaller bandwidths (for example, 100 KHz) and operating multiple CDMA channels, as many as needed based on traffic conditions. Such a channelized CDMA can be implemented within a multi-carrier system with a suitable waveform such as UFMC [15] to suppress inter-channel interference.

Scheduled Transmission

Optimal scheduled strategy. For given K devices, the optimal strategy is the one where all the devices transmit simultaneously over all the resources and the receiver uses a weakest-last successive interference cancellation (SIC) strategy [4]. Under this strategy, the receiver first decodes the device with the highest channel gain, assuming interference from the K-1 other devices. Using the decoded bits, the received signal for this device is reconstructed and subtracted from the received signal. Devices are decoded and cancelled successively in order from highest to

lowest channel gains. The transmit power needed to communicate L bits in the given resource slice using this strategy was derived in [4].

Suboptimal Scheduled Strategies. The optimal strategy discussed above is sensitive to channel estimation errors. As was the case with the RACH transmission above, we consider more practical strategies using FDMA with either optimal or equal bandwidth allocation strategies [4]. Under optimal FDMA bandwidth allocation, W Hz bandwidth is allocated among the K devices to minimize the sum power. Under the equal bandwidth allocation, each device is allocated bandwidth W/K Hz.

Figure 2 shows the peak (95th percentile) power for the FDMA and the optimal SIC strategies, using the same system assumptions as the RACH simulations (L = 500 bits, W = 100 KHz, T = 1 second, cell radius 2 km, pathloss exponent 3.7). Comparing equal and optimal bandwidth allocation for FDMA, we note that equal allocation is near-optimal except at very high loading. While the optimal SIC strategy shows significant performance gains versus FDMA, it is important to note that its performance shown in Fig. 2 assumes perfect channel estimates for all the devices, which may not be realistic, especially at high arrival rates. In practice, the SIC performance gains are unlikely to be significant if we account for these impairments. This motivates the use of equal bandwidth FDMA as the preferred scheme for minimizing transmit power for the scheduled transmission. Note that FDMA could be implemented in practice as OFDMA in a wide band system.

While the scheduled strategies are able to achieve significantly higher packet arrival rates for a given transmit power constraint, it should be noted that control signaling is required to indicate a need and provide a grant for the scheduled transmission. We next compare the random access and scheduled strategies taking the overhead into account.

One-Stage Design vs. Two-Stage Design: In this section, we study the performance of M2M communication at the system level by making a distinction between control information (i.e., device ID) and device-specific information (e.g., sensor measurement). As shown in Fig. 3, we consider two design options that build on the RACH and scheduled strategies.

- **One-stage design.** Both control and device-specific information are transmitted on an uplink RACH. The base sends an acknowledgement on the downlink to indicate a successful reception.
- **Two-stage design.** Control information is first sent on an uplink RACH to indicate a service request by a specific device. After successful processing of the request, the base sends a resource allocation message to the device, and this is followed by a second uplink transmission on the scheduled resource.

In Fig. 4, we show the device throughput in terms of the average arrival rate as a function of the device-specific payload for a given set of frequency resources (10 KHz), a given target latency requirement (1 second) and a given outage probability (0.1). Results are shown for different one-stage and two-stage designs, assuming a cell radius of 2 km and a pathloss exponent of 3.7. For the one-stage design, we consider either the optimal RACH strategy or the FDMA RACH strategy. For the two-stage design, we assume FDMA RACH transmission for the first stage and scheduled FDMA for the second. For both the one-stage and two-stage options, the control information is assumed to be 20 bits. For the two-stage option, the downlink resource allocation is 64 bits, and the downlink spectral efficiency is 2 bps/Hz. Other details about the simulation assumptions can be found in [14].

As we observe from Fig. 4, for smaller payload sizes, the supportable arrival rates for the one-stage designs are higher than the two-stage design because the cost of the two-stage downlink overhead is greater than the relative inefficiency of the random access transmission. For larger payloads, the overhead becomes negligible,

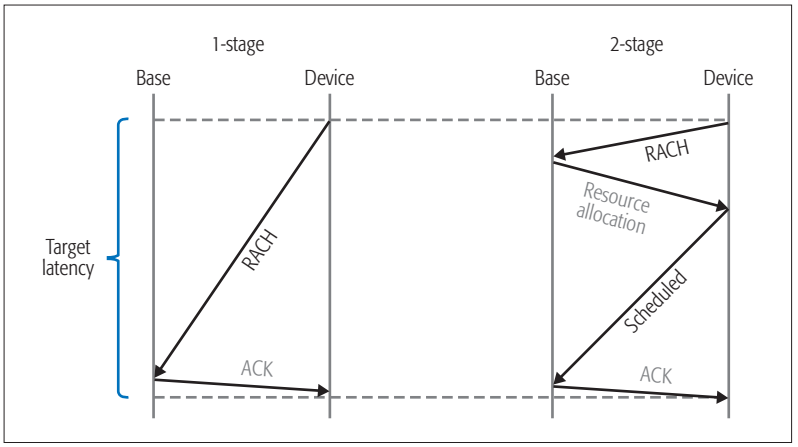


Figure 3. Schematics of one-stage and two-stage protocols. In the one-stage protocol, the device ID and data payload are transmitted on the uplink RACH, possibly resulting in a collision. In the two-stage protocol, the device ID is transmitted on the RACH. Following a downlink resource allocation, the data payload is transmitted on scheduled resources.

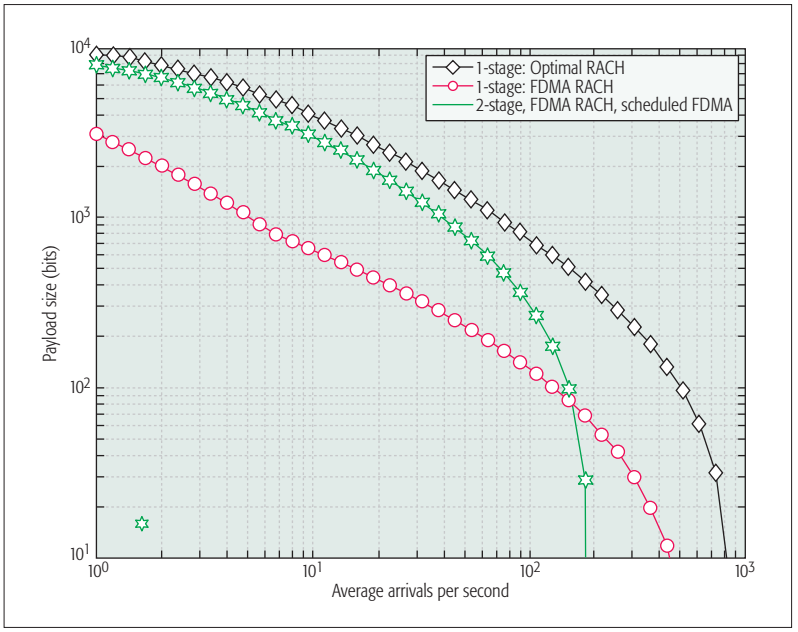


Figure 4. Payload size (device-specific information) versus arrival rate performance for one-stage and two-stage transmission strategies, 10KHz bandwidth, 1 second latency.

To provide additional functionality for an Internet of Mobile Things, the 5G standard could also be natively designed to enable ubiquitous localization and tracking for low-cost devices, to complement existing techniques such as GPS and RF fingerprinting.

and we expect the two-stage strategy to be more efficient. Indeed, there is a crossover in performance between the two-stage and FDMA RACH one-stage designs, so that for payloads larger than about 100 bits, the two-stage throughput is larger. This crossover threshold depends on the two-stage downlink overhead message size. On the other hand, the one-stage performance with the optimal RACH is uniformly better than the two-stage performance. However, as discussed in [14], the optimal RACH performance characterization should be thought of as an optimistic bound as a result of ideal assumptions and the impractical complexity of its implementation.

Conclusions

Current cellular networks are optimized for high-rate broadband access for sophisticated smartphone devices. To enable wide-area wireless communications for the future Internet of Things, the networks must also accommodate orders-of-magnitude more devices that communicate at orders-of-magnitude lower data rates. While evolving 2G and 4G cellular standards and dedicated networks for long-range M2M communication could address these needs in the near term, a future 5G cellular standard could present a unified solution that jointly optimizes the access network for both broadband and M2M communications. Our analysis suggests that for small payloads, a random access strategy with code multiplexing of transmissions within narrow bandwidth channels reduces transmit power and provides flexibility to allocate resources for such transmissions based on demand. For larger payloads, a scheduled transmission strategy carefully designed to minimize the amount of control overhead is recommended. To provide additional functionality for an Internet of Mobile Things, the 5G standard could also be natively designed to enable ubiquitous localization and tracking for low-cost devices, to complement existing techniques such as GPS and RF fingerprinting.

References

[1] J. A. Stankovic, "Research Directions for the Internet of Things," *IEEE Internet of Things J.*, vol. 1, no. 1, Feb. 2014, pp. 3–9.

[2] P. C. Evans and M. Annunziata, "Industrial Internet: Pushing the Boundaries of Minds and Machines," GE White Paper, Nov 26, 2012, http://goo.gl/BLXCPq.

[3] M. Z. Shafiq *et al.*, "Large-Scale Measurement and Characterization of Cellular Machine-to-Machine Traffic," *IEEE/ACM Trans. Net.*, vol. 21, no. 6, Dec. 2013, pp. 1960–73.

[4] H. S. Dhillon *et al.*, "Power-Efficient System Design for Cellular-based Machine-to-Machine Communications," *IEEE Trans. Wireless Commun.*, vol. 12, no. 11, Nov. 2013, pp. 5740–53.

[5] S.-Y. Lien, K.-C. Chen, and Y. Lin, "Toward Ubiquitous Massive Accesses in 3GPP Machine-to-Machine Communications," *IEEE Commun. Mag.*, vol. 49, no. 4, Apr. 2011, pp. 66–74.

[6] Cellular System Support for Ultra Low Complexity and Low Throughput Internet of Things, 3GPP TR 45.820, 2015.

[7] System Improvements for Machine-Type Communications, 3GPP TR 23.888, 2012.

[8] Further LTE Physical Layer Enhancements for MTC, 3GPP RP-141865, 2014.

[9] Machine-Type Communications (MTC) User Equipments (UEs) Based on LTE, 3GPP TR 36.888, 2013.

[10] 3GPP TSG GERAN #65, GP-150057, "C-UNB Technology for Cellular IoT — Physical Layer," *SigFox Wireless*, Mar. 2015.

[11] 3GPP TSG GERAN #65, GP-150075, "Combined Narrow-Band and Spread Spectrum Physical Layer Coverage and Capacity Simulations," *Semtech*, Mar. 2015.

[12] R. Quinnell, "Low Power Wide-Area Networking Alternatives for the IoT," *EDN Network*, edn.com, Sept. 2015.

[13] K. Wang, J. Alonso-Zarate, and M. Dohler, "Energy-Efficiency of LTE for Small Data Machine-to-Machine Communications," *IEEE ICC*, Budapest, Hungary, June 2013.

[14] H. S. Dhillon *et al.*, "Fundamentals of Throughput Maximization with Random Arrivals for M2M Communications", *IEEE Trans Commun.*, vol. 62, no. 11, Nov. 2014, pp. 4094–109.

[15] G. Wunder *et al.*, "5GNOW: Non-Orthogonal, Asynchronous Waveforms for Future Mobile Applications," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 97–105.

Biographies

Harpreet Dhillon (hdhillon@vt.edu)received the B.Tech. degree in electronics and communication engineering from IIT Guwahati in 2008, the M.S. degree in electrical engineering from Virginia Tech in 2010, and the Ph.D. degree in electrical engineering from the University of Texas at Austin in 2013. After a postdoctoral year at the University of Southern California, he joined Virginia Tech in 2014, where he is currently an assistant professor. His current research interests include communication theory, stochastic geometry, heterogeneous cellular networks, and Internet of Things. He is a co-author of five best paper award recipients, including the 2016 IEEE ComSoc Heinrich Hertz Award, the 2015 IEEE ComSoc Young Author Best Paper Award, and the 2014 IEEE ComSoc Leonard G. Abraham Prize.

Howard Huang [F] received a B.S. in electrical engineering from Rice University in 1991 and a Ph.D. in electrical engineering from Princeton University in 1995. Since graduating, he has been a research engineer at Bell Labs, where he has worked on multiple antenna techniques and their application in cellular networks. More recently, he has worked on cellular IoT communications, and he currently leads the Bell Labs Localization Research Project for enabling scalable tracking of IoT devices. He has taught as an adjunct professor at Columbia University and is a Fellow of the IEEE.

Harish Viswanathan [F] received the B. Tech. degree from the Department of Electrical Engineering, Indian Institute of Technology, Chennai, India, and the M.S. and Ph.D. degrees from the School of Electrical Engineering, Cornell University, Ithaca, NY. Since joining Bell Labs in October 1997, he has worked on multiple antenna technology for cellular wireless networks, network optimization, network architecture, and IoT. He is currently the head of the Radio Systems Research group in the Mobile Radio lab, and also leads the 5G radio access research project within Bell Labs. He has published extensively with over 100 publications. He is a Fellow of Bell Labs.

CALL FOR PAPERS

IEEE COMMUNICATIONS MAGAZINE

EMERGING TRENDS, ISSUES, AND CHALLENGES IN BIG DATA AND ITS IMPLEMENTATION TOWARD FUTURE SMART CITIES

BACKGROUND

The world is experiencing a period of extreme urbanization. Cities in the 21st century will account for nearly 90% of the global population growth, 80% of wealth-creation and 60% of total energy consumption. The world urbanization continues to grow, and the global population is expected to double by 2050. Smart Cities are emerging as a priority for research and development across the world. In general, Smart cities integrate multiple Internet of Things (IoT) and emerging communication technologies such as fifth generation (5G) solutions in a secure fashion to manage a city's assets, such as transportation systems, hospitals, water supply networks, waste management. The goal of building a smart city is to improve the quality of life by using technology to improve the efficiency of services and meet residents' needs.

Smart cities' economic growth and large-scale urbanization drive innovation and new technologies. Technology is driving the way city officials interact with the community and the city infrastructure. The rapid progress in smart cities research is posing enormous challenge in terms of large amounts and various types of data at an unprecedented granularity, speed, and complexity are increasingly produced by the sensors of IoT via emerging communication technologies. Meanwhile, the accumulation of huge amounts of data can be used to support smart city components to reach the required level of sustainability and improve living standards. Smart cities have become data-driven, thus effective computing and utilization of big data such as distributed and parallel computing, artificial intelligence and cloud/fog computing are key factors for success in future smart cities. The use of big data can certainly help create cities where infrastructure and resources are used in a more efficient manner.

Any smart city project willing to use big data will need to capture, store, process and analyze a large amount of data generated by several sources to transform the data into useful knowledge that is applicable to a decision-making process. For example, with the help of big data and its Implementation, citizens could rapidly find available parking slots in large urban areas; big data can contribute in the city's efforts to reduce pollution through the deployment of street sensors. These sensors can measure traffic flows at different times as well as total emissions. The government can implement actions to divert traffic to less congested areas in a move to reduce carbon emissions in a particular area.

This Feature Topic (FT) is intended to encourage high-quality researchers in big data and its Implementation for future smart cities, and push the theoretical and practical research forward for a deeper understanding of future smart city constructions and operations.

In this FT, we would like to try to answer some (or all) of the following questions:

How to analyze the mass data that IOT devices produce by future smart cities? How to design the algorithm to process the mass data? How to utilize the machine learning and artificial intelligence techniques to improve the quality of life for future smart cities? How to utilize the "big data" to improve the QoS for future smart cities? How to guarantee the security and the privacy when mass data generated by IOT devices of future smart cities? How to diagnose the fault among the mass IOT devices of future smart cities? How to design the hardware to be suitable to process the mass data, among others?

Topics of interest include, but are not limited to:

- Distributed and parallel algorithms for big data in smart cities
- Big data analytics in data processing center for smart cities
- Cloud/fog computing in data processing center for smart cities
- The application of mobile cloud/fog computing for smart cities
- Fault tolerance, reliability and survivability in smart cities
- E-health and connected healthcare systems in smart cities
- Cyber-physical and social computing and networks in smart cities
- Environmental and urban monitoring in smart cities
- QoS and QoE of systems, applications, and services for smart cities
- Safety, security, privacy and trust in applications and services for smart cities
- Other topics related to big data and its implementations for smart cities

SUBMISSIONS

Articles should be tutorial in nature, with the intended audience being all members of the communications technology community. They should be written in a style comprehensible to readers outside the specialty of the article. Mathematical equations should not be used (in justified cases up to three simple equations are allowed). Articles should not exceed 4500 words (from introduction through conclusions). Figures and tables should be limited to a combined total of six. The number of references is recommended not to exceed 15. In some rare cases, more mathematical equations, figures, and tables may be allowed if well-justified. In general, however, mathematics should be avoided; instead, references to papers containing the relevant mathematics should be provided. Complete guidelines for preparation of the manuscripts are posted at http://www.comsoc.org/commag/paper-submission-guidelines. Please send a PDF (preferred) or MSWORD formatted paper via Manuscript Central (http://mc.manuscriptcentral.com/commag-ieee). Register or log in, and go to Author Center. Follow the instructions there. Select "December 2017 / Emerging Trends, Issues and Challenges in Big Data and Its Implementation towards Future Smart Cities" as the Feature Topic category for your submission.

IMPORTANT DATES

- Manuscript Submission Deadline: April 1, 2017
- Decision Notification: August 1, 2017
- Final Manuscript Due Date: September 15, 2017
- Publication Date: December 2017

GUEST EDITORS

Guangjie Han
Hohai University, China
hanguangjie@ieee.org

Jaime Lloret
Universidad Politecnica de Valencia, Spain
jlloret@dcom.upv.es

Liangtian Wan
Nanyang Technological University, Singapore
wan.liangtian.2015@ieee.org

Sammy Chan
City University of Hong Kong, Hong Kong, China
eeschan@cityu.edu.hk

Mohsen Guizani
University of Idaho, USA
mguizani@ieee.org

Wael Guibene
Intel Labs, Ireland
wael.guibene@intel.com

ACCEPTED FROM OPEN CALL

Overview of Full-Dimension MIMO in LTE-Advanced Pro

Hyoungju Ji, Younsun Kim, Juho Lee, Eko Onggosanusi, Younghan Nam, Jianzhong Zhang, Byungju Lee, and Byonghyo Shim

The authors provide an overview of the FD-MIMO system, with emphasis on the discussion and debate conducted on the standardization process of Release 13. They present key features for FD-MIMO systems, a summary of the major issues for the standardization and practical system design, and performance evaluations for typical FD-MIMO scenarios.

ABSTRACT

Multiple-input multiple-output (MIMO) systems with a large number of base station antennas, often called massive MIMO, have received much attention in academia and industry as a means to improve the spectral efficiency, energy efficiency, and processing complexity of next generation cellular systems. The mobile communication industry has initiated a feasibility study of massive MIMO systems to meet the increasing demand of future wireless systems. Field trials of the proof-of-concept systems have demonstrated the potential gain of the Full-Dimension MIMO (FD-MIMO), an official name for the MIMO enhancement in the 3rd generation partnership project (3GPP). 3GPP initiated standardization activity for the seamless integration of this technology into current 4G LTE systems. In this article, we provide an overview of FD-MIMO systems, with emphasis on the discussion and debate conducted on the standardization process of Release 13. We present key features for FD-MIMO systems, a summary of the major issues for the standardization and practical system design, and performance evaluations for typical FD-MIMO scenarios.

INTRODUCTION

Multiple-input multiple-output (MIMO) systems with a large number of base station antennas, often referred to as *massive MIMO systems*, have received much attention in academia and industry as a means to improve spectral efficiency, energy efficiency, and processing complexity [1]. While massive MIMO technology is a promising technology, there are many practical challenges and technical hurdles down the road to its successful commercialization. These include the design of low-cost and low-power base stations with acceptable antenna space, improvement in the fronthaul capacity between the radio and control units, the acquisition of high dimensional channel state information (CSI), and many others. Recently, the 3rd generation partnership project (3GPP) standards body initiated standardization activity to employ tens of antennas at the base station, with a goal to satisfy the spectral efficiency requirement of future cellular systems [2, 3]. Considering the implementation cost and complexity, and also the timeline to real deployment, 3GPP decided to use tens of antennas with a two dimensional (2D) array structure as a starting point. Full-Dimension MIMO (FD-MIMO), the official name for the MIMO enhancement in 3GPP, targets the system utilizing up to 64 antenna ports at the transmitter side. Recently, field trials of proof-of-concept FD-MIMO systems have been conducted successfully [4]. A study item, a process done before a formal standardization process, was completed in June 2015, and the follow-up work item process will be finalized soon for the formal standardization of Release 13 (Rel. 13).[1]

The purpose of this article is to provide an overview of FD-MIMO systems with an emphasis on the discussion and debate conducted on the standardization process of Rel. 13. We note that preliminary studies addressed the feasibility of 2D array antenna structures and performance evaluation in ideal pilot transmission and feedback scenarios [2, 3]. This work is distinct from those in the sense that we put our emphasis on describing realistic issues in the standardization process, including TXRU architectures, beamformed CSI-RS, 3D beamforming, details of CSI feedback, and performance evaluation in realistic FD-MIMO scenarios with new feedback schemes.

KEY FEATURES OF FD-MIMO SYSTEMS

In this section, we discuss key features of FD-MIMO systems. These include a large number of base station antennas, 2D active antenna arrays, 3D channel propagation, and new pilot transmission with CSI feedback. In what follows, we will use LTE terminology exclusively: enhanced node-B (eNB) for the base station, user equipment (UE) for the mobile terminal, and reference signal (RS) for the pilot signal.

INCREASE THE NUMBER OF TRANSMIT ANTENNAS

One of the main features of FD-MIMO systems distinct from the MIMO systems of the current LTE and LTE-Advanced standards is the use of a large number of antennas at the eNB. In theory, as the number of eNB antennas N_T increases, the cross-correlation of two random channel realizations goes to zero [1] so that the inter-user interference in the downlink can be controlled via a simple linear precoder. However, such a benefit can be realized only when the perfect CSI is available at the eNB. While the CSI acquisition in time division duplex (TDD) systems is relatively simple due to the channel reciprocity, such is not the case for frequency division duplex (FDD) systems. Note that in the FDD systems, time variation and frequency response of the channel are measured via the downlink RSs and then sent back to the eNB after the quantization. Even in TDD mode, one

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government(MSIP) (2014R1A5A1011478) and the ICT R&D program of MSIP/IITP. [B0717-16-0023, Fundamental technology development of transmission, modulation, and coding techniques for low-power and low-complexity 5G-based IoT environments with massive connectivity.

[1] LTE-Advanced Pro is the LTE marker that is used for the specifications from Release 13 onwards by 3GPP.

Digital Object Identifier: 10.1109/MCOM.2016.1500743RP

Hyoungju Ji,Younsun Kim, and Juho Lee, are with Samsung Electronics, Korea; Eko Onggosanusi, Younghan Nam, and Jianzhong Zhang are with Samsung Research America; Byungju Lee is with Purdue University; Byonghyo Shim is with Seoul National University.

0163-6804/16/$25.00 © 2017 IEEE

cannot solely rely on channel reciprocity because the measurement at the transmitter does not capture the downlink interference from neighboring cells or co-scheduled UEs. As such, downlink RSs are still required to capture the channel quality indicator (CQI) for the TDD mode, and thus the downlink RS and the uplink CSI feedback are essential for both duplex modes. Therefore, identifying the potential issues of CSI acquisition and developing the proper solutions are of great importance for the successful commercialization of FD-MIMO systems. Before we go into detail, we briefly summarize two major problems related to CSI acquisition.

Degradation of CSI Accuracy: One well known problem for MIMO systems, in particular for FDD-based systems, is that the quality of CSI is affected by the limitation of feedback resources. As CSI distortion increases, the quality of the multiuser MIMO (MU-MIMO) precoder to control the inter-user interference is degraded, and so will be the performance of the FD-MIMO systems. In general, the amount of CSI feedback, determining the quality of CSI, needs to be scaled with N_T to control the quantization error so that the overhead of CSI feedback increases in FD-MIMO systems.

Increase of Pilot Overhead: An important problem related to CSI acquisition at the eNB, yet to be discussed separately, is the pilot overhead problem. UE performs channel estimation using the RS transmitted from the eNB. Since RSs need to be assigned in an orthogonal fashion, RS overhead typically grows linearly with N_T. For example, if N_T = 64, RS will occupy approximately 48 percent of resources, using up a substantial amount of downlink resources for data transmission.

2D Active Antenna System (AAS)

Another interesting feature of the FD-MIMO system is an introduction of the active antenna with 2D planar array. In the active antenna-based systems, gain and phase are controlled by the active components, such as the power amplifier (PA) and low noise amplifier (LNA), attached to each antenna element. In the 2D structured antenna array, one can control the radio wave on both vertical (elevation) and horizontal (azimuth) direction so that the control of the transmit beam in 3D space is possible. This type of wave control mechanism is also referred to as *3D beamforming*. Another important benefit of 2D AAS is that it can accommodate a large number of antennas without increasing the deployment space. For example, when 64 linear antenna arrays are deployed in a horizontal direction, under the common assumption that the antenna spacing is half wavelength ($\lambda/2$) and the system is using the LTE carrier frequency (2 GHz), it requires a horizontal room of 3m. Due to the limited space on a rooftop or mast, this space would be burdensome for most cell sites. In contrast, when antennas are arranged in a square array, a relatively small space is required for a 2D antenna array (e.g., 1.0 × 0.5m with dual-polarized 8 × 8 antenna array).

The 3D Channel Environment

When the basic features of the FD-MIMO systems are determined, the next step is to design a system that maximizes performance in terms of throughput, spectral efficiency, and peak data rate in the realistic channel environment. There are various issues to consider in the design of practical systems, such as the investigation and characterization of the realistic channel model for performance evaluation. While conventional MIMO systems consider the propagation in the horizontal direction only, FD-MIMO systems employing 2D planar arrays should consider the propagation in both the vertical and horizontal directions. To do so, the geometric structure of the transmitter antenna array and the propagation effect of the 3D positions between the eNB and UE should be reflected in the channel model. The main features of 3D channel propagation obtained from real measurements are as follows [5]:

- Height and distance-dependent line-of-sight (LOS) channel condition: LOS probability between the eNB and UE increases with the UE's height, and also increases when the distance between eNB and UE decreases.
- Height-dependent path loss: the UE experiences less path loss on a higher floor (e.g., 0.6dB/m gain for a macro cell and 0.3dB/m gain for a micro cell).
- Height and distance-dependent elevation spread of departure angles (ESD): When the location of the eNB is higher than the UE, ESD decreases with the height of the UE. It is also observed that the ESD decreases sharply as the UE moves away from the eNB.

RS Transmission for CSI Acquisition

From LTE to LTE-Advanced, there has been substantial improvement in the RS scheme for MIMO systems (Fig. 1a). From the common RS (CRS) to channel state information RS (CSI-RS), various RSs to perform CSI acquisition have been introduced. While these are common to all users in a cell and thus un-precoded, the demodulation RS (DM-RS) is UE-specific (i.e., dedicated to each UE) so that it is precoded by the same weight applied for data transmission. Since the DM-RS is present only on time/frequency resources where the UE is scheduled, this cannot be used for CSI measurements [6].

One of the new features of FD-MIMO systems is the use of a beamformed RS, called beamformed CSI-RS, for CSI acquisition. Beamformed RS transmission is a channel training technique that uses multiple precoding weights in a spatial domain. In this scheme, the UE picks the best weight among those transmitted and then feeds back its index. This scheme provides many benefits over non-precoded CSI-RS, in particular when N_T is large. Some of the benefits are summarized as follows:

Less Uplink Feedback Overhead: In order to maintain a rate comparable to the case with perfect CSI, feedback bits used for channel vector quantization should be proportional to N_T [7], whereas the amount of feedback for the beamformed CSI-RS scales logarithmically with the number of RSs N_B since this scheme only feeds back an index of the best beamformed CSI-RS. Thus, as depicted in Fig. 1b, the benefit of beamformed CSI-RS is pronounced when N_T is large.

Less Downlink Pilot Overhead: When the non-precoded CSI-RS is used, pilot overhead increases with N_T, resulting in a substantial loss of the sum capacity in the FD-MIMO regime (Fig. 1c), whereas pilot overhead of the beamformed CSI-RS is proportional to N_B and independent of N_T so that the rate loss of the beamformed CSI-RS is marginal even when N_T increases.

One of the new features of FD-MIMO systems is the use of a beamformed RS, called beamformed CSI-RS, for CSI acquisition. Beamformed RS transmission is a channel training technique that uses multiple precoding weights in a spatial domain. In this scheme, the UE picks the best weight among those transmitted and then feeds back its index.

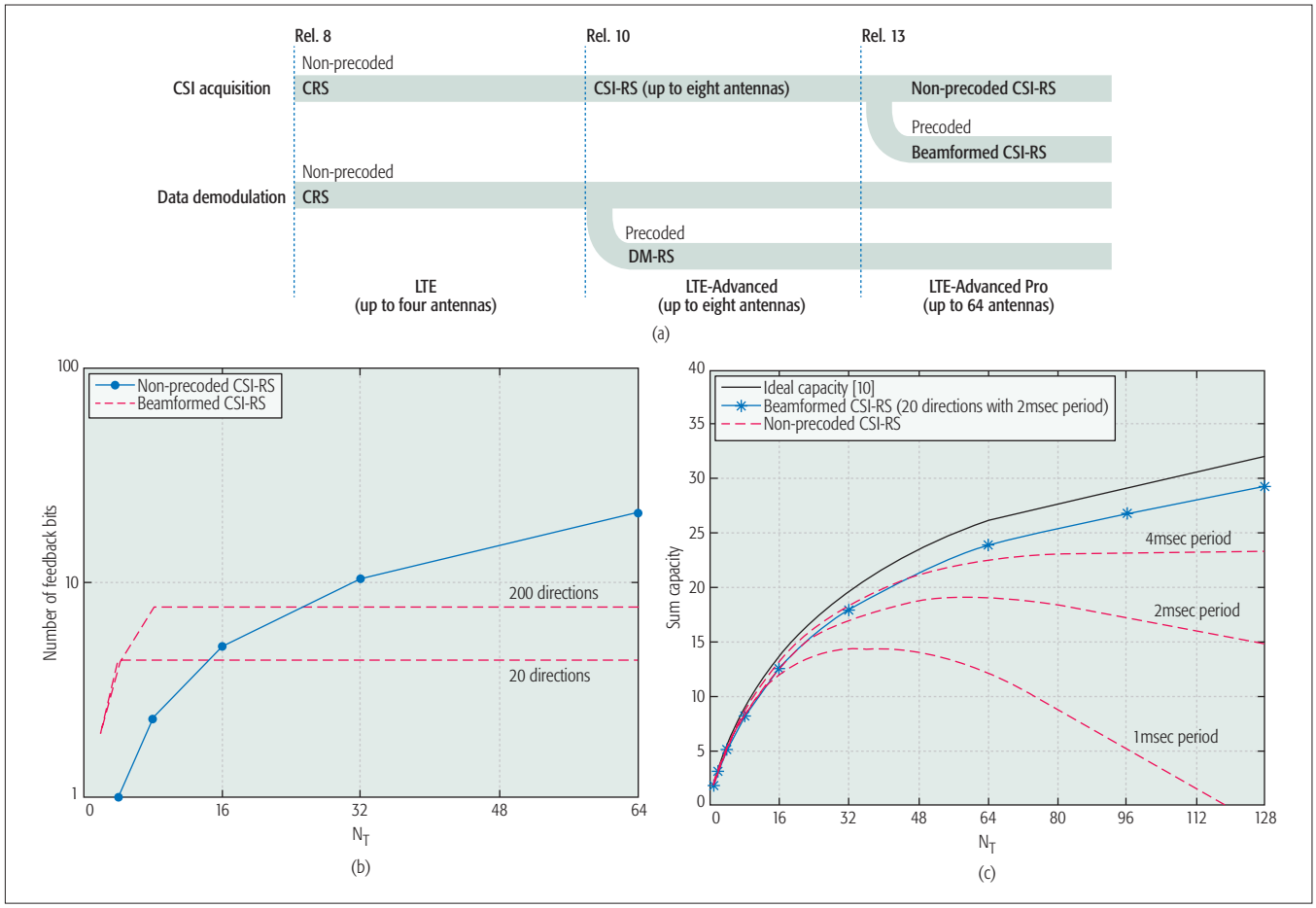


Figure 1. MIMO evaluation: a) RS evolution in LTE systems; b) uplink feedback overhead (SNR = 10dB [7]); c) MU-MIMO capacity with considering CSI-RS overhead (ideal CSI and ZFBF MU-precoding with 10 UEs and SNR = 10dB).

Higher Quality in RS: If the transmit power is P watt, P/N_T watt is needed for each non-precoded CSI-RS transmission, while P/N_B watt is used for the beamformed CSI-RS. For example, when N_T = 32 and N_B = 12, beamformed CSI-RS provides 4.3dB gain in signal power over the non-precoded CSI-RS.[2]

In order to support the beamformed CSI-RS scheme, a new transmitter architecture called the transceiver unit (TXRU) architecture has been introduced. By TXRU architecture, we mean a hardware connection between the baseband signal path and antenna array elements. Since this architecture facilitates the control of phase and gain in both the digital and analog domains, more accurate control of the beamforming direction is possible. One thing to note is that the conventional codebook cannot measure the CSI of the beamformed transmission so that a new channel feedback mechanism supporting the beamformed transmission is required.

System Design and Standardization of FD-MIMO Systems

The main purpose of the Rel. 13 study item is to identify key issues to support up to 64 transmit antennas placed in the form of a 2D antenna array. Standardization of systems supporting up to 16 antennas is an initial target of Rel. 13, and issues to support more than 16 antennas will be discussed in subsequent releases. In the study item phase, there has been extensive discussion to support 2D array antennas, elaborated TXRUs, enhanced channel measurement and feedback schemes, and also an increased number of co-scheduled users (up to eight users). Among these, an item tightly coupled to standardization is the CSI measurement and feedback mechanism. In this subsection, we discuss the deployment scenarios, antenna configurations, TXRU structure, new RS strategy, and feedback mechanisms.

Deployment Scenarios

For the design and evaluation of FD-MIMO systems, a realistic scenario in which the antenna array and UEs are located at different heights is considered. To this end, two typical deployment scenarios, viz., a 3D urban macro scenario (3D-UMa) and a 3D urban micro scenario (3D-UMi), are introduced (Fig. 2). In the former case, transmit antennas are placed over the rooftop, and in the latter case, they are located below the rooftop. In the case of 3D-UMa, diffraction over the rooftop is a dominant factor for propagation so that down-tilted transmission in the vertical direction is desirable (Fig. 2b). In fact, by transmitting beams with different steering angles, the eNB can separate channels corresponding to multiple UEs. In the 3D-UMi scenario, on the other hand, the location of users is higher than the height of the antenna so that a direct signal path is dominant (Fig. 2c). In this sce-

[2] In 3D channel model, the typical number of multi-paths (clusters) is 12 [5].

[3] Note that the total number of antenna elements in this setup is the same as that of 8Tx antennas in conventional systems and thus FD-MIMO eNB can provide backward compatibility [8]. The vertical configuration is to ensure the same cell coverage and the horizontal configuration is for the conventional MIMO operation for LTE.

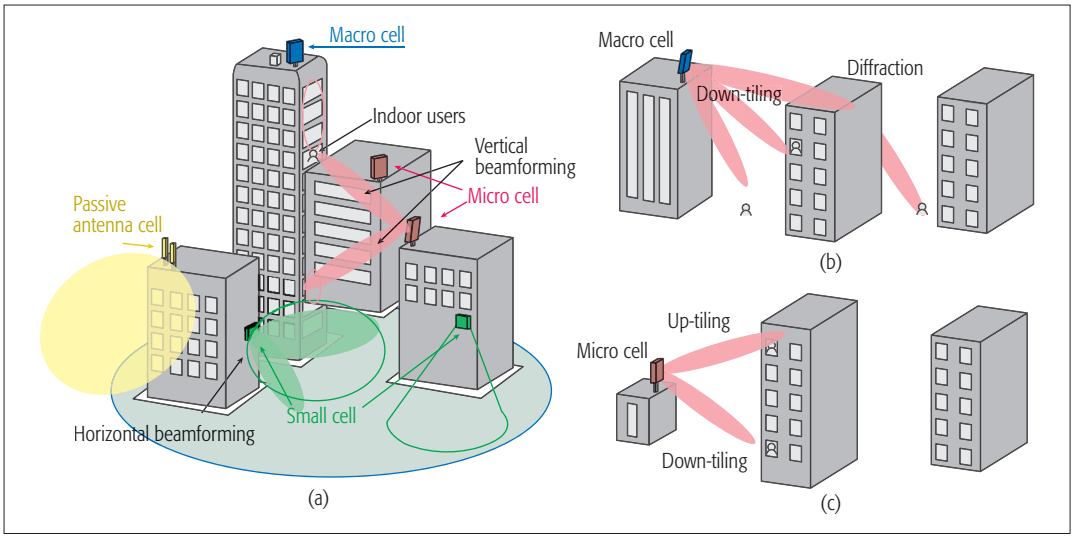


Figure 2. FD-MIMO deployment scenarios: a) 3D macro cell site (placed over the rooftop) and 3D micro cell site (placed below the rooftop) with small cell; b) beamforming for 3D macro cell; and c) beamforming in 3D micro cell.

Unlike conventional MIMO systems relying on a passive antenna, systems based on an active antenna can dynamically control the gain of an antenna element by applying the weight of low-power amplifiers attached to each antenna element. Since the radiation pattern depends on the antenna arrangement, the antenna system should be modeled in an element-level.

nario, both up-tilting and down-tilting can be used to schedule UEs on different floors. Since the cell radius of the 3D-UMi scenario is typically smaller than that of the 3D-UMa scenario, the LOS channel condition is predominant, and thus more UEs can be co-scheduled without increasing inter-user interference [5]. Although not as strong as the 3D-UMi scenario, LOS probability in the 3D-UMa scenario also increases when the distance between the eNB and UE decreases.

Antenna Configurations

Unlike conventional MIMO systems relying on a passive antenna, systems based on an active antenna can dynamically control the gain of an antenna element by applying the weight of low-power amplifiers attached to each antenna element. Since the radiation pattern depends on the antenna arrangement, such as the number of antenna elements and antenna spacing, the antenna system should be modeled in an element-level. As shown in Fig. 3a, there are three key parameters characterizing the antenna array structure (M, N, P): the number of elements M in the vertical direction; the number of elements N in the horizontal direction; and the polarization degree P (P = 1 is for co-polarization and P = 2 is for dual-polarization). As a benchmark setting, a 2D planar array using a dual polarized antenna (P = 2) configuration with M = 8 (0.8λ spacing in the vertical direction) and N = 4 (0.5λ spacing in the horizontal direction) is suggested.[3] In this setting, null direction, an angle to make the magnitude of the beam pattern equal to zero, for the elevation beam pattern is 11° and for the horizontal beam pattern is 30° (Fig. 3c). Since the null direction in the vertical domain is much smaller than that of the horizontal domain, scheduling UEs in the vertical domain is more effective in controlling the inter-user interference. Also, a tall or fat array structure ($M >> N$ or $M << N$) is favorable since it will generate a sharp beam, but it might be less flexible in the situation where the surrounding environment is changed. Further, large antenna spacing is not always a desirable option since it can increase inter-cell interference due to the narrow beamforming for cell edge UEs (this phenomenon is called the *flash-light effect*). For this reason, in a real deployment scenario, the design parameters should be carefully chosen by considering various factors such as user location, cell radius, building height, and antenna height.

TXRU Architectures

As mentioned earlier, one interesting feature of active antenna systems is that each TXRU contains PA and LNA so that the eNB can control the gain and phase of an individual antenna element. In order to support this, a power feeding network between TXRUs and antenna elements called the *TXRU architecture* is introduced [9]. The TXRU architecture consists of three components: TXRU array, antenna array, and radio distribution networks (RDN). A role of the RDN is to deliver the transmit signal from the PA to the antenna array elements, and the received signal from the antenna array to the LNA. Depending on the CSI-RS transmission and feedback strategy, two representative options, *array partitioning* and *array connected architecture*, are suggested. The former is for the conventional codebook scheme and the latter is for the beamforming scheme.

In the array partitioning architecture, antenna elements are divided into multiple groups and each TXRU is connected to one of them (Fig. 3d), whereas in the array connected structure, the RDN is designed such that the RF signals of multiple TXRUs are delivered to the single antenna element. To mix RF signals from multiple TXRUs, additional RF combining circuitry is needed, as shown in Fig. 3e. The difference between the two can be better understood when we discuss the transmission of the CSI-RS. In the array partitioning architecture, N_T antenna elements are partitioned into L groups of TXRUs, and an orthogonal CSI-RS is assigned for each group. Each TXRU transmits its own CSI-RS so that the UE measures the channel h from the CSI-RS observation $y = hx + n$. In the array connected architecture, each antenna element is connected to L' (out of L) TXRUs and an orthogonal CSI-RS is assigned for

[3] Note that the total number of antenna elements in this setup is the same as that of 8Tx antennas in conventional systems and thus FD-MIMO eNB can provide backward compatibility [8]. The vertical configuration is to ensure the same cell coverage and the horizontal configuration is for the conventional MIMO operation for LTE.

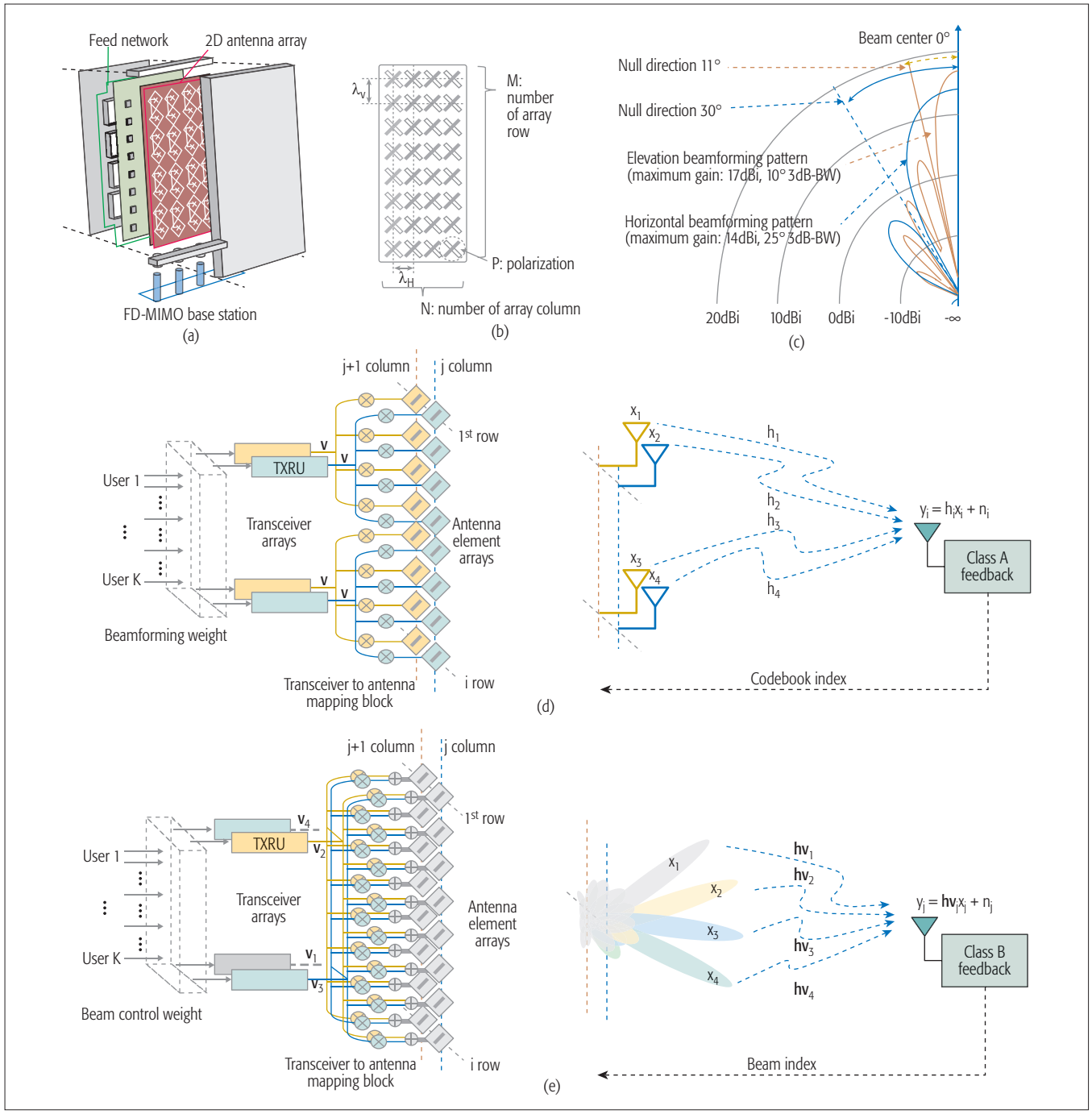


Figure 3. FD-MIMO systems: a) concept of FD-MIMO systems; b) 2D array antenna configuration; c) vertical and horizontal beamforming patterns; d) array partitioning architecture with the conventional CSI-RS transmission; and e) array connected architecture with beamformed CSI-RS transmission.

each TXRU. Denoting $\mathbf{h} \in \mathbb{C}^{1\times N_C}$ as the channel vector and $\mathbf{v} \in \mathbb{C}^{N_C\times 1}$ as the precoding weight for each beamformed CSI-RS, the beamformed CSI-RS observation is $y = \mathbf{hv}x + n$ and the UE measures the precoded channel $\mathbf{hv}$ from this. Due to the narrow and directional CSI-RS beam transmission with a linear array, the SNR of the precoded channel is maximized at the target direction.[4]

New CSI-RS Transmission Strategy

In the standardization process, two CSI-RS transmission strategies, i.e., extension of the conventional non-precoded CSI-RS and the beamformed CSI-RS, are suggested. In the first strategy, UE observes the non-precoded CSI-RS transmitted from each of the partitioned antenna arrays (Fig. 3d). By sending the precoder maximizing the properly designed performance criterion to the eNB, the UE can adapt to the channel variation. In the second strategy, the eNB transmits multiple beamformed CSI-RSs (we call it *beam* for simplicity) using the connected arrays architecture. Among these, the UE selects the preferred beam and then feeds back its index. When the eNB receives the beam index, the weight corresponding to the selected beam is used for data transmission.

[4] $$\mathrm{SNR} = \frac{|\mathbf{hv}(\phi)|^2}{\sigma^2},$$

where ϕ is the beam direction and σ^2 is the noise power.

Category	Class-A CSI feedback (conventional CSI-RS)	Class-B CSI feedback (beamformed CSI-RS)
Feedback design	Need to design codebook for 2D antenna layout and feedback mechanism for adapting channel variation.	Need to devise a method to feedback beam index for adapting both weight changes and channel variation.
UL feedback overhead	Depend on resolution of codebook and the number of antennas.	Depend on the number of operating beams N_B.
CSI-RS overhead	Require N_T CSI-RS resources.	Scale linearly with the number of beams N_B.
Backward compatibility	Supportable with virtualization between TXRUs and antenna ports.	Supportable with vertical 1D beamforming weight.
Forward compatibility	Scalable to larger TXRU system if CSI-RS resources are allowed.	Scalable to larger TXRU system if long-term channel statistics are acquired.

Table 1. Comparison between CSI-RS transmission and CSI feedback classes.

In the array partitioning architecture, antenna elements are divided into multiple groups and each TXRU is connected to one of them, whereas in the array connected structure, the RDN is designed such that the RF signals of multiple TXRUs are delivered to the single antenna element.

Overall downlink precoder for data transmission $\mathbf{W}_{\text{data}}$ and CSI-RS transmission $\mathbf{W}_{\text{rs}}$ can be expressed as

$$\mathbf{W}_{\text{data}} = \mathbf{W}_T\mathbf{W}_{\text{rs}} \quad \text{and} \quad \mathbf{W}_{\text{rs}} = \mathbf{W}_P\mathbf{W}_U, \tag{1}$$

where $\mathbf{W}_T \in \mathbb{C}^{N_T \times L}$ is the precoder between the TXRU and the antenna element, $\mathbf{W}_P \in \mathbb{C}^{L \times N_P}$ is the precoder between the CSI-RS port and the TXRU (N_P is the number of antenna ports), and $\mathbf{W}_U \in \mathbb{C}^{N_P \times r}$ is the precoder between the rank-r data channel and the CSI-RS port.

In the following, we summarize details of two strategies.

Conventional CSI-RS Transmission: One option to maximize capacity is to do one-to-one mapping of the TXRU and the CSI-RS resource (i.e., $\mathbf{W}_P = \mathbf{I}_{N_{TXRU}}$). To achieve the same coverage for each CSI-RS resource, an identical weight $\mathbf{v}$ is applied to L groups.[5] Each UE measures the CSI-RS resources and then chooses the preferred codebook index i^* maximizing the channel gain for each subband:

$$i^* = \arg\max_i \left\| \bar{\mathbf{h}}^H \mathbf{W}_U^i \right\|_2^2, \tag{2}$$

where

$$\bar{\mathbf{h}} = \mathbf{h} / \left\| \mathbf{h} \right\|_2$$

is the estimated channel direction vector, and $\mathbf{W}_U^i$ is the ith precoder between the data channel and CSI-RS ports. This scheme is called the class-A CSI feedback.

Beamformed CSI-RS Transmission: In order to acquire the spatial angle between the eNB and UE, the eNB transmits multiple beamformed CSI-RSs. Let N_B be the number of CSI-RSs, then we have $\mathbf{W}_T = [\mathbf{v}_1\ \mathbf{v}_2\ \ldots\ \mathbf{v}_{N_B}]$ where $\mathbf{v}_i \in \mathbb{C}^{N_T \times 1}$ is the 3D beamforming weight for the ith beam. For example, when rank-1 beamforming is applied, we have $\mathbf{W}_P = 1_{N_B}$ and $\mathbf{W}_U = 1$. Among all possible beams $\mathbf{v}_1, \ldots, \mathbf{v}_{N_B}$, UE selects and feeds back the best beam index j^* maximizing the received power:

$$j^* = \arg\max_j \left| \bar{\mathbf{h}}^H \mathbf{v}_j \right|^2. \tag{3}$$

This scheme is called the class-B CSI feedback. Under the rich scattering environment, dominant paths between the eNB and UE depend on the direction and width of the transmit signal. In the multiple-input single-output (MISO) channel, for example, the channel vector in an angular domain is expressed as $\mathbf{h} = \Sigma_i e_r \mathbf{e}_t(\phi_i)^*$, where $e_r = 1$ and $\mathbf{e}_t(\phi_i) = [1\ e^{-j2\pi\gamma\phi_i}\ \ldots\ e^{-j2\pi(N_T-1)\gamma\phi_i}]^T$ is the spatial signature of the transmitter (ϕ_i is the direction of the ith path and γ is the normalized antenna spacing) [10]. When the RS is transmitted in a direction ϕ_j, the beamforming weight would be $\mathbf{v} = \mathbf{e}_t(\phi_j)$ so that the resulting beamformed channel is readily expressed as one or at most a few dominant taps $(\mathbf{e}_t(\phi_i)^T\mathbf{e}_t(\phi_j) \approx 0$ when $i \neq j)$. In fact, by controlling the weight applied to the CSI-RS, the effective dimension of the channel vector can be reduced so that the feedback overhead can be reduced substantially. In Table 1, we summarize two CSI-RS transmission schemes discussed in FD-MIMO.

CSI Feedback Mechanisms for FD-MIMO Systems

In the study item phase, various RS transmission and feedback schemes have been proposed. As shown in Fig. 1, capacity and overhead of the class-A and class-B feedback schemes are more or less similar in the initial target range (N_t = 16) so that Rel. 13 has decided to support both classes. In this subsection, we briefly describe the CSI feedback schemes associated with the TXRU architectures. Among various schemes, composite codebook and beam index feedback have received much attention as main ingredients for class-A and class-B CSI feedback. The rest will be considered in a future release.

Composite Codebook: In this scheme, the overall codebook is divided into two codebooks (vertical and horizontal) and thus the channel information is separately delivered to the eNB. By combining two codebooks (e.g., the Kronecker product of two codebooks $\mathbf{W}_U = \mathbf{W}_{U,V} \otimes \mathbf{W}_{U,H}$), the eNB reconstructs the entire channel information. Considering that the angular spread of the vertical direction is smaller than that of the horizontal direction, one can reduce the feedback overhead by setting a relatively long reporting period to the vertical codebook.

Beam Index Feedback: To obtain the UE's channel direction information (CDI) from the beamformed CSI-RSs, the eNB needs to transmit multiple beamformed CSI-RSs. When the channel rank is one, feedback of a beam index and corresponding CQI is enough, whereas when the channel rank is two with dual-polarized antennas,

[5] In this paper, we assume that discrete Fourier transform (DFT) weights are used as $\mathbf{W}_T$ for mapping between TXRU and antenna elements for simplicity. For example, $\mathbf{W}_T$ can be expressed as $\mathbf{W}_T = [\mathbf{v}\ \mathbf{v}; \mathbf{v}\ \mathbf{v}]$ in Fig. 3d.

co-phase information is additionally required for adapting channel orthogonalization between layers. For example, once the eNB obtains the CDI, this can be used for the beamforming vector of a two-port CSI-RS, and each CSI-RS port is mapped to the different polarized antennas. The UE then estimates and feeds back short-term co-phase information between two ports.

Other CSI Feedback Schemes: In the *partial CSI-RS transmission*, the CSI-RS overhead can be reduced by partitioning the 2D antenna array into horizontal and vertical ports, e.g., N_H ports in the row and N_V ports in the column. In doing so, the total number of CSI-RSs can be reduced from $N_H \times N_V$ to $N_H + N_V$. Overall channel information can be reconstructed by exploiting the spatial and temporal correlation among antenna elements [11]. In the *adaptive CSI feedback* scheme, the benefits of the beamformed and non-precoded CSI-RS transmission can be combined. First, in order to acquire long-term channel information, the eNB transmits N_T non-precoded CSI-RSs. After receiving sufficient long-term channel statistics from the UE, the eNB determines the spatial direction roughly and then transmits the beamformed CSI-RSs used for short-term and subband feedbacks. The *flexible codebook* scheme can support various 2D antenna layouts without increasing the number of codebooks. In this approach, one master codebook is designed for a large number of TXRUs, e.g., 16 TXRUs, and the specific codebook (e.g., (2 × 8), (4 × 4), or (1 × 16)) is derived based on this. To support this, the eNB needs to send the layout information via separate signaling.

Parameter	Value
Duplex method	FDD
Bandwidth	10 MHz
Center frequency	2GHz/3.5GHz
Inter-site distance	500m for 3D-UMa, 200m for 3D-UMi
Network synchronization	Synchronized
Cellular layout	3D hexagonal grid, 19 eNBs, three cells per site
Users per cell	10 (uniformly located in 3D space)
Downlink transmission scheme	$N_T \times 2$ MU-MIMO SLNR precoding with rank adaptation with two layers per UE
Downlink scheduler	Proportional fair scheduling in the frequency and time domain.
Downlink link adaptation	CQI and PMI 5ms feedback period 6ms delay total (measurement in subframe n is used in subframe n + 6) Quantized CQI, PMI feedback error: 0% MCSs based on LTE transport formats
Downlink HARQ	Maximum three re-transmissions, IR, no error on ACK/NACK, 8ms delay between re-transmissions
Downlink receiver type	MMSE : based on demodulation reference signal (DM-RS) of the serving cell
Channel estimation	Non-ideal channel estimation on both CSI-RS and DM-RS
Antenna configuration	$(M, N, P) = (8, 4, 2)$
TXRU configuration ($N_H \times N_V$)	1 × 8, 2 × 8, 4 × 8, and 8 × 8, with X-pol (0.5λ, 0.8λ antenna spacing for vertical and horizontal)
Control channel overhead, acknowledgments etc.	Control channel: three symbols in a subframe Overhead of DM-RS: 12 RE/RB/subframe Overhead of CSI-RS: in maximum 16 REs of CSI-RS every 5ms per RB (this is, in eight Tx antenna case, eight REs/RB per 10ms) Overhead of CRS: two-port CRS
Channel model	3D urban macro and micro channel model [5] with 3km/h UE speed
Inter-cell interference modeling	57 intercell interference links are explicitly considered.
Maximum number of layers	Four
Traffic model	Full buffer and non-full buffer (FTP model) with 0.5 MBytes packet and various arrival rates

Table 2. System simulation assumptions.

Performance of FD-MIMO System

In order to observe the potential gain of FD-MIMO systems, we perform system-level simulations under a realistic multicell environment. In our simulations, we test two typical deployment scenarios (3D-UMa and 3D-UMi) with a 2-tier hexagonal layout. As a performance metric, we use spectral efficiency for cell average and cell edge. Detailed simulation parameters are provided in Table 2. We first investigate the system performance of FD-MIMO systems with two types of antenna configurations. For type I and type II configurations, $(M, N, P) = (8, 4, 2)$ and $(M, N, P) = (32, 4, 2)$ are used, respectively. In the type II configuration, antenna spacing is set four times larger than the spacing of type I. To investigate the effect of the antenna structure, the ideal feedback under the full buffer traffic model (each user has an unlimited amount of data to transmit) is used. In Fig. 4a, we plot the throughput of the conventional LTE systems with 8Tx ($N_V \times N_H = 1 \times 8$), and FD-MIMO systems with 16, 32, and 64Tx ($N_V \times N_H = 2 \times 8$, 4×8, and 8×8), where N_V and N_H are the number of CSI-RSs in the vertical and horizontal dimensions, respectively. This result shows that both antenna configurations provide a large gain over the conventional 8Tx in LTE-A, resulting in a 105 percent (type I) and a 484 percent (type II) gain at the cell edge, respectively. Due to the sufficient antenna spacing, cross-correlation between channels becomes negligible, and thus the spectral efficiency of type II increases linearly with the TXRU, resulting in a 30 percent gain (cell average) and 70 percent gain (cell edge) when the number of TXRUs is doubled. However, due to the insufficient antenna spacing, the spectral efficiency of the type I configuration does not scale linearly with the number of TXRUs.

We next investigate the system performance under the finite traffic model (e.g., the FTP model) where each UE with distinct arrival time receives a file with finite size. As a performance metric, we use a user packet throughput, the number of successively received packets during the transmission period. In order to support backward compatibility and also perform a fair comparison among the schemes under test, we employ the conventional MMSE-based channel estimation. In our simulations, the following CSI feedback strategies are considered.

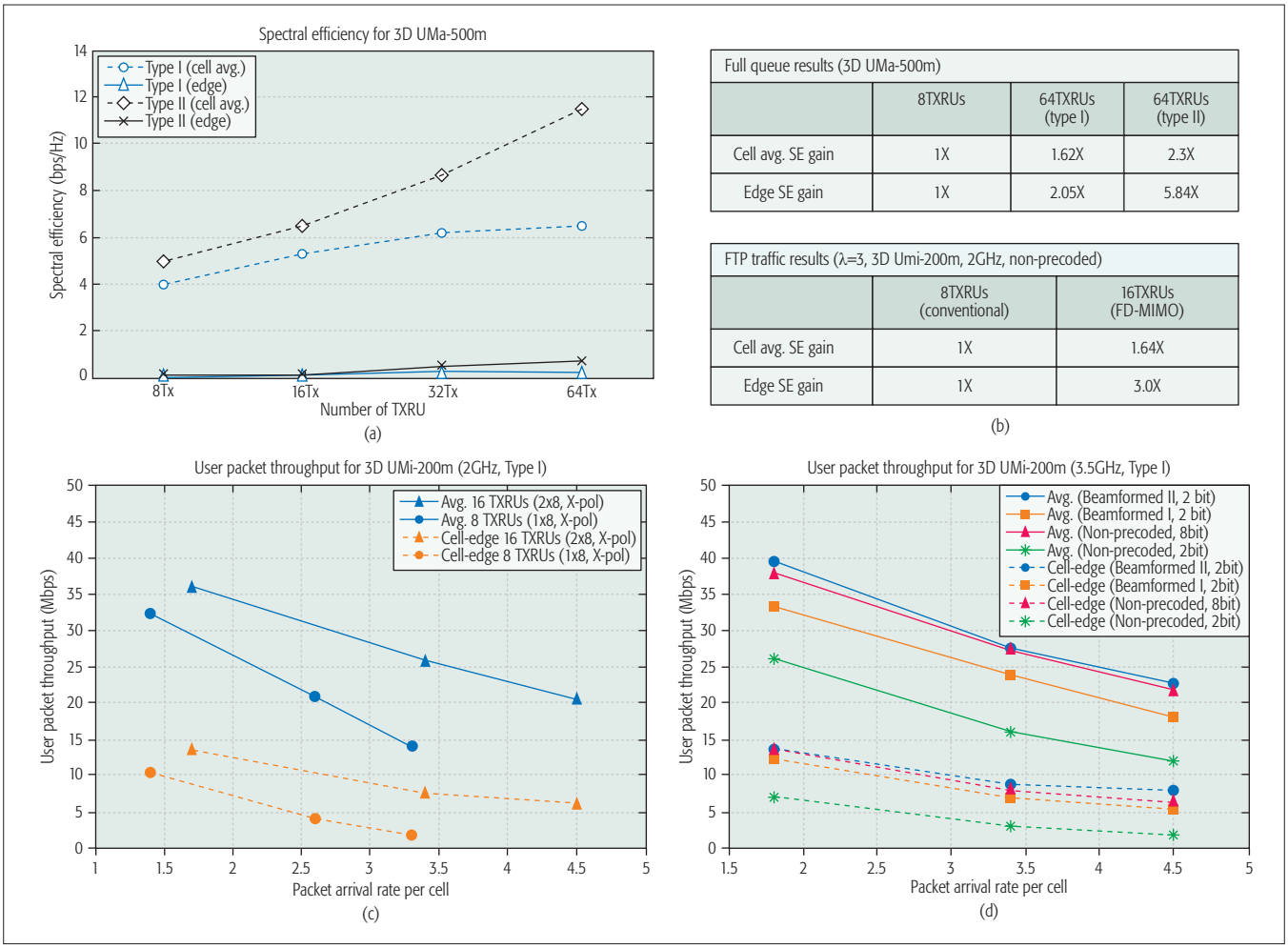


Figure 4. System level performance results and comparison with full buffer and FTP traffic model.

- **Conventional 8Tx LTE Systems**: Rel. 10 LTE-A feedback mechanism using 8TX codebook is used. The implicit feedback (RI, horizontal and vertical PMIs, CQI) is used for the CSI feedback.
- **FD-MIMO systems with**:

Non-precoded CSI-RS: A composite codebook of horizontal and vertical codebooks is used. In the case of 16Tx with ($N_V \times N_H = 2 \times 8$) antenna configuration, the codebook is generated via the Kronecker product of 2Tx and 8Tx LTE codebooks. The implicit feedback is used for the CSI feedback.

Beamformed CSI-RS scheme I: Beam index feedback is used. Four beams are used to represent the vertical angles ($N_B = 4$). Each UE reports the best beam index (BI) and corresponding CQI.

Beamformed CSI-RS scheme II: The eNB transmits both non-precoded and beamformed CSI-RSs. The UE feeds back long-term CSI (RI, long-term PMI) using the non-precoded CSI-RSs and reports the short-term CSI (BI, CQI) using the beamformed CSI-RSs ($N_B = 4$). The precoding weight of the beamformed CSI-RS is changed based on the long-term PMI.

In Fig. 4c, we plot the user throughput of the finite traffic model as a function of the packet arrival rate. Note that when the packet arrival rate is high, co-scheduled users need to be increased and thus the intercell and multiuser interference will also increase. In this realistic scenario, FD-MIMO systems outperform the conventional MIMO systems by a large margin, achieving 1.5× improvement in cell average and 3× improvement in edge user packet throughput. Note that in low network loading (low interference scenario) gain of the FD-MIMO systems is coming from the 3D beamforming. In medium to high network loading (high interference scenario), this gain is mainly due to the multiuser precoding of the 2D active antenna array. Figure 4d summarizes the throughput of various CSI feedback frameworks. With the same feedback overhead (2-bit), the beamformed CSI-RS scheme I outperforms the non-precoded scheme by a large margin, because the number of codewords for the channel feedback is only four so that the channel state information at the eNB is very coarse. Since the beamformed CSI-RS scheme II can adapt weights of the beamformed CSI-RS to generate an accurate CDI, it performs the best during the tests. It is worth mentioning that the non-precoded CSI-RS scheme requires a large amount of feedback overhead (approximately 128 quantization levels) to achieve performance comparable to the CSI-RS scheme I. From this observation, we clearly see that beamformed CSI-RS transmission is effective in controlling the precoding weights (in time, frequency, and space), feedback overhead, and pilot resource overhead.

Although our work focused primarily on the standardization in Rel. 13, there are still many challenges in the successful deployment of FD-MIMO systems in the future, including pilot overhead reduction, beam adaptation and optimization, and advanced channel estimation exploiting time and angular domain sparsity.

Concluding Remarks

In this article, we have provided an overview of FD-MIMO systems in 3GPP LTE (recently named LTE-Advanced Pro) with an emphasis on the discussion and debate conducted during the Rel. 13 phase. We discussed key features of FD-MIMO systems and main issues in the standardization of the system design, such as channel model, transceiver architectures, pilot transmission, and CSI feedback scheme. To make the most of a large number of eNB antennas in a cost effective and space effective manner, new key features, distinct from MIMO systems in conventional LTE-A, should be introduced in the standardization, system design, and transceiver implementation. These include a new transmitter architecture (array connected architecture), a new RS transmission scheme (beamformed CSI-RS transmissions), and enhanced channel feedback (beam index feedback). Although our work focused primarily on the standardization in Rel. 13, there are still many challenges in the successful deployment of FD-MIMO systems in the future, including pilot overhead reduction, beam adaptation and optimization, and advanced channel estimation exploiting time and angular domain sparsity [12].

Acknowledgments

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP-2014R1A5A1011478), and the ICT R&D program of MSIP/IITP (B0717-16-0023).

References

[1] T. L. Marzetta, "Non Cooperative Cellular Wireless with Unlimited Numbers of Base Station Antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, 2010, pp. 3590–600.
[2] Y. Kim *et al.*, "Full Dimension MIMO (FD-MIMO): The Next Evolution of MIMO in LTE Systems," *IEEE Wireless Commun.*, vol. 21, issue 3, 2014.
[3] Y. H. Nam *et al.*, "Full-Dimension MIMO (FD-MIMO) for Next Generation Cellular Technology," *IEEE Commun. Mag.*, vol. 51, issue 6, 2014.
[4] W. Zhang *et al.*, "Field Trial and Future Enhancements for TDD Massive MIMO Networks," *Proc. 26th Int'l. Symp. on Personal, Indoor, and Mobile Radio Comm. (PIMRC) Wksp. Advancements in Massive MIMO*, 2015, pp. 1114–18.
[5] 3GPP Technical Reports TR36.873, "Study on 3D Channel Model for LTE."
[6] C. Lim *et al.*, "Recent Trend of Multiuser MIMO in LTE-Advanced," *IEEE Commun. Mag.*, vol. 51, no. 3, 2014.
[7] N. Jindal, "MIMO Broadcast Channels with Finite-Rate Feedback," *IEEE Trans. Info. Theory*, vol. 52, issue 11, 2006.
[8] 3GPP Technical Reports TR36.897, "Study on Elevation Beamforming/Full-Dimension (FD) MIMO for LTE."
[9] 3GPP Technical Reports TR36.847, "E-UTRA and UTRA; Radio Frequency (RF) Requirement Background for Active Antenna System (AAS) Base Station (BS)."
[10] D. Tse and P. Viswanath, *Wireless Communication*, Cambridge University Press, 2005.
[11] B. Lee *et al.*, "Antenna Grouping based Feedback Compression for FDD-based Massive MIMO Systems," *IEEE Trans. Commun.*, vol. 63, no. 9, Sept. 2015, pp. 3261–74.
[12] J. Choi *et al.*, "Compressive Sensing for Wireless Communications: Useful Tips and Tricks," submitted to *IEEE Commun. Survey and Tutorials.*

Biographies

HYOUNGJU JI is currently working toward the Ph.D. degree at the School of Electrical and Computer Engineering, Seoul National University, Seoul, Korea. He joined Samsung Electronics in 2007, and has been involved in 3GPP RAN1 LTE technology developments and standardization. His current interests include multi-antenna techniques, massive connectivity, machine type communications, and IoT communications.

YOUNSUN KIM received B.S. and M.S. degrees in electronic engineering from Yonsei University, Korea, and his Ph.D. degree in electrical engineering from the University of Washington, in 1996, 1999, and 2009, respectively. He joined Samsung Electronics in 1999 and has been working on the standardization of wireless communication systems such as cdma2000, HRPD, and recently LTE/LTE-A. His research interests include multiple access schemes, coordination schemes, multiple antenna techniques, and advanced receivers for next generation systems.

JUHO LEE is currently a master (technical VP) with Samsung Electronics and is in charge of research on standardization of wireless communications. He received his B.S., M.S., and Ph.D. degrees in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Korea, in 1993, 1995, and 2000, respectively. He joined Samsung Electronics in 2000 and has been working on the standardization of mobile communications for 3G and 4G such as WCDMA, HSDPA, HUSPA, LTE, and LTE-Advanced, and is also actively working on research and standardization for 5G. He was a vice chairman of TSG RAN WG1 from February 2003 through August 2009, chaired LTE/LTE-Advanced MIMO sessions, and served as the rapporteur for the 3GPP LTE-Advanced Rel-11 CoMP work item.

EKO ONGGOSANUSI is currently a director of standards at the Standardization and Multimedia Innovation (SMI) Lab of Samsung Dallas. Prior to joining Samsung in 2014, he was a manager at Texas Instruments, working on cellular standards and algorithm/system designs, especially HSPA and LTE systems. Having been a 3GPP RAN1 delegate since 2005, he has contributed to numerous components of the LTE physical layer specification. He was the 3GPP rapporteur of the EBF/FD-MIMO study and work items and is currently the 3GPP rapporteur of the Enhanced FD-MIMO work item. He received his Ph.D. in electrical engineering from the University of Wisconsin-Madison (2000). He has authored a number of papers in conferences and peer-reviewed journals, and is an inventor of numerous patents in wireless communications.

YOUNGHAN NAM is currently a senior staff engineer at Samsung Research America, Richardson, TX. He has been engaged in standardization, design, and analysis of the 3GPP LTE, LTE-Advanced, and 5G NR since 2008. He is currently a study item rapporteur of the above 6 GHz channel models (3GPP TR38.900). He received a Ph.D. in electrical engineering from the Ohio State University, Columbus, OH, in 2008, and received his M.S. and B.S debrees from Seoul National University, Korea, in 2002 and 1998, respectively. His research interests include MIMO, cooperative communications, and channel modeling.

JIANZHONG ZHANG [F] is a VP and head of the Standards and Mobility Innovation Lab with Samsung Research America, where he leads research and standards activities for 5G cellular systems and next generation multimedia networks. He received his Ph.D. degree from the University of Wisconsin, Madison. From August 2009 to August 2013 he served as the Vice Chairman of the 3GPP RAN1 working group and led the development of LTE and LTE-Advanced technologies such as 3D channel modeling, UL-MIMO and CoMP, carrier aggregation for TD-LTE, etc. Before joining Samsung he was with Motorola from 2006 to 2007 working on 3GPP HSPA standards, and with the Nokia Research Center from 2001 to 2006 working on the IEEE 802.16e (WiMAX) standard and EDGE/CDMA receiver algorithms.

BYUNGJU LEE received the B.S. and Ph.D. degrees at the School of Information and Communication, Korea University, Seoul, Korea, in 2008 and 2014, respectively. He is now with the School of Electrical and Computer Engineering at Purdue University, West Lafayette, IN, USA, as a postdoctoral scholar. From 2014 to 2015 he was a postdoctoral fellow at Seoul National University, Seoul, Korea. His research interests include information theory and signal processing for wireless communications.

BYONGHYO SHIM (bshim@snu.ac.kr) is an associate professor in the Department of Electrical and Computer Engineering at Seoul National University, and a director of the Information System Laboratory. He received B.S. and M.S. degrees in control and instrumentation engineering from Seoul National University in 1995 and 1997, respectively, and an M.S. degree in mathematics and a Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign in 2004 and 2005, respectively. From 2005 to 2007 he worked for Qualcomm Incorporated, and from 2007 to 2014 he was with Korea University. His current research focuses on 5G wireless communications (physical layer system design) and big data signal processing.

ACCEPTED FROM OPEN CALL

Application of Non-Orthogonal Multiple Access in LTE and 5G Networks

Zhiguo Ding, Yuanwei Liu, Jinho Choi, Qi Sun, Maged Elkashlan, Chih-Lin I, and H. Vincent Poor

ABSTRACT

As the latest member of the multiple access family, non-orthogonal multiple access (NOMA) has been recently proposed for 3GPP LTE and is envisioned to be an essential component of 5G mobile networks. The key feature of NOMA is to serve multiple users at the same time/frequency/code, but with different power levels, which yields a significant spectral efficiency gain over conventional orthogonal MA. The article provides a systematic treatment of this newly emerging technology, from its combination with MIMO technologies to cooperative NOMA, as well as the interplay between NOMA and cognitive radio. This article also reviews the state of the art in the standardization activities concerning the implementation of NOMA in LTE and 5G networks.

INTRODUCTION

Non-orthogonal multiple access (NOMA) has recently been recognized as a promising multiple access (MA) technique to significantly improve the spectral efficiency of mobile communication networks [1–4]. For example, multiuser superposition transmission (MUST), a downlink version of NOMA, has been proposed for Third Generation Partnership Project LTE-Advanced (3GPP-LTE-A) networks [5]. Furthermore, the use of NOMA has also been envisioned as a key component in fifth generation (5G) mobile systems in [6, 7].

The key idea of NOMA is to use the power domain for MA, whereas the previous generations of mobile networks have relied on the time/frequency/code domain. Take the conventional orthogonal frequency-division MA (OFDMA) used by 3GPP-LTE as an example. A main issue with this orthogonal MA (OMA) technique is that its spectral efficiency is low when some bandwidth resources, such as subcarrier channels, are allocated to users with poor channel conditions. On the other hand, the use of NOMA enables each user to have access to all the subcarrier channels, and hence the bandwidth resources allocated to users with poor channel conditions can still be accessed by users with strong channel conditions, which significantly improves the spectral efficiency. Furthermore, compared to conventional opportunistic user scheduling, which only serves the users with strong channel conditions, NOMA strikes a good balance between system throughput and user fairness. In other words, NOMA can serve users with different channel conditions in a timely manner, which provides the possibility to meet the demanding 5G requirements of ultra-low latency and ultra-high connectivity [6].

In this article, we first provide an introduction to the basics of NOMA, such as typical NOMA power allocation policies, the use of successive interference cancellation (SIC), and the relationship between NOMA and conventional information theoretic concepts, where a simple example with two users is used to illustrate the benefit of NOMA. This introduction is then followed by a detailed overview of the recent developments in NOMA. We begin by considering the combination of NOMA and multiple-input multiple-output (MIMO) technologies. Various MIMO-NOMA designs are introduced to achieve different trade-offs between reception reliability and data rates, since spatial degrees of freedom can be used to improve either the receive signal-to-noise ratio (SNR) or the system throughput. The concept of cooperative NOMA is then described, where employing user cooperation in NOMA is a natural choice since some users in NOMA systems know the information sent to the others and hence can be used as relays. In addition, cooperative NOMA has the potential to exploit the heterogeneous nature of future mobile networks, in which some users might have better capabilities (e.g., more antennas) than others. Therefore, the reception reliability of users with poor capabilities can be improved by requesting the ones with strong capabilities to act as relays. The interplay between NOMA and cognitive radio (CR) technologies, which have also been viewed as a key component of next-generation mobile networks, are further discussed, and standardization activities to implement NOMA in LTE and 5G networks are also reviewed. Finally, research challenges and some promising future directions for designing spectrum- and energy-efficient NOMA systems are provided, followed by concluding remarks.

NOMA BASICS

In order to better illustrate the concept of NOMA, we take NOMA downlink transmission with two users as an example. As shown in Fig. 1, the two users can be served by the base station (BS) at the same time/code/frequency, but with different power levels. Specifically, the BS will send a superimposed mixture containing two messages

The authors provide a systematic treatment of this newly emerging technology, from its combination with MIMO technologies to cooperative NOMA, as well as the interplay between NOMA and cognitive radio. They also review the state of the art in the standardization activities concerning the implementation of NOMA in LTE and 5G networks.

This article was prepared in part under support of the U.K. EPSRC under grant number EP/L025272/1, and the U. S. National Science Foundation under Grants CNS-1456793 and ECCS-1343210.

Zhiguo Ding and H. Vincent Poor are with Princeton University; Zhiguo Ding is also with Lancaster University; Yuanwei Liu and Maged Elkashlan are with Queen Mary University of London; Jinho Choi is with the Gwangju Institute of Science and Technology; Qi Sun and Chih-Lin I are with China Mobile Research Institute.

Digital Object Identifier: 10.1109/MCOM.2017.1500657CM

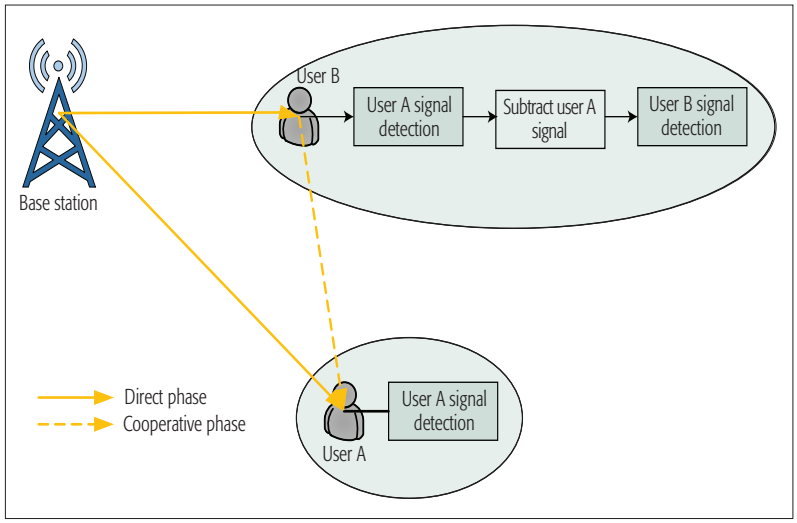


Figure 1. Illustration of a two-user NOMA network that involves non-cooperative NOMA transmission without considering the cooperative phase illustrated by the dashed line.

for the two users, respectively. Recall that conventional power allocation strategies, such as water filling strategies, allocate more power to users with strong channel conditions. Unlike these conventional schemes, in NOMA, users with poor channel conditions get more transmission power. In particular, the message to the user with the weaker channel condition is allocated more transmission power, which ensures that this user can detect its message directly by treating the other user's information as noise. On the other hand, the user with the stronger channel condition needs to first detect the message for its partner, then subtract this message from its observation, and finally decode its own information. This procedure is called SIC (as shown in Fig. 1).

The performance gain of NOMA over conventional OMA can easily be illustrated by carrying out high SNR analysis assuming an additive white Gaussian noise channel. With OMA, the achievable data rates for the two users are $1/2 \log_2(1 + \rho|h_A|^2)$ and $1/2 \log_2 (1 + \rho|h_B|^2)$, respectively, where 1/2 is due to the fact that the bandwidth resources are split between two users, ρ denotes the transmit SNR, and h_A and h_B denote the channel gains for user A and user B, respectively. Following Fig. 1, we assume $|h_A|^2 < |h_B|^2$. At high SNR (i.e., $\rho \to \infty$), the sum rate of OMA can be approximated as $1/2 \log_2 (\rho|h_A|^2) + 1/2 \log_2 (\rho|h_B|^2)$. By using NOMA, the achievable rates are

$$\log_2\left(1+\frac{\rho a_A|h_A|^2}{1+\rho a_B|h_A|^2}\right)$$

and $\log_2 (1 + \rho a_B|h_B|^2)$, respectively, where a_A and a_B are the power allocation coefficients. Therefore, the high SNR approximation for the NOMA sum rate is $\log_2 (\rho|h_B|^2)$, which is much larger than that of OMA, particularly if the channel gain of user B is much larger than that of user A. In other words, the reason for the performance gain of NOMA is that the effect of the factor 1/2 outside of the logarithm of the OMA rates, which is due to splitting bandwidth resources among the users, is more damaging than that of the factors inside of the logarithm of the NOMA rates, which are for power allocation. It is worth pointing out that NOMA suffers some performance loss at low SNR, compared to OMA, if the strict NOMA power allocation policy is used.

Downlink and uplink NOMA can be viewed as special cases of multiple access channels (MACs) and broadcast channels (BCs), and therefore the rate regions achieved by NOMA are bounded by the capacity regions of the corresponding MACs and BCs. Compared to existing information theoretic works, which mainly focus on the maximization of system throughput, a key feature of NOMA is to realize a balanced trade-off between system throughput and user fairness. Again take the two-user downlink case as an example. If the system throughput is the only objective, all the power will be allocated to the user with strong channel conditions, which results in the largest throughput, but user A is not served at all. The feature of NOMA is to yield a throughput larger than OMA, and also ensure that users are served fairly. This feature is particularly important to 5G, since 5G is envisioned to support the functionality of the Internet of Things (IoT) to connect very large numbers of devices. With OMA, connecting thousands of IoT devices, such as vehicles in vehicular ad hoc networks for intelligent transportation, requires thousands of bandwidth channels; however, NOMA can serve these devices in a single channel use. An important phenomenon in NOMA networks is that some users with poor channel conditions will experience low data rates. The reason for this is that these users cannot remove their partners' messages completely from their observations, which means that they will experience strong co-channel interference, and therefore their data rates will be quite small. In the context of IoT, this problem is not an issue, since many IoT devices need to be served with only small data rates.

MIMO NOMA Transmission

The basic idea of NOMA can be extended to the case in which a BS and users are equipped with multiple antennas, which results in MIMO NOMA. Of course, for downlink transmissions, the BS could use its multiple antennas either for beamforming to improve the signal-to-interference-plus-noise ratio (SINR) [10] or for spatial multiplexing to increase the throughput [11]. We discuss these two options in the following sections.

NOMA with Beamforming

NOMA with beamforming (NOMA-BF) can exploit the power domain as well as the spatial domain to increase the spectral efficiency by improving the SINR. To see this, we consider a system of four users, as shown in Fig. 2. There are two clusters of users. User 1 and user 3 belong to cluster 1, while user 2 and user 4 belong to cluster 2. In each cluster, the users' spatial channels should be highly correlated so that one beam can be used to transmit signals to the users in the cluster. For example, we can assume that $\mathbf{h}_3 = c\mathbf{h}_1$ for cluster 1, where $\mathbf{h}_k$ is the channel vector from the antenna array at the BS to user k, and for cluster 2, we have $\mathbf{h}_2 = c'\mathbf{h}_4$, where c and c' are constants. Furthermore, we assume that the beam to cluster 1 is orthogonal to the channel vectors of the users in cluster 2, and vice versa. That is, $\mathbf{w}_1$

$\perp \mathbf{h}_2$, $\mathbf{h}_4$ and $\mathbf{w}_2 \perp \mathbf{h}_1$, $\mathbf{h}_3$, where $\mathbf{w}_m$ denotes the beam to cluster m.

Due to BF, the signals from one cluster to another are suppressed. Thus, at a user in cluster 1, the received signal would be a superposition of x_1 and x_3, while a user in cluster 2 receives a superposition of x_2 and x_4, where x_k is the signal to user k. As shown in Fig. 2, if user 3 is closer to the BS than user 1, user 3 would first decode x_1 and subtract it to decode x_3 using SIC. User 1 decodes x_1 with the interference, x_3. Clearly, conventional NOMA of two users can be applied in each cluster. In [8], this approach is studied to support $2N$ users in the same frequency and time slot with N beams that are obtained by zero-forcing (ZF) BF to suppress the inter-cluster interference.

A two-stage BF approach is proposed using the notion of multicast BF in [9]. In [10], it is assumed that the users have multiple receive antennas. Thus, receive BF can be exploited at the users to suppress the inter-cluster interference. In this case, the BS can employ a less restrictive BF approach than ZF BF.

NOMA with Spatial Multiplexing

Unlike NOMA-BF, the purpose of NOMA with spatial multiplexing (NOMA-SM) is to increase the spatial multiplexing gain using multiple antennas. In NOMA-SM, each transmit antenna sends an independent data stream. Thus, the achievable rate can be increased by a factor of the number of transmit antennas. This requires multiple antennas at the users as well. In [11], the achievable rate is studied for NOMA-SM. In principle, NOMA-SM can be seen as a combination of MIMO and NOMA. Recall that the achievable rate of MIMO channels grows linearly with the minimum of the numbers of transmit and receive antennas under rich scattering environments, and therefore, this scaling property of MIMO should also be valid in NOMA with spatial multiplexing. Figure 3 shows the achievable rate results of NOMA-SM and OMA with different numbers of antennas (denoted by M) under a rich scattering environment. It is assumed that the number of antennas at the BS is the same as that at each user. The power of the channel gain of the weak user is four times less than that of the strong user. The total powers allocated to the strong and weak users are 3 and 6 dB, respectively. For OMA, we consider time-division MA (TDMA) with equal time slot allocation. Thus, each user's achievable rate in OMA is the same as that of conventional MIMO. However, since a given time slot is equally divided between two users, each user's achievable rate is halved. On the other hand, in NOMA, each user can use a whole time slot and have a higher achievable rate that could be two times higher than that in OMA, as shown in Fig. 3.

Cooperative NOMA Transmission

The basic idea of cooperative NOMA transmission is that users with stronger channel conditions act as relays to help users with weaker channel conditions. Again, take the two-user downlink case illustrated in Fig. 1 as an example. A typical cooperative NOMA transmission scheme can be divided into two phases, the direct transmission phase and cooperative transmission phase, respectively. During the direct transmission phase, the BS broadcasts a combination of messages for user A (weaker channel condition)

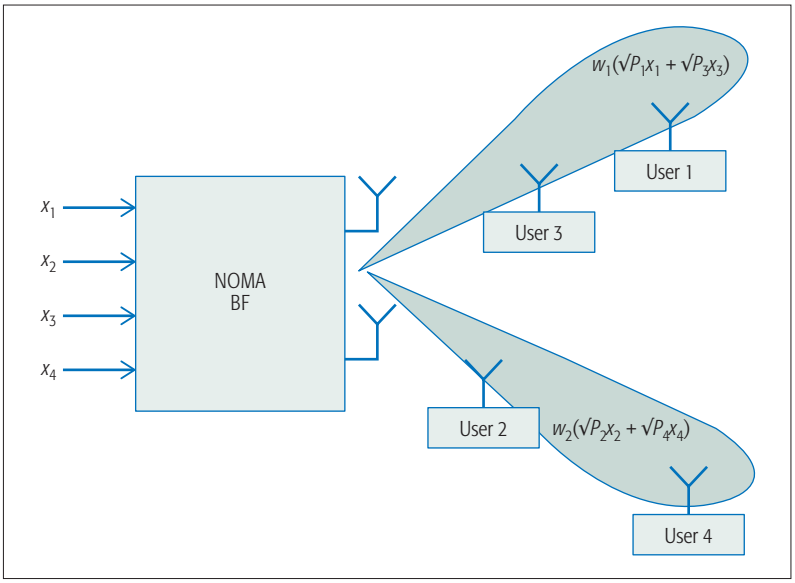


Figure 2. An illustration of NOMA with beamforming.

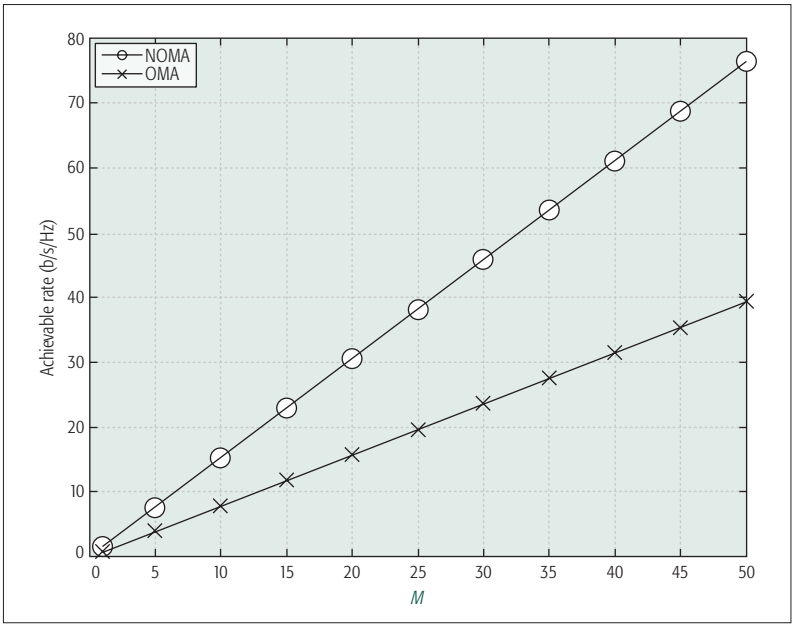


Figure 3. Scaling property of NOMA with spatial multiplexing.

and user B (stronger channel condition). During the cooperative transmission phase, after carrying out SIC at user B for decoding user A's message, user B acts as a relay to forward the decoded information to user A. Therefore, two copies of the messages are received at user A through different channels. A more sophisticated and general cooperative NOMA scheme involving K users was introduced in [12]. The advantages of cooperative NOMA transmission is that since SIC is employed at receivers in NOMA systems, the messages to the users with weaker channel conditions have already been decoded by the users with stronger channel conditions. Hence, it is natural to recruit the users with stronger channel conditions as relays. As a consequence, the reception reliability of the users with weaker channel conditions is significantly improved. The performance improvement of cooperative NOMA is illustrated in Fig. 4. Particularly, Figs. 4a and 4b demonstrate that cooperative NOMA outperforms non-coop-

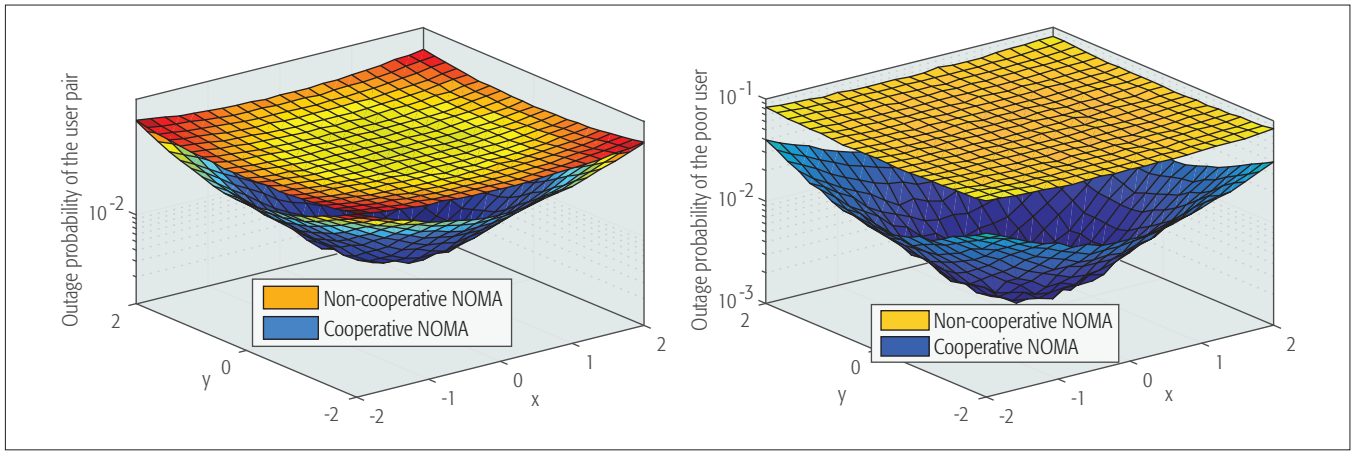


Figure 4. Performance of cooperative NOMA transmission – an example. The BS is located at (0, 0). User A is located at (5*m*, 0). The *x*-*y* plane denotes the location of User B. A bounded path loss model is used to ensure all distances are greater than one. The path loss exponent is 3. The transmit signal-to-noise ratio (SNR) is 30 dB. The power allocation coefficient for user A and user B are (a_A, a_B) = (4/5, 1/5). The targeted data rate is 0.5 bits per channel use (BPCU): a) outage probability of the user pair; and b) outage probability of the weaker user.

erative NOMA in terms of the outage probability of the user pair and the outage probability of the weaker user, respectively. In addition, in [12], it is demonstrated that cooperative NOMA achieves a larger outage probability slope than non-cooperative NOMA, which is due to the fact that the former can achieve the maximum diversity gain for all users.

It is worth pointing out that complexity is an important consideration when implementing cooperative NOMA. For example, it is not realistic to combine all users to perform cooperative NOMA. The main challenges are:

- Coordinating multi-user networks will require a significant amount of system overhead.
- User cooperation will consume extra time slots.

To overcome these issues, a hybrid MA system incorporating user pairing/grouping has been proposed and is viewed as a promising solution to reduce the system complexity of cooperative NOMA. In particular, users in one cell can be first divided into multiple pairs/groups; then cooperative NOMA is implemented within each pair/group, while OMA is implemented among pairs/groups. The performance of user pairing was investigated in [14], which demonstrates that pairing users with distinctive channel conditions yields a significant sum rate gain.

Furthermore, it is important to point out that power allocation coefficients have been recognized to have a great impact on the performance of non-cooperative NOMA [1], and thus, investigating optimal power allocation to further improve the performance of cooperative NOMA systems is an important research topic. There are other promising research directions based on cooperative NOMA. For example, considering simultaneous wireless information and power transfer (SWIPT), the NOMA user with the stronger channel condition can be used as an energy harvesting relay to help the user with the weaker channel condition without draining the latter's battery. A class of cooperative SWIPT NOMA protocols is proposed in [13], and its performance is evaluated by applying stochastic geometry.

Interplay between Cognitive Radio and NOMA

NOMA can be viewed as a special case of CR. For example, consider the two-user scenario shown in Fig. 1. User A can be viewed as a primary user in a CR network. If OMA is used, the orthogonal bandwidth allocated to user A cannot be accessed by other users, despite the fact that user A has a poor connection to the BS; that is, the bandwidth resource allocated to this user cannot be used efficiently. The use of NOMA is equivalent to the application of the CR concept. Specifically, user B, a user with stronger channel condition, is introduced to the channel occupied by user A. Although user B causes extra interference at user A and hence reduces user A's rate, the overall system throughput will be increased significantly since user B has a strong connection to the BS, as can be observed from Fig. 5a.

The analogy with CR not only yields insight into the performance gain of NOMA, but also provides guidance for the design of a practical NOMA system [14]. For example, NOMA seeks to strike a balanced trade-off between system throughput and user fairness. However, user fairness can be measured by many different metrics. By using the CR concept, an explicit power allocation policy can be obtained to meet the users' predefined quality of service (QoS). An example of such CR inspired NOMA networks is illustrated in the following by considering the two-user case shown in Fig. 1. Consider that user A (i.e., the user with weaker channel condition) has a targeted data rate R_1. Here, the CR inspired power allocation coefficient a_A needs to satisfy

$$\log_2\left(1+\frac{a_A^2|h_A|^2\rho}{\left(1-a_A^2|h_A|^2\right)\rho+1}\right)\geq R_1.$$

The aim of this CR inspired power allocation policy is to ensure that the QoS requirements at the primary user, user A, are strictly met, and the BS can explore the degrees of freedom in the power domain to serve user B opportunistically, as shown in Fig. 5b.

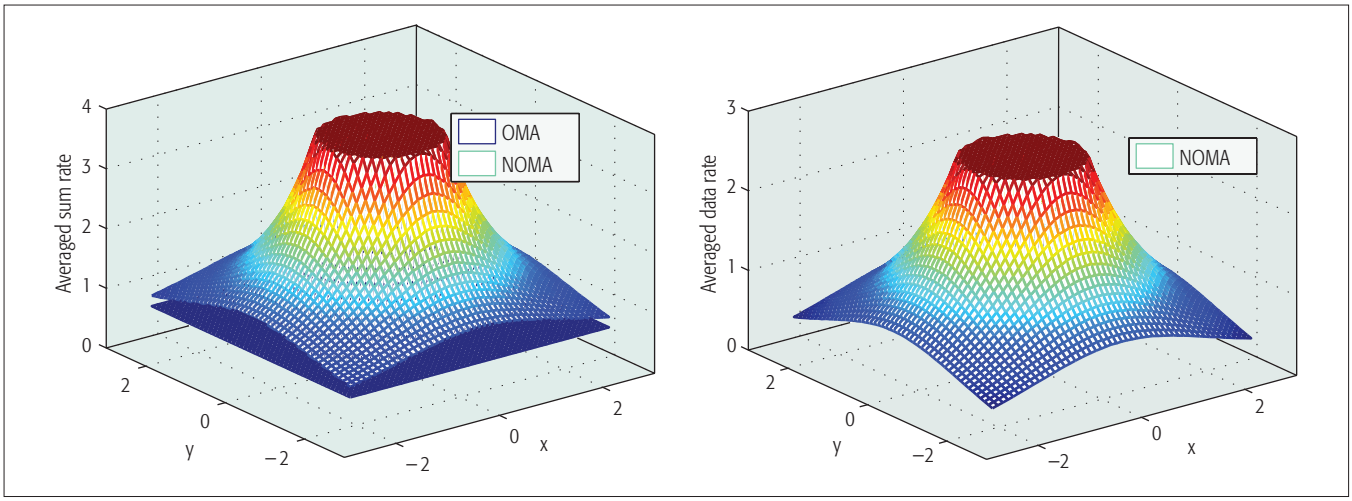


Figure 5. Performance of downlink NOMA transmission — an example. The transmit SNR is 20 dB. A fixed power allocation policy, (a_A, a_B) = (7/8, 1/8), is used in the first subfigure. The power allocation coefficient used in the second figure needs to satisfy a targeted data rate of 0.5 BPCU at the primary user. The other parameters are set the same as in Fig. 4: a) fixed power allocation; b) cognitive radio power allocation.

It is worth pointing out that this CR inspired NOMA is particularly useful in the MIMO scenario, where it is difficult to order users according to their channel conditions and hence challenging to find an appropriate power allocation policy [8].

The interplay between CR and NOMA is bidirectional, where NOMA can also be applied in CR networks to significantly increase the chance of secondary users to be connected. For example, without using NOMA, separate bandwidth resources are required to serve different secondary users, which can potentially introduce a long delay for secondary users to be served. The use of NOMA can ensure that multiple secondary users are served simultaneously, which effectively increases the connectivity of the secondary users. Power allocation at the secondary transmitters is critical to the application of NOMA in CR networks. Specifically, it is important to ensure that the secondary users are served without causing too much performance degradation at the primary receiver, where the total interference observed at the primary receiver is an important criterion. Furthermore, the power control policy used also needs to ensure that interference among the secondary users is carefully controlled in order to meet the secondary users' QoS requirements.

State of the Art for NOMA in 3GPP LTE and 5G

There have been a number of standardization activities related to the implementation of NOMA in next-generation mobile networks. In particular, 3GPP initiated a study item on downlink MUST for LTE in Rel-13, focusing on multiuser non-orthogonal transmission schemes, advanced receiver designs, and related signaling schemes [5]. Various non-orthogonal transmission schemes have been proposed and studied in the MUST study item. Based on their characteristics, they can generally be divided into three categories [15], and examples of transmitter processing for these three categories are shown in Fig. 6.

Category 1: Superposition transmission with an adaptive power ratio on each component constellation and non-Gray-mapped composite constellation

Category 2: Superposition transmission with an adaptive power ratio on component constellations and Gray-mapped composite constellation

Category 3: Superposition transmission with a label-bit assignment on composite constellation and Gray-mapped composite constellation

To characterize the gains of the non-orthogonal transmission schemes studied in MUST quantitatively, the initial link-level and system-level evaluation has been provided by various companies. It is envisioned that almost 20 percent cell-average and cell-edge throughput gains can be obtained [15]. Other topics supporting non-orthogonal transmission, such as channel state information (CSI) reporting schemes, retransmission schemes, hybrid automatic repeat request (HARQ) process design, and the signaling schemes associated with the advanced receiver, are still under active discussion.

In addition to MUST, there are other forms of non-orthogonal multiple access schemes, for example, sparse code multiple access (SCMA), pattern-division multiple access (PDMA), and multiuser shared multiple access (MUSA), which have also been actively studied as promising MA technologies for 5G [6, 7]:

- SCMA is proposed as a multi-dimensional constellation codebook design based on the non-orthogonal spreading technique, which can be overloaded to enable massive connectivity and support grant-free access. SCMA directly maps the bitstreams to different sparse codewords, and different codewords for all users are multiplexed over shared orthogonal resources (e.g., OFDM subcarriers). At the receiver, a low-complexity message passing algorithm is utilized to detect the users' data.
- The uplink MUSA scheme is based on the enhanced multi-carrier CDMA (MC-CDMA) scheme. Equipped with advanced low correlation spreading sequences (e.g., I/Q data randomly taking {–1, 0, 1} values at the transmitter), linear processing, and SIC techniques at the receiver, MUSA can achieve remarkable gains in system performance, especially when the user overloading factor is high (e.g., larger than 300 percent).

While most of the examples provided in this article consider two-user downlink scenarios, it is important to point out that NOMA can be applied to general uplink and downlink scenarios with more than two users. However, the use of superposition coding and SIC can cause extra system complexity

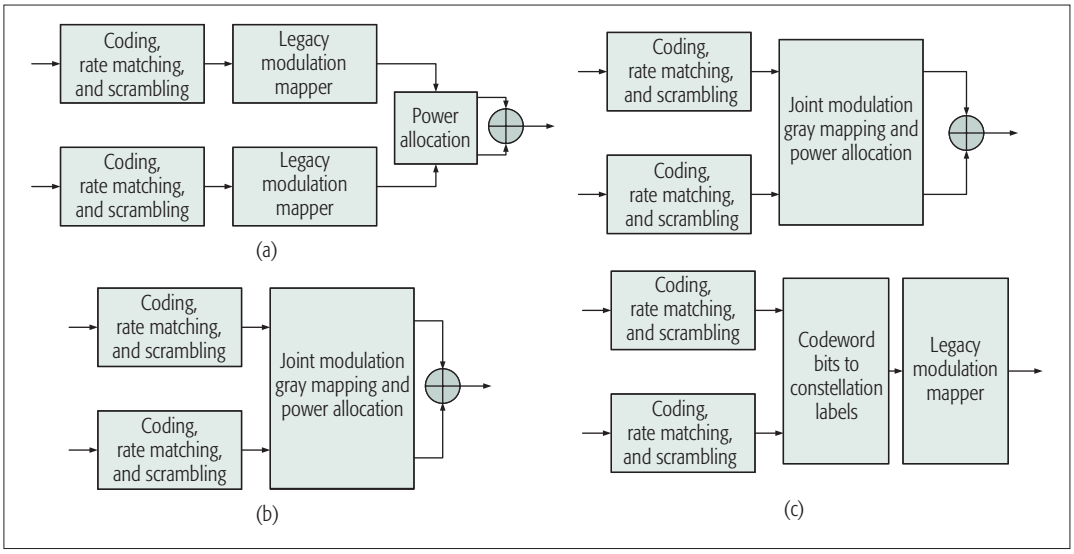


Figure 6. Examples of transmitter processing of candidate MUST schemes: a) MUST Category 1; b) MUST Category 2; c) MUST Category 3.

•PDMA employs multiple-domain non-orthogonal patterns, which are realized by maximizing the diversity and minimizing the overlaps among multiple users. The multiplexing can be realized in the code, power, and spatial domains or their combinations, which enables high flexibility for coding and decoding processing. PDMA can promote one to two times increase of the system spectral efficiency, decrease data transmission delay, and enhance quality of experience (QoE) of user access.

It is worth pointing out that these aforementioned MA candidates are closely related to the fundamental principle of NOMA, which is to serve multiple users at the same channel use. Take SCMA as an example. The term *sparsity* refers to the fact that each user can occupy only a small number of orthogonal channel uses, such as subcarriers, but there is always more than one user occupying each of the subcarriers. Therefore, at each subcarrier, SCMA can be viewed as NOMA, since multiple users are sharing the same bandwidth resource. In other words, SCMA can be built by combining NOMA with advanced strategies for subcarrier allocation, coding, and modulation.

Research Challenges

User Pairing/Clustering

While most of the examples provided in this article consider two-user downlink scenarios, it is important to point out that NOMA can be applied to general uplink and downlink scenarios with more than two users. However, the use of superposition coding and SIC can cause extra system complexity, which motivates the use of user pairing/clustering, an effective approach to reduce the system complexity since fewer users are coordinated for the implementation of NOMA. However, in cluster-based NOMA systems, it is very challenging to determine how best to dynamically allocate users to a fixed/dynamic number of clusters with different sizes. It is important to point out that the resulting combinatorial optimization problem is in general NP-hard, and performing an exhaustive search for an optimal solution is computationally prohibitive. Therefore, it is important to propose new low-complexity algorithms to realize optimal user clustering. Note that the performance of the cluster-based NOMA system can be further improved by opportunistically performing power allocation among different users in each cluster.

Hybrid Multiple Access

It has been envisioned that future cellular networks will be designed using more than one MA technique, and this trend has also been evidenced by the recent application of NOMA to 3GPP-LTE (MUST). Particularly, MUST is a hybrid MA scheme between OFDMA and NOMA, where NOMA is to be used when users have very different CSI (e.g., one user close to the BS and the other at the cell edge). Therefore, it is important to study how to combine NOMA with other types of MA schemes, including not only the conventional OMA schemes but also the newly developed 5G MA techniques. Advanced game theoretic approaches can be applied to optimize the use of bandwidth resources in the power, frequency, time, and code domains.

MIMO-NOMA

In NOMA-BF, there are still various issues and challenges. For example, optimal joint user allocation and BF schemes have not been considered as their computational complexity would be prohibitively high. Joint transmit and receive BF is also an important topic that has not been well investigated yet. The main difficulty for NOMA with spatial multiplexing is the complexity of the receivers of the users. A strong user needs to jointly detect multiple signals twice, which might be computationally demanding. The extension of NOMA with spatial multiplexing to more than two users with multiple carriers also requires user clustering and resource allocation in a multi-dimensional space (i.e., frequency, time, spatial, and power domains), which is an analytical and computational challenge.

Imperfect CSI

Most existing work on NOMA has relied on the perfect CSI assumption, which is difficult to realize, since sending more pilot signals to improve the accuracy of channel estimation reduces the spec-

tral efficiency. Therefore, it is important to study the impact of imperfect CSI on the reception reliability in NOMA systems. Another example of the strong CSI assumption is that many NOMA protocols require the CSI at the transmitter, which can cause significant system overhead. The use of only a few bits of feedback is a promising solution in NOMA systems, since obtaining the ordering of users' channel conditions is sufficient for the implementation of NOMA in many applications.

Cross-Layer Optimization

Cross-layer optimization is important to maximize the performance of NOMA in practice and meet the diversified demands of 5G, such as spectral efficiency, energy efficiency, massive connectivity, and low latency. For example, practical designs of coding and modulation are important to realize the performance gain of NOMA at the physical layer, and it is crucial to study how to feed these gains from the physical layer to the design of upper-layer protocols. This cross-layer optimization is particularly important to NOMA, which, unlike conventional MA schemes, takes user fairness into consideration, which means that the issues related to user scheduling and pairing, power allocation, pilot, and retransmission scheme design need to be jointly optimized.

Conclusions

In this article, the concept of NOMA has been first illustrated by using a simple scenario with two single-antenna users. Then various forms of MIMO-NOMA transmission protocols, the design of cooperative NOMA, and the interplay between two 5G technologies, NOMA and cognitive radio, are discussed. The recent industrial efforts for the standardization of NOMA in LTE and 5G networks have been conclusively identified, followed by a discussion of research challenges and potential solutions.

References

[1] Y. Saito *et al.*, "Non-Orthogonal Multiple Access ({NOMA}) for Cellular Future Radio Access," *Proc. IEEE VTC-Spring.*, Dresden, Germany, June 2013.

[2] J. Choi, "On Multiple Access Using H-ARQ with SIC Techniques for Wireless Ad Hoc Networks," *Wireless Personal Commun.*, vol. 69, 2013, pp. 187–212.

[3] J. Choi, "Non-Orthogonal Multiple Access in Downlink Coordinated Two-Point Systems," *IEEE Commun. Letters*, vol. 18, no. 2, Feb. 2014, pp. 313–16.

[4] Z. Ding *et al.*, "On the Performance of Non-Orthogonal Multiple Access in 5G Systems with Randomly Deployed Users," *IEEE Signal Processing Letters*, vol. 21, no. 12, Dec. 2014, pp. 1501–05.

[5] 3GPP TD RP-150496, "Study on Downlink Multiuser Superposition Transmission."

[6] White Paper, "Rethink Mobile Communications for 2020+," *FuTURE Mobile Communication Forum 5G SIG*, Nov. 2014. http://www.future-forum.org/dl/141106/whitepaper.zip.

[7] L. Dai *et al.*, "Non-Orthogonal Multiple Access for 5G: Solutions, Challenges, Opportunities, and Future Research Trends," *IEEE Commun. Mag.*, vol. 53, no. 9, Sept. 2015, pp. 74–81.

[8] Z. Ding, F. Adachi, and H. V. Poor, "The Application of MIMO to Non-Orthogonal Multiple Access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, Jan. 2016, pp. 537–52.

[9] J. Choi, "Minimum Power Multicast Beamforming with Superposition Coding for Multiresolution Broadcast and Application to NOMA Systems," *IEEE Trans. Commun.*, vol. 63, no. 3, Mar. 2015, pp. 791–800.

[10] K. Higuchi and A. Benjebbour, "Non-Orthogonal Multiple Access (NOMA) with Successive Interference Cancellation for Future Radio Access," *IEICE Trans. Commun.*, vol. E98.B, no. 3, 2015, pp. 403–14.

[11] Q. Sun, S. Han, C. I, and Z. Pan, "On the Ergodic Capacity of MIMO NOMA Systems," *IEEE Wireless Commun. Lett.*, vol. 4, Aug. 2015, pp. 405–08.

[12] Z. Ding, M. Peng, and H. V. Poor, "Cooperative Non-Orthogonal Multiple Access in 5G Systems," *IEEE Commun. Lett.*, vol. 19, no. 8, Aug. 2015, pp. 1462–65.

[13] Y. Liu *et al.*, "Cooperative Non-Orthogonal Multiple Access with Simultaneous Wireless Information and Power Transfer," *IEEE JSAC*, vol. 34, no. 4, Apr. 2016, pp. 938–53.

[14] Z. Ding, P. Fan, and H. V. Poor, "Impact of User Pairing on 5G Non-Orthogonal Multiple Access," *IEEE Trans. Vehic. Tech.*, vol. 65, no. 8, Aug. 2016, pp. 6010–23.

[15] 3GPP R1-154999, "TP for Classification of MUST Schemes," TSG-RAN WG1 #82, Beijing, China, Aug. 24–28, 2015.

Biographies

Zhiguo Ding (z.ding@lancaster.ac.uk) received his Ph.D. degree from Imperial College London in 2005, and is currently a chair professor at Lancaster University, United Kingdom. His research interests include 5G communications, MIMO and relaying networks, and energy harvesting. He serves as an Editor for several journals including *IEEE Transactions on Communications*, *IEEE Communication Letters*, *IEEE Wireless Communication Letters*, and *Wireless Communications and Mobile Computing*.

Yuanwei Liu (yuanwei.liu@qmul.ac.uk) is currently working toward his Ph.D. degree in electronic engineering at Queen Mary University of London. Before that, he received his M.S. and B.S. degrees from Beijing University of Posts and Telecommunications in 2014 and 2011, respectively. His research interests include non-orthogonal multiple access, wireless energy harvesting, massive MIMO, HetNets, D2D communication, cognitive radio, and physical layer security. He received the Exemplary Reviewer Certificate of *IEEE Wireless Communication Letters* in 2015. He has served as TPC member for IEEE conferences such as IEEE GLOBECOM.

Jinho Choi (jchoi0114@gist.ac.kr) received his Ph.D. degree from Korea Advanced Institute of Science and Technology in 1994, and is with Gwangju Institute of Science and Technology Korea, as a professor. His research interests include statistical signal processing and wireless communications. He authored two books, *Adaptive and Iterative Signal Processing in Communications* and *Optimal Combining and Detection*, published by Cambridge University Press, and currently serves as an Editor for *IEEE Transactions on Communications*.

Qi Sun (sunqiyjy@chinamobile.com) received her Ph.D. degree in information and communication engineering from Beijing University of Posts and Telecommunications in 2014. After graduation, she joined the Green Communication Research Center of the China Mobile Research Institute. Her research interest focuses on 5G communications, including new waveforms, non-orthogonal multiple access, massive MIMO, and full duplex.

Maged Elkashlan (maged.elkashlan@qmul.ac.uk) received his Ph.D. degree in electrical engineering from the University of British Columbia in 2006. From 2007 to 2011, he was with the Wireless and Networking Technologies Laboratory at Commonwealth Scientific and Industrial Research Organization, Australia. In 2011, he joined the School of Electronic Engineering and Computer Science at Queen Mary University of London. He serves as an Editor of *IEEE Transactions on Wireless Communications*, *IEEE Transactions on Vehicular Technology*, and *IEEE Communications Letters*.

Chih-Lin I (icl@chinamobile.com) is CMCC Chief Scientist of Wireless Technologies, launched 5G R&D in 2011, and leads C-RAN, Green ,and Soft initiatives. She received the *IEEE Transactions on Communications* Stephen Rice Best Paper Award and the IEEE ComSoc Industrial Innovation Award. She was on the IEEE ComSoc Board, GreenTouch Executive Board, and WWRF Steering Board, and was M&C Board Chair and WCNC SC Founding Chair. She is on the IEEE ComSoc SPC and EDB, ETSI/NFV NOC, and Singapore NRF SAB.

H. Vincent Poor (poor@princeton.edu) is with Princeton University, where his interests are in wireless networking and related fields. He is a member of the National Academy of Engineering and the National Academy of Sciences, and a foreign member of the Royal Society. He received the IEEE ComSoc Marconi and Armstrong Awards in 2007 and 2009, respectively, and more recently the 2016 John Fritz Medal and honorary doctorates from several universities.

This cross-layer optimization is particularly important to NOMA which, unlike conventional MA schemes, takes the user fairness into consideration, which means that the issues related to user scheduling and pairing, power allocation, pilot, and retransmission scheme design need to be jointly optimized.

ACCEPTED FROM OPEN CALL

Mobile Edge Computing Empowered Fiber-Wireless Access Networks in the 5G Era

Bhaskar Prasad Rimal, Dung Pham Van, and Martin Maier

Given the importance of scaling up research in the area of network integration and convergence in support of MEC toward 5G, the authors explore the possibilities of empowering integrated fiber-wireless (FiWi) access networks to offer MEC capabilities. More specifically, envisioned design scenarios of MEC over FiWi networks for typical RAN technologies are investigated, accounting for both network architecture and enhanced resource management.

ABSTRACT

The expected stringent requirements of future 5G networks such as ultra-low latency, user experience continuity, and high reliability will drive the need for highly localized services within RANs in close proximity to mobile subscribers. In light of this, the mobile edge computing (MEC) concept has emerged, which aims to unite telco, IT, and cloud computing to deliver cloud services directly from the network edge. To facilitate better understanding of MEC, this article first discusses its potential service scenarios and identifies design challenges of MEC-enabled networks. Given the importance of scaling up research in the area of network integration and convergence in support of MEC toward 5G, the article explores the possibilities of empowering integrated fiber-wireless (FiWi) access networks to offer MEC capabilities. More specifically, envisioned design scenarios of MEC over FiWi networks for typical RAN technologies (i.e., WLAN, 4G LTE, LTE-A HetNets) are investigated, accounting for both network architecture and enhanced resource management. The performance of MEC over Ethernet-based FiWi networks in terms of delay, response time efficiency, and battery life of edge devices is then analyzed. The obtained results demonstrate the feasibility and effectiveness of the proposed MEC over FiWi concept.

INTRODUCTION

Future fifth generation (5G) networks are expected to be characterized by massive capacity and connectivity, seamless heterogeneity, high flexibility, and adaptability. 5G will be highly integrative and convergent with a focus on increasing integration of cellular and wireless local area network (WLAN) technologies. Typical requirements of 5G networks include diverse quality of service (QoS) levels such as ultra-low latency and ultra-high reliability, reduced costs, low energy consumption, and support of different types of devices and applications [1]. However, to render the 5G vision a reality, huge challenges need to be addressed. Among those, capacity-limited backhaul links are identified as one of the key challenges, especially when considering the extreme densification and diversity of small cells. Network integration is another critical challenge that requires efficient merging and coordination of various types of networks (e.g., wired and wireless access networks). In 5G, the coexistence of different types of traffic will further diversify communication characteristics and requirements. Since both conventional human-to-human (H2H) traffic (e.g., voice, data, and video) and emerging types of traffic (e.g., machine-centric communications) coexist, it is challenging to design unified resource management schemes to support such coexistence to ensure that critical traffic is not affected by other coexistent traffic.

Unlike conventional centralized clouds, local clouds (cloudlets) are pushing the frontier of computing away from central nodes toward the network edge to enhance the availability and reachability of cloud services, while minimizing wide area network (WAN) latencies [2]. In light of this, mobile edge computing (MEC) [3] has recently emerged, which offers cloud capabilities (e.g., computing, storage, and caching) at the edge of networks in close proximity to mobile devices, thereby enriching users' broadband mobile experience. MEC transforms base stations (BSs, e.g., 3G, 4G) into intelligent service hubs that are capable of delivering highly personalized services. Further, MEC helps provide the backhaul with real-time information about radio access network (RAN) and traffic requirements, and thus facilitates coordination between the backhaul and RAN segments, which has not been fully realized so far [4]. Such coordination is required when, for example, radio networks need less bandwidth but the backhaul is not aware of it, and vice versa. From a business viewpoint, the emergence of MEC allows network operators, independent software vendors, and web service and content providers to create new value chains [3].

By offering attractive features including extremely low latency and high throughput, which 5G networks necessitate, MEC creates a pathway to 5G. For instance, a class of 5G applications called mission-critical Internet of Things (IoT) and Tactile Internet [5], which require extremely low latency and carrier-grade reliability (99.999 percent availability), are expected to rely on MEC [3]. Therefore, MEC is recognized by the European 5G Infrastructure Public Private Partnership (5G PPP) research body as one of the key emerging technologies for 5G networks [3]. In addition, backed by industry leaders (e.g., Intel, Nokia, Huawei, and Vodafone) who participate in the European Telecommunication Standards Institute (ETSI) MEC Industry Specification Group (ISG), MEC is expected to provide a standard-based

Digital Object Identifier:
10.1109/MCOM.2017.1600156CM

Bhaskar Prasad Rimal and Martin Maier are with the Institut National de la Recherche Scientifique; Dung Pham Van is with the KTH Royal Institute of Technology

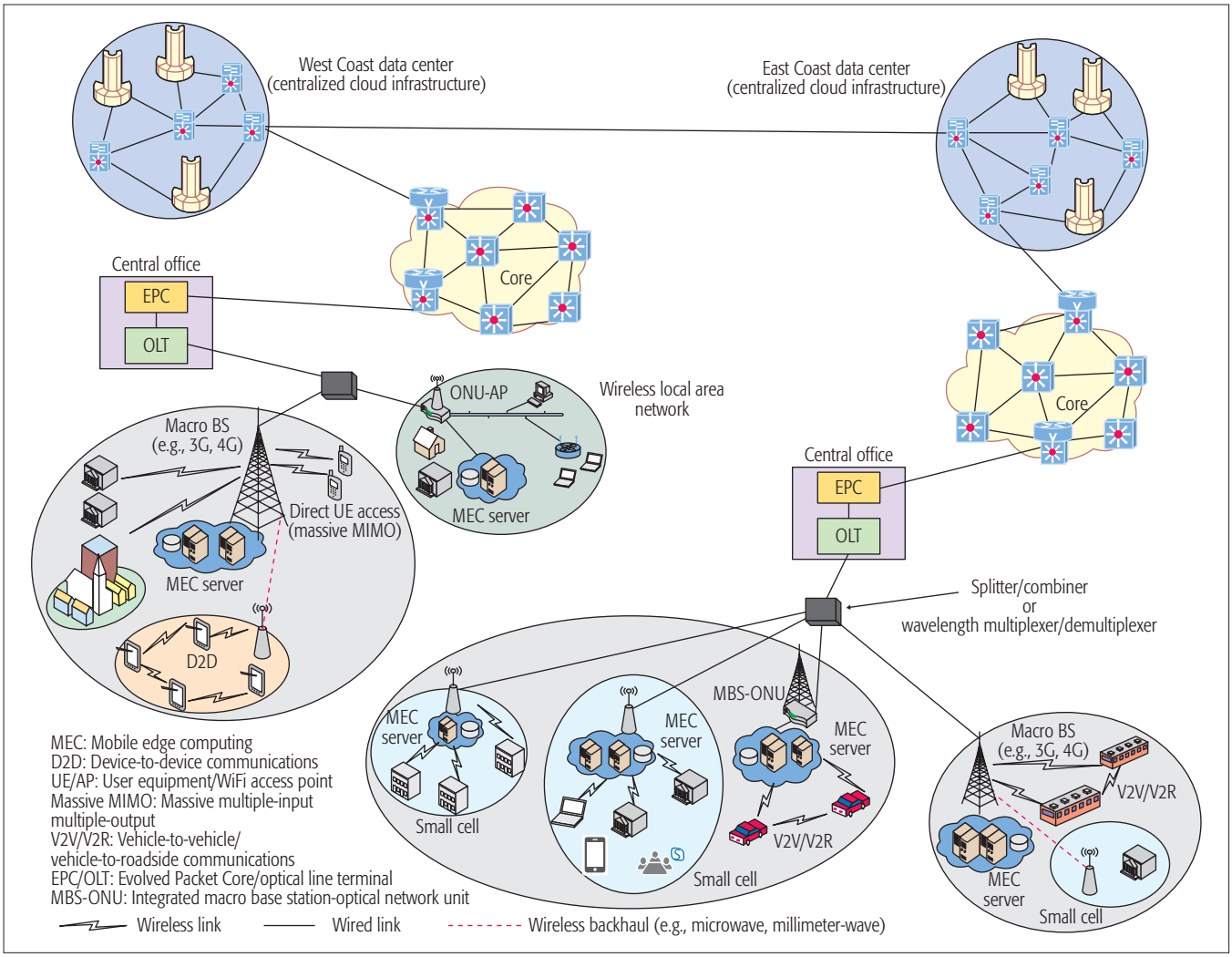


Figure 1. Illustration of mobile edge computing and cloudification of 5G networks.

approach for significant progress toward 5G. Given the growing momentum in MEC, this article aims to investigate MEC in the context of 5G.

Figure 1 provides an overview of potential deployment scenarios of MEC together with cloudification in 5G networks, including optical backhaul (e.g., Ethernet passive optical network [EPON] and 10G-EPON) and wireless backhaul technologies (e.g., millimeter-wave, microwave, and sub-6 GHz unlicensed/licensed). As shown in Fig. 1, MEC servers can be deployed in various scenarios such as at WiFi access points, cellular macro base stations (e.g., 3G, 4G/LTE-A), small cell aggregation points, and central offices in coexistence with conventional centralized clouds.

Ethernet technology is ubiquitously deployed due to its cost effectiveness, interoperability, and backward compatibility. For example, recently, it was advocated that Ethernet should be used to transport common public radio interface (CPRI) frames, the most commonly used standard in cloud-RAN (C-RAN) [6]. In this regard, Ethernet-based integrated fiber-wireless (FiWi) access networks present a compelling solution for not only broadband access but also mobile backhaul by combining the reliability and capacity of Ethernet-based optical backhaul (e.g., IEEE 802.3ah EPONs) and the extended coverage and flexibility of Ethernet-based wireless front-end (e.g., IEEE 802.11 WLANs). Recently, FiWi networks were integrated with LTE-Advanced heterogeneous networks (HetNets) [7] to further support conventional cloud computing and cloudlets [8]. Because of to their salient features, in this work, FiWi networks are further enhanced with MEC capabilities. MEC servers are integrated at the edge of FiWi networks, that is, access points or BSs collocated with optical network units (ONUs), giving rise to *MEC over FiWi networks*. Since MEC is considered one of the key emerging technologies for 5G networks [3], the introduced MEC over FiWi concept has to tackle not only the aforementioned 5G challenges, but also its own and unique challenges. Among those, the integration of MEC with existing network infrastructures (both wired and wireless), coexistence between cloudlets and conventional clouds, and enhanced resource management in consideration of backhaul/RAN coordination are of primary importance.

Research in the area of MEC is still in its infancy. Even though there is growing interest in MEC from both academia and industry, to the best of the authors' knowledge, the MEC over FiWi design scenarios and unified resource management scheme presented in this article are the first that take the network integration, H2H/MEC

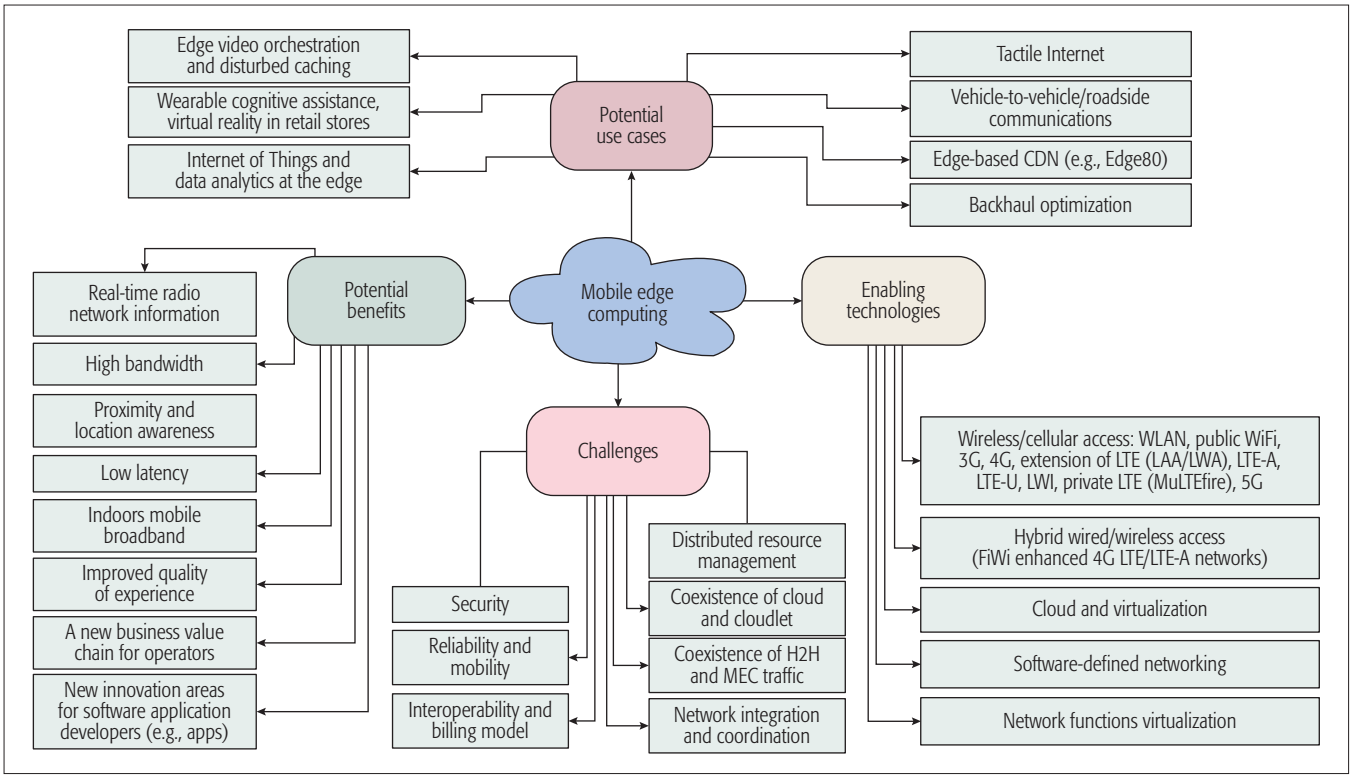


Figure 2. Overview of mobile edge computing: benefits, potential use cases, enabling technologies, and challenges (LWA: LTE and WiFi aggregation, LAA: licensed assisted access, LTE-U: LTE Unlicensed, LWI: LTE/WiFi interworking, CDN: content delivery network).

coexistence, and resource management issues into consideration. This article aims to provide a comprehensive inquiry into the design of MEC over FiWi networks. It first elaborates on the concept of MEC by studying typical service scenarios. Key technical challenges of implementing MEC are then identified and discussed in detail. Given that WLAN and 4G LTE/LTE-A are among the most common RAN technologies, three typical foreseeable design scenarios are considered:

1. MEC over WLAN-based FiWi
2. MEC over 4G LTE-based FiWi
3. Coexistence of MEC and C-RAN in FiWi enhanced LTE-A HetNets

To examine achievable network performance gains, for the first time, this article proposes a novel unified resource management scheme that jointly allocates bandwidth for transmissions of both conventional broadband traffic and MEC data in a time-division multiple access (TDMA) fashion.

The rest of the article is structured as follows. We discuss potential service scenarios of MEC and its design challenges. We propose the network architecture for MEC over FiWi. We present the resource management scheme and its performance evaluation for the MEC over WLAN-based FiWi network. Finally, we conclude the article.

Mobile Edge Computing: Typical Service Scenarios and Design Challenges

An overview of MEC, including its benefits, potential use cases, enabling technologies, and challenges, is illustrated in Fig. 2. Further, this section presents potential service scenarios and challenges in designing MEC-enabled networks for 5G in great detail.

Typical Service Scenarios

Edge Video Orchestration and Distributed Caching: Mobile edge platforms may provide edge video orchestration services, whereby visual content may be produced and consumed at a location close to subscribers in densely populated areas. Typical examples include ultra-high-definition (4K/8K) video streaming, multiple viewpoints, and live streaming events, where a large number of edge devices access the content with high required minimum bit rates. In addition, the most popular content consumed in the geographical area may be cached at an MEC server. Caching content or delivering services directly from the network edge helps backhaul and core networks reduce traffic loads. MEC therefore reduces network delay and increases throughput, thereby enriching users' broadband experience.

Backhaul Optimization: Due to the dense deployments of small cells and the increased reliance on unlicensed spectrum resources, network optimization should account for multiple RAN technologies and localized service information. MEC services such as traffic and performance monitoring help provide the backhaul segment with real-time information about RAN and traffic requirements [4]. Such information may be used for advanced backhaul optimization methods. This type of real-time optimization (e.g., rerouting traffic) allows for more efficient use of network resources, better user experience, and advanced traffic congestion management.

Vehicle-to-Vehicle/Roadside (V2V/V2R) Com-

munications: Connected vehicles, one typical 5G use case, are a set of networked systems with hundreds of sensors (e.g., Google self-driving cars and high-speed trains). While processing and storing V2V/V2R data centrally may be effective in some cases, it is not practical in many other cases, where real-time information (e.g., about traffic accidents) with low latency and low jitter is required. For instance, V2R services usually have a latency requirement of 10 ms or below. MEC helps provide low-latency and location-based information in real time. This will enable a nearby car to receive data on the order of milliseconds, thus allowing the driver to immediately switch lanes, slow down, or change his/her route. For example, recently, Deutsche Telekom has deployed MEC to enable V2V communications with latencies under 20 ms, compared to the traditional 100 ms.

Internet of Things Services: The IoT ecosystem, a network of billions of physical objects or things, provides very diverse types of applications and services with a wide range of QoS requirements. It is expected that most of the network intelligence will reside closer to IoT sensors, and many IoT applications will have stringent latency, reliability, and security requirements [9]. For instance, telesurgery applications require latency below 10 ms. Further, resource constrained IoT devices may offload their data onto MEC servers, thereby extending battery life, reducing latency, and enabling applications to respond in real time. Also, MEC is expected to provide reliable intercity communications by moving computing and processing resources closer to sensor nodes in support of capillary networks in large-scale IoT deployments (e.g., smart cities).

Design Challenges

Network Integration and Coordination: The key challenge is to seamlessly integrate multiple network technologies (i.e., wired, wireless, cellular) in order to support MEC capabilities. Indeed, MEC servers should be compatible with underlying network architectures, interfaces, and functionalities for improved network performance [3]. Given the diverse potential deployment scenarios of MEC over multiple RANs (e.g., WLAN, LTE), network integration should be considered in the first place at both the architectural and protocol levels. Further, the coordination between backhaul and front-end segments of converged networks in 5G is another important issue. Although MEC holds promise to realize such coordination, it is challenging to incorporate the coordination and synchronization into the MEC design.

Distributed Resource Management: It is important to ensure that different and diverse edge devices (e.g., moving users, mobile devices, and connected vehicles) have access to network resources (e.g., bandwidth, storage) at the edge [10]. Since the design complexity increases when a shared but limited amount of resources must be allocated to accommodate dynamic needs of such devices at the edge, designing a resource management scheme with QoS guarantees for MEC networks is more challenging than in conventional networks.

Coexistence of H2H and MEC Traffic: The coexistence of H2H and MEC traffic further diversifies communication characteristics and requirements. The integration of MEC and existing access networks will need to cope with the coexistence of broadband access and MEC traffic. While both H2H and MEC support voice calls, video streaming, web surfing, and social networking, MEC applications have their own diverse QoS requirements such as real-time response, and location and mobility awareness. Importantly, MEC also includes control communications (e.g., emerging Tactile Internet applications). Such applications have very stringent end-to-end latency requirements of about 1 ms to avoid perceivable lags to remotely control actuators (e.g., robots) [5].

Cloud and Cloudlet Coexistence: Centralized clouds and distributed cloudlets may coexist and be complementary to each other, and thus support a more diverse set of emerging applications and services in 5G networks. However, determining where an application is executed, at either a cloudlet or a conventional cloud, is a nontrivial task. It depends on the available infrastructure and application requirements, as well as willingness of users to pay. Some applications or parts of an application may be executed at the edge device itself, cloudlets, or centralized clouds. A fundamental question is how to identify which part of the application to offload onto clouds/cloudlets and which not, given that partitioning applications for offloading increases the complexity and overhead of the MEC design. Further investigation is required to find smart strategies for coexistent cloud and cloudlet systems under realistic network conditions. On the other hand, from a business model point of view, given that clouds and cloudlets may be owned by different operators, interoperability and billing issues may arise. It is challenging to coordinate with individual cloud service providers, each having their own interfaces. To this end, a common deployment and management platform for a multi-cloud environment is desirable to optimize network performance and minimize costs.

Reliability and Mobility: Given that mission-critical applications (e.g., telecontrol of heavy machinery) may rely on MEC, meeting the required level of reliability and resilience of such applications is challenging. For instance, providing carrier-grade reliability (99.999 percent availability) at the network edge will create a new dimension of complexity for designing robust and highly optimized protocols. Further, MEC should provide service continuity, application and virtual machine (VM) mobility, and application-specific user-related information. In general, mobility management (e.g., behavior of mobile users' mobility and its predictability [10] and handover optimization) is nontrivial. When a user moves, his/her VM should be seamlessly transferred between MEC servers. Since VM mobility is sensitive to various factors such as data volume, processing speed, compression ratio, and bandwidth, it is challenging to do such migration of VMs in a seamless manner, without degrading the quality of experience (QoE).

MEC over FiWi Network Architecture

This section presents different design scenarios of MEC over FiWi networks from the architectural perspective.

The key challenge is to seamlessly integrate multiple network technologies, i.e., wired, wireless, cellular, in order to support MEC capabilities. Indeed, MEC servers should be compatible with underlying network architectures, interfaces, and functionalities for an improved network performance.

	Conventional D-RAN	Cloud-RAN	Emerging cloudlet enhanced D-RAN
Base stations	Standard complexity and high cost	Lower complexity, conceivably cheaper	Lower complexity and cheaper
Diversity gains	Per-BS diversity gains	Multiplexing and computational (multi-user) diversity gains	Both multiplexing and computational diversity gains at the network edge
Hardware resources	Dedicated digital signal processors (DSPs) or application-specific integrated circuit implementations	High-volume commodity hardware (e.g., general-purpose processors)	High-end rack servers, cloudlets, Nokia Siemens Networks' radio applications cloud server
Backhauling	Backhaul links of up to several tens of kilometers, leading to high latencies in the range of several tens of milliseconds	1) Backhaul links with critically higher throughput required and latency on the order of a few milliseconds; 2) CPRI is widely used for fronthaul interface	1) Significantly lower round-trip latency and real-time information offered within RAN; 2) Ethernet-based (e.g., IEEE 1904.3) and virtual networking techniques in the fronthaul in support of variable bit rates
Flexibility	Hardware driven	Software driven	1) Driven by both hardware and software (e.g., Nokia Liquid Application, RACS, OpenStack++); 2) multipoint-to-multipoint communication
Programmability	Based on DSP	Based on general-purpose processor	Supports both DSP and general-purpose processor

Table 1. Comparison of conventional D-RAN, C-RAN, and emerging cloudlet enhanced D-RAN.

MEC over FiWi Networks

FiWi networks are realized by integrating optical and wireless technologies, thereby forming a powerful platform to provide future-proof connectivity for existing and emerging applications and services [8]. FiWi networks can be built using any optical access networks, for example, EPON and next-generation PON 1&2 in the backhaul segment, and RAN technologies in the front-end segment, such as Gigabit-class IEEE 802.11ac very high throughput (VHT) WLAN, and 4G LTE/LTE-A.

An Ethernet-based FiWi network may rely on the emerging cloudlet enhanced distributed RAN (D-RAN) [11] based on so-called radio-and-fiber (R&F) technologies [8]. In cloudlet enhanced D-RAN, the functionalities of remote radio heads (RRHs) and baseband units (BBUs) are split, whereby RRHs and BBUs are linked via an Ethernet interface and the baseband processing is done at a MEC server. Alternatively, a FiWi network can be realized via radio over fiber (RoF) technologies such as C-RAN. In C-RAN, BBUs that connect a number of macro BSs or small cells (i.e., femto-, picocell) are centralized with pool baseband processing (i.e., BBU pool), while RF signaling is digitized and transmitted over optical fiber for fronthauling (i.e., between RRHs and BBUs). Further, the digitalized RF signal received at the RRH is then converted to an analog signal before being transmitted to its associated edge devices in the downlink transmission. CPRI is the currently used standard transmission technology in the fronthaul. However, as fronthaul and backhaul in future mobile networks will converge, CPRI may be mapped into Ethernet frames, as specified in the emerging IEEE 1904.3 standard.

Table 1 summarizes the architectural evolution of cloud enabled RAN technologies. To better understand the evolution, typical features of cloudlet enhanced D-RAN, C-RAN, and conventional DRAN [12] are summarized in Table 1. Note that, unlike the other two concepts, conventional DRAN separates the processing of control and user plane processes. The RAN server is responsible for the management of RAN-specific functions, (e.g., mobility, bearer services, and paging), whereas radio and cell-specific functions (e.g., data bearers, scheduling) are decentralized to the base station.

MEC over FiWi networks can be realized by enabling cloud computing capabilities using, for example, a powerful rack server or cloudlets directly connected to the integrated ONU-mesh portal point/access point (AP) (i.e., at the edge of FiWi networks). Even though MEC servers could be deployed at different locations in FiWi networks such as at a central office or anywhere along the backhaul segment, the most important design principle is that MEC servers should be in close proximity to subscribers.

MEC over FiWi networks support the coexistence of conventional clouds and cloudlets (Fig. 3) in order to provide both centralized and distributed cloud services. Centralized clouds are suitable for stateless services with limited data transmissions, such as web and batch processing. Indeed, transmitting large volumes of data streams to a remote cloud is not only costly but also incurs higher latencies. Conversely, MEC is suitable for real-time applications (e.g., interactive collaboration, augmented reality, and gaming) that require low latency and location-aware data processing. The coexistence of conventional clouds and MEC in FiWi networks provides a powerful hybrid cloud platform for a wide range of applications and helps meet the QoS requirements of both human and machine generated traffic in 5G networks. Further, from an *economical perspective*, MEC over FiWi helps reduce capital and operational expenditures (CAPEX/OPEX) due to sharing of existing fiber and wireless infrastructures. From a *technical viewpoint*, a unified resource management mechanism can be designed to efficiently operate such a highly integrated network.

It is widely agreed that WLAN and LTE/LTE-A HetNets represent two major RAN technologies for 5G networks [1]. Therefore, in this work, three design scenarios with WLAN, LTE, and HetNets deployed at the front-end of MEC over FiWi networks are considered in detail in the following.

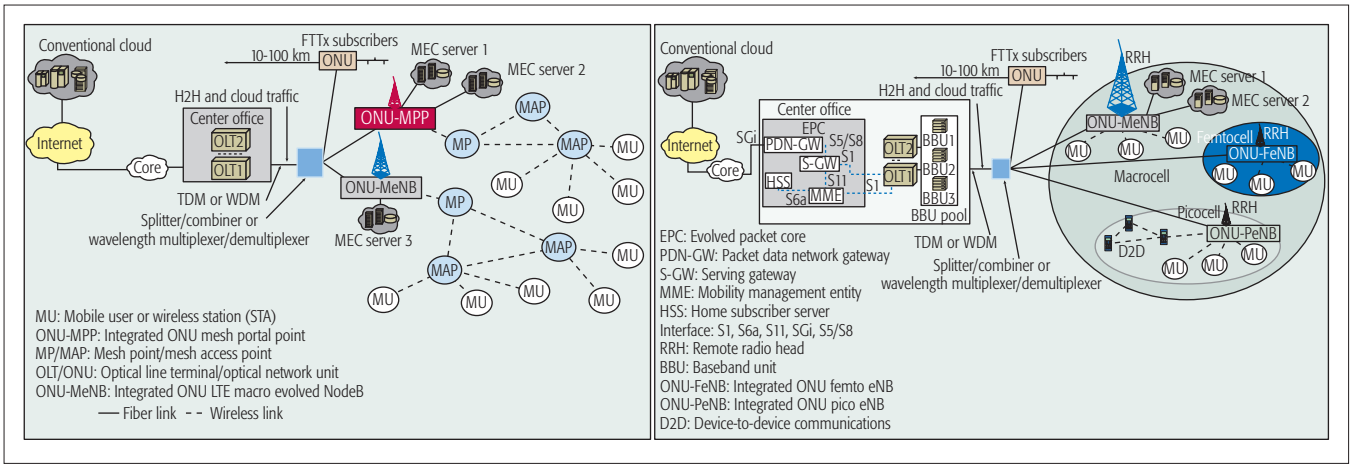


Figure 3. MEC over FiWi network architectures: a) MEC over Ethernet-based FiWi networks and MEC over 4G LTE-based FiWi networks; b) coexistence of MEC and C-RAN over FiWi enhanced 4G LTE HetNets.

MEC over Ethernet-Based FiWi Networks

This design scenario is based on the integration of EPON or 10G-EPON in the optical backhaul and a wireless Ethernet LAN (WLAN) in the front-end. Figure 3a illustrates MEC over Ethernet-based FiWi network architecture as a shared communication platform for both broadband H2H and MEC services. The backhaul consists of an OLT located at the central office that serves a single or multiple ONUs at the customer premises. A subset of ONUs is located at the premises of residential or business subscribers, providing fiber to the X (FTTx) services (e.g., fiber to the home) to a single or multiple wired subscribers. The second subset of ONUs is equipped with a mesh portal point (MPP) to interface with the WiFi mesh network, consisting of mesh points (MPs) and mesh APs (MAPs), each serving mobile users within their coverage area. The collocated ONU-APs/MPPs are realized by using R&F technologies [8]. MEC servers are connected to ONU-APs/MPPs through optical fiber point-to-point links. The resource management scheme and its performance evaluation for this design scenario are explained in greater detail later.

This design scenario is considered for medium- and small-scale MEC applications, where each ONU-AP/MPP zone covers a small number of edge devices. It may be suitable for Tactile Internet applications and 5G-enabled robots.

MEC over 4G LTE-Based FiWi Networks

In this scenario, MEC servers are deployed at 4G LTE macro BSs (MeNBs), as shown in Fig. 3a. Bandwidth allocation is based on a pool/request/grant mechanism. The OLT schedules transmissions and allocates bandwidth to each ONU-MeNB in a centralized fashion. Upon a granted bandwidth, each ONU-MeNB makes local decisions to schedule transmissions and allocate the bandwidth to its associated edge devices and MEC servers in a fair and distributed manner. Depending on the QoS requirements, an ONU-MeNB forwards the packet to either the OLT or the MEC server.

MEC over 4G LTE-based FiWi networks are especially suitable for edge video orchestration, fast moving users (e.g., train passengers), and wide-area applications (e.g., smart cities). Further, it will significantly accelerate the deployment of V2R communications, where mobility and low latency are highly desirable.

Coexistence of MEC and C-RAN in FiWi Enhanced LTE-A HetNets

C-RAN and cloudlet enhanced D-RAN may coexist in HetNets. Figure 3b depicts a way to realize such coexistence. Given the ability to overlay multiple channels, wavelength-division multiplexing (WDM)-PON technologies may be deployed in this scenario without upgrading the optical infrastructure, where C-RAN and cloudlet enhanced D-RAN may use different wavelength channels for baseband and RF transmission. This helps reduce CAPEX, since WDM-PON provides a substantial reduction in the number of fibers used. The collocated ONU-FeNB (femtocell BS) and ONU-PeNB (picocell BS) may rely on WDM-based C-RAN, while an ONU-MeNB may rely on cloudlet enhanced D-RAN, as shown in Fig. 3b.

Since MEC servers are connected to the ONU-MeNB, and the ONU-MeNB may rely on cloudlet enhanced D-RAN, the scheduling and bandwidth allocation should be handled by the ONU-MeNB. In the coordination of BBUs, the OLT is fully responsible for scheduling transmissions and allocating bandwidth to each ONU-FeNB and ONU-PeNB in a centralized fashion. However, due to not only the heterogeneity of constituting network components but also the presence of different radio access technologies in the same network, designing a unified resource management scheme is more complex than the previous two scenarios.

Another possibility of deploying MEC over LTE-A HetNet is at the aggregation level. When multiple BSs are located close to each other, it is effective to deploy an MEC server there. Serving several BSs with a single MEC server not only centralizes computing resources but also reduces CAPEX/OPEX of network operators. Further, device-to-device (D2D) links may be employed between edge devices to improve spectrum efficiency and reduce backhaul traffic loads. Deploying MEC servers in LTE-A HetNets and at the aggregation levels pose several challenges. Among others, advanced resource management schemes are needed to jointly coordinate and synchronize a large number of BSs.

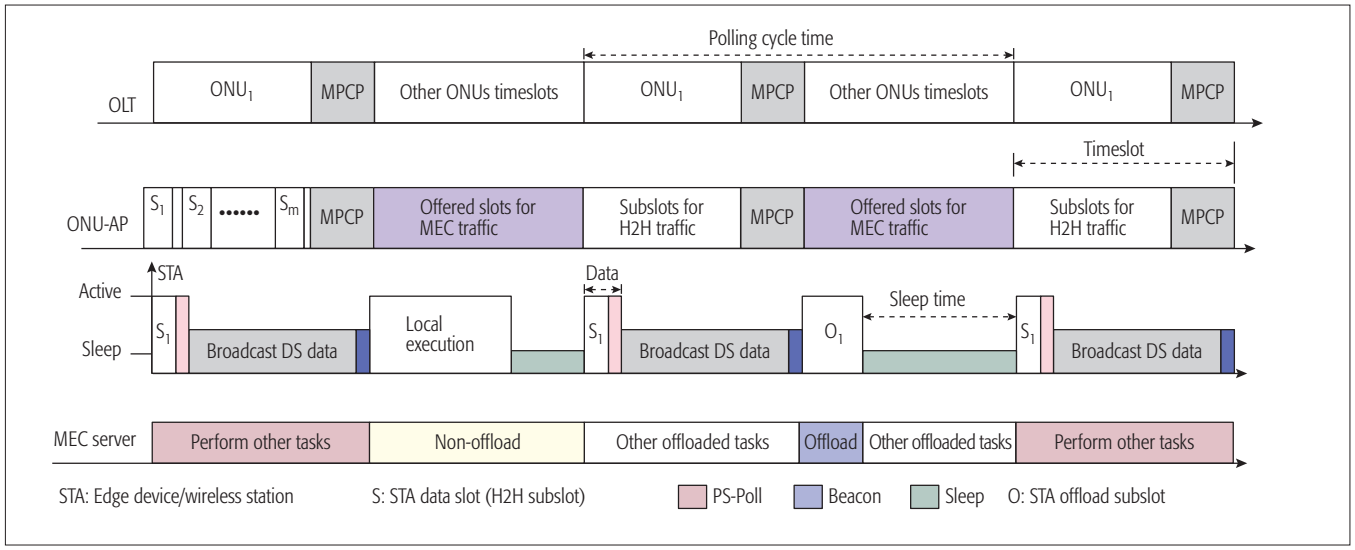


Figure 4. Illustration of TDMA-based unified resource management scheme and sleep mode for MEC over WLAN-based FiWi networks.

This design scenario is suitable for networks with ultra-high densification, fast moving users (e.g., V2V and drones), and mission-critical IoT applications.

Resource Management and Performance Evaluation

Resource Management Scheme for MEC over Ethernet-Based FiWi Networks

This subsection proposes a TDMA-based unified resource management scheme for MEC over Ethernet-based FiWi networks. TDMA scheduling offers advantages such as reduced energy consumption and collision-free transmission compared to contention-based protocols. In this work, a conventional WLAN topology is considered, where stations (STAs) with no mobility connect directly to APs that are collocated at ONUs.

Figure 4 illustrates the unified resource management scheme. The system is based on a two-layer time-division multiplexing (TDM) design. The first layer is designed for the optical backhaul, where the OLT schedules time slots and allocates bandwidth to ONU-APs via multipoint control protocol (MPCP) messages (`GATE` and `REPORT`). In the second layer, the ONU-AP assigns bandwidth in subslots and schedules transmissions of both H2H and MEC traffic for its associated STAs. The transmissions of MEC offloaded traffic and computation results are scheduled outside the H2H subslots within an EPON polling cycle time (Fig. 4) to allow for H2H/MEC coexistence without degrading H2H network performance. The ONU-AP receives offloaded traffic from its associated STAs. It then immediately relays the traffic to the MEC server using a dedicated point-to-point fiber communication link. When the ONU-AP receives the computation results from the MEC server, it broadcasts them to its STAs. A given STA sends its H2H traffic to the ONU-AP within its assigned H2H subslots.

The ONU-AP allocates subslots to its STAs by means of WLAN `Beacon` and `PS-Poll` frames. These frames are extended by using their optional bits to include subslot parameters and bandwidth requests, respectively. The ONU-AP broadcasts a `Beacon` to its STAs containing an uplink H2H subslot map, whereby each STA sends a `PS-Poll` at the end of its own H2H subslot. The ONU-AP aggregates the requested bandwidth and reports it to the OLT via a `REPORT` at the end of its time slot. The STA transmits offloaded traffic to and receives computation results from the MEC server in the offload subslot. Further, a power-saving method is employed using a similar approach as in [13] to extend the battery life of edge devices. The general idea is to schedule sleep mode for the STA in a PON cycle if it is idle after the completion of both its H2H transmission and MEC offloading subslots. For network synchronization, the timestamp mechanism specified in the EPON standard is adopted, where all network devices assign their local clocks to the OLT global clock.

Performance Evaluation

This subsection discusses results and findings obtained from an analytical evaluation of the considered scenario. Computation offloading, that is, offloading compute-intensive tasks to an MEC server connected to an ONU-AP, should be performed if the time to execute a task on the edge device locally is longer than the response time of offloading that task onto an MEC server. This response time difference is called offload gain. The response time efficiency is defined as the ratio of the offload gain and the response time of a task that is locally executed on edge devices. Packet delay is the time a packet waits in a data buffer. The battery life of an edge device is computed based on its battery capacity and the average power consumption.

The TDMA-based resource management scheme is analyzed assuming a polling system with M/G/1 queues [14]. For computation offloading, a face detection application is considered using OpenCV.[1] An image size of 500 × 426 pixels greyscale mode is converted into kilobytes and used as input data for face detection. The data load of a computation task is assumed to be fragmented into packets of fixed size, and the application is divided into a number of fine-grained tasks, similar to [15]. The application is modeled as a

[1] Open source computer vision library (OpenCV), https//www.willowgarage.com/pages/software/opencv

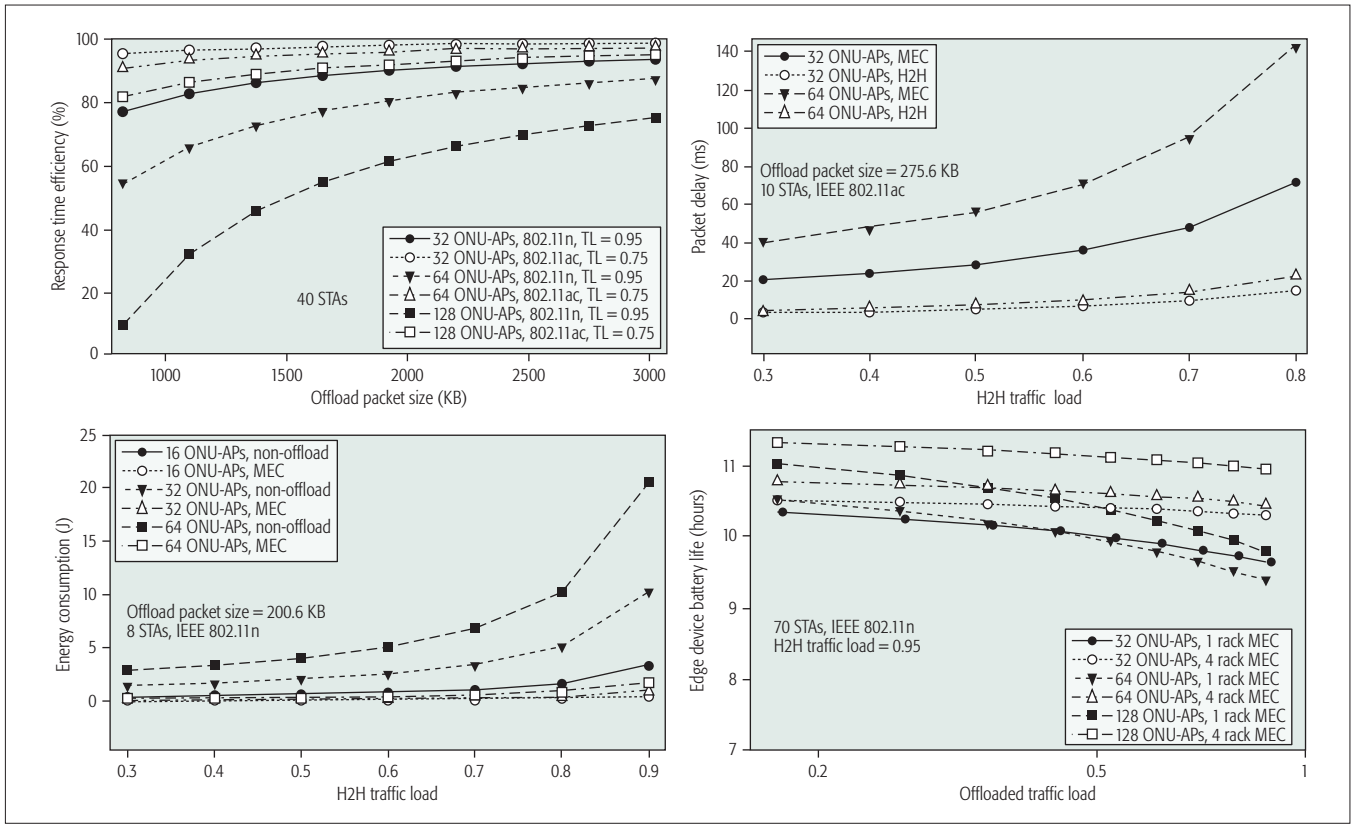


Figure 5. a) response time efficiency vs. offload packet size; b) mean packet delay of co-existing H2H and MEC traffic; c) energy consumption vs. H2H traffic load for MEC and non-MEC scenarios; d) battery life of edge devices vs. offloaded traffic load. (TL: H2H traffic load).

call graph (directed acyclic graph) with computational components, each characterized by size of methods, energy consumption, and number of instructions to perform the computations.

In the evaluation, an IEEE 802.3ah EPON is considered, and the H2H traffic load (intensity) is varied from 0.05 to 0.9 with Poisson distribution and average packet transmission time of 5.09 μs as in [14]. MEC servers are assumed to be 1 rack server (100 CPUs) or 4 rack servers with a clock speed of 3.2 GHz. The considered edge devices are HP iPAQ PDAs with a 400 MHz clock speed and a battery capacity of 1000 mAh Lithium-Ion with power levels in active and sleep states of 0.9 W and 0.3 W, respectively. Maximum data rates of 300 Mb/s and 6900 Mb/s are considered for the wireless front-end based on IEEE 802.11n and IEEE 802.11ac VHT WLANs, respectively.

Figure 5a depicts the maximum achievable response time efficiency for different offload packet sizes. For increasing offload packet sizes, the average response time efficiency asymptotically approaches 100 percent. For instance, for a typical case of 32 ONU-APs and an offload packet size of 825.60 kB, the average response time efficiency equals 77.10 and 95.36 percent for the two considered WLANs, respectively. This translates into a response time reduction of 77.10 and 95.36 percent with respect to the response time obtained in a non-offload scenario. Figure 5a also reveals that MEC over VHT WLAN-based FiWi helps increase the maximum achievable response time efficiency significantly. This set of results verifies that MEC over FiWi-based computation offloading is a promising solution for improving users' QoE in support of a wide range of future 5G applications.

Figure 5b shows the packet delay performance of coexisting H2H and MEC traffic under varying H2H traffic loads. Both curves have similar behavior. However, the offload packet delay is higher than H2H delay because offloaded traffic waits longer before being transmitted due to a longer resultant PON polling cycle time. An H2H mean packet delay of below 23 ms is achieved for all values of H2H traffic loads. Importantly, for a typical scenario of 32 ONU-APs and a H2H traffic load of 0.3, an MEC mean packet delay of 20.50 ms is obtained. Further, even with a H2H traffic load of 0.7, the MEC delay remains below 95 ms, which is not feasible with typical centralized clouds. This means that many delay-sensitive applications can be offloaded on the MEC over an Ethernet-based FiWi network. However, as shown in Fig. 5b, the offloading may not be efficient when H2H traffic load is greater than 0.8 because of significantly longer delay.

Figure 5c compares the average energy consumption between the MEC scenario and local execution scenario (i.e., non-offload). The energy consumption is a function of H2H traffic load, polling cycle time, and number of ONU-APs and STAs. It is shown in Fig. 5c that MEC-over-FiWi-based offloading significantly reduces the energy consumption of edge devices. For instance, for 32 ONU-APs and an H2H traffic load of 0.8, the energy consumption in the MEC scenario is 10.13 times less than in the non-offload scenario. Finally, Fig. 5d shows the battery life of edge devices as a function of offloaded traffic load and

While TDMA scheduling has been shown to be effective in MEC over Ethernet-based FiWi networks, resource management schemes for other two design scenarios and an in-depth performance analysis, especially with wireless backhaul technologies would be an interesting topic for future research.

number of MEC servers. Remarkably, by employing the proposed unified resource management scheme and sleep mode scheduling, up to 11.035 h and 11.302 h of battery life can be obtained with a 1-rack and 4-rack MEC server, respectively. This verifies that an MEC-over-Ethernet-based FiWi network with the proposed scheme helps prolong the battery life of edge devices significantly. Note that due to the TDMA nature of the proposed scheme, employing backhaul links with higher capacity (e.g., 10G-EPON) in the proposed MEC over FiWi network would not affect the obtained findings.

Conclusions

This article introduces the novel concept of MEC over FiWi networks. Besides several benefits of MEC, a number of interesting research challenges in terms of network integration and coordination, distributed resource management, coexistence of H2H and MEC traffic, cloud and cloudlet coexistence, reliability, and mobility are discussed. Three envisioned design scenarios for MEC over FiWi networks are studied, followed by a novel unified resource management scheme proposed for MEC-over-Ethernet-based FiWi networks. The obtained results show the significant benefits of MEC over FiWi networks. For instance, for a typical scenario, a response time efficiency of 95.36 percent can be achieved. Importantly, the mean MEC packet delay of 20.50 ms is obtained for the considered scenario, while allowing for efficient H2H/MEC coexistence without degrading network performance. Moreover, the battery life of edge devices is prolonged up to 11.30 h by employing the proposed solution. While TDMA scheduling has been shown to be effective in MEC over Ethernet-based FiWi networks, resource management schemes for the other two design scenarios and an in-depth performance analysis, especially with wireless backhaul technologies (e.g., millimeter-wave, microwave) would be an interesting topic for future research.

Acknowledgment

This work was supported by the Fonds de recherche du Québec — Nature et Technologies (FRQNT) MERIT Doctoral Research Scholarship Program.

References

[1] J. G. Andrews *et al.*, "What Will 5G Be?" *IEEE JSAC*, vol. 32, no. 6, June 2014, pp. 1065–82.
[2] M. Satyanarayanan *et al.*, "An Open Ecosystem for Mobile-Cloud Convergence," *IEEE Commun. Mag.*, vol. 53, no. 3, Mar. 2015, pp. 63–70.
[3] ETSI ISG, "Mobile-Edge Computing — A Key Technology Towards 5G," White Paper, no. 11, Sept. 2015, pp. 1–16.
[4] —, "Mobile-Edge Computing (MEC); Technical Requirements, Group Specification," ETSI GS MEC 002, v1.1.1, Mar. 2016, pp. 1–40.
[5] G. Fettweis and S. Alamouti, "5G: Personal Mobile Internet Beyond What Cellular Did to Telephony," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 140–45.
[6] N. J. Gomes *et al.*, "Invited Paper: Fronthaul Evolution: From CPRI to Ethernet," Optical Fiber Technology, vol. 26, Dec. 2015, pp. 50–58.
[7] H. Beyranvand *et al.*, "FiWi Enhanced LTE-A HetNets with Unreliable Fiber Backhaul Sharing and WiFi Offloading," *Proc. IEEE INFOCOM*, Apr. 2015, pp. 1275–83.
[8] M. Maier and B. P. Rimal, "Invited paper: The Audacity of Fiber-Wireless (FiWi) networks: Revisited for Clouds and Cloudlets," *China Commun.*, vol. 12, no. 8, Aug. 2015, pp. 33–45.
[9] S. Andreev *et al.*, "Understanding the IoT Connectivity Landscape: A Contemporary M2M Radio Technology Roadmap," *IEEE Commun. Mag.*, vol. 53, no. 9, Sept. 2015, pp. 32–40.
[10] S. Davy *et al.*, "Challenges to Support Edge-As-A-Service," *IEEE Commun. Mag.*, vol. 52, no. 1, Jan. 2014, pp. 132–39.
[11] Nokia Networks, "Intelligent Base Stations," White Paper, Feb. 2013.
[12] V. Suryaprakash, P. Rost, and G. Fettweis, "Are Heterogeneous Cloudbased Radio Access Networks Cost Effective?," *IEEE JSAC*, vol. 33, no. 10, Oct. 2015, pp. 2239–51.
[13] D. Pham Van *et al.*, "Machine-to-Machine Communications over FiWi Enhanced LTE Networks: A Power-Saving Framework and End-to-End Performance," *IEEE/OSA J. Lightwave Tech.*, vol. 34, no. 4, Feb. 2016, pp. 1062–71.
[14] B. P. Rimal, D. Pham Van, and M. Maier, "Mobile-Edge Computing vs. Centralized Cloud Computing in Fiber-Wireless Access Networks," *Proc. IEEE INFOCOM Wksps. 5G & Beyond*, Apr. 2016, pp. 595–600.
[15] E. Cuervo *et al.*, "MAUI: Making Smartphones Last Longer with Code Offload," *Proc. ACM MobiSys*, 2010, pp. 49–62.

Biographies

Bhaskar Prasad Rimal [S'13] (bhaskar.rimal@emt.inrs.ca) received his M.Sc. degree in information systems from Kookmin University, Seoul, South Korea. He is currently pursuing his Ph.D. degree in telecommunications at the Optical Zeitgeist Laboratory, Institut National de la Recherche Scientifique (INRS), Montréal, Quebec, Canada. His research interests include MEC, FiWi enhanced networks, Tactile Internet, Internet of Things, smart grids, and game theory. He is also a Student Member of the ACM and OSA. He has received the Certificate of Outstanding Contribution in Reviewing in 2014 and the Certificate of Reviewing in 2015 from *Computer Communications* (Elsevier). He was the recipient of the Doctoral Research Scholarship from the Québec Merit Scholarship Program for foreign students of Fonds de Recherche du Québec — Nature et Technologies (FRQNT), the Korean Government Information Technology (IT) Fellowship, the Kookmin University IT Scholarship, and the Kookmin Excellence Award as an Excellent Role Model Fellow.

Dung Pham Van [M'14] (dungpham@kth.se) received his B.Sc. degree in information technology from Hong Duc University, Thanh Hóa, Vietnam, in 2003, his M.Sc. degree in ICT from Waseda University, Tokyo, Japan, in 2009, and his Ph.D. degree (cum laude) in telecommunications from Scuola Superiore Sant'Anna, Pisa, Italy, in 2014. He is a postdoctoral researcher with the Optical Networks Laboratory (ONLab), KTH Royal Institute of Technology, Stockholm, Sweden. From January to August 2015, he was a postdoctoral researcher with the Optical Zeitgeist Laboratory, INRS. He was a visiting researcher at the University of Melbourne, Australia, in the first half of 2014. He has authored more than 30 papers in international journals and conference proceedings. His research interests include converged fiber-wireless networks, 5G backhaul, Internet of Things, energy efficiency, and data center networks. He was the recipient of the Distinguished Student Paper Award presented at the OptoElectronics and Communication Conference and Australian Conference on Optical Fibre Technology 2014 (OECC/ACOFT), the Best Student Paper Award (first class) presented at the Asia Communications and Photonics Conference 2013 (ACP), and the IEEE Standards Education Grant for the project "FPGA-Based Design and Evaluation of an Energy-Efficient 10G-EPON."

Martin Maier (maier@emt.inrs.ca) [M'04,- SM'09] is a full professor with INRS. He was educated at the Technical University of Berlin, Germany, and received M.Sc. and Ph.D. degrees (both with distinction) in 1998 and 2003, respectively. In the summer of 2003 he was a postdoc fellow at MIT, Cambridge. He was a visiting professor at Stanford University, California, from October 2006 through March 2007. Further, he was a Marie Curie IIF Fellow of the European Commission from March 2014 through February 2015. He is a co-recipient of the 2009 IEEE Communications Society Best Tutorial Paper Award and Best Paper Award presented at the International Society of Optical Engineers (SPIE) Photonics East 2000-Terabit Optical Networking Conference. He is the founder and creative director of the Optical Zeitgeist Laboratory (www.zeitgeistlab.ca). He currently serves as the Vice Chair of the IEEE Technical Subcommittee on Fiber-Wireless Integration. He is the author of the book *Optical Switching Networks* (Cambridge University Press, 2008), which was translated into Japanese in 2009, and the lead author of the book *FiWi Access Networks* (Cambridge University Press, 2012).

ACCEPTED FROM OPEN CALL

Licensed-Assisted Access to Unlicensed Spectrum in LTE Release 13

Hwan-Joon Kwon, Jeongho Jeon, Abhijeet Bhorkar, Qiaoyang Ye, Hiroki Harada, Yu Jiang, Liu Liu, Satoshi Nagata, Boon Loong Ng, Thomas Novlan, Jinyoung Oh, and Wang Yi

ABSTRACT

Exploiting the unlicensed spectrum is considered by 3GPP as one promising solution to meet ever-increasing traffic growth. As a result, one major enhancement for LTE in Release 13 has been to enable its operation in the unlicensed spectrum via licensed-assisted access (LAA). In this article, we provide an overview of the Release 13 LAA technology including motivation, use cases, LTE enhancements for enabling the unlicensed band operation, and the coexistence evaluation results contributed by 3GPP participants.

INTRODUCTION

To cope with ever-increasing traffic demand, the 3rd Generation Partnership Project (3GPP) has been continuously endeavoring to increase the network capacity by improving the spectral efficiency of the Long Term Evolution (LTE) system through the introduction of higher order modulations, advanced multi-input multi-output (MIMO) antenna technologies, and multi-cell coordination techniques, to name a few. Another fundamental approach to improve network capacity is to expand the system bandwidth, but newly available spectrum in the lower frequency bands, which have traditionally been individually allocated to each mobile network operator, has become very scarce. This is the main rationale behind the recent study item (SI) and work item (WI) in 3GPP Release 13 to enable the operation of an LTE system in unlicensed spectrum. Since 3GPP considers unlicensed spectrum as supplemental to licensed spectrum, this new feature is called licensed-assisted access (LAA) to unlicensed spectrum, often referred to as LAA. One important consideration for operating LTE in unlicensed spectrum is to ensure fair coexistence with the incumbent systems such as wireless local area networks (WLANs),[1] which have been the principal focus of LAA standardization.

The purpose of this article is to provide an overview of the LAA technology developed during the LTE Release 13. This article is organized as follows. We give a brief overview of the relevant unlicensed spectrum bands and their associated regulatory requirements. This is followed by a discussion of the deployment scenarios and use cases. We summarize the LAA standardization activities in 3GPP. We then take a deep-dive into the key technical features of the LAA. After that, we present a summary of evaluation results verifying the coexistence between LAA and Wi-Fi, presented by a number of companies at 3GPP meetings. Finally, we draw conclusions.

UNLICENSED SPECTRUM AND REGULATIONS

The initial LAA deployments are expected to be limited to globally available 5 GHz unlicensed spectrum (Table 1). Although the 5 GHz spectrum is generally designated as unlicensed spectrum, radio equipment operating in the spectrum must abide by the regulatory requirements, which vary by regions, as summarized in Table 1. In addition to various requirements such as indoor-only use, maximum in-band output power, in-band power spectral density, and out-of-band and spurious emissions, LTE operation in some unlicensed spectrum should also implement dynamic frequency selection (DFS) and transmit power control (TPC) depending on the operating band to avoid interfering with radars.

SCENARIOS AND USE CASES

SCENARIOS AND USE CASES

The introduction of carrier aggregation in LTE-Advanced required the distinction between a primary cell (PCell) and a secondary cell (SCell). The PCell is the main cell with which a user equipment (UE) communicates and maintains its connection with the network. One or more SCells can be allocated and activated to the UEs supporting carrier aggregation for bandwidth extension. Since the unlicensed carrier is shared by multiple systems, it can never match the licensed carrier in terms of mobility, reliability, and quality of service. Hence, in LAA the unlicensed carrier is considered only as a supplemental downlink (DL) SCell assisted by a licensed PCell via carrier aggregation. LAA deployment scenarios encompass scenarios with and without macro coverage, both outdoor and indoor small cell deployments, and both co-located and non-co-located (with ideal backhaul) cells operating in licensed and unlicensed carriers. Figure 1 captured from 3GPP Technical Report (TR) 36.889 shows four considered LAA deployment scenarios [1].

- Scenario 1: Carrier aggregation between licensed macro cell (F1) and unlicensed small cell (F3).

One major enhancement for LTE in Release 13 has been to enable its operation in the unlicensed spectrum via Licensed-Assisted Access (LAA). The authors provide an overview of the Release 13 LAA technology including motivation, use cases, LTE enhancements for enabling the unlicensed band operation, and the coexistence evaluation results contributed by 3GPP participants.

[1] Throughout the article, we use the term Wi-Fi interchangeably with WLAN.

Hwan-Joon Kwon, Jeongho Jeon, Abhijeet Bhorkar, and Qiaoyang Ye are with Intel Corporation; Hiroki Harada, Yu Jiang, Liu Liu, and Satoshi Nagata are with NTT DOCOMO; Boon Loong Ng, Thomas Novlan, Jinyoung Oh, and Wang Yi are with Samsung Electronics

Digital Object Identifier: 10.1109/MCOM.2016.1500698CM

Region	5.15–5.25 GHz	5.25–5.35 GHz	5.47–5.725 GHz	5.725–5.85 GHz
USA	–	DFS/TPC		–
	FCC Part 15 Rules (max EIRP, emission mask, etc.)			
EU[a]	Indoor only		Indoor/outdoor	
	–	DFS/TPC		
	ETSI Harmonized European Standards (LBT, max EIRP, emission mask, etc.)			
China	Indoor only		TBD	Indoor/outdoor
	–	DFS/TPC		–
	Max EIRP, emission mask, etc.			Max EIRP, emission mask, etc.
Japan	Indoor only		–	Band not available
	–	DFS/TPC		
	LBT, max burst length (4ms), max EIRP, emission mask, etc.			
Korea[b]	–	DFS/TPC		–
	Max EIRP, emission mask, etc.			

N/A: Not applicable
LBT: Listen-before-talk
DFS: Dynamic frequency selection
TPC: Transmit power control
EIRP: Equivalent isotropically radiated power
[a] EU band 4 is 5.725–5.875 GHz, where wireless access systems (WAS) are not operating in.
[b] Korea band 3 is 5.47–5.65 GHz.

Table 1. Unlicensed band regulations by region.

- Scenario 2: Carrier aggregation between licensed small cell (F2) and unlicensed small cell (F3) without macro cell coverage.
- Scenario 3: Licensed macro cell and small cell (F1), with carrier aggregation between licensed small cell (F1) and unlicensed small cell (F3).
- Scenario 4: Carrier aggregation between licensed small cell (F2) and unlicensed small cell (F3). An ideal backhaul between macro cell and small cell can enable carrier aggregation between macro cell (F1), licensed small cell (F2), and unlicensed small cell (F3).

The carrier aggregation between non-co-located cells was mainly motivated by the hotspot scenario where a macro cell behaves as an anchor cell to provide robust connection management, while each small cell behaves as a booster cell to offer higher throughput. Such a hotspot scenario is the main use case of the small cells.

Comparison of LAA with Other LTE-based Unlicensed Technologies

There are two competing LTE-based unlicensed technologies to LAA: LTE unlicensed (LTE-U) and LTE-WLAN aggregation (LWA). LTE-U is based on the 3GPP Release 12 LTE technology to be used in the unlicensed spectrum. LTE-U uses an adaptive on/off duty cycle as a mechanism to share the medium with existing Wi-Fi networks. As will be seen in the following sections, the channel access mechanism of LAA largely resembles that of Wi-Fi and, thereby, it is natural to expect that LAA will provide better coexistence with existing Wi-Fi networks. On the other hand, LWA is developed as part of Release 13 WI, which enables the simultaneous LTE and Wi-Fi connectivity via dual connectivity where the data traffic is aggregated at the eNB and routed to an operator core network.

When compared to the current implementation of Wi-Fi offloading, which prefers Wi-Fi connection regardless of the Wi-Fi link condition, the LTE-based unlicensed technologies anchored in the licensed carrier can provide a better user experience thanks to reliable connection management and optimized link selection/activation. On the other hand, compared to LWA, LAA can provide a tighter integration of the licensed and unlicensed spectrum by performing lower-layer aggregation, thereby improving the overall efficiency and especially the quality of delay-sensitive applications.

LAA in Release 13 and Beyond

The standardization of LAA in Release 13 was conducted in two phases; the first phase was the SI phase [2]; the second phase was the WI phase [3]. The goal of the SI phase was to study the feasibility of LTE enhancement to enable LAA operation in unlicensed spectrum while coexisting with other incumbent systems and fulfilling the regulatory requirements. The SI concluded that it is feasible for LAA to fairly coexist with Wi-Fi and other LAA networks, if an appropriate channel access scheme is adopted such as listen-before-talk (LBT) [1], which is explained in detail later.

The main objective of the LAA WI is to specify LTE enhancements for operation in unlicensed spectrum, which is limited to support for LAA SCells operating with only DL transmissions, under the design criteria of a single global solution framework, fair coexistence between Wi-Fi and LAA, and fair coexistence between different LAA networks. The detailed objectives of the WI are to specify the support for the following functionalities: channel access framework, including clear channel assessment; discontinuous transmission with limited maximum transmission duration; UE support for carrier selection; UE support for radio resource management (RRM) measurements, including cell identification; time and frequency synchronization; and channel-state information (CSI) measurement. The LAA WI was completed by the end of 3GPP Release 13 in late 2015.

In Release 14, 3GPP is defining an uplink (UL) access scheme for LTE operation in unlicensed spectrum in addition to the already defined DL access scheme in Release 13. The detailed scope of the enhanced LAA for LTE (eLAA) WI includes the support for physical uplink shared channel (PUSCH) and sounding reference signal (SRS), along with the UL channel access mechanism that is largely based on the Release 13 design. The support for physical uplink control channel (PUCCH) and physical random access channel (PRACH) is unlikely given the current status of the WI. At the time of this writing, the eLAA WI was planned to be concluded by September 2016. Further enhancement of LAA is expected to be discussed in 3GPP.

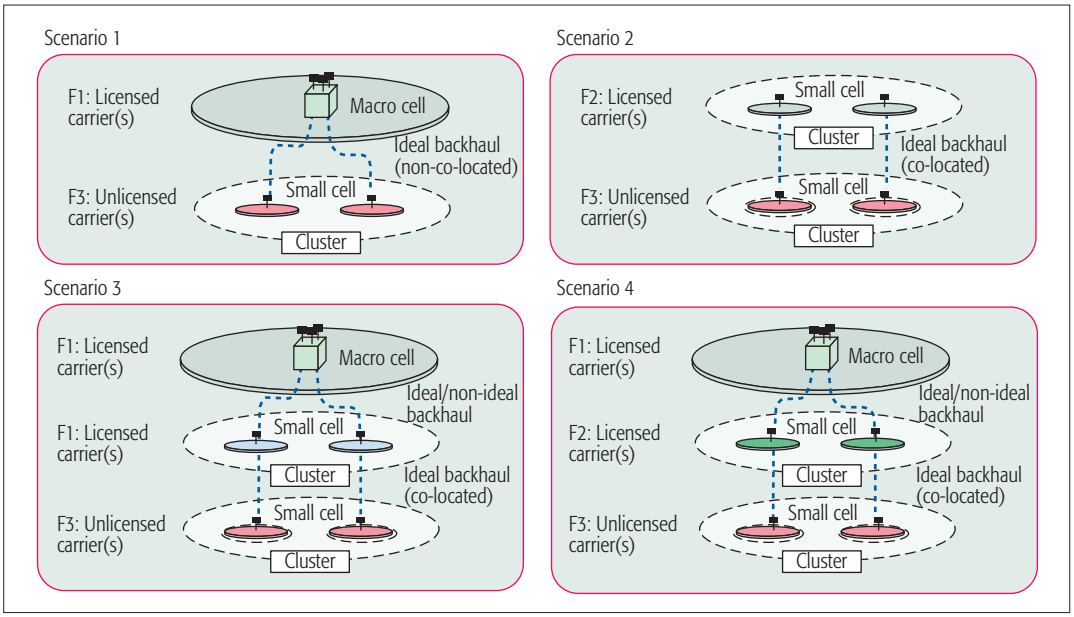


Figure 1. LAA deployment scenarios from 3GPP TR 36.889 [1].

Apart from regulatory requirements, LBT is highly beneficial for fair and friendly coexistence with incumbent systems in the unlicensed spectrum and with other LAA networks. The main incumbent systems in the 5 GHz band are the WLANs based on IEEE 802.11n/ac technologies, which are widely deployed both by individuals and operators for data offloading.

Key Technical Features of the Release 13 LAA

LBT and Overall DL Data Transmission

LBT is a procedure whereby radio transmitters first sense the medium and transmit only if the medium is sensed to be idle, which is also called clear channel assessment (CCA). The CCA utilizes at least energy detection (ED) to determine the presence of signals on a channel. Recall from earlier that LBT in 5 GHz unlicensed spectrum is required in Europe and Japan but not in the U.S., China, and Korea. However, the adoption of LBT is necessary for LAA to become a single global solution that complies with any regional regulatory requirements. Apart from regulatory requirements, LBT is highly beneficial for fair and friendly coexistence with incumbent systems in the unlicensed spectrum and with other LAA networks. The main incumbent systems in the 5 GHz band are the WLANs based on IEEE 802.11n/ac technologies, which are widely deployed both by individuals and operators for data offloading. The WLAN employs a contention-based channel access mechanism, called carrier sense multiple access with collision avoidance (CSMA/CA) [4]. A WLAN node that intends to transmit first performs CCA before transmission. An additional backoff mechanism is designed for the collision avoidance aspect to cope with the situation when more than one node senses that the channel is idle and transmits at the same time. The backoff counter is drawn randomly within the contention window size (CWS), which is increased exponentially upon the occurrence of a collision and reset to the minimum value when the transmission succeeds.

The LBT mechanism designed for LAA fundamentally resembles the CSMA/CA of a WLAN. The specified LBT procedure for LAA DL transmission bursts containing PDSCH[2] is illustrated in Fig. 2. The size of the LAA contention window is variable between X and Y extended CCA (ECCA) slots, which are the minimum and maximum CWSs, respectively. The ECCA slot duration is at least 9 μs, which is exactly the same as a WLAN slot.

An illustration of an LAA DL burst transmission is given in Fig. 3, where MCOT stands for maximum channel occupancy time. 3GPP has introduced four different priority classes for DL LBT with random backoff, where the smaller the LBT priority class number, the higher the priority. Release 13 supports at least priority class 3, and best effort traffic shall not use a priority class with higher priority than the priority class 3. 3GPP has differentiated the MCOT according to the LBT priority classes. For priority classes 3 and 4, MCOT is 10 ms, if the absence of any other technology sharing the carrier can be guaranteed on a long term basis. Otherwise, it is 8 ms. For LAA operation in Japan, the E-UTRAN NodeB (eNB) may need to sense the channel to be idle for an additional single continuous interval of duration 34 μs after every 4 ms of transmission if the DL transmission burst is longer than 4 ms.

ED Threshold: An important component of LBT design is the choice of ED threshold, which determines the level of sensitivity to declare the existence of ongoing transmissions. 3GPP considers the mechanism to adapt the ED threshold. For instance, if the absence of any other technology sharing the carrier cannot be guaranteed on a long term basis (e.g., by level of regulation), the maximum energy detection threshold used by LAA for category 4 LBT is

$$\mathrm{TH} = \max(-72\ \mathrm{dBm}\ (20\mathrm{MHz}), \min(T_{\max}, T_{\max} - 10\ \mathrm{dB} + (P_H - P_{TX}))),$$

where P_H is a reference power equaling 23 dBm, P_{TX} is the configured maximum transmit power for the carrier in dBm, and it is given by $T_{\max} = -75$ dBm/MHz+ 10*log10($\mathrm{BW}_{\mathrm{MHz}}$) where $\mathrm{BW}_{\mathrm{MHz}}$ is the channel bandwidth in MHz. In a nutshell, the ED threshold can be raised if the bandwidth $\mathrm{BW}_{\mathrm{MHz}}$ becomes wider and/or the transmit power P_{TX} is lowered.

CWS Adaptation: The CWS is adapted based on the hybrid automatic repeat request (HAR-

[2] PDSCH is the physical downlink shared channel in LTE for the transmission of unicast user data and paging information.

[3] HARQ is a mechanism that works at the physical layer to deal with the errors in the reception of transmitted data. Unlike the ARQ, the retransmissions in response to the occurrence of errors are different redundancy versions of the original coded block.

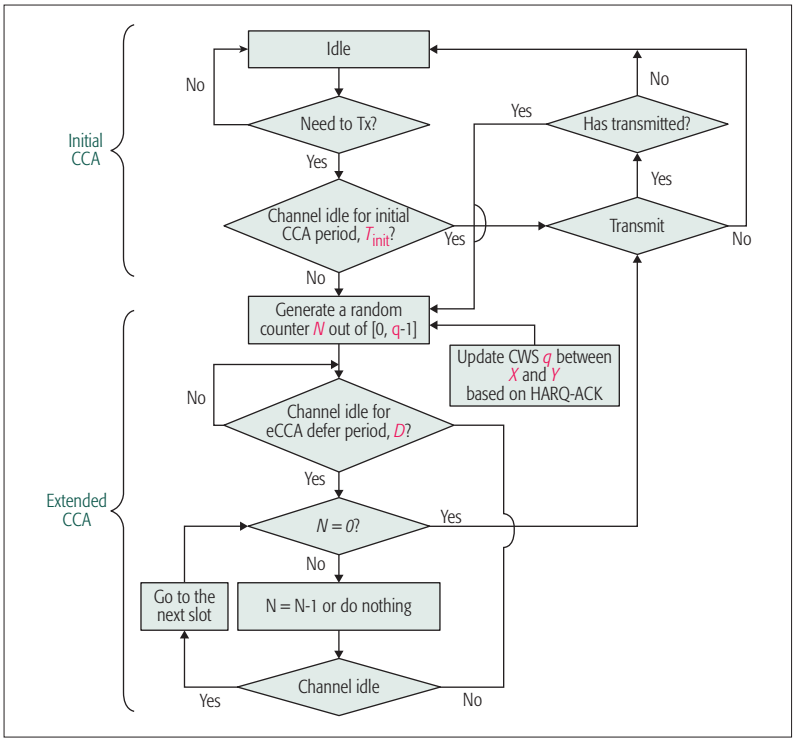


Figure 2. Flowchart of the recommended DL LBT procedure by 3GPP.

Q)-ACK feedback.[3] HARQ-ACK feedback can take a value from ACK, NACK, and DTX, where ACK refers to the situation of correct reception, NACK refers to the situation where control information (i.e., PDCCH[4]) is correctly detected but there is an error in the data (i.e., PDSCH) reception, and DTX refers to the situation when a UE misses the control message containing scheduling information (i.e., PDCCH), rather than the data itself (i.e., PDSCH). The set of CWSs for LAA DL transmission bursts containing PDSCH for priority class 3 is {15, 31, 63}, i.e., X and Y in Fig. 2 are 15 and 63, respectively. The X and Y values are set differently for different LBT priority classes. The CWS is increased if at least 80 percent of the HARQ-ACK feedback values for the first subframe of a DL burst are NACK. The CWS increase is in an exponential manner as in Wi-Fi. Otherwise, the CWS is reset to the minimum value. DTX is considered as NACK except when the UEs were not actually scheduled by eNB or the scheduling information was sent through the licensed PCell.

Multicarrier LBT: LAA supports two alternative solutions for multi-carrier LBT. In the first option, the eNB is required to designate a carrier requiring LBT with random backoff, as illustrated in Fig. 2, and the eNB can sense other configured carriers with single interval LBT only if the eNB completes the LBT with random backoff on the designated carrier. In the second option, the eNB performs LBT with random backoff on more than one unlicensed carriers and is allowed to transmit on the carriers that have completed the LBT with potential self-deferral to align transmissions over multiple carriers.

LBT for LAA Discovery Reference Signal

In LTE Release 12, a discovery reference signal (DRS) was introduced to facilitate fast transition of small cells from the OFF state to the ON state by transmitting low duty cycle signals for RRM measurement during the OFF state. During the OFF period, DRS, consisting of synchronization signals and reference signals, is transmitted to allow UEs to discover and measure the dormant cell. The RRM measurement details are further explained in a later Section. LAA DRS is the same as the first twelve OFDM symbols of the Release 12 DRS in frame structure type 1[5] (frame structure defined for frequency division duplexing (FDD)).

DRS can be transmitted within a periodically occurring time window called the DRS measurement timing configuration (DMTC) window, which has a duration of 6 ms and a configurable period of 40/80/160 ms. The transmission of DRS is also subject to LBT. A DL transmission burst containing DRS without PDSCH follows a single idle observation interval of at least 25 μs. Due to LBT, the DRS may not be transmitted as frequently as scheduled. To increase the DRS transmission opportunity so as to improve the performance of functionalities (e.g., synchronization, RRM measurement) relying on DRS, DRS can be transmitted by the network once in any subframe within the DMTC occasion.

LTE Enhancement to Support LAA

Frame Structure and Partial Subframe: For LAA, the LBT procedure can be completed at any time. Moreover, DL transmission may not start/end at the subframe boundary. To support such flexible operation for LAA, a new frame structure, called type 3 has been introduced in Release 13 for which UE considers each subframe as empty unless DL transmission is detected in that subframe.

Note that other neighboring systems can take advantage of the transmission opportunity while the LAA eNB is awaiting the next subframe boundary unless a reservation signal is transmitted after successful LBT. To efficiently utilize radio resources, a partial subframe has been introduced for LAA SCell, where DL transmission, excluding reservation signal, can start at the first or second slot boundaries of a subframe, as illustrated in Fig. 3. Depending on starting position of DL transmission and due to MCOT limitation, DL transmission may not end at the subframe boundary. To utilize the ending partial subframe with minimal specification efforts, the existing downlink pilot time slot (DwPTS) structure is reused, where DwPTS is the DL portion of the special subframe of the frame structure type 2 for time division duplexing (TDD). With the existing DwPTS configuration, the duration of the last subframe of a DL transmission burst can be one of {3, 6, 9, 10, 11, 12} OFDM symbols or a whole subframe consisting of 14 OFDM symbols. A common control signal in LAA SCell is used to indicate the number of OFDM symbols in the current and the next subframe for DL transmission.

RRM Measurement: RRM measurement is required for proper LAA SCell selection/reselection. RRM measurement is based on the reception of DRS containing CRS/CSI-RS[6], and the reporting consists of reference signal received power (RSRP) and reference signal received quality (RSRQ).[7] Due to the dynamically changing channel condition in unlicensed spectrum, legacy RRM measurement reporting may not be sufficient to reflect load conditions, interference

[4] PDCCH refers to the Physical Downlink Control Channel used to convey DL control information, including DL/UL scheduling grants.

[5] The 1 ms LTE subframe consists of two slots of 0.5 ms each. Each slot contains either six or seven OFDM symbols, depending on the cyclic prefix (CP) length. LAA supports only normal CP, which corresponds to seven OFDM symbols per slot.

[6] CRS and CSI-RS refer to cell-specific reference signal and channel state information reference signal, respectively.

[7] RSRQ is calculated as the ratio of RSRP to reference signal strength indicator (RSSI), indicating the ratio of the received signal power to the total received power, including its own signal power, interference, and noise.

outside DL burst, and potential hidden nodes in the unlicensed channel. In this regard, LAA UEs can be configured to report average RSSI and channel occupancy as a part of RRM measurements. Average RSSI provides an estimation of load conditions and captures the overall interference on LAA SCells. The channel occupancy is defined as the percentage of time when the channel is sensed to be busy, i.e., when the measured RSSI sample is above a predefined threshold. It is important that the layer 1 (L1) averaging duration of UE-reported RSSI measurements should roughly be of the same order as the minimum transmission granularity on an unlicensed carrier. For example, Wi-Fi ACK duration can typically be less than 100 μs. As a result, the L1 averaging duration is one LTE OFDM symbol. In addition, multiple consecutive L1 RSSI samples can be aggregated to produce measurement durations ranging from 1 ms to 5 ms.

Cell Detection and Synchronization: Cell detection and synchronization rely on the reception of the synchronization signals such as primary and secondary synchronization signal (PSS/SSS) and CRS. Specifically, PSS/SSS can be used for physical-layer cell identity (ID) detection, and CRS can be used to further improve the performance of cell ID detection, for example, to confirm cell detection. PSS/SSS and CRS can also be used to acquire coarse and fine time/frequency synchronization, respectively. Thanks to the multiple DRS transmission opportunities within a DMTC occasion, large time/frequency drift between two successive DL bursts is unlikely. Thus, the synchronization based on DRS in LAA systems can achieve reliable performance. On the other hand, the DL subframe presence detection by UE is needed as the eNB does not always transmit. The exact detection method employed is left to the UE implementation.

CSI Measurement and Reporting: LAA supports transmission modes (TMs) with CRS-based CSI feedback, including TM1, TM2, TM3, TM4, and TM8, and those with CSI-RS based CSI feedback, including TM9 and TM10[8]. CSI-RS/CSI-IM[9] for CSI measurement is present in the configured periodic CSI-RS/CSI-IM subframes within DL transmission bursts. Similar to the legacy LTE systems, both periodic and aperiodic CSI reports are supported. Unlike the legacy LTE system whereby the CRS/CSI-RS transmission power, or energy per resource element (EPRE) is fixed, CRS/CSI-RS transmission power on LAA SCells is only fixed within a DL transmission burst, while it can vary across DL transmission bursts. As a result, UE should not average CRS/CSI-RS measurements across transmission bursts. UE could either rely on CRS detection or common control signaling to differentiate DL bursts.

Scheduling and HARQ: LTE supports two different scheduling approaches, cross-carrier scheduling and self-scheduling. With cross-carrier scheduling, the control information, including scheduling indication, i.e., PDCCH, and the actual data transmission, i.e., PDSCH, take place on different carriers, whereas they are transmitted on the same carrier in the case of self-scheduling. Due to the uncertainty of channel access opportunities on unlicensed carriers, the synchronous HARQ protocol[10] with fixed time relation between retransmissions is difficult to use for LAA. Thus, the existing asynchronous HARQ protocol can be used for LAA DL/UL. For LAA UL, in particular, UEs would need to rely on the UL grant from eNB for UL (re)transmissions.

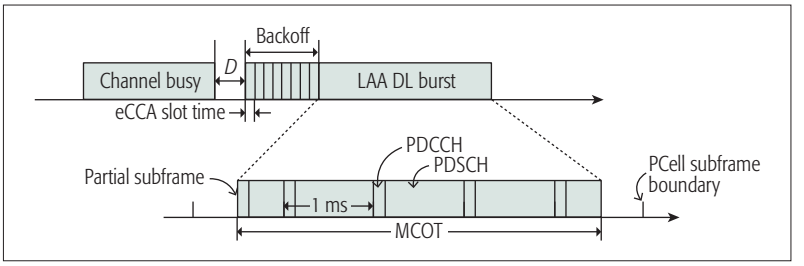


Figure 3. Illustration of an LAA DL burst transmission.

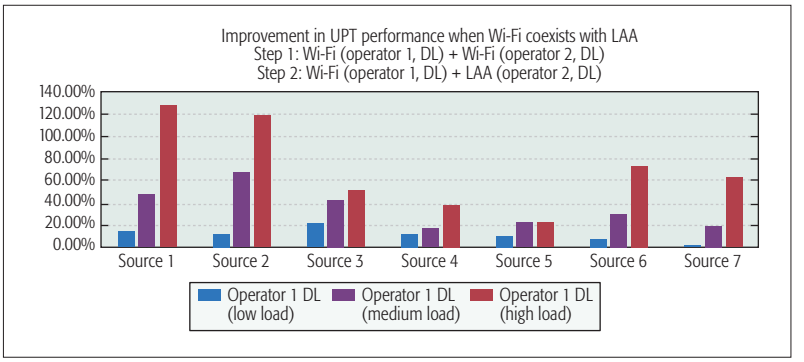


Figure 4. Improvement in the UPT for the DL only Wi-Fi network (Sources 1-7 are from 3GPP contributions R1-150694, R1-152732, R1-151821, R1-152863, R1-153384, R1-153426, and R1-153629, respectively.)

Coexistence Performance

Evaluation Methodology

The first and foremost goal of LAA design is to ensure fair coexistence with other incumbent systems operating in the same unlicensed spectrum. This is captured in the LAA design target in terms of fair sharing metrics; an LAA network should not impact Wi-Fi services more than an additional Wi-Fi network on the same carrier. In this section, we highlight the extensive evaluation efforts contributed by numerous sources during the LAA SI phase [1].

A 3GPP defined indoor scenario consists of four equally spaced LAA eNBs and/or Wi-Fi APs deployed by each operator in a single story building serving 10 uniformly distributed LAA UEs and/or Wi-Fi STAs operating on the same unlicensed carrier. The distance between the two closest nodes from two operators is random. The set of small cells for both operators is centered along the longer dimension of the building. The outdoor scenario considers a hexagonal grid with three sectors per site and inter-site distance of 500m. Clusters of small cells are uniformly random within a macro geographical area. Within each cluster, there are four small cells per operator, randomly dropped within the cluster area. Ten UEs are randomly dropped within the coverage area of a small cell in unlicensed spectrum. 3GPP also considered various traffic models such as File Transfer Protocol (FTP) traffic, and mixed FTP and Voice over Internet Protocol (VoIP) traffic. The Wi-Fi network with DL-only traffic and both DL and UL traffic were

[8] The TMs differ in terms of number of antennas, MIMO mode, and number of spatial streams, etc.

[9] CSI-IM refers to channel-state information – interference measurement, whose resource configuration is based on zero-power CSI-RS configuration.

[10] The HARQ protocol can be categorized into synchronous and asynchronous HARQ based on the flexibility in the time domain. With synchronous HARQ, re-transmission occurs at a fixed time, while with asynchronous HARQ, re-transmission can occur at any time.

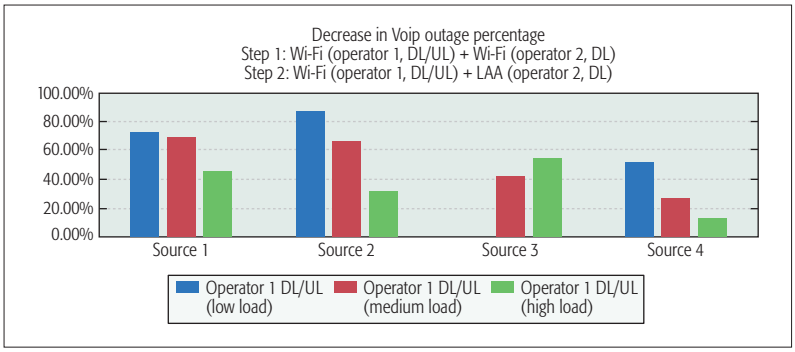


Figure 5. Decrease in VoIP outage for the DL/UL Wi-Fi network (Sources 1-4 are from 3GPP contributions R1-152326, R1-152642, R1-152937, and R1-153343, respectively.)

considered as well. To verify the coexistence, a two-step methodology is used. In step 1, the performance of two coexisting Wi-Fi networks is evaluated as a benchmark; in step 2, a Wi-Fi network is replaced with an LAA network, and the performance of the non-replaced Wi-Fi network is compared against step 1.

Evaluation Results from 3GPP

During the discussion in 3GPP, it was identified that ensuring coexistence for the indoor scenario is more difficult than for the outdoor scenario due to the close proximity between LAA eNBs and Wi-Fi access points (APs)/stations (STAs). It is also apparent that restricting LAA eNB to transmit data only in the unlicensed carrier is more challenging to prove fair coexistence because the licensed carrier given to the LAA eNB is an additional resource that can be exploited to alleviate the transmission demand on unlicensed spectrum in step 2, resulting in a more friendly environment for fair coexistence. The results captured in this section from [1] are thus focused on the most demanding scenarios in proving fair coexistence. IEEE 802.11ac technology is assumed for Wi-Fi networks.

The user perceived throughput (UPT) is considered by 3GPP as an important performance measure for networks serving non-full-buffer traffic. The UPT is defined as the amount of data over the actual time spent for downloading, excluding idle time waiting for files to arrive. Fig. 4 shows the improvement in the UPT for the non-replaced DL-only Wi-Fi network in step 2 compared to step 1 with different loading conditions. Buffer occupant time of 15–30 percent, 35–50 percent, and 60–80 percent averaged over the APs of the non-replaced Wi-Fi network in step 1 is considered as low, medium, and high load, respectively. From Fig. 4, it can be observed that Wi-Fi UPT performance is improved when the Wi-Fi network coexists with an LAA network rather than another Wi-Fi network. This is mainly because LTE has higher spectral efficiency than Wi-Fi due to the better link adaptation based on explicit CSI feedback, while the control messages such as CSI feedback can go through a licensed carrier. Consequently, the interference from Operator 2 to Operator 1 is reduced in step 2, thereby improving Wi-Fi performance in step 2. Figure 5 shows the coexistence performance when Operator 1's Wi-Fi network serves bidirectional, i.e., both DL and UL, mixed FTP and VoIP traffic. From the figure, it is shown that VoIP outage for the non-replaced Wi-Fi network can be reduced significantly when it coexists with an LAA network. This leads to the conclusion that the 3GPP LAA design can indeed ensure coexistence with incumbent Wi-Fi networks for both non-real-time and real-time traffic.

Finally, we make a note on the observations made during the early SI phase. Multiple sources identified that Wi-Fi performance can be significantly degraded when it coexists with LAA. These observations were the main motivation behind the adoption of the LBT algorithm based on the exponential backoff as in Wi-Fi. The simulation results summarized here are those following the LBT algorithm, which was finally agreed.

Conclusion

This article gave an overview of 3GPP Release 13 LAA technology. The LAA supplements a licensed primary carrier with unlicensed secondary carriers via carrier aggregation. The 3GPP aimed at meeting the regulatory requirements as well as ensuring fair coexistence with existing Wi-Fi networks. These design goals have led to significant changes at the LTE physical layer for LAA. Based on the evaluations contributed to 3GPP provided from a wide spectrum of sources, there is a consensus that LAA can fairly coexist with Wi-Fi networks serving various traffic types.

References

[1] Study on Licensed-Assisted Access to Unlicensed Spectrum, 3GPP TR 36.889, May 2015.
[2] RP-141664 "Study on Licensed-Assisted Access Using LTE," Ericsson, Qualcomm, Huawei, Alcatel-Lucent, 3GPP TSG RAN Meeting #65, Sept. 2014.
[3] RP-151045 "New Work Item on Licensed-Assisted Access to Unlicensed Spectrum," Ericsson, Huawei, Qualcomm, Alcatel-Lucent, 3GPP TSG RAN Meeting #68, June 2015.
[4] Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, IEEE Std 802.11TM-2012, Mar. 2012.

Biographies

Hwan-Joon Kwon (eddy.kwon@intel.com) is the 3GPP RAN prime at Intel Corporation, where he leads Intel's standardization efforts on LTE-Advanced Pro and 5G. Before joining Intel, he worked at Samsung Electronics in Suwon, South Korea, on standardization development of 3GPP, 3GPP2, and ETSI DVB-T2. He has authored more than 13 research papers and holds over 100 U.S. patents. He received his B.E. degree and M.S. degree in electrical communications engineering from Hanyang University, Seoul, and his Ph.D. degree in electrical engineering from the University of California, San Diego.

Jeongho Jeon (jeongho.jeon@intel.com) received the Ph.D. degree in electrical engineering from the University of Maryland, College Park, MD, USA, in 2013. Since then he has been with Intel Corporation, where he conducts research and standardization of next generation wireless technologies. He is a recipient of a National Institute of Standards and Technology (NIST) Fellowship from 2011 to 2013, and the 14th Samsung Humantech Thesis Prize in 2008.

Abhijeet Bhorkar (abhijeet.bhorkar@intel.com) received B.Tech. and M.Tech. degrees in electrical engineering from the Indian Institute of Technology, Bombay, both in 2006. He received a Ph.D. degree from the University of California, San Diego in 2012. He is currently working at Intel Corporation as a system engineer. His research interests are primarily in the areas of wireless networks, protocol design, stochastic control and estimation theory, information theory, and their applications in the optimization of wireless communication systems.

Qiaoyang Ye (qiaoyang.ye@intel.com) received her B.S. degree in information science and communication engineering from Zhejiang University (China) in 2010, and the M.S. and Ph.D.

degrees in electrical and computer engineering from the University of Texas at Austin, in 2013 and 2015, respectively. She is currently a wireless system engineer with Intel Corp., Santa Clara, CA, USA. She has held summer internships at Huawei Technologies in Plano, TX, and DOCOMO Innovations in Palo Alto, CA. She was an exemplary reviewer for *IEEE Wireless Communications Letters* in both 2014 and 2015.

Hiroki Harada (hiroki.harada.sv@nttdocomo.com) received B.E., M.E., and Ph.D. degrees in electrical engineering from Yokohama National University in 2003, 2005, and 2008, respectively. He was a research fellow at the Japan Society for the Promotion of Science at Yokohama National University from 2005 to 2008. He joined NTT DOCOMO, Inc. in 2008, and has been engaged in 3GPP LTE standardization, including small cell enhancements and LAA. He received the IEEE VTS Japan VTC2006-Fall Student Paper Award in 2006, and IEICE Young Researcher's Award in 2011.

Yu Jiang (jiangy@docomolabs-beijing.com.cn) received the B.E and M.E. degrees from Xidian University, China, in 2005 and 2008, respectively. In 2010 he joined DOCOMO Beijing Communication Laboratories Co., Ltd. and worked on system level evaluation and research of wireless access technologies for LTE and LTE-Advanced. His research interests include MIMO, CoMP, and LAA.

Liu Liu (liul@docomolabs-beijing.com.cn) received a B.E. degree from ZheJiang University in 2005, and an M.E. degree from Tsinghua University in 2008. In 2008 she joined DOCOMO Beijing Communication Laboratories Co., Ltd. as a researcher, and engaged in the investigation of next generation wireless access technologies. She has been involved with 3GPP LTE/LTE-A standardization since 2009. Her research interests include radio resource management, carrier aggregation, and LAA.

Satoshi Nagata (nagatas@nttdocomo.com) received his B.E. and M.E. degrees from Tokyo Institute of Technology, Japan. He joined NTT DOCOMO, INC., and worked on the research and development of wireless access technologies for LTE and LTE-Advanced. He is currently involved with 5G and 3GPP standardization. He has contributed to 3GPP for many years, and contributed to 3GPP TSG-RAN WG1 as a vice chairman. He has been the chairman of 3GPP TSG-RAN WG1 since 2013.

Boon Loong Ng (b.ng@samsung.com) received the bachelor of engineering (electrical and electronic) degree and the Ph.D. degree in engineering from the University of Melbourne, Australia, in 2001 and 2007, respectively. He is currently a senior staff engineer with Samsung Research America – Standards & Mobility Innovation (SMI) Lab in Dallas, Texas. He is also the group leader for the New Communications Technology Group, where the main R&D focus is on the system designs of next generation communication systems. He has contributed to 3GPP standards in RAN working groups since LTE Release 8, and holds numerous patents on LTE/LTE-Advanced.

Thomas Novlan (t.novlan@samsung.com) is a staff research engineer with the Standards and Mobility Innovation Lab of Samsung Research America, currently working on 5G air-interface research and system design. He joined Samsung in 2012, working on LTE-Advanced standardization for device-to-device and small cell-related technologies, including LTE-U and licensed assisted access (LAA). He received his B.S. degree with high honors, and the M.S. and Ph.D. degrees in electrical engineering from The University of Texas at Austin in 2007, 2009, and 2012, respectively.

Jinyoung Oh (jy81.oh@samsung.com) is a senior engineer with Samsung Electronics. He received a Ph.D. degree in wireless communication from Korea Advanced Institute of Science and Technology (KAIST) in 2013. He joined Samsung Electronics in 2013, and has been working on standardization of LTE and 5G air-interface research and system design.

Yi Wang (y0917.wang@samsung.com) is a senior engineer with Samsung Electronics. She received a Ph.D. degree in wireless communication from Beijing University of Posts and Telecommunications, China, in June 2009. She has been working on LTE since 2010, including 3GPP standardization, LTE physical layer algorithm research and development, and algorithm implementation support.

ACCEPTED FROM OPEN CALL

Enhancing the Robustness of LTE Systems: Analysis and Evolution of the Cell Selection Process

Mina Labib, Vuk Marojevic, Jeffrey H. Reed, and Amir I. Zaghloul

The authors analyze the effect of different levels of RF spoofing applied to LTE. RF spoofing affects LTE devices during the initial cell selection process, where a strong nearby cell can impede access to a serving LTE network. This is a serious threat and can be caused unintentionally, in the case of dense and uncoordinated LTE deployment in unlicensed spectrum, or intentionally, where an adversary sets up a fake LTE cell in either licensed or unlicensed LTE spectrum.

ABSTRACT

The commercial success of LTE makes it the primary standard for 4G cellular technology, and its evolution paves the path for 5G technology. Furthermore, LTE Unlicensed has been proposed recently to allow cellular network operators to offload some of their data traffic to LTE component carriers operating in the unlicensed band. Hence, it is critical to ensure that the LTE system performs effectively even in harsh signaling environments in both licensed and unlicensed spectrum. This article analyzes the effect of different levels of RF spoofing applied to LTE. RF spoofing affects LTE devices during the initial cell selection process, where a strong nearby cell can impede access to a serving LTE network. This is a serious threat and can be caused unintentionally, in the case of dense and uncoordinated LTE deployment in unlicensed spectrum, or intentionally, where an adversary sets up a fake LTE cell in either licensed or unlicensed LTE spectrum. This article analyzes and experimentally demonstrates the severity of these threats for the evolution of LTE and proposes effective mitigation techniques to prevent denial of service. These mitigation techniques improve the cell selection process at the LTE user equipment, and are backward-compatible with existing LTE networks. We recommend that these modifications be enforced in future releases for increasing the availability and scalability of LTE.

INTRODUCTION

Since the introduction of smartphones in 2007, the mobile data demand has been growing tremendously. During that time, Long Term Evolution (LTE) has been in the process of being standardized by the Third Generation Partnership Project (3GPP). Cellular network operators' attention was drawn toward LTE as the enabling technology to meet service demands. The LTE specifications were finalized by the 3GPP in March 2009 (LTE Rel-8), and the first LTE commercial network was launched in Sweden in late 2009. LTE soon became the primary standard for fourth generation (4G) wireless communications. The 3GPP offered several enhancements in subsequent releases. In June 2011, the 3GPP finalized the first specifications for LTE-Advanced (LTE-A) in Rel-10, which adds features such as advanced modes for multiple-input multiple-output (MIMO) systems and carrier aggregation. In Rel-11, coordinated multipoint (CoMP) transmission modes were defined. In Rel-12, features such as LTE-wireless LAN integration and machine type communications (MTC) were introduced. Several new features have been introduced in Rel-13, such as narrowband Internet of Things (IoT), enhancing LTE device-to-device (D2D) services and using LTE in unlicensed spectrum.

LTE offers better coverage, enhanced system capacity, higher spectral efficiency, lower latency, and higher data rates than its predecessors in a cost-effective manner. Currently, there are 494 commercially LTE networks in 162 countries, of which 127 operators have commercially launched LTE-A carrier aggregation in 61 countries [1]. The huge amount of research and development (R&D) that has been invested in the development and deployment of LTE makes it an ideal candidate for non-commercial service, too. It promises to become the standard for a unified broadband public safety network for providing better awareness of and faster recovery from emergency situations [2]. LTE is also considered for supporting mission-critical operations, inter-vehicle communications, machine-to-machine (M2M) communications, and many other applications. LTE/LTE-A is unarguably the primary standard for 4G cellular and is expected to play a big role in the development of 5G technology [3].

The success of LTE has pushed cellular network operators to strive for innovative and scalable solutions to keep pace with the steadily growing service demand. LTE Unlicensed has been proposed recently to operate in the 5 GHz band and will allow cellular network operators to offload some of their data traffic to unlicensed bands, which can lead to a significant increase in data rates offered for LTE users [4]. Currently, there are three proposed variants of LTE Unlicensed [5]. The first is called LTE-U and is developed by the LTE-U Forum to work with the existing 3GPP Releases 10/11/12. LTE-U is designed to operate in countries, such as the United States and China, that do not mandate implementing the listen-before-talk (LBT) technique. The second variant is called licensed assisted access (LAA) and is being standardized by the 3GPP in Rel-13. The major design target for LAA is to have a single unified global

Digital Object Identifier:
10.1109/MCOM.2017.1500706CM

Mina Labib, Vuk Marojevic, Jeffrey H. Reed, and Amir I. Zaghloul are with Virginia Tech; Amir I. Zaghloul is also with the U.S. Army Research Laboratory.

framework that complies with all the regulatory requirements in the different regions of the world. Accordingly, several functionalities need to be supported for an LAA system such as LBT (which is mandated in Europe and Japan) and dynamic frequency selection (DFS) to avoid causing interference to radar systems. These functionalities are in addition to the requirement to extend the physical layer capabilities to support operation in the 5 GHz frequency band, with system bandwidths not less than 5 MHz. These new functionalities will require modifying several functions that are performed by the different layers of the protocol stack [6]. In Rel-13, the downlink operation for LAA is defined, and the uplink operation will be added in subsequent releases. Both variants, LTE-U and LAA, propose to use the licensed spectrum as the primary carrier for signaling (control channels) and to carry data of users with high quality of service (QoS) requirements. Carrier aggregation will be used to add secondary component carriers in the unlicensed spectrum to deliver data to users with best effort QoS requirements. The third variant of LTE Unlicensed is called MulteFire and is proposed by Qualcomm as a standalone version of LTE for small cells. This variant will use only the unlicensed spectrum in the 5 GHz band as the primary and only LTE carrier.

Since LTE will keep playing a dominant role for broadband communications for years to come, reliability becomes an important aspect for the evolution of LTE. LTE security-related questions have been posed recently and investigated by several researchers. Despite the fact that LTE provides higher security than previous generations of cellular systems, such as UMTS (3G) or GSM (2G) [7], LTE is still vulnerable to natural or unintentional as well as intentional interference [8]. Unintentional interference has been broadly analyzed, and interference mitigation, coordination, and cancellation techniques have been proposed. Enhanced interference cancellation techniques are deployed today and continuously improved to allow for network densification as a means of increasing capacity. LTE was originally designed for use in licensed spectrum and has mechanisms for dealing with low signal-to-noise ratio. However, LTE lacks mechanisms for protecting control channels from intentional interference. Intentional interference at the physical layer can be in the form of jamming or RF spoofing. Jamming refers to intentional RF interference to a target signal. The work in [9] provides an overview of the physical layer resiliency of orthogonal frequency-division multiplexing (OFDM), the air interface of LTE. Jamming can be categorized under barrage jamming, partial-band jamming, pilot-tone jamming, and protocol-aware jamming. Protocol-aware jamming refers to interference generated to a specific subsystem by using knowledge of the different physical channels and signals, their locations, and their roles in effective system operation. The potential threat of LTE protocol-aware jamming was brought to light in a letter to the U.S. Department of Commerce [10]. As opposed to jamming, RF spoofing does not generate a noise-like signal that interferes with the target signal, but rather regenerates specific control signals that impede the user from attaching to the regular network and receiving communications service [11].

This article discusses the following aspect of the evolution of LTE: the need to ensure service availability to satisfy the growing dependence on 4G LTE services for different types of applications, including commercial and mission-critical. We identify an emerging threat that can slow down the evolution of LTE. This threat is in the form of RF spoofing, which affects the initial cell selection process as a result of natural or international interference. After providing the necessary background, this article analyzes the problem, proposes effective solutions, and discusses their impact.

The Initial Cell Selection Process in LTE

During initialization, the user equipment (UE) performs the cell selection process and acquires the basic network information. The UE then performs the random access procedure to access the network and set up a dedicated connection with the eNodeB. Once the connection is established, the UE requests to attach to the network, and the authentication procedure follows. In each of these stages, protocol-specific messages are exchanged between the different protocol layers of the UE and their counterparts at the eNodeB or the evolved packet core network. Figure 1 summarizes the main steps that the UE performs as part of the cell selection process. The protocol layers involved are the physical (PHY), radio resource control (RRC), and non-access stratum (NAS) layers. At power up, the UE tries to find a cell to camp on. *Camping on a cell* means tuning to the control channels of that cell and enables the UE to receive broadcast messages transmitted by the eNodeB. These are a series of messages that are collectively called *system information* messages. Some of these system information messages comprise information regarding the cell and its configuration, and enable the UE to access the cell and establish a connection with the network.

In order to select a cell, a public land mobile network (PLMN) should first be selected by the NAS layer, which then requests the RRC layer to select a cell of the selected PLMN (or its equivalents), that is, a cell that broadcasts in its system information messages that it belongs to the selected PLMN (or its equivalents) [12].

During the cell selection process, the UE sequentially scans the bands that it supports. This band scanning enables the UE to find the active RF channels on the supported LTE bands. (An RF channel is considered active if the received signal strength indicator, or RSSI, exceeds a certain threshold.) In the case where there is more than one LTE cell on an active RF channel, the UE selects the strongest cell as per the LTE 3GPP specifications, which state that "the UE needs only search for the strongest cell" at any given frequency [13]. The reason for this is to prevent UEs from creating uplink interference by choosing a cell other than the strongest. The UE acquires timing and frequency synchronization with the help of the cell's primary and secondary synchronization signals (PSS and SSS). More precisely, the PHY layer down-converts and digitizes the received signal on a carrier frequency and correlates it with three locally generated primary synchronization sequences in the time domain to find the strongest cell, that is, the cell that provides the highest correlation result. Based on the correlation results,

There are a series of messages in LTE that are collectively called *system information* messages. Some of these system information messages comprise information regarding the cell and its configuration, and enable the UE to access the cell and establish a connection with the network.

RF spoofing in LTE can have two forms: intentional and unintentional. Intentional spoofing involves an attacker that creates a partial or full LTE downlink frame (fake cell) trying to deceive UEs and prevent them from camping on a legal cell. Unintentional spoofing happens when cells are densely deployed in an uncoordinated way.

the UE determines the cell's physical layer identity and acquires time and frequency synchronization. The PHY layer then detects the SSS, which is at a known position with respect to the PSS and carries the physical cell identity group. The physical cell identity group together with the physical layer identity provides the unambiguous physical cell identity (PCI). The UE also learns about the cyclic prefix (CP) type and the frequency-/time-division duplex (FDD/TDD) mode used by the cell.

Once the UE is time- and frequency-aligned with the LTE downlink frame, it looks for the reference signal (RS) at known locations for a cell quality check and for channel equalization, which is needed to decode the master information block (MIB). The MIB is part of the broadcast control channel (BCCH) carried over the physical broadcast channel (PBCH) and contains the essential LTE system access parameters, such as the LTE system bandwidth and the system frame number.

The UE next acquires the *SystemInformationBlockType1* or SIB1 message, which is also part of the BCCH, but is carried over the physical downlink shared channel (PDSCH). By decoding this message, the UE can complete its check if the cell

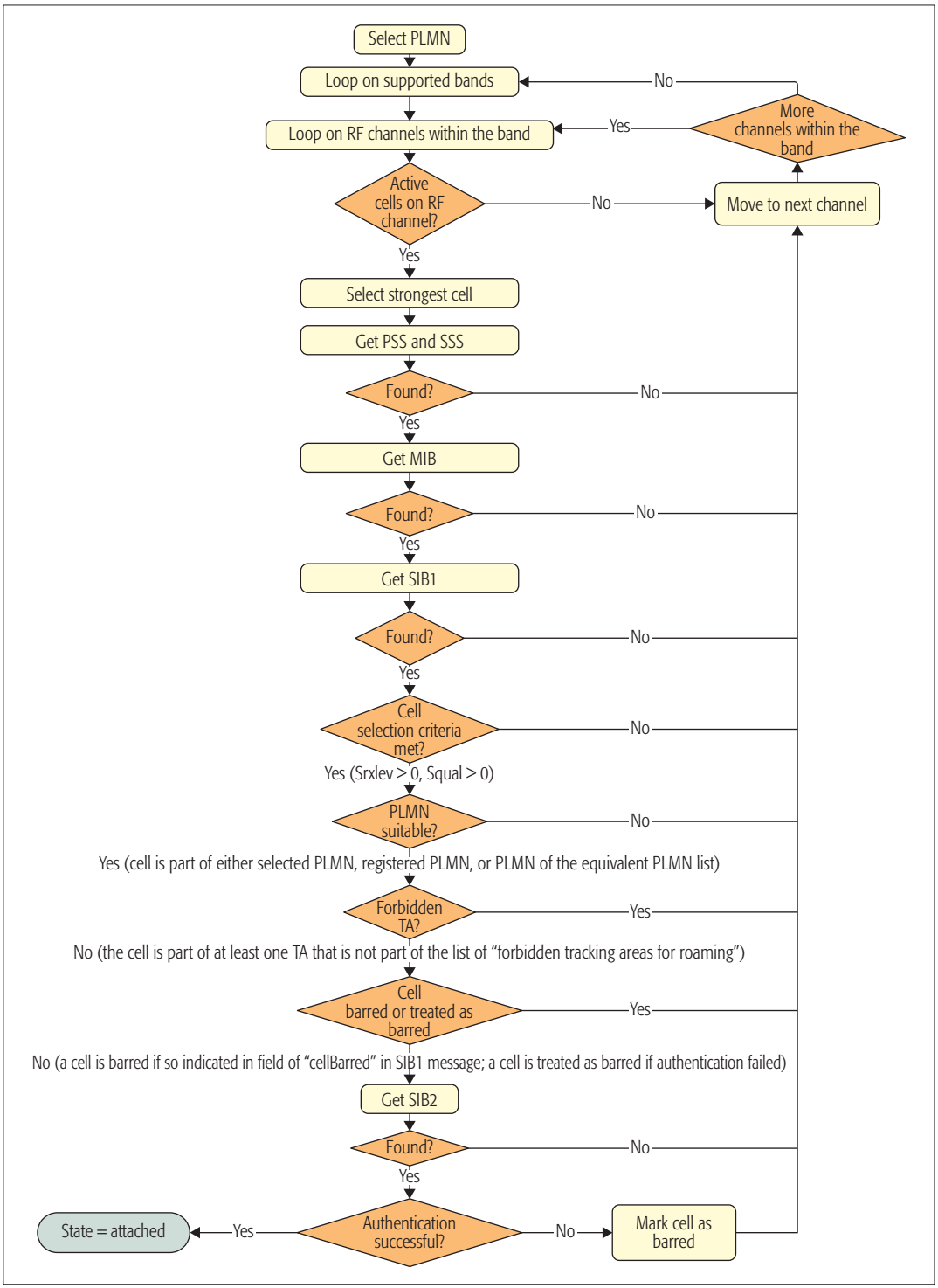


Figure 1. Initial cell selection process for the UE.

is suitable for camping. A cell is suitable for camping if it satisfies the following criteria:

- The *S-criterion*, meaning that the cell has a good power level and quality, measured in terms of reference signal received power (RSRP) and reference signal received quality (RSRQ).
- The cell is part of either the selected PLMN, the registered PLMN, or a PLMN from the equivalent PLMN list. The selected PLMN or its equivalents are provided by the NAS layer to the RRC layer when requesting to select a cell. The PLMN to which the cell belongs is advertised in the SIB1 message.
- The cell is part of at least one tracking area (TA) that is not part of the list of *forbidden tracking areas for roaming*. The *forbidden tracking areas for roaming list* is provided by the NAS layer. The SIB1 message contains the tracking area code (TAC) to which the cell belongs.
- The cell is not barred or treated as a barred cell. A cell is barred if it is indicated as such in the SIB1 message field *cellBarred*. A cell is treated as barred when the mutual authentication between the UE and the network fails.

The SIB1 message also contains other important information, such as the cell identity and the *intraFreqReselection* flag, which indicates whether the UE is allowed to choose the second strongest cell at the same frequency in case the strongest cell is barred or to be treated as barred by the UE.

If all the conditions for camping are satisfied, the RRC layer considers the cell suitable for camping and instructs the PHY layer to acquire the *SystemInformationBlockType2* or SIB2 message. Like the SIB1 message, the SIB2 message is also part of the BCCH and is carried over the PDSCH. The UE is now camped on the cell, and the initial cell selection procedure terminates. Otherwise, if the UE finds that the strongest cell is not suitable for camping (one or more suitable cell criteria are not met), it will try to camp on the strongest cell on another active carrier in a given band. If no suitable cell is found on all active RF channels in a band, the UE will continue searching for a suitable cell on another band that is supported by the UE. The network is acquired when a suitable cell for camping is found.

After network acquisition, the UE starts the connection establishment procedure in order to attach to the network. Attaching to the network means that the UE registers itself with the network, followed by the mutual authentication process. If the UE successfully authenticates with the network, the attachment procedure is completed successfully. If the mutual authentication between the UE and the network fails, the UE will treat the cell as barred. The UE shall exclude the barred cell, or the cell that is treated as barred, for 300 s [13].

The initial cell selection process was first standardized in 3GPP Rel-8 and was not changed through Rel-12. All LTE service consumers need to go through the initial cell selection process every time the UE is turned on or returns from being out of coverage. Whereas making it simple and flexible was the primary objective when LTE was first launched, the massive deployment of LTE in dense urban areas, the need for providing extensive new applications, and the LTE evolution into the unlicensed spectrum require revisiting the cell selection process and analyzing its robustness against unintentional and intentional RF interference.

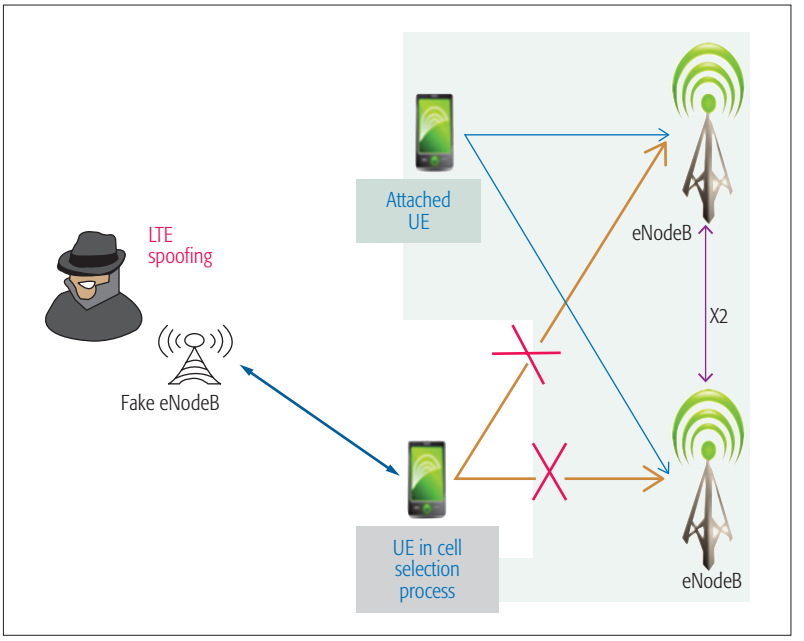


Figure 2. Intentional LTE RF spoofing.

RF Spoofing in LTE

RF spoofing in LTE can have two forms: intentional and unintentional. Intentional spoofing involves an attacker that creates a partial or full LTE downlink frame (fake cell) trying to deceive UEs and prevent them from camping on a legal cell. Unintentional spoofing happens when cells are densely deployed in an uncoordinated way.

Intentional RF Spoofing

RF spoofing in LTE was introduced in [11] and refers to setting up a fake eNodeB that transmits some of the LTE signals and higher-layer control messages (which is why it can be called LTE control channel spoofing), but does not have the authentication keys or offer any service. If this fake eNodeB appears as the strongest cell at the UE for a given frequency channel, the UE will try to camp on this fake cell and will not be able to select any other LTE cell in that channel. The different levels of spoofing range from creating a fake LTE frame that contains only the PSS/SSS to creating a fake LTE frame that contains most of the LTE downlink control signals and channels. The concept of intentional RF spoofing is illustrated in Fig. 2.

Synchronization Signal Spoofing: The simplest way of RF spoofing is PSS & SSS spoofing, which is a PHY layer signaling attack. A PSS along with an SSS are created according to the LTE specifications. That is, the fake eNodeB arbitrarily chooses the synchronization sequences (PSS and SSS) and periodically transmits them asynchronous to, but on, the LTE carrier frequency of the legal eNodeB. When the PHY layer of the UE reports to the RRC layer that it has received the PSS and SSS, the RRC will expect to receive the MIB message next. Since no PBCH is transmitted, the RRC layer

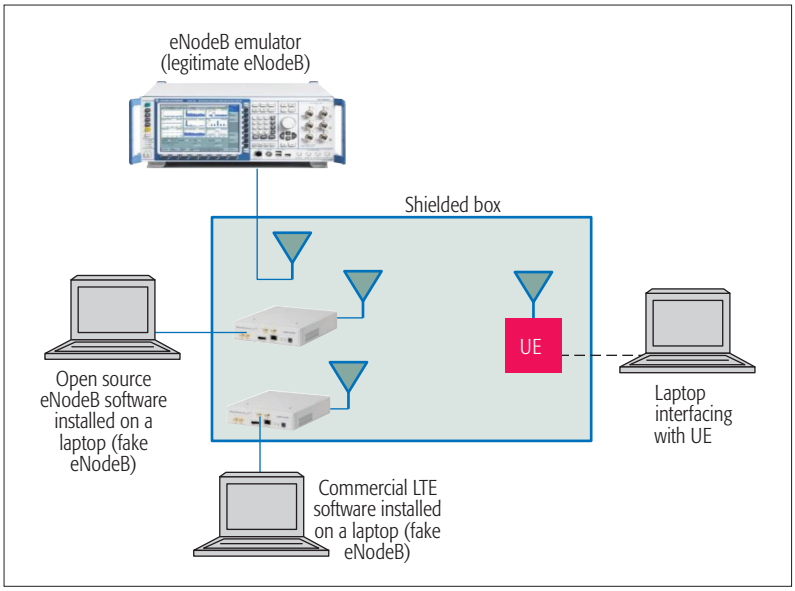


Figure 3. Block diagram of the testbed.

eventually instructs the PHY layer to search for another cell at another frequency.

The 3GPP specifications for RRC [14] state that if the RRC is in the idle mode and does not receive either the MIB or SIB1 message, the UE shall treat this cell as barred and perform barring as if the *intraFreqReselection* flag is set to *allowed*. Hence, the UE may select the second strongest cell. However, many UE manufacturers may overlook the importance of choosing the second strongest cell for the sake of simplifying the interface between the PHY and RRC layers. The implications of this would be allowing the RRC to instruct the PHY to scan only a specific frequency, and the PHY is programmed to deliver only the strongest cell and does not have the ability to distinguish between a barred and a non-barred cell, as demonstrated in [15].

Partial LTE Downlink Frame Spoofing: A more sophisticated way of spoofing is when the fake eNodeB transmits a partial LTE frame using the PLMN of the legal eNodeB. This frame contains the PSS, SSS, RS, PBCH, and the PDSCH's SIB1 message, but not SIB2. The 3GPP specifications state that if the UE does not receive the SIB2 message, it shall treat this cell as barred and shall refer to the received SIB1 message to learn if it is allowed to select another cell within the same frequency by reading the *intraFreqReselection* flag in the SIB1 message. The attacker simply needs to set this flag to *notAllowed* to prevent the UE from camping on a cell at this frequency. This type of vulnerability is not implementation-specific, but rather standard-specific [15].

If the attacker sends a SIB2 message as well, the UE initiates the mutual authentication process after decoding the SIB2 message. The authentication process fails since the fake eNodeB does not have the valid keys. As a result, the UE treats this cell as barred, and the RRC layer starts a new cell selection process. During this new cell selection process, the PHY layer will again report the strongest cell; the RRC layer will find that the cell is to be treated as barred and consequently instruct the PHY layer to resume the cell search on a different frequency [11].

System Information Message Spoofing: The fake eNodeB transmits specific parameters in the SIB messages to cause denial of service. The *cellBarred* field in the SIB1 message was introduced by the 3GPP to enable mobile operators to perform testing and maintenance on any cell before allowing users to actually access it. If the *cellBarred* field is set to *True*, the particular cell will be barred. In this case, and according to the 3GPP specifications, the UE will exclude the barred cell for 300 s. If the fake cell transmits the same PCI as the legal eNodeB, the legal eNodeB will be wrongly excluded by the UE for 300 s, even if the fake cell is turned off in the meanwhile. The fake cell could operate at a low duty cycle and still permanently prevent a UE from camping on the legal cell by repeatedly barring it [11].

Unintentional RF Spoofing

MulteFire entails multiple private small cell eNodeBs transmitting exclusively in the unlicensed band. If there are two eNodeBs belonging to different networks that happen to transmit on the same frequency, an effect similar to RF spoofing may occur. When a UE receives an LTE frame from a neighbor eNodeB with higher power than its own eNodeB, it will try to access this network. The authentication process will fail because the UE does not belong to the network it tries to access. It is important to point out that this may happen only if the two eNodeBs have the same PLMN, and the PLMN consists of a Mobile Country Code (MCC), which is 3 digits, and a Mobile Network Code (MNC), which is 2 or 3 digits. Given the fact that we are considering the case where MulteFire will be operated in an uncoordinated way for private and dense small cells, and a PLMN can be set up by the user during the initial eNodeB setup and configuration, this scenario is not unlikely. Furthermore, the authentication process for MulteFire has not been identified yet, and there is a discussion about removing the need for a SIM card at the UE to keep it as simple as WiFi.

In this case, once the UE synchronizes to the strongest cell, it will not be able to resynchronize to a weaker cell, as explained above and demonstrated later. Hence, denial of service will occur. Given that for MulteFire the cell is not expected to offer access through another type of wireless technology (e.g., 3G or 2G), the UE will not be able to receive any service at all.

Experimental Assessment

We have created a lab-scale testbed in order to validate the effect of LTE control channel spoofing on the UE. Figure 3 shows the block diagram of the testbed. The test results consistently show that RF spoofing impedes UEs that are in the initial cell selection process from attaching to the legal eNodeB. The UE is able to attach to the legal eNodeB after the fake or uncoordinated cell is turned off. The test procedures and results are summarized in Fig. 4.

For LTE and LTE-A systems, a fake eNodeB that transmits part of the LTE frame can enforce the UE to search for an LTE signal in a different RF channel or band. Given that most LTE networks are based on a single frequency, this will cause

permanent denial of service for the UE to the LTE network. In this case, the UE will have to downgrade its service to a 3G or 2G system, which offers a much slower data rate and is vulnerable to other types of security attacks. It is worth mentioning that rebooting the UE device will not solve the problem. The UE is not able to attach to the LTE network as long as the UE receives the fake synchronization signals at a higher power level than those that are transmitted from the legal eNodeB.

Mitigation

We propose simple modifications to the LTE cell selection process that would effectively mitigate the effect of RF spoofing. Figure 5 indicates the proposed modifications. They can be summarized as follows:

- The PHY layer should report to the RRC layer a list of all cells that are detected along with their corresponding PCI and received power levels. The RRC layer should save this list.
- The RRC layer should create two separate flags, one for a barred cell as received in the SIB1 message, and one for a cell to be treated as barred to ensure that the UE will handle these two cases differently. The cell should be treated as barred in the following cases:
  - –When the mutual authentication process between the UE and the network fails at the UE side
  - –When the UE does not receive any of the control messages, specifically the MIB, SIB1, or SIB2 messages, within a specific time frame, which also needs to be specified
- The RRC layer should first check if the cell is to be treated as barred, before checking if the cell is barred.
- When the cell is to be treated as barred, the UE should be allowed to select the second strongest cell at the same frequency.
- If the *cellBarred* flag in SIB1 is *True*, the RRC layer should check for any duplicate PCI at the same frequency. If the RRC finds that *cellBarred* is *True*, and the PCI of that cell is not unique, the RRC layer should flag this cell to be treated as barred.

To elaborate further, we propose allowing the UE to select the second strongest cell at the same frequency only for the case when the strongest cell is to be treated as barred. The reason for the 3GPP specifications mandating the UE to select the strongest cell is to not create excessive uplink interference by the UE to other UEs when communicating with a farther cell. The proposed modifications allow selecting the second strongest cell if and only if the strongest cell is to be treated as barred. In the case of licensed spectrum, when the cell is flagged as "to be treated as barred," there are no UEs attached to it. Hence, the UE will not create any uplink interference if it selects the second strongest cell. This proposed modification will not affect the current UE procedure in case of roaming or handover, as these only affect the UE initialization process when the authentication fails at the UE side, and it will be an effective mitigation technique against partial LTE downlink frame spoofing. In the case of unlicensed spectrum, when the strongest cell is marked as "to be treated as barred," this means that cell is not the serving eNodeB for this UE. In this case, the UE should keep searching for its own eNodeB by allowing it to select the second strongest cell.

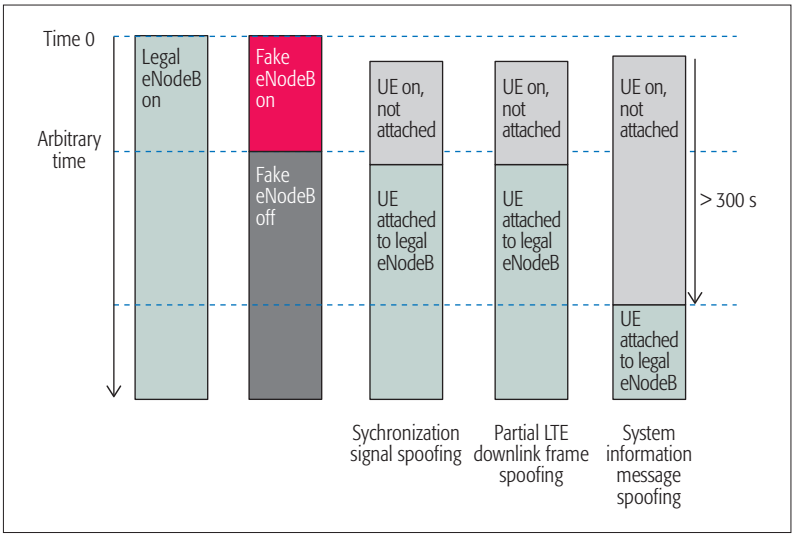


Figure 4. Summary of test results for the different types of RF spoofing.

We also propose to create a timer at the RRC layer for receiving any of the essential LTE control messages (e.g., the MIB, SIB1, or SIB2 messages), and when the timer expires, the RRC marks this cell as "to be treated as barred." This then allows the UE to select the second strongest cell in case of failing to receive any of the required control messages. If the RRC layer does not receive the expected message within the specified time, the RRC can mark this cell to be treated as barred. Hence, the UE would mark a cell to be treated as barred when the mutual authentication fails or when the UE does not receive a control message within the specified timeframe. This is our proposed mitigation method against synchronization signal spoofing.

By creating two separate flags, one for a cell to be treated as barred and one for a barred cell, we ensure that the UE will handle the two cases differently. In the case where the fake eNodeB broadcasts itself as barred in the SIB1 message, we propose that the UE checks if this PCI is duplicated in the list of the received PCIs at that frequency. If so, we recommend that the UE treats this cell as barred. This would allow the UE to select the second strongest cell rather than automatically search on a different frequency. The network can eventually hand over the UE to the strongest cell, which was mistakenly barred as a result of RF spoofing, and so avoid denial of service caused by system information message spoofing.

For the case of unintentional spoofing, after the UE fails to authenticate the network that is transmitting the strongest cell, the proposed modifications will enable the UE to select the second strongest cell, which belongs to its own network.

The proposed modifications collectively provide an effective mitigation method against the different types of intentional and unintentional RF spoofing discussed in this article. These changes ensure backward compatibility with the deployed LTE devices and networks, and do not

With the advent of LTE operating in unlicensed bands, the availability of the LTE network can become increasingly compromised the more networks are deployed. This is not only because of the classical inter-cell interference, but also because of the cell selection process that does not scale well for uncoordinated operation in unlicensed bands.

introduce excessive uplink interference. Hence, the proposed modifications will not lead to any performance degradation when compared to the current LTE system deployments and operation.

Conclusions

We have shown how RF spoofing, whether intentional or unintentional, can prevent LTE/LTE-A users from accessing the network and obtaining 4G services. As the number of users grows and new LTE technologies appear, this can become a serious threat that will impede scaling LTE to fulfill the emerging needs.

With the advent of LTE operating in unlicensed bands, the availability of the LTE network can become increasingly compromised the more networks are deployed. This is not only because of the classical inter-cell interference, but also because of the cell selection process that does not scale well for uncoordinated operation in unlicensed bands. Future mission-critical networks likewise can suffer from this. Public safety and mil-

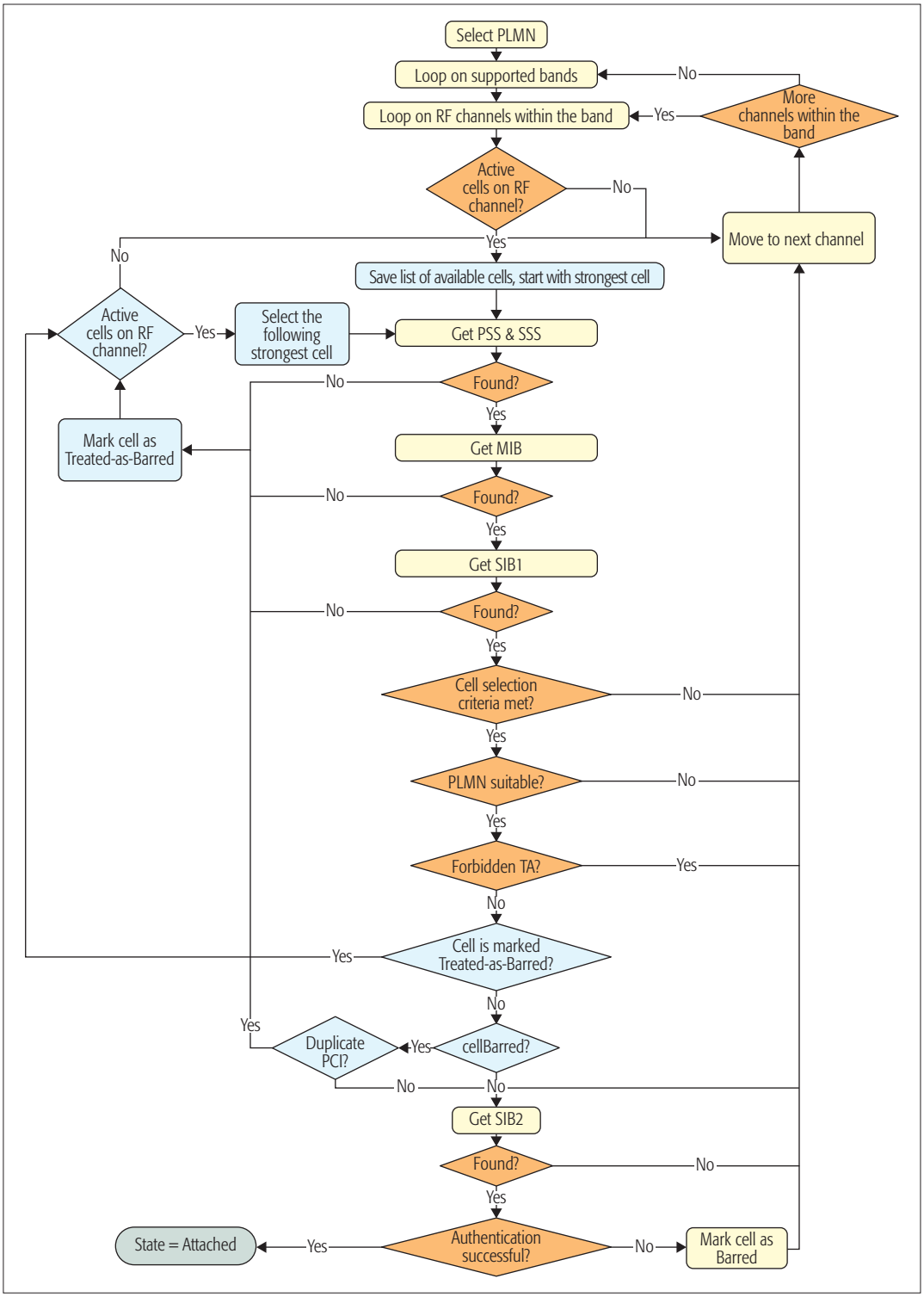


Figure 5. Proposed cell selection process.

itary systems will use LTE Rel-8 or higher. Hence, we recommend adopting these simple fixes in future releases of LTE-A and use these releases for networks that will offer mission-critical services.

Beyond LTE-A, next generation networks need to be more reliable by providing robust network access. Research is needed to redesign the network access procedure and help avoid potential problems that can arise from exploiting the openness of standards and control channel dependency.

Acknowledgment

Portions of this work are covered under U.S. Patent Application 62/156377. This work was supported in part by Vencore Inc., the Defense University Research Instrumentation Program (DURIP) contract numbers W911NF-14-1-0553/0554 through the Army Research Office, and the National Science Foundation (NSF) contract number CNS-1642873.

References

[1] GSA, "Evolution to LTE Report: 4G Market and Technology Update," Global Mobile Suppliers Assn., tech. rep., Apr. 2016; http://www.gsacom.com.
[2] T. Doumi *et al.*, "LTE for Public Safety Networks," *IEEE Commun. Mag.*, vol. 51, no. 2, Feb. 2013, pp. 106–12.
[3] J. Andrews *et al.*, "What Will 5G Be?," *IEEE JSAC*, vol. 32, no. 6, June 2014, pp. 1065–82.
[4] R. Zhang *et al.*, "LTE-Unlicensed: The Future of Spectrum Aggregation for Cellular Networks," *IEEE Wireless Commun.*, vol. 22, no. 3, June 2015, pp. 150–59.
[5] D. R. Brenner and J. W. Kuzin, "Before the Federal Communications Commission: Reply Comments of Qualcomm Incorporated," June 2015; http://apps.fcc.gov/ecfs/document/view?id=7022130311.
[6] 3GPP, "Study on Licensed-Assisted Access to Unlicensed Spectrum (Release 13)," TS 36.889, 2015; http://www.3gpp.org/dynareport/36889.htm.
[7] J. Cao *et al.*, "A Survey on Security Aspects for LTE and LTE-A Networks," *IEEE Commun. Surveys & Tutorials*, vol. 16, no. 1, 1st qtr. 2014, pp. 283–302.
[8] R. Jover, "Security Attacks against the Availability of LTE Mobility Networks: Overview and Research Directions," *2013 16th Int'l. Symp. Wireless Personal Multimedia Commun.*, June 2013, pp. 1–9.
[9] C. Shahriar *et al.*, "PHY-Layer Resiliency in OFDM Communications: A Tutorial," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 1, 1st qtr., 2015, pp. 292–314.
[10] J. H. Reed and M. Lichtman, "FirstNet Conceptual Network NOI — Comments of Wireless @ Virginia Tech," before the Dept. of Commerce, Docket No. 120928505250501, RIN 0660XC002, Nov 2012; http://www.ntia.doc.gov/files/ntia/va_tech_response.pdf.
[11] M. Labib, V. Marojevic, and J. H. Reed, "Analyzing and Enhancing the Resilience of LTE/LTE-A Systems to RF Spoofing," *IEEE Conf. Standards for Commun. and Net. Proc.*, Oct. 2015, pp. 160–65.
[12] S. Sesia, I. Toufik, and M. Baker, *LTE — The UMTS Long Term Evolution: From Theory to Practice*, 2nd ed., Wiley, 2011.
[13] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) Procedures in Idle Mode (Release 12)," TS 36.304, Mar. 2015; http://www.3gpp.org/dynareport/36304.htm
[14] —, "Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC) (Release 12)," TS 36.331, Mar. 2015; http://www.3gpp.org/dynareport/36331.htm
[15] M. Labib *et al.*, "How to Enhance the Immunity of LTE Systems against RF Spoofing," *Int'l. Conf. Computing, Networking and Commun.*, Feb. 2016.

Biographies

Mina Labib (mlabib@vt.edu) received his B.S. degree from Ain Shams University, Cairo, Egypt, in electronics and communications engineering, and his M.Sc. degree from Carleton University, Ottawa, Ontario, Canada, in systems and computer engineering. He is currently working toward his Ph.D. degree at the Bradley Department of Electrical and Computer Engineering at Virginia Tech within the Wireless@VirginiaTech research group. He has wide industrial experience, especially in the field of physical and MAC layer design. His current research interests are in the broad areas of wireless communications, with a particular emphasis on LTE systems, enhancing the security of wireless communication systems, LTE-Unlicensed, spectrum sharing, and game theory.

Vuk Marojevic(maroje@vt.edu) received his M.S. from the University of Hannover, Germany, and his Ph.D. from the Universidad Politècnica de Catalunya, both in electrical engineering. He joined Wireless@Virginia Tech in 2013. His research interests are in software-defined radio, spectrum sharing, 4G/5G cellular technology, wireless testbeds and testing, and resource management with application to public safety and mission-critical networks and unmanned aircraft systems.

Jeffrey H. Reed [F] is the founder of Wireless@Virginia Tech, and served as its director until 2014. He is the founding faculty member of the Ted and Karyn Hume Center for National Security and Technology, and served as its interim director when founded in 2010. His book *Software Radio: A Modern Approach to Radio Design* was published by Prentice Hall ,and his latest textbook, *Cellular Communications: A Comprehensive and Practical Guide*, was published by Wiley-IEEE in 2014. He is co-founder of Cognitive Radio Technologies (CRT), a company commercializing cognitive radio technologies; Federated Wireless, a company developing spectrum sharing technologies; and PFP Cybersecurity, a company specializing in security for embedded systems. In 2005, he became a Fellow of the IEEE for contributions to software radio and communications signal processing and for leadership in engineering education. In 2013 he was awarded the International Achievement Award by the Wireless Innovations Forum. In 2012 he served on the President's Council of Advisors of Science and Technology Working Group, which examines ways to transition federal spectrum for commercial use. He is a past member of CSMAC, a group that provides advice to the NTIA on spectrum issues.

Amir I. Zaghloul [LF] is an ARL Fellow with the Sensors and Electron Devices Directorate (SEDD) of the U.S. Army Research Laboratory (ARL), Adelphi, Maryland. After 24 years at COMSAT Laboratories performing and directing R&D efforts on satellite communications and antennas, he joined Virginia Tech in 2001 as a professor in the Electrical and Computer Engineering Department. In 2008, he was assigned as an IPA from Virginia Tech to the ARL, and subsequently switched to full-time at ARL in 2012, maintaining his affiliation with Virginia Tech as a research professor. He is also affiliated with the University of Delaware (2012–present). He has held positions at the University of Waterloo, Canada (1968–1978), the University of Toronto, Canada (1973–1974), Aalborg University, Denmark (1976), and Johns Hopkins University, Maryland (1984–2001). He is a Fellow of the Applied Computational Electromagnetics Society (ACES) and an Associate Fellow of the American Institute of Aeronautics and Astronautics. He is also the International Vice-Chair of Commission C of the International Union of Radio Science (URSI), and has held several positions at the IEEE, URSI, and ACES. He has received several research and patent awards, including the Exceptional Patent Award at COMSAT and the 1986 Wheeler Prize Award for Best Application Paper in *IEEE Transactions on Antennas and Propagation*. He received his Ph.D. and M.A.Sc. degrees from the University of Waterloo in 1973 and 1970, respectively, and his B.Sc. degree (Honors) from Cairo University, Egypt, in 1965, all in electrical engineering. He also received an M.B.A. degree from George Washington University in 1989.

Beyond LTE-A, next generation networks need to be more reliable by providing robust network access. Research is needed to redesign the network access procedure and help avoid potentail problems that can arise from exploiting the openness of standards and control channel dependency.

ACCEPTED FROM OPEN CALL

Service Function Chaining in Next Generation Networks: State of the Art and Research Challenges

Ahmed M. Medhat, Tarik Taleb, Asma Elmangoush, Giuseppe A. Carella, Stefan Covaci, and Thomas Magedanz

The authors introduce a service function chaining taxonomy that considers architecture and performance dimensions as the basis for the subsequent state-of-the-art analysis. The article concludes with a gap analysis of existing solutions and the identification of future research challenges.

ABSTRACT

Service function chaining is a network capability that provides support for application-driven-networking through the ordered interconnection of service functions. The lifecycle management of service function chains is enabled by two recently emerged technologies, software defined networking and network function virtualization, that promise a number of efficiency, effectiveness, and flexibility gains. This article introduces a service function chaining taxonomy that considers architecture and performance dimensions as the basis for the subsequent state-of-the-art analysis. The article concludes with a gap analysis of existing solutions and the identification of future research challenges.

INTRODUCTION

Network resource management and service differentiation according to user requirements and network constraints are crucial elements of the business and operations support systems of any telecommunications operator. These two key capabilities are particularly challenging considering the steady increase in the number of services/applications, their heterogeneous quality of service (QoS) requirements, and the overall traffic that the network has to provide. Service function chaining (SFC) is an enabling technology for the flexible management of specific service/application traffic, providing solutions for classifying flows and enforcing adequate policies along the flow routes according to the service requirements and considering the availability status of the network. SFC is defined as a chain-ordered set of service functions (SFs) that handles the traffic of the delivery (data plane), control, and monitoring (control plane) of a specific service/application.

Recently, SFC has made use of the new technology called software defined networking (SDN). Architecturally seen, SDN decouples the control plane from the data plane and introduces appropriate programming abstractions exploited in SFC for the dynamic control of the topology of SFCs and the traffic steering across SFs. Network function virtualization (NFV) is related to the telco initiative of adopting cloud-computing technology enabling the virtualization of software-implemented network functions (SFs in SFC terminology). NFV is adopted by SFC to provide efficient and effective deployment and orchestration of SFs.

The SFC architecture specifications are addressed by the IETF SFC working group (RFC 7665) and the Open Network Foundation (ONF). SFC becomes particularly relevant in the new emerging value chains involving multiple data centers (central, edge, fog), access-, core- and transit-networks, and application service providers. As such, SFC has attracted much attention within the community of researchers as well as among network operators and network equipment vendors (e.g., Juniper, QOSMOS, and Huawei). Numerous open source tools enabling SFC are also available. Notable examples are OpenDaylight, OPNFV, ONOS, OpenContrail, and OpenStack's Neutron/Service Insertion and Chaining.

The main contributions of this article are twofold. First, the article explores the limitations of current SFC approaches in next generation networks in terms of architectural and conceptual research work by providing a brief analysis of each solution in the state of the art. The limitations are explored with reference to the SFC IETF specification. Second, the article draws some new research directions. To the best knowledge of the authors, this article is the second research work highlighting the limitations of SFC approaches and the first work to provide a detailed overview of the SFC state of the art and evaluation. The work introduced in [1] was the first in defining the new research directions and challenges of SFC. The authors in [1] presented SFC design considerations and requirements with use cases that show the advantages of adopting SFC. Their main contribution is to explore the research challenges during the exemplary lifecycle of an SFC in an applicable telco network, covering SFC definition, deployment, programming, and security concerns.

The rest of the article is organized as follows fashion. We illustrate the SFC standardized architectures, as defined by the IETF SFC and ONF working groups, and discuss how the ETSI NFV architecture provides SFC. We highlight previous research work conducted on the SFC architectural concepts and implementations. SFC challenges and limitations are discussed. Finally, the article concludes.

Digital Object Identifier:
10.1109/MCOM.2016.1600219RP

Ahmed M. Medhat, Asma Elmangoush, Giuseppe A. Carella, and Stefan Covaci are with Technical University Berlin; Tarik Taleb is with Sejong University and Aalto University; Thomas Magedanz is with Fraunhofer FOKUS

0163-6804/17/$25.00 © 2017 IEEE

SFC Standardized Architectures

According to the IETF SFC specifications (draft-ietf-sfc-control-plane-06), a typical SDN-based SFC architecture consists of components grouped into two layers, the control plane and data plane, as shown in Fig. 1. The control plane is responsible for the SFC management, SF instances management, mapping SFC to a specific service function path (SFP), installing and administering forwarding rules on the service function forwarding (SFF) components of the data plane, and adjusting the SFP in terms of SF instances and overlay links as a result of their status (i.e., overloaded, active, inactive, failed, etc.). The SFC control plane components interact with the SFC data plane components via four interfaces. The first interface C1 is responsible for pushing the SFC classification rules defined by the SFC control plane into the SFC classifiers. The SFFs report the connectivity status of their attached SFs to the SFC control plane. Interface C3 is between the NSH-aware SFs and the SFC control plane. It is used to collect some packet-processing statistics (e.g., SFs' load update) from the SFs. For NSH-unaware SFs, a SFC proxy is provided for collecting statistics (e.g., SF processing latency and workload) and transmitting this information over the C4 interface to the SFC control plane. The SFC control plane uses these statistics (received through interfaces C2, C3, and C4) to dynamically adjust the SFPs.

The main components of the SFC data plane, as shown in Fig. 1, are the SFC classifier, SFF, SF, and SFC proxy. The SFC classifier differentiates the incoming traffic into flows, based on the target application and other predefined requirements. The SFC classifier tags each flow by adding an SFC header containing a service function path (SFP) ID to each flow packet header. The path ID is related to an SFC and identifies the ordered set of abstract SFs which must be performed to the particular flow. The SFP is the real path (the exact SFFs/SFs) that packets traverse.

An SF executes a particular set of actions on incoming packets (e.g., deep packet inspection or firewall functions) and can process packets belonging to several SFPs. An SF can be present with multiple, distributed instances in the network (e.g., for scalability reasons). An SFF is in charge of sending the incoming traffic to SFs and/or other SFFs, according to the defined SFPs. To this purpose, the SFF uses and inserts SFP-specific information in an additional packet-header (SFP packet encapsulation). The IETF SFC working group does not standardize a particular SFF, but instead the SFC special header, called the network service header (NSH) (draft-ietf-sfc-nsh-02). An SFC proxy may become required between SFF and SFs as the majority of SFs do not recognize the SFC packet headers (NSH). The SFC proxy performs SFC packet de-capsulation for the packets forwarded to the NSH-unaware SFs and encapsulates these packets before sending them to the SFF (IETF SFC RFC 7665).

In the SDN context, the Open Networking Foundation (ONF) also proposed another model for the L4-L7 SFC architecture, based on the SDN/OpenFlow controller (ONF TS-027). The ONF SFC system is based on the IETF SFC specification in that it specifies the SFF by an extended OpenFlow switch version supporting NSH.

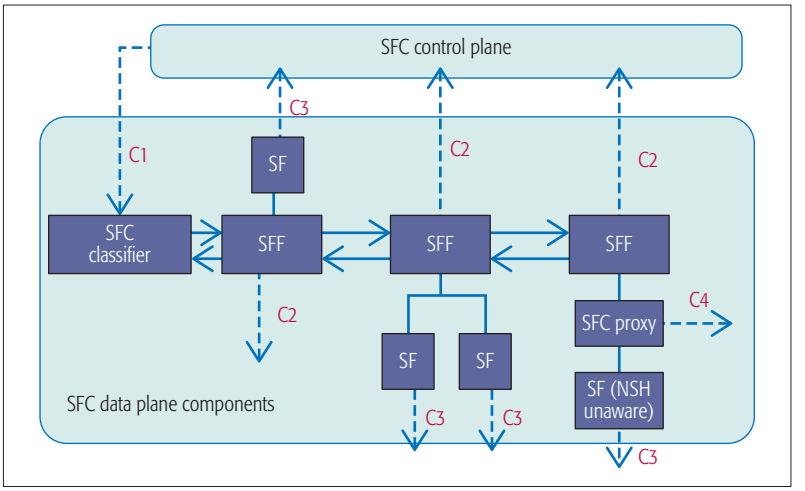


Figure 1. A typical SFC architecture (draft-ietf-sfc-control-plane-06).

An SFC control plane functional architecture is addressed by the ETSI NFV architecture (ETSI GS NFV-MAN 001 V1.1.1) (see Fig. 2). The main components of the ETSI NFV architecture are: NFV orchestrator (NFVO), virtual network function manager (VNFM), and virtualized infrastructure manager (VIM).

NFVO is responsible for the end-to-end management and orchestration of network services (NS) provided by an administrative domain. Each NS is specified by a network service descriptor (NSD). An NS may span multiple network domains belonging to the same or server different administrations. Each network domain contains a network level manager called the network controller that is responsible for network connectivity management. In the case of an NS spanning multiple administrative domains, the overall end-to-end management of the NSs is realized by co-operation of the participating NFVOs, either in a hierarchical or in a peer-to-peer manner. In the case of the hierarchical arrangement, an additional NFVO is introduced in the architecture. Each virtualized infrastructure domain is managed by the so called VIM (e.g., in the case of OpenStack, the virtual network infrastructure manager is the neutron component). NFVO is also concerned with instantiating/updating/terminating of SFCs (i.e., life cycle management of the SFC) and their constituent VNFs (instantiation, update, scaling, migration, and termination) in coordination with VNFMs. The VNFM is responsible for VNFs life cycle management such as VNFs instantiation, update/upgrade, scaling, and termination. The VIM is concerned with controlling and managing the NFV infrastructure (NFVI) compute, storage, and network resources such as providing a "Network as a Service" northbound interface to the higher layers (NFVO and VNFM) and invoking the NFVI network southbound interfaces (network controller or/and VNFs/PNFs) to construct the service within the domain. Each NS contains at least one VNF forwarding graph (VNFFG) that describes the network topology of the NS or a portion of the NS by referencing the VNFs, PNFs, network forwarding path (NFP) that provides the order of involved VNFs or PNFs in the VNFFG, and the virtual links that connect them. In SFC terminology, the VNFFG is considered as the SFC,

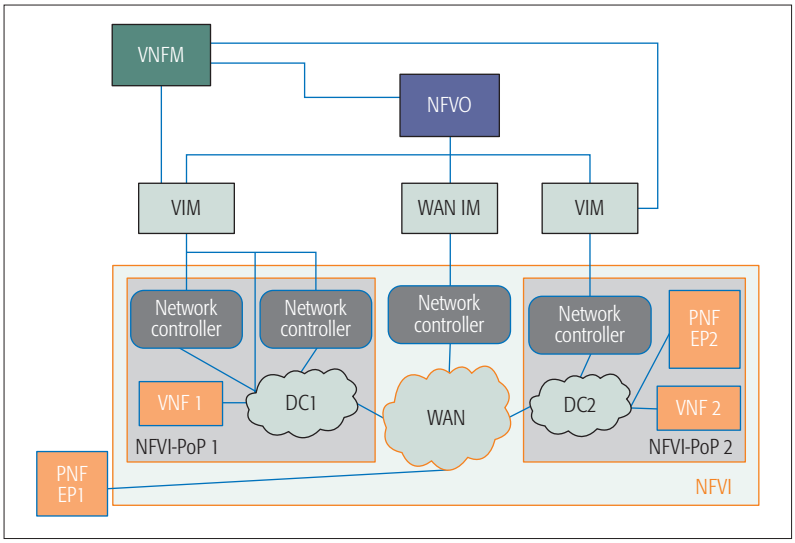


Figure 2. ETSI NFV Architecture (ETSI GS NFV-MAN 001 V1.1.1).

VNFs or PNFs are the SFs, NFPs are the SFPs, and Virtual Links are implemented by one or different SFFs. Fig. 3 shows an example of two VNFFGs (SFCs) imbedded in the same virtual network infrastructure.

Recent research works, such as the work presented in [2], exemplify how SDN/NFV-based SFC standardized solutions can be applied to solving severe congestion in mobile networks (access and core networks) caused by the exchanged user generated content of mobile social media applications through mobile devices.

State of the Art of SFC Concept and Implementations

A wide range of research work has been conducted proposing new frameworks, concepts, and implementations of SFC. These approaches can be classified into two categories based on the adopted technology (i.e., SDN and NFV). Different SFC solutions are investigated, compared, and evaluated in this section, discussing their limitations and defining the research directions that should be considered in the future to improve them. The comparison is made according to the architecture (SFC control and data planes) and the approaches' performance. The key points of comparison in the SFC control plane are:

- Implementation: Shows the technologies used to implement the SFC solution's control plane.
- SFP Adjustment: A dynamic SFP computing in the run-time phase with an approach such as SFP-fail over, SFP with better latency, traffic engineered SFP, and SF/SFP load balancing.
- Orchestrator-based: Shows if the SFC approach's control plane depends on an orchestrator or not.
- QoS/Policy Engine: Shows if the SFC solution's control plane has the capability of enforcing QoS and policies into the network.

The key points of comparison in the SFC data plane are:

- SFF: Explores the scheme applied by the SFC solution on the SFFs in order to steer traffic through the chains.
- SFC Classifier: Shows how to classify incoming traffic.

The key points of comparison in the approaches' performance are:

- Flexibility: Shows the level of flexibility in the SFC approach. The flexibility level is based on the efficiency of the traffic steering scheme implemented in the SFC solution.
- Scalability: Defines the level of scalability in the SFC approach. The scalability level is based on the number of rules needed to apply traffic steering for one chain.

SDN-based SFC Solutions

In [3], the NIMBLE system proves the potential of SDN to simplify and improve the existing middle-box management deployments, addressing challenges relevant to middle-box composition, load balancing, and packet modifications. The proposed NIMBLE system permits network operators to abstract the logical view of the middle-box policy and automatically pushes the forwarding rules into the switches. It considers the network topology, switches' capacities, and middle-box resource constraints. The NIMBLE design implies three main ideas. The first idea consists of the support of the middle-box composition by an efficient data plane that has tunnels between switches and pushes tags to packet headers using the SDN capabilities in order to know the processing status of each packet. The second idea is to provide resource management in a practical unified way and optimization using information on the switches' capacities and load balancing based on traffic fluctuations. The third key idea is to let the middle-box act dynamically by reporting the capabilities of SDN switches to design lightweight flow correlation schemes. A proof-of-concept of NIMBLE is showcasing the improvements achieved in terms of middle-box load balancing. The results also demonstrated the speed of the network bootstrap, and the high responsiveness of the system to network dynamics and load rebalancing.

In [4], a Squid-based FlowTags architecture is proposed whereby middle-boxes add tags to transmitted packets to communicate the necessary middle-box context (e.g., source hosts or internal cache state). Switches and middle-boxes can utilize these tags to provide consistent policy enforcement. An SDN controller is responsible for pushing the actions to the switches and middle-boxes in order to use the proposed tags in the packet header. FlowTags modify the architecture of the interface between the controller and switches by providing a new southbound interface for the flow tagging configuration process and for communication establishment with the FlowTags-aware middle-boxes. The modification takes part in three dimensions. First, FlowTags-aware middle-boxes are assumed to have the ability to process the incoming tags and add new tags based on the context. These tags are used by switches to steer the traffic. Second, a new FlowTags interface is proposed between the SDN controller and the FlowTags-aware middle-boxes. Third, a new control application is assumed to be used for the configuration of tags at switches and middle-boxes that is ultimately for the enforcement and verification of policies. In [4], the authors also provided a proof-of-concept

implementation to show how they modified Squid to support FlowTags and to also demonstrate the capability of a new policy enforcement process. This work seems promising, but there are still significant challenges to tackle. These challenges are related to the scalability and flexibility of the approach.

The position paper in [5] introduced a high-level concept and architecture to support SFC based on OpenFlow in a telecommunication network environment. The architecture supports the assignment of multiple subscribers to a single service while conserving the desired information about subscriber identification by the SF. The proposed model facilitates the SF instances deployment operation by reducing the network configuration modifications needed during deployment. In addition, the architecture instantiates many SF instances with low overhead. The SF instances are dedicated to one SFC only at a time, which provides an isolated network environment. Therefore, the model does not need to do packet matching conditions at the classifier. The separation of SF instances avoids using a consistent network addressing approach that crosses the whole service chaining system. The forwarding of traffic between the SF instances and switches is MAC address based. A proof-of-concept implementation for a relevant use case was provided to evaluate the feasibility of the model.

The work in [6] provides a proof-of-concept implementation of the SDN-based SFC approach presented in [5]. The approach merges common SFs without knowing their chain details. The proposed conception of SF instance separation facilitates the instantiation of SFs and provides a high degree of flexibility. The prototype's feasibility is tested over a hardware device that hosts a group of SF instances. These SFs were used to instantiate SFC for two types of applications (web traffic and video streaming). The demonstration showed the dynamicity of allocating users to new service classes.

A service-oriented SDN controller is proposed in [7] that deploys a programmable data delivery route by setting up multiple chains of VNFs existing in different locations of an OpenFlow-enabled network within the framework of service overlay networks. It also provides network service control, orchestration, and SDN network control functions in order to cope with the "extended QoS" requirements, and provides context-aware delivery of application service data. Moreover, the SFC context-aware architecture provides a realistic differentiation feature known as class-based forwarding. This feature simplifies the scalability issues, resulting in a decreased number of flow entries installed in the network switches. The authors have validated their proposed controller experimentally. The results proved that network optimization could be reached by assigning a specified number of crossed SF instances.

The model proposed in [8] presents a software architecture to dynamically instantiate network function-flow graphs (NF-FGs) beginning from a high level description of the targeted graphs and the existence of a specific incident (e.g., a new user is connected to the node), ending with common traffic steering provided by the SDN architecture. SDN technology is used to dynamically reset the network paths inside the network unit. Traffic forwarding among the nodes of the NF-FG is based on eXtensibleOpenFlowDatapath daemon (xDPd). xDPd is a software switch that creates multiple software Openflow switches in a dynamic way, called logical switch instances (LSIs). These LSIs can be tied together to physical interfaces and to NFs. A three-fold process occurs when the orchestrator node receives a new NF-FG description. First, it calls each required NF implementation and installs it. Second, it instantiates a user-LSI on xDPd, and then attaches it to the suitable NFs and to the classifier. Third, it instantiates an OF controller per each tenant that provides the insertion of appropriate rules into the flow tables of the LSIs.

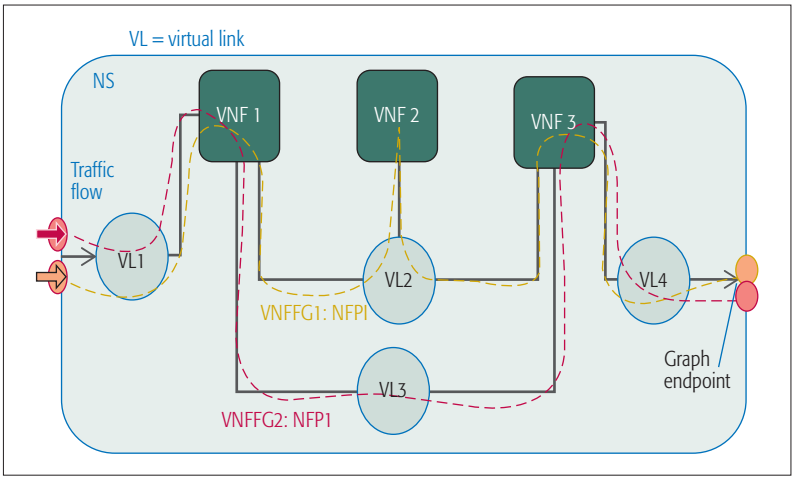


Figure 3. NS with two VNFFGs with different NFPs (ETSI GS NFV-MAN 001 V1.1.1).

The StEERING framework was introduced in [9] to provide an SFC model supporting dynamic traffic routing. StEERING uses a simple central controller that can adjust the traffic steering of various flow types through the targeted chain of middle-boxes. Moreover, it supports high scalability at the level of users and application policies. Scalability is offered through three dimensions. First, the rules at switches can be scaled linearly with the number of users and applications by using multiple tables to convert a single policy space into a multi-dimensional space. Second, it facilitates the integration of various types of policies by specifying the ordered group of service functions that each flow crosses as one type of metadata, so every table can work on the service functions separately. Third, the model provides the classification and header editing rules at the gateways only once within the network. The authors have provided a prototype to check the feasibility of their implementation and show its efficiency in providing flexible routing.

SIMPLE [10] is an efficient routing model for connecting SF instances and an approach to load balancing the SF instances. SIMPLE explicitly considers the inclusion of legacy SF instances. SIMPLE permits allocation of a logical middle-box steering policy and directly transposes this into forwarding rules that consider the network topology, switch capacities, and SF instances resource constraints. In the SIMPLE design, a particular SF instance is chosen to run within the limits of existing SDN capabilities (e.g., OpenFlow) and there is no need

Management of resource utilization is also required in the SFC framework to ensure high-speed communication for delivering ready-to-use media-optimized applications in SDN networks [17]. Such features are deemed important to enhance QoS provisioning to the users and applications as well.

to reconfigure SFs implementations. This article provides an approach to track packets when processed by SFs that modifies the packet header information. The approach relies on correlating packets before processing by an SF instance and afterward, which does not require modifications or even detailed knowledge of the SF instance. However, the approach needs the system to collect packets for the correlation analysis. In addition, the approach is rather complex as it needs packet matching with high accuracy to perform the classification function.

SDN and NFV-Based SFC Approaches

The MIDAS architecture is proposed in [11] to solve the problems of simultaneously detecting middle-boxes and selecting among multiple network function (NF) providers. MIDAS is based on a central controller per each NF provider to support coordination of traffic steering installation among all NF providers. MIDAS has the capabilities of middle-box signaling, controller chaining, and multi-party computation (MPC), which support on-path installing setup. MPC is used for NF provider assignment. MPC is characterized by privacy conservation, so it is used for middle-box usages over the NF providers. The MIDAS architecture featured with multiple NF providers cooperates for consolidated middle-box (CoMBs) detection over the traffic path and CoMB selection while preserving confidential information. The proposed architecture is based on three units: the CoMBs; a logical centralized controller per each NF provider; and the network processing client (NPCL) that provides the client's network service requests (CNSR). The authors also proposed a heuristic selection algorithm called the Intra-Provider Middle-box selection algorithm for NF allocations to CoMBs in the right place with an objective of load balancing provisioning over the CoMBs. They analyzed the applicability of MIDAS using the implemented prototype by delay measurement afforded during flow installing setup among all existing NF providers, middle-boxes, and CNSR arrival rates. The results showed that MPC does not have scalability problems as the MPC delay is not elongated by the CNSR arrival rate and the number of NF providers, which does not override the average internal path length. Simulation outcomes showed that utilizing MPC with the proposed middle-box selection algorithm shows good load balancing results and high request acceptance rates.

The ESCAPE prototype system, introduced in [12], is a developing and testing system for different nodes of the service function chaining framework. This model is applied to the UNIFY architecture. It is based on Mininet, Click, POX, and NetCONF tools integrated together in the ESCAPE framework. In addition, an orchestrator layer is added to allow SFC configuration, allocating VNFs into the physical resources, flow routing across the VNFs based on policies, and provisioning live management information on operating VNF instances. ESCAPE adopts VNF deployment by implementing a simple Mininet-based API where chain paths are constructed from available VNFs. These VNFs can be deployed and examined automatically. Moreover, a compact set of VNFs implemented in Click constructs a VNF catalog inside the ESCAPE system. The article presents a demo to show each unit of the architecture in a joint GUI. The demo includes:

- VNF containers and topology specification.
- Usage of a service graph to create chains.
- Service graph allocation to network resources.
- Traffic generation using standard tools.
- Monitoring the VNFs using Click.

The work in [13] introduced a new architecture that provides policy-based network management and has the ability to orchestrate and simplify fast deployment of various VNFs within an SDN/NFV environment. The architecture allows the selection of VNFs from available VNF instances using a policy engine staying in the NFV orchestrator. This NFVO provides various stitched VNFs, using them to build OSS/BSS applications. Moreover, the architecture addresses VNF life cycle management and service chaining among these different VNFs sent to large scale customers. The proposed architecture features:

- The ability to separate hardware elements, VNFs, services, and orchestration.
- Abstraction of network resources and network functions through predefined information models.
- Policy-based management allowance for singular VNFs and orchestration of NFV service chains.
- The ability to deploy NFV services ruled by policies.

The authors have deployed a prototype to provide an evaluation for their proposed architecture. They presented the use case of a telecom operator who instantiates VNFs on-desire for the management of network traffic outgoing from the content delivery network (CDN) caching nodes of CDN providers positioned inside the operator's sites. They implemented a policy-based traffic engineering service by supporting VNF deployment, virtual links assignment to the physical topology, flow monitoring, and orchestration.

The authors in [14] focus on SFC implementation in a cloud-based edge data center network where all SFs are software applications operating in virtual machines within these data centers. The main target of this work is to prove that this new softwarized environment permits a high level of flexibility and dynamicity of SFC in comparison with the traditional hardware-based architectures. To reach these flexible and dynamic SFCs for Layer 2 and Layer 3 edge network function implementations, the SDN control plane is used to provision the forwarding rules into OpenFlow switches. They also provided a proof-of-concept using Mininet emulation in order to evaluate their approach under a feasible scenario. The results showed that they can provide dynamic SFC and flexible traffic routing.

The work in [15] shows how telecom operators benefit from the NFV and SDN paradigms to improve the management of SFs and construct new business models. The article targets two major sides. The first side is how telco infrastructure deploys this new paradigm. The second issue is orchestration and management of SFs in distributed telco cloud environments by presenting the Cloud4NFV platform. The approach of modeling SFs in the cloud infrastructure is highlighted in that work, and the ability to perform SFC provisioning

SFC solutions	Architecture						Performance	
	SFC control plane				SFC data plane			
	Implementation	SFP adjustment	Orchestrator based	QoS/policy engine	SFF	SFC classifier	Flexibility	Scalability
NIMBLE [3]	SDN	Dynamic SFP with load balancing approach			Tags based	Packet matching	Low	Medium
FlowTags [4]	SDN	Static SFP		✓	Tags based	Packet matching	Low	Medium
SDN-based SFC [5, 6]	SDN	Static SFP			MAC address based	N/A	Medium	Medium
Context-aware SFC [7]	SDN	Static SFP		✓	MAC address based	Class-based forwarding	Medium	High
User-specific SFC [8]	SDN	Dynamic SFP	✓		xDPd	Packet matching	Low	Low
StEERING [9]	SDN	Dynamic SFP		✓	MAC address based	Packet matching	Medium	High
SIMPLE [10]	SDN	Dynamic SFP with load balancing approach		✓	MAC address based	Complex packet matching	Medium	Low
MIDAS [11]	SDN & NFV	Static SFP with load balancing approach			MAC address based	N/A	Medium	Low
ESCAPE [12]	SDN & NFV	Dynamic SFP	✓		MAC address based	Policy based	High	Medium
Policy-based SFC [13]	SDN & NFV	Static SFP	✓	✓	N/A	Policy based	High	High
Cloud-based SFC [14]	SDN & NFV	Dynamic SFP			MAC address based	Packet matching	Medium	Low
Cloud4NFV [15]	SDN & NFV	Static SFP	✓		N/A	Packet matching	High	High
Optical SFC [16]	SDN & NFV	Static SFP	✓		Optical circuit switches	N/A	High	High

Table 1. Taxonomy of the prior research work relevant to the SFC concept and its implementations.

is demonstrated as one of the essential features of SF composition. The Cloud4NFV platform is constructed over cloud, SDN, and WAN technologies to provide SF as a service. The Cloud4NFV platform also provides service monitoring and deployment, and optimized WAN and cloud resources for SFs support. A proof-of-concept is presented to evaluate some practical examples of the possible advantages of the proposed platform and the given standards in a telco environment.

In [16], an optical SFC architecture is proposed. The proposed architecture steers functionality into the datacenters for SFC using wavelength switching. The authors set up a packet/optical hybrid datacenter architecture to steer large volumes of flows in an optical steering network. They introduced such a solution to cope with the limitations of packet-switched SFC, such as complicated configuration of flow matching rules when the number of flows increases, which may lead to high operational cost, inefficient power consumption, and performance degradation due to scalability. The architecture consists of an operations support system/business support system (OSS/BSS) module, connected to an SDN controller and a NFV manager. The SFC configuration is done at the OSS/BSS module. Furthermore, the OSS/BSS module enforces the operator's policies. The SDN controller and NFV manager are responsible for resource allocation. The optical steering layer, including the network nodes, is placed on the southbound side of the SDN controller, which uses the OpenFlow v.1.4 protocol with an extension for optical circuit configuration to communicate with the optical circuit switches in the data plane layer. The proposed architecture shows its advantages, compared to packet-based routing, in terms of flexibility, scalability, reduced operational complexity, and energy efficiency.

Comparison and Evaluation

Table 1 shows some of the taxonomy used in the above mentioned approaches. The taxonomy shows that the approaches that adopt SDN and NFV technologies together alongside the orchestrator layer provide higher SFC scalability and flexibility than others. This comparison shows that most of the SFC approaches did not involve QoS and policy enforcement and neglect the load balancing functionality. Most of the frameworks use MAC address and OpenFlow functionality to apply traffic steering among the SFs without NSH support, as specified by the IETF SFC group. The usage of MAC address and/or OpenFlow protocols without NSH support has limited scalability and is more complex than using them with

There are two standards for SFC: one by IETF SFC WG and one by ONF. These standards impose the requirements that should exist in each SFC architecture, design, or implementation. These requirements are used to define the gaps and limitations in the previous SFC-related research work.

NSH support. There are some approaches that use tags instead of NSH. The work presented in [16] defines a solution to this limitation that applies optical steering of the data plane, using optical circuit switching devices that enhance the scalability and flexibility in the SFC domain network.

Challenges and Limitations

This section highlights the common limitations in the previous work and summarizes the open challenges relevant to the SFC concept and architecture. NSH capability in switches is one important challenge. Indeed, there is a lack of NSH-supported switches. As a countermeasure, some previous research works consider instead the use of tags or MAC addresses to steer traffic among the SFs. The trend is also toward supporting NSH in virtualized switches such as Open vSwitch (OvS). SFs do not have NSH capability either. Consequently, an SFC proxy must be used to encapsulate and de-encapsulate the packets travelling to and from SFs. However, the SFC proxy process may impact network performance, which can be alleviated only by equipping SFs with NSH support.

Traffic-engineered (TE) SFC is needed to provide an optimized SFC network with short computational latency. The literature provides limited concepts of traffic engineering in SFC. Some research works provide QoS-aware SFC paths to meet user and application requirements; other research works aim at maximizing the available data rate on the network links or cost savings. TE SFC needs to support all these features. This will be possible only by improving the SFC architecture through a well synchronized monitoring system to collect the required information from the network, QoS probes to test the reliability and performance of the existing SFPs, and a TE system that has the ability to instantiate TE-SFPs when needed. In terms of programmability, an efficient scheme is required to provide the optimized TE-SFP that satisfies the QoS requirements and network performance requirements.

The placement of SFs is a challenge and not sufficiently investigated in the literature. Furthermore, to the best knowledge of the authors, it was never investigated in the case of a network bottleneck scenario. There are two options in this scenario. The first option is to migrate the SF instance to a new location in the network; the second option is to instantiate a new SF instance. The choice between the two options adds a new challenge and must be investigated. The best location for the migrated or new instantiated SF must also be investigated, and novel optimal placement schemes must be proposed. There is also no previous work in the literature that discusses the use of SFC under different SLAs to support different classes of service.

Management of resource utilization is also required in the SFC framework to ensure high-speed communication to deliver ready-to-use media-optimized applications in SDN networks [17]. Such features are deemed important to enhance QoS provisioning to users and applications as well.

Conclusion

The delivery of end-to-end service requires various service functions to be provisioned in a SFC. This article introduces a survey of all existing SFC architectures, and conceptual approaches that are based on SDN and NFV. Research works are presented, compared, and evaluated. There are two standards for SFC: one by IETF SFC WG and one by ONF. These standards impose the requirements that should exist in each SFC architecture, design, or implementation. These requirements are used to define the gaps and limitations in the previous SFC-related research work.

Finally, the open challenges are discussed. Policy-based SFC, Cloud4NFV, and Optical SFC architectures exhibit high performance in terms of SFC orchestration, scalability, and flexibility to provide SFC in cloud environments, making use of SDN and NFV technologies.

Acknowledgment

This work has been funded with the support of the European Commission. This article reflects the view of the authors only. The European Commission cannot be held responsible for any use that may be made of the information contained therein.

References

[1] W. John *et al.*, "Research Directions in Network Service Chaining," *IEEE SDN for Future Networks and Services (SDN-4FNS)*, 2013, pp. 1–7.
[2] T. Taleb *et al.*, "Coping with Emerging Mobile Social Media Applications Through Dynamic Service Function Chaining," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, 2016, pp. 2859–71.
[3] Z. Qazi *et al.*, "Practical and Incremental Convergence between SDN and Middleboxes," *Open Network Summit*, Santa Clara, CA, 2013.
[4] S. Fayazbakhsh *et al.*, "FlowTags: Enforcing Network-wide Policies in the Presence of Dynamic Middlebox Actions," *Proc. 2nd ACM SIGCOMM Wksp. Hot Topics in Software Defined Networking*, 2013, pp. 19–24.
[5] J. Blendin *et al.*, "Position Paper: Software-Defined Network Service Chaining," *3rd European Wksp. Software Defined Networks*, 2014, pp. 109–14.
[6] J. Blendin *et al.*, "Demo: Software-Defined Network Service Chaining," *3rd European Wksp. Software Defined Networks*, 2014, pp. 139–40.
[7] B. Martini *et al.*, "SDN Controller for Context-Aware Data Delivery in Dynamic Service Chaining," *1st IEEE Conf. Network Softwarization (NetSoft)*, 2015, pp. 1–5.
[8] I. Cerrato *et al.*, "User-Specific Network Service Functions in an SDN-enabled Network Node," *3rd European Wksp. Software Defined Networks*, 2014, pp. 135–36.
[9] Y. Zhang *et al.*, "Steering: A Software-defined Networking for Inline Service Chaining," *21st IEEE Int'l. Conf. Network Protocols (ICNP)*, 2013, pp. 1–10.
[10] Z. A. Qazi *et al.*, "SIMPLE-fying Middlebox Policy Enforcement Using SDN," *ACM SIGCOMM Comp. Commun. Rev.*, vol. 43, no. 4, 2013, pp. 27–38.
[11] A. Abujoda and P. Papadimitriou, "MIDAS: Middlebox Discovery and Selection for On-path Flow Processing," *7th Int'l. Conf. Communication Systems and Networks (COMSNETS)*, 2015, pp. 1–8.
[12] A. Csoma *et al.*, "ESCAPE: Extensible Service Chain Prototyping Environment using Mininet, Click, Netconf and POX," *ACM SIGCOMM Computer Commu. Rev.*, vol. 44, no. 4, 2015, pp. 125–26.
[13] K. Giotis, Y. Kryftis, and V. Maglaris, "Policy-based Orchestration of NFV Services in Software-defined Networks," *1st IEEE Conf. Network Softwarization (NetSoft)*, 2015, pp. 1–5.
[14] F. Callegati *et al.*, "Dynamic Chaining of Virtual Network Functions in Cloud-based Edge Networks," *1st IEEE Conf. Network Softwarization (NetSoft)*, 2015, pp. 1–5.
[15] J. Soares *et al.*, "Toward a Telco Cloud Environment for Service Functions," *IEEE Commun. Mag.*, vol. 53, no. 2, 2015, pp. 98–106.
[16] M. Xia *et al.*, "Optical Service Chaining for Network Function Virtualization," *IEEE Commun. Mag*, vol. 53, no. 4, 2015, pp. 152–58.
[17] F. Pop *et al.*, "Adaptive Scheduling Algorithm for Media-optimized Traffic Management in Software Defined Networks," *Computing*, vol. 98, no. 1–2, 2016, pp. 147–68.

Biographies

Ahmed M. Medhat (a.hassan@campus.tu-berlin.de) is currently a Ph.D. candidate and a research assistant on the faculty of electrical engineering and computer sciences, Technical University of Berlin, Germany. He received his B. Sc. degree in information engineering and technology, and his M.Sc. degree in communication engineering from German University in Cairo (GUC), Egypt in 2010 and 2011, respectively. His research interests are in the field of next generation network infrastructures (related topics to SDN and NFV) with a focus on service function chaining and its quality of service enhancement solutions.

Tarik Taleb [M] (talebtarik@ieee.org, tarik.taleb@aalto.fi) is currently a professor at the School of Electrical Engineering, Aalto University, Finland. He has worked as a senior researcher and 3GPP standards expert at NEC Europe Ltd. Prior to his work at NEC, until March 2009, he worked as an assistant professor at the Graduate School of Information Sciences, Tohoku University, Japan, in a lab fully funded by KDDI. He received his B.E. degree in information engineering with distinction, and his M.Sc. and Ph.D. degrees in information sciences from Tohoku University in 2001, 2003, and 2005, respectively. His research interests lie in the field of architectural enhancements to mobile core networks (particularly 3GPP's), mobile cloud networking, mobile multimedia streaming, and social media networking. He has also been directly engaged in the development and standardization of the Evolved Packet System. He is a member of the IEEE Communications Society Standardization Program Development Board and serves as Steering Committee Chair of the IEEE Conference on Standards for Communications and Networking. He has received many awards for his many contributions in the area of mobile networking.

Asma Elmangoush (asma.elmangoush@alumni.tu-berlin.de, asma.a.elmangoush@campus.tu-berlin.de) received her B.E and MS.c. in computer engineering from the College of Industrial Technologies-Misurata, Libya, and received her Ph.D. degree from the Technical University Berlin in 2016. She is currently a lecturer at the College of Industrial Technologies-Misurata, Libya.

Giuseppe A. Carella (giuseppe.a.carella@tu-berlin.de) is a senior researcher at the Fraunhofer FOKUS and at the Technische Universität Berlin (TUB). He received his M.Sc. in engineering of computer science from the Alma Mater Studiorum University of Bologna in 2011. During his studies he focused on next generation network infrastructure, especially in IMS services, such as presence and messaging. In 2012 he joined the Next Generation Networks (AV) team at the Technical University Berlin, where he started investigating topics related to SDN and NFV in the context of his Ph.D. studies. His strong background in cloud computing is the basis of his research and contributed to the virtualization of the software-based network functions developed at Fraunhofer FOKUS, namely OpenEPC and Open5GCore. He is currently leading the team developing the Open Baton toolkit, an open source platform providing the means for building a comprehensive NFV environment.

Stefan Covaci (stefan.covaci@tu-berlin.de) is a senior solutions architect for future Internet services platforms on the computer sciences and electrical engineering faculty of Technical University of Berlin, Institute for Telecommunication Systems. Between 1990 and 2007 he was working at GMD FOKUS (today Fraunhofer Institute FOKUS) acting as director of the competence center Intelligent Mobile Agents. He co-initiated and participated in the agent standardization work of OMG (Object Management Group) and FIPA (Foundation of Intelligent Physical Agents), and was one of the co-authors of the first agent technology standard, the OMG-MASIF (Mobile Agent System Interoperability Facility). He has extensive experience in the management of a variety of projects and study contracts for the European Commission, German Agencies (BMBF, BMWF), and other industrial national and international organizations in the areas of IT and telecommunication networks and services. His recent work is in the areas of next generation network infrastructures and service delivery platforms for the Future Internet, with a focus on interoperability and management solutions. He is member of the Architectural Board of the European Future Internet Public-Private Partnership (FI-PPP), and technical coordinator of the FI-STAR project, applying Future Internet technology in the healthcare sector. He is the technical coordinator of the DAAD–UNIFI project enabling future Internet academic teaching and research in developing countries. He has published more than 90 papers and was a chair or member of the program committee of many international conferences.

Thomas Magedanz [SM] (magedanz@ieee.org, Thomas.magedanz@fokus.fraunhofer.de) is a full professor on the electrical engineering and computer sciences faculty at the Technical University of Berlin, Germany, leading the chair for Next Generation Networks. In addition, he is the director of the "Software-based Networks" division of the Fraunhofer Institute FOKUS. In 2006, he was named an Extraordinary Professor in the Department of Electrical Engineering of the University of Cape Town, South Africa. Since 2007, he has also been a visiting professor in the Department of Mathematics, Physics and Computing at the Waterford Institute of Technology in Ireland. For more than 20 years he has been working in the convergence field of fixed and mobile telecommunications, the Internet, and information technologies, which resulted in many international R&D projects centered around next generation service delivery platforms prototyped in a set of globally recognized open technology testbeds. In 2007 he joined the European FIRE (Future Internet Research and Experimentation) Expert Group. In the course of his research activities he has published more than 250 technical papers/articles. In addition, he is a senior member of the IEEE, and serves on the editorial board of several journals. He received his diploma and his Ph.D. in computer sciences from the Technical University of Berlin, Germany, in 1988 and 1993, respectively. In 2000 he finished his postdoctoral lecture qualification in applied computer sciences at the Technical University of Berlin, Germany.

ADVERTISING SALES OFFICES

Closing date for space reservation: 15th of the month prior to date of issue

NATIONAL SALES OFFICE
Mark David
Sr. Manager Advertising & Business Development
EMAIL: m.david@ieee.org

NORTHERN CALIFORNIA
George Roman
TEL: (702) 515-7247
FAX: (702) 515-7248
EMAIL: George@George.RomanMedia.com

SOUTHERN CALIFORNIA
Marshall Rubin
TEL: (818) 888 2407
FAX: (818) 888-4907
EMAIL: mr.ieeemedia@ieee.org

MID-ATLANTIC
Dawn Becker
TEL: (732) 772-0160
FAX: (732) 772-0164
EMAIL: db.ieeemedia@ieee.org

NORTHEAST
Merrie Lynch
TEL: (617) 357-8190
FAX: (617) 357-8194
EMAIL: Merrie.Lynch@celassociates2.com

Jody Estabrook
TEL: (77) 283-4528
FAX:: (774) 283-4527
EMAIL: je.ieeemedia@ieee.org

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
Christian Hoelscher
TEL: +49 (0) 89 95002778
FAX: +49 (0) 89 95002779
EMAIL: Christian.Hoelscher@husonmedia.com

TOPICS PLANNED FOR THE MARCH ISSUE

ENABLING MOBILE AND WIRELESS TECHNOLOGIES FOR SMART CITIES

SUSTAINABLE INCENTIVE MECHANISMS FOR MOBILE CROWDSENSING

INTERNET OF THINGS

NETWORK TESTING AND ANALYTICS

RADIO COMMUNICATIONS

IEEE GLOBECOM™

IEEE ComSoc™ IEEE Communications Society

IEEE Global Communications Conference

4-8 December 2017 // Singapore • Global Hub: Connecting East and West

CALL FOR PAPERS

The 2017 IEEE Global Communications Conference (GLOBECOM) will feature a comprehensive technical program including 13 symposia, tutorials, workshops and an industrial program featuring prominent keynote speakers, technology and industry forums and vendor exhibits.

TECHNICAL SYMPOSIA

- Ad Hoc and Sensor Networks
- Cognitive Radio and Networks
- Communication and Information System Security
- Communication QoS, Reliability and Modeling
- Communication Software, Services and Multimedia Applications
- Communication Theory
- Green Communications Systems and Networks
- Mobile and Wireless Networks
- Next-Generation Networking and Internet
- Optical Networks and Systems
- Signal Processing for Communications
- Wireless Communication
- Selected Areas in Communications
  - Access Networks and Systems
  - Big Data
  - Data Storage
  - e-Health
  - Internet of Things
  - Molecular, Biological, and Multi-scale Communication
  - Power Line Communications
  - Satellite and Space Communications
  - Smart Grid Communications
  - Social Networks

Please address questions regarding the Technical Symposia to Technical Program Committee (TPC) Chair: Ying-Chang Liang (liangyc@ieee.org), and TPC Co-Chairs: Teng Joon Lim (eleltj@nus.edu.sg) and Chengshan Xiao (xiaoc@mst.edu). Accepted and presented technical and workshop papers will be published in the IEEE GLOBECOM 2017 Conference Proceedings and submitted to IEEE Xplore®. See the website for author requirements of accepted authors. Full details of submission procedures are available at www.ieee-globecom.org.

TUTORIALS

Proposals are invited for half-day tutorials in communications & networking. Please address questions regarding tutorials to Tutorial Chair: **Rui Zhang (elezhang@nus.edu.sg)**.

WORKSHOPS

Submissions are sought for workshops on the latest technical and business issues in communications and networking. Please address questions on workshops to Workshop Chair: **Tony Quek (tonyquek@sutd.edu.sg)**.

IMPORTANT DATES

Symposia Papers	Workshop Proposals	Tutorial Proposals
1 April 2017	**15 February 2017**	**15 March 2017**

ORGANIZING COMMITTEE

General Chair
Dim-Lee Kwong (I2R, Singapore)

General Vice Chairs
Shiang Long Lee
(Singapore Technologies, Singapore)
Pak Lum Mock (Starhub, Singapore)

Executive Chair
Lawrence Wong (NUS, Singapore)

Executive Vice Chairs
Ying-Chang Liang
(UESTC, China & I2R, Singapore)
Sumei Sun (I2R, Singapore)

TPC Chair
Ying-Chang Liang
(UESTC, China & I2R, Singapore)

TPC Co-Chairs
Teng Joon Lim (NUS, Singapore)
Chengshan Xiao (Missouri S&T, USA)

Tutorial Chair
Rui Zhang (NUS, Singapore)

Tutorial Co-Chairs
Lingyang Song (PKU, China)
Stefano Bregni (Politecnico Milano, Italy)

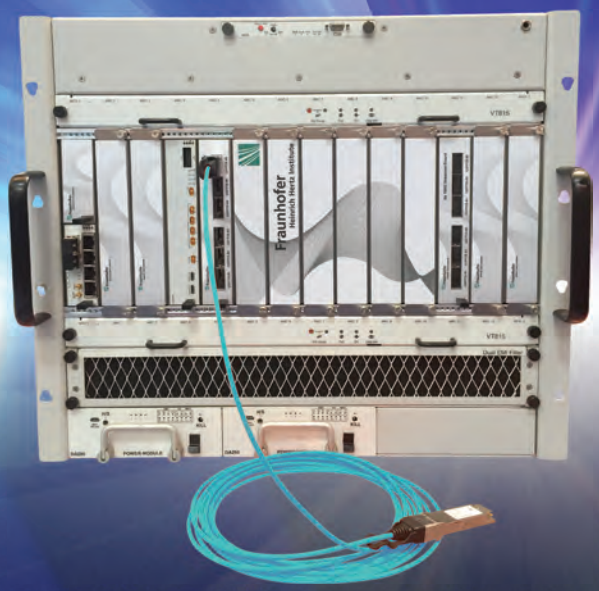
Workshop Chair
Tony Quek (SUTD, Singapore)

Workshop Co-Chairs
Wei Zhang (UNSW, Australia)
Gang Wu (UESTC, China)

For more information about IEEE GLOBECOM 2017, visit www.ieee-globecom.org.

Fraunhofer
Heinrich Hertz Institute

HIGH-SPEED DSP PLATFORM



Flexible Platform for 100G Real-time Digital Signal Processing

Key Features

- Multi-100G real-time data access
- Compact and flexible solution, based on Micro-TCA chassis and plug-in boards
- 4-channel 56-GSa/s, 8-bit ADCs
- 4-channel 65-GSa/s, 8-bit DACs
- High-performance Virtex Ultrascale FPGAs
- Multi-100G QSFP28 interface to external hardware
- Multi-purpose, built-in Gigabit Ethernet (GbE) connection
- Ready-to-use synchronization, interface and control IP cores included

Applications

- Test of digital signal processing algorithms in real-time
- Online transmission performance evaluation of optical transmission systems
- Realization of FPGA-based real-time transceivers

E-Mail: products-pn@hhi.fraunhofer.de
Web: www.hhi.fraunhofer.de