• Wireless Technologies for Development
• Network and Service Management
• Ad Hoc and Sensor Networking
• LTE-Advanced Pro

If your 5G idea works here...

50,000 soccer fans streaming live-action mobile video. All at once.


HARDWARE + SOFTWARE + PEOPLE = 5G INSIGHTS
In the race to design next-generation wireless technologies, research teams must rely on platforms and tools that accelerate their productivity. Using the NI software defined radio platform and LabVIEW Communications, leading researchers are innovating faster and building 5G proof-of-concept systems to demonstrate new technologies first.

Accelerate your innovation at ni.com/5g
IEEE TECHNOLOGY INITIATIVES AND RELATED COMSOC STANDARDS ACTIVITIES

In the three previous issues of this magazine, we described the three new technology areas in which the Communications Society (ComSoc) is playing a leading role. This President’s message illustrates ComSoc’s standards activities in relation to these emerging technology initiatives. To address the standards’ aspects of the initiatives, we invited Dr. Alexander Gelman, ComSoc’s Director of Standardization Programs Development, to write this article.

Alex is a veteran industrial researcher and a seasoned ComSoc volunteer. He holds Masters and Ph.D. degrees from the City University of New York. Currently he is the CTO of the NETovations Consulting Group. From 1998 to 2007 he was Chief Scientist of the Panasonic Princeton Laboratory, leading research in various areas of consumer communications and networking. And from 1984 to 1998 he held various positions at Bellcore, the last one as Director of Internet Access Architectures Research. At Bellcore, Alex pioneered multimedia bridging, Video on Demand, and XDSL Internet Access. Alex initiated ComSoc’s Standards Board, several standardization projects, and architected ComSoc’s Standards Activities Council. He also served on the IEEE Standards Association Board of Governors as well as several terms on the IEEE-SA Standards Board and its committees.

ComSoc has a unique approach to standardization of emerging technologies. This approach is explained though four phases of the technology evolution: conceptual, pre-competitive, competitive, and deployment, as depicted in Fig. 1. The conceptual phase is the phase when only a concept of a particular technology exists. The pre-competitive phase assumes that the technology is being researched, but there are still no business models associated with it. The competitive phase assumes that viable business models have been developed and companies are engaged in R&D activities at full-speed. Finally, these efforts result in deployment.

It is important to note that, as illustrated in Fig. 1, the strength of intellectual property created by industrial and academic researchers (patents, copyrights, etc.) related to the evolving technology is reduced with time as more (e.g., fundamental) patent applications are filed. Meanwhile the relevance (e.g., to standards) of patents filed is growing in time as it becomes more clear which solutions are making it to market. These two intellectual property parameters, the strength and the relevance, determine the degree of material interest in technology on the part of the industry and the degree of professional (or professed, as we call it in ComSoc Standards Activities policy documents) interest on the part of industrial and academic researchers.

ComSoc Technical and Standards Activities serve our membership, the industry, and humanity throughout all phases of the technology evolution by providing a platform for technical discussions and global information dissemination. For technologies within the technical scope of ComSoc we enable the innovation process to happen. ComSoc products and services positioned on Fig. 1 in relation to these phases range from workshops in early technology evolution stages to conferences, publications, and finally to standards.

Our early workshops discuss technologies in the conceptual phase, and our conferences feature papers that often belong to both the conceptual and the pre-competitive phases. Following this are ComSoc’s publications. Standards activities historically start at the latter part of the competitive R&D phase. The reduced professed interest in the later phases of technology evolution, as indicated above, to a significant degree explains the lack of participation on the part of researchers in standards activities. ComSoc’s objective is to change this situation.

The typical standards activities’ modus operandi has been to evaluate standards project proposals as they are submitted by industry and then to accept or reject them. ComSoc’s Standards Activities Council pioneered a novel approach to proactively discover standardization opportunities. Instrumental in this process are IEEE initiatives that offer a platform and the resources necessary to set up an IEEE infrastructure for discussion of emerging technologies. From this we can identify needed publications and conferences as well as standardization opportunities.

ComSoc’s innovative approach to standardization for emerging technologies is based on pre-standardization activities that allow not only discovery of standardization opportunities, but also setting up research projects that help to resolve unsolved issues that stand in the way of clearly defining standardization projects. The key to our approach is to move the standardization activities into the area of the conceptual technology phase as far as possible (Fig. 1). However, to achieve this it is critical to engage academic and industrial researchers.

ComSoc’s pre-standardization process is kicked off by a Rapid Reaction Standardization Activity (RRSA) event in the form of an expert brainstorming session. This is preceded by a formulation of position statements on the global standardization situation in the particular emerging technology or family of technologies. Specific standardization project proposals are also solicited from potential participants. From this we generate a strategy going forward for ComSoc’s standards activities role in the standardization of the technology.

RRSA participants are typically recommended by ComSoc’s relevant Technical Committees and selected based on their position statements. The outcomes of the RRSA sessions differ depending on the degree of maturity of the proposed standardization projects. The proposal evaluation process is illustrated in Fig. 2. If the attendees come to a
consensus that a particular proposal is mature enough to define a standardization project, then a group of supporters is chartered with the task of generating an IEEE Project Authorization Request (PAR) that is to be submitted to the ComSoc Standards Development Board. If more time is needed to figure out the project scope, a Study Group is proposed to be formed under the auspices of the ComSoc Standards Development Board. If any open issues are identified, a research group is formed under the umbrella of the ComSoc Standardization Programs Development Board.

This proactive approach to the standardization of emerging technologies has produced significant results. ComSoc has enriched its standards portfolio with exciting projects in emerging technology areas that also include scholarly standards produced by industrial and academic researchers.

**IoT Initiative**

The IoT Initiative RRSA session initiated the following Projects.
- Research Group on IoT Architectures. Chair: JaeSeung Song, Sejong University, South Korea.
- Research Group on IoT Services. Chair: Yacine Ghamri-Doudane, University of La Rochelle, France.
- Research Group on IOT Communications and Networking Infrastructure. Chair: Stefano Giordano, University of Pisa, Italy.

Research groups are chartered to resolve issues that arise on the way to standardization. While working on this, they produce outputs in the form of white papers as well as various types of content for ComSoc products. These include conference sessions and panels, feature topics for the Communications Standards Supplement and for the new Communications Standards Magazine, which is scheduled to launch in 2017.

The ComSoc Standards Development Board initiated the

---

**Figure 1. ComSoc’s Standards Activities approach.**

**Figure 2. Rapid Reaction Standardization Activities methodology.**

Study Group on IOT API and Interfaces with Ming-Jye Sheng of Princeton University as the chair.

Presently the ComSoc Standards Development Board sponsors two standardization projects:
In the May issue of this magazine, a ComSoc 5G perspective was addressed and the global standardization map was mentioned. IEEE is the best venue for the standardization of core technologies, and in particular, 5G technology with its much broader scope than previous wireless generations. IEEE will play a significant role in 5G standardization.

The ComSoc RRSA has produced results after its very first session. Two standardization projects were proposed, and after passing the ComSoc and IEEE-SA approval phases, are now in operation. These are:


In addition, a Research Group on Channel Modelling was formed under the leadership of Kevin Lu.

**FOG COMPUTING**

While this initiative is in its infancy, ComSoc has taken some quite aggressive positioning steps. IEEE has joined the Open Fog Consortium, and two ComSoc lead volunteers, Doug Zuckerman and Russel Hsing, are on its Board of Directors representing IEEE. Another prominent ComSoc volunteer, Tao Zhang, ComSoc’s CIO, is on the Board as well as representing Cisco. ComSoc’s Standards Activities team’s intention is to help the consortium in the standardization area.

ComSoc’s Strategic Planning Committee and the Emerging Technology Committee help identify future technologies and opportunities. “Network Softwarization” and “Smart Cities” are two that are looming on the horizon. It is our intent to apply our RRSA methodology to identify standardization roles for ComSoc in these and other areas as they arise.

---

**ComSoc 2016 Election**

**Take Time to Vote**

Ballots were e-mailed and/or postal mailed 27 May 2016 to all ComSoc members (excluding Student Members, Associate Members, and Affiliates) whose memberships were effective prior to 1 May 2016. You must have an e-ballot or paper ballot before you can vote.

VOTE NOW using the URL below. You will need your IEEE account user name/password to access the ballot. If you do not remember your password, you may retrieve it on the voter login page.


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

If you do not receive a ballot by 30 June, but you feel your membership was valid before 1 May 2016, you may e-mail ieee-comsocvote@ieee.org or call +1 732 562 3904 to check your member status. (Provide your member number, full name, and address.)

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

Voting for this election closes 22 July 2016 at 4:00 p.m. EDT! Please vote!
A new procedure for contesting the validity of patents was instituted under the American Invents Act (AIA) of 2011. In the U.S. Patent and Trademark Office, the Patent Trial and Appeal Board can conduct a proceeding for the review on the validity of a patent instituted by a party other than the patent owner. The proceeding is called an inter partes review proceeding. While it may be considered a review, nevertheless it is a trial before the Patent Trial and Appeal Board and is litigation focused on the patent validity issues.

The Patent Trial and Appeal Board proceeding is an administrative hearing, and there is no jury trial as might be found in a U.S. District Court. In comparison, a trial in U.S. District Court would include all issues, not only the validity or invalidity of the patent, but also infringement and damages. A jury often is used as the finder of facts.

Timing is very important from the filing of a lawsuit in the Patent Trial and Appeal Board. The Patent Trial and Appeal Board has six months to institute the trial and inter partes review proceeding, and 18 months (18 months in exceptional circumstances) to conduct the trial. This is a possibility of up to two years before completion of a case before the Patent Trial and Appeal Board. A decision from the Patent Trial and Appeal Board can be appealed to the Court of Appeals of the Federal Circuit. This appeal could take another two years.

Now what happens if a patent infringement lawsuit is filed in U.S. District Court? Assume the plaintiff of the lawsuit is the patent owner. The business that makes the product is the defendant. The Patent Trial and Appeal Board may resolve the matter to two years before completion of a case before the Patent Trial and Appeal Board. A decision from the Patent Trial and Appeal Board can be appealed to the Court of Appeals of the Federal Circuit. This appeal could take another two years.

As a result, the patent troll’s patent may become a zombie patent. These lawsuits typically are on a contingency fee basis. The lawyers for the patent trolls do not get paid until and unless they recover money from the businesses with products that infringe the patents. There are many defenses to a patent infringement lawsuit. Some of them are:

1. I do not infringe your patent.
2. Your patented invention is not patentable because it was not patentable subject matter, for example, a method of doing business [2].
3. Your patented invention is not patentable because it is anticipated by the prior art [3].
4. Your invention is not patentable because it is obvious in view of the prior art [4].

There are many other defenses. But only the defenses concerning patentability can be raised before the Patent Trial and Appeal Board in an inter partes review proceeding. Other defenses, such as I do not infringe your patent, or even if I did infringe your patent, I did not cause any damages, must be raised in the lawsuit in the U.S. District Court. What we are talking about here is jurisdiction, and the Patent Trial and Appeal Board has jurisdiction only to hear issues on patentability that are normally brought before the U.S. Patent and Trademark Office during the process of obtaining a patent. The U.S. District Court has general jurisdiction over all patent matters, including patentability, infringement, damages, and other issues and defenses.

By instituting an inter partes review proceeding before the Patent Trial and Appeal Board, the accused infringer of the patent may gain an advantage over the patent troll on several grounds. As a result, the patent troll’s patent may become a zombie patent. If the patent troll fails at the Patent Trial and Appeal Board, their entire enforcement strategy in the U.S. District Court fails. A patent declared invalid by the Patent Trial and Appeal Board, and held invalid on appeal at the Court of Appeals for the Federal Circuit, ends the trial in the U.S. District Court. The defendant business has won its case, and the case is over.

The delay of two to four years in the Patent Trial and Appeal Board and possible appeal to the Court of Appeals for the Federal Circuit is costly in time and money to the patent troll. All this costs money to the contingency fee lawyers, and the proceedings can be very expensive. A patent troll may not have budgeted for a proceeding before the Patent Trial and Appeal Board, which is typically done on a fee for service basis, not a contingency fee basis. To engage good, competent counsel to represent them before the Patent Trial and Appeal Board in the inter partes review proceeding usually is very expensive.

There is a possibility, or a high likelihood, that the Court of Appeals for the Federal Circuit will confirm decisions of the Patent Trial and Appeal Board. If the finding at the Patent Trial and Appeal Board did not end the lawsuit in the U.S. District Court, the holding on appeal at the Court of Appeals for the Federal Circuit likely will end the patent troll’s lawsuit. Thus, patents belonging to patent trolls can become zombie patents through an inter partes review proceeding before the Patent Trial and Appeal Board.

There is a possibility, or a high likelihood, that the Court of Appeals for the Federal Circuit will confirm decisions of the Patent Trial and Appeal Board. If the finding at the Patent Trial and Appeal Board did not end the lawsuit in the U.S. District Court, the holding on appeal at the Court of Appeals for the Federal Circuit likely will end the patent troll’s lawsuit. Thus, patents belonging to patent trolls can become zombie patents through an inter partes review proceeding before the Patent Trial and Appeal Board.

REFERENCES


BIOGRAPHY

David B. Newman, Jr. (SM ’02) (dnewman@lig.com) received his Ph.D. in electrical engineering from the Pennsylvania State University and his J.D. from American University. He served as Feature Editor for Communications and the Law and for IEEE Communications Magazine in the 1980s. He is a former associate professor of electrical engineering and professional lecturer of law at George Washington University, and previously an engineer with the CIA. He is a patent attorney in the Washington, DC metropolitan area.
A TRIBUTE TO MARVIN KENNETH SIMON
BY M.-S. ALOUNI, E. BIGLIERI, D. DIVSALAR, S. DOLINAR, A. GOLDSMITH, AND L. MILSTEIN

MARV’S LIFE AND WORK

It is a measure of the importance and profundity of Marvin Kenneth Simon’s contributions to communication theory that a tribute to his life and work is of current research relevance in spite of the continually accelerating rate of evolution in this area. Marv, as the entire community affectionately knew him, was one of the most prolific and influential communications researchers of his generation. Moreover, he laid the foundation for many of the techniques used in communication systems today. Marv’s tragic death on September 23, 2007 continues to engender pangs not only of sadness at the passing of a great friend to many in our community, but also of regret that he is no longer with us to help resolve the many challenges facing communication systems today.

Marv was born on September 10, 1939, in New York City, NY. He attended Bronx High School of Science, and received the B.S.E.E. degree from the City College of New York in 1960, the M.S.E.E. from Princeton University in 1961, and the Ph.D. degree from New York University in 1966. From 1963 to 1966 he was an instructor in Electrical Engineering at New York University, teaching graduate courses in circuit theory. From 1961 to 1963 and 1966 to 1968, he was a Member of the Technical Staff at AT&T Bell Laboratories, where he conducted theoretical studies of digital communication systems with a focus on pulse-coded modulation. In 1968 Marv joined the Jet Propulsion Laboratory (JPL), a division of the California Institute of Technology (Caltech) in Pasadena, CA, where he spent the rest of his career, ultimately as a Principal Scientist. In 1978, concurrent with his JPL appointment, Marv began an affiliation with the Electrical Engineering Department at Caltech as a visiting lecturer, which ended in 1996. In this role he created and taught the department’s three-quarter sequence of first-year graduate courses on random processes and digital communication. In addition to teaching this sequence a total of ten times over several decades, Marv also mentored many of his Caltech students in their research and careers.

Over four decades starting in 1966, Marv performed seminal research in a wide variety of areas within digital communications and left a deep and broad legacy in each area. For each of these areas, before moving on to the next one, Marv authored or co-authored an influential textbook on the subject, including the fruits of his own research. Several of these went on to become the definitive book for students and researchers working in that area. Marv’s specific contributions to synchronization, digital modulation, spread spectrum, trellis coding, differential detection, fading channel performance evaluation, and software-defined radios (SDRs) are deep and fundamental. They have had, and will continue to have, a lasting impact on space, terrestrial, satellite, and mobile communication system analysis and design. These technical contributions are described in more detail in a tutorial on Marv’s work that appears in IEEE Communications Surveys and Tutorials.

Marv’s broad and deep contributions to the field of digital communications and leadership in advancing the discipline garnered him much recognition and many awards from both the IEEE and from NASA. He was elevated to IEEE Fellow in 1978, and was awarded the IEEE Centennial Medal in 1984 and the IEEE Third Millennium Medal in 2000. In 1997 Marv received the IEEE Communications Society Edwin H. Armstrong Achievement Award, the society’s highest honor, recognizing his seminal contributions spanning three decades in the design and analysis of novel coherent digital communication systems. For his outstanding contributions to their deep space and near-earth missions, NASA awarded him its 1979 Exceptional Service Medal and its 1995 Exceptional Engineering Achievement Medal. Marv was also the co-recipient of the IEEE 1999 Vehicular Technology Conference Paper Award for his work on trellis coded differential detection systems, and the co-recipient of the 1990 IEEE Transactions on Vehicular Technology Paper Award for his work on trellis coded differential detection systems, and his paper co-authored with Dariush Divsalar, entitled “Multiple Symbol Differential Detection of Multiple Phase-Shift-Keying,” which appeared in the March 1990 issue of the IEEE Transactions on Communications, was selected for inclusion in a book celebrating the 50th anniversary of the founding of the IEEE Communications Society. This book, entitled The Best of the Best: Fifty Years of Communications and Networking Research, contains the 50 research papers that, during the previous five decades, were the most significant and frequently cited in the evolution of communications systems design and analysis over that period. Marv was also the co-recipient of the 1990 IEEE Transactions on Vehicular Technology Paper Award for his work on trellis coded differential detection systems, and the co-recipient of the 1999 IEEE Vehicular Technology Conference Paper Award for his work on the performance analysis of dual-branch switched diversity systems over fading channels.

A prolific writer of technical books, Marv authored or co-authored seven: Telecommunications Systems Engineering (with Lindsey; Prentice-Hall, 1973, and Dover Press, 1991); Spread Spectrum Communications, Volumes I, II, and III (with Omura, Scholtz, and Levitt; Computer Science Press, 1984, and McGraw Hill, 1994, 2nd ed, 2002); Introduction to Trellis Coded Modulation with Applications (with Biglieri, Divsalar, and McLane; MacMillan, 1991); Digital Communication Techniques: Signal Design and Detection (with Hindle and Lindsey; Prentice Hall, 1994); Digital Communication over
Marv's professional life are provided in the next section.

Marv’s works were highly cited, often long after publication. Indeed, one reviewer of Marv’s 1998 IEEE Proceedings paper on performance analysis of digital communications in generalized fading, jointly authored with Slim Alouini, stated, “This paper is a groundbreaking piece of work (that) will have a long and useful life.” This was a prophetic statement as Marv’s co-authored textbook on this technique is his most cited work, with over 6000 citations in Google scholar. Marv’s research as applied to the design of NASA’s deep space and near-earth missions resulted in the issuance of 12 U.S. patents, 29 NASA Tech Briefs, and four NASA Space Act Awards. In addition to over 200 published research papers, Marv also contributed to over 80 JPL technical publications. Marv’s overall body of work has been cited more than 24,000 times in Google Scholar, with an h-index of 57. Since 2010, five years after his passing, Marv’s work has been cited almost 10,000 times, a remarkable tribute to the longevity of his contributions and impact.

Marv was very active in the IEEE Communications Society (ComSoc), and a frequent and welcomed presence at its symposia and workshops. Perhaps his most beloved technical event was the IEEE Communication Theory Workshop. Without any published proceedings and with its relaxed format and prestigious participants, this workshop allowed for a free and creative exchange of ideas as well as in-depth analysis and debate. Marv was a constant fixture at the workshop, presenting thought-provoking talks, asking probing questions of other presenters, and taking part in the pervasive informal technical discussions throughout the day and sometimes late into the night.

In addition to being a leader in the field Marv was, for those who had the good fortune to interact with him, a researcher, professor, collaborator, and mentor. Memories of Marv bespeak a man who took intense interest both in the problems he wished to solve and the people with whom he worked. As a researcher, Marv was intrigued by interesting problems, and smart enough, creative enough, and persistent enough to solve them. As a teacher, Marv provided his students with the fundamental knowledge they needed to form the foundation for their own research. As a collaborator, Marv was an eternal graduate student, always hungry for new discoveries and new results and eager to investigate them deeply. Reflections on these many dimensions of Marv’s professional life are provided in the next section.

Marv’s intellect encompassed topics well beyond research. In his early years Marv was a talented accordion player and performed regularly on a TV variety show called The Horn and Hardart Children’s Hour, as well as in the Catskills with his band, “Marvin and the Mellowtones.” He eventually chose engineering over music for his career (much to the benefit of our research community), but music remained his avocation. He frequently shared his musical talents with his peers, entertaining his colleagues on the piano with remarkable finesse and talent at workshops, in their homes, or on the concert-sized grand piano that occupied most of his living room. In 2004 Marv accompanied Robert McEliece in his rendition of “Thank you very much” for one of the most memorable Claude E. Shannon Award Lectures ever presented. One of Marv’s other beloved hobbies was magic; he would perform tricks for colleagues and at his kids’ birthday parties under the moniker “Marvin the Magnificent.” Marv was also an award-winning photographer and a prolific author of non-technical books. In the 1980s he became one of the world’s experts on computer adventure games, writing several best-selling books on the topic, including Keys to Solving Computer Adventure Games and Hints, Maps, and Solutions to Computer Adventure Games.

Marv was also an extraordinary husband and father. He met his future wife Anita at a dance while he was working toward his Ph.D. (which to Anita indicated great husband potential).
Marv and Anita were constant companions and deeply in love throughout more than 41 years of marriage, perfectly counterbalancing each other’s character traits. Sadly, Anita passed away just eight months after Marv’s death. Marv was also a doting and loving father, and his pride in his children was palpable. His constant bragging about his daughter Brette, a highly accomplished corporate lawyer in Southern California, and his son Jeffrey, a nutritional consultant and successful entrepreneur. While Marv’s death struck a huge and tragic blow to the research community, the loss to his family was infinitely greater. He is missed by all.

Marv’s dominant technical contributions spanned the areas of synchronization, spread spectrum, interference cancellation, trellis coding, differential detection, fading channel performance evaluation, and SDRs. The significance of these contributions cannot be overstated. In particular, Marv’s early pioneering work on carrier and symbol synchronization, tracking, and acquisition earned him a reputation as one of the world’s leading authorities on this subject. Moreover, Marv’s contributions to time and frequency synchronization of direct sequence and frequency-hopped communications served as the basis for synchronization design and analysis in many military and commercial systems. Marv and Divsalar were the first to give criteria for designing trellis codes to operate over channels with multipath fading. These criteria have since been employed by many others, most recently for the design of space-time trellis codes. Marv’s research on multiple symbol differential detection, also jointly with Divsalar, demonstrated the ability to obtain, with a simple implementation modification, significantly improved performance over the classical (two-symbol observation) method. The reduction to practice of this theoretical work resulted in the issuance of a landmark patent.

Marv’s discovery with Alouini in the late 1990s of the moment-generating function (MGF)-based unified approach to the performance analysis of communication over fading channels stimulated a vast amount of research throughout the communications community. This novel approach to performance analysis drastically simplified previous approaches both analytically and computationally and led to new results that heretofore had resisted solution. Marv’s last significant contributions were in the area of SDR. In that work, he developed and tested a variety of classifiers, estimators, and synchronization modules that, in combination, allow a radio receiver to, on its own, adaptively reconfigure itself to its environment based solely on knowledge obtained from the received signal. This technology is significant because it transforms an ordinary SDR, which is simply reprogrammable, or reconfigurable, into something that is completely autonomous and which can react intelligently to whatever kind of telecommunications signal is sent to it. This is extremely important for deep space assets, which are difficult to reprogram or reconfigure because of the distances involved and low uplink data rates. These many profound technical contributions by Marv are elaborated on in a tutorial article submitted to IEEE Communications Surveys and Tutorials that complements this tribute (http://arxiv.org/ftp/arxiv/papers/1601/1601.01923.pdf).

**Reflections on Marv Simon**

The professional accomplishments of Marv described in the previous section let us know the caliber of the man. The stories in our memories — of Marv as a researcher, collaborator, mentor, professor, and leader in the field — let us know the quality of the person. This section provides collective reflections on these many dimensions of Marv Simon.

**Marv as a Researcher**

Marv was a consummate researcher, always hungry for new discoveries and results. Interesting ideas would come to him constantly, and he would investigate all of them. His outstanding computational skills made it possible for him to tackle extremely difficult performance analysis problems. Moreover, he strove to obtain closed-form expressions for his results, following Hamming’s philosophy that “the purpose of computing is insight, not numbers.” To do so, he would generate page after page of complex analysis in his own unique stylistic script. He also spent a considerable amount of time “playing” with any newly derived equations that he worked on to put them in the most compact and elegant form. Marv believed that mathematical elegance was important. His objective was not just to end up with an expression that could be computed but rather an expression that was written and presented in such a way that made it easy to understand and allowed the reader to see easily the dependence of a particular performance metric on the key system parameters.

One of Marv’s most unique characteristics as a researcher was that he would not only solve the problem at hand, but his creative imagination would spawn a related problem, then another and another, and he’d develop a common theoretical framework that could treat all of them. Pretty soon he’d have a book-length rounded subject matter rather than a collection of random problems and solutions. And he’d always bequeath to his fellow researchers a thorough documentation of his analyses in the form of a book or multiple peer-reviewed papers. He also loved elegance of style, and was hence an early fan of the Mac computer, using it to create some of his later books in camera-ready form.

**Marv as a Collaborator and Mentor**

The diversity of Marv’s collaborations is quite extraordinary, spanning a broad swath of topics and people. In many of his collaborative projects, Marv served as the inspiration and chief motivator. He also worked to ensure that all his joint publications adhered to his high standards of eloquence and style. Marv’s speed in writing a paper once the results were obtained was amazing; many times, it took no more than a single day. Though quick, he always took care with his expressions. Marv synthesized ideas quickly, wrote as naturally as he spoke, even on the most technical of topics, and demonstrated a beautiful talent with words. Few people could match Marv’s speed and elegance in writing, and he served as a benchmark and golden role model for all his collaborators in this regard.

Marv took great pleasure in teaching and mentoring. Whether at Caltech or JPL, he would sit down with novice students and engineers, explaining with great patience and expertise the fundamentals of the problem at hand. Marv worked closely with several of these young researchers, co-authoring important publications with them that helped to launch their careers. In fact, one of his landmark textbooks was written with a researcher just starting his academic career. It is common advice to Assistant Professors that they should concentrate on writing and publishing papers, which will help them establish their research credentials and earn tenure; authoring books, on the other hand,
Marv served frequently as Chairman and Session Organizer for technical sessions in IEEE conferences, including the National Electronics Conference (NEC), the National Telecommunications Conference (NTC), the International Conference on Communications (ICC), the International Symposium on Information Theory (ISIT), and the Global Telecommunications Conference (GLOBECOM).

Marv as a Professor

In addition to his position at JPL, over 10 years Marv served as a visiting lecturer in the department of Electrical Engineering at Caltech. In this role he developed a year-long course on Digital Communication Theory. This was a classic course and a special experience for students in the field at Caltech. Each of the students was challenged by Marv from the first day of class. Although these students were very accomplished with a strong background in communications, Marv showed them what they did not know by going back to basic principles. He taught the course from his own notes rather than a textbook, and relied heavily on probability theory, random processes, and detection and estimation fundamental principles. His lectures were always thought provoking and very lively.

Marv also had an amazing encyclopedic knowledge of the field and, off the top of his head, could tell many anecdotes about a topic, including when it was presented and discussed for the first time, the researchers and groups that first worked on the topic, which groups were collaborating in the area, the evolution of the topic over the years, and practical implementations of the results in NASA/JPL missions. Through his many examples, Marv demonstrated the importance of knowing the work in a field well in order to build on what came before and also see where the field was heading.

All of Marv’s assignments and exams were take home, which took away the stress of sitting through an exam but actually required huge amounts of time and creativity. As the students carried the questions home, they knew that they would be challenging but very interesting. In fact, Marv designed many of the problems he posed on the assignments and exams himself, out of recently published papers in the IEEE Transactions on Communications. Marv taught not only the material students were supposed to learn but, more importantly, how to be good researchers and how to tackle open problems in the digital communications field. Though he had high expectations and was always demanding, Marv was an excellent teacher who focused on the fundamentals to provide a foundation for his students’ research.

Marv as a Leader in the Field

Beyond impacting the lives of individuals, Marv also personally impacted the field, not just through his research but also by his professional service to the community and as a vocal member of research groups and societies. Marv cared a great deal about the well-being of the Communication Theory field. He was a lynchpin of ComSoc’s Communication Theory Technical Committee (CTTC), which he joined in 1973 and participated in thereafter, serving as its Chairman from 1977 to 1980, a critical period of resurgent interest in the field. In this role he was responsible for leading the organization and technical contributions of CTTC in all of ComSoc’s conferences and workshops. He was also an Editor of the IEEE Transactions on Communications in the area of Communication Theory from 1973 to 1976. Marv served frequently as Chairman and Session Organizer for technical sessions in IEEE conferences, including the National Electronics Conference (NEC), the National Telecommunications Conference (NTC), the International Conference on Communications (ICC), the International Symposium on Information Theory (ISIT), and the Global Telecommunications Conference (GLOBECOM).

In all of these roles Marv strove to ensure that meetings and publications around the topic of Communication Theory would adhere to the strictest standards of principle and quality.

Marv was a conscientious editor and reviewer, exerting a very strong influence on papers under his purview. When he agreed to review a paper, he typically invested much effort into double-checking all of its mathematical details. He would also devote a significant amount of time to improving the technical content of the paper by suggesting alternative mathematical proofs or asking for specific clarifications. He also often sent the authors of papers he reviewed his own marked-up version of the paper in which he made language-related comments, suggestions for better phrasing and grammatical corrections to improve the presentation and readability of the paper. This was extremely unusual, as most paper reviewers prefer to remain anonymous. While Marv’s standard for what was acceptable as a paper was very high, he always helped the authors improve their work. Moreover, when he was impressed with a particular research paper, he was very generous in his praise of both the paper and the authors.

Marv was known for his high level of integrity and strong sense of fairness. In particular, he would exert significant effort and time to form a detailed and honest appraisal of any work or colleague he was asked to evaluate. Marv was once asked to write a letter of recommendation for the early promotion of a younger colleague. Though he was not completely familiar with this person’s work, he took the time to read many of his papers and then scheduled a phone meeting with him to ask him detailed questions about his work before writing his letter. Marv was typically quick with praise when merited, and his criticisms were not only fair, but also often coupled with thoughtful and detailed advice on means of improvement.

CLOSING REMARKS

Communications technology continues to evolve at a rapid pace, building on many of the foundational ideas developed by Marv during his four decades of research. Indeed, many of the techniques Marv developed are experiencing a resurgence to address new challenges of emerging systems. In particular, millimeter wave communications requires far more demanding synchronization techniques than at lower frequencies, renewing the challenge of developing fast and accurate synchronization algorithms with low complexity. Diversity and range benefits of “Massive multiple-input multiple-output (MIMO) systems,” whereby large antenna arrays are deployed at cellular base stations, are being studied and implemented for millimeter wave systems as well as at lower frequencies. Marv’s diversity analysis techniques provide a powerful mechanism to understand the performance gains of such systems. The “Internet of Things” promises billions of additional devices sharing existing wireless spectrum, demanding new forms of multiple access and interference cancellation that will build on the groundbreaking work of Marv and others in spectrum. Finally, the standard for the next generation of cellular communication, 5G, is looking at new waveforms for data transmission, which will require the rigorous performance analysis techniques Marv developed for today’s waveforms that helped cement them into existing systems. While Marv cannot help us solve these emerging challenges, his legacy of deep, rigorous, and creative research as outlined in this tribute will serve as an inspiration and model for the next generation of communication theorists in addressing emerging wireless system design challenges.

ACKNOWLEDGMENTS

The authors gratefully acknowledge Norm Beaulieu, Jon Hamkins, William Lindsey, Andreas Polydoros, and Robert Scholtz for their contributions to and comments on the tribute. They would also like to thank Marv’s daughter, Brette Simon, for her comments as well as for sharing many of the pictures that appear in the tribute.
2016

AUGUST

Waikoloa, HI
http://icccn.org/icccn16/

St. Petersburg, Russia
http://ismw-fruct.spbu.ru/#general

SEPTEMBER

Valencia, Spain
http://www.ieee-pimrc.org/

Vienna, Austria
http://edoc2016.univie.ac.at/

Palma De Mallorca, Spain
http://www.asmsconference.org/

Würzburg, Germany
http://itc28.org/

Munich, Germany
http://iee ehealthcom2016.com/call-for-submission

Newark, NJ
http://sites.ieee.org/sarnoff2016/

Poznan, Poland
http://iswcs2016.org/

Jaipur, India
http://icacci-conference.org/2016/home

Split, Croatia
http://marjan.fesb.hr/softcom2016/cfp.html

Montreal, Canada
http://networks2016.etsmtl.ca

Aachen, Germany
http://www.itzwit-aachen.de/WISEE2016

OCTOBER

Pisa, Italy
http://cloudnet2016.ieee-cloudnet.org

Kaiserslautern, Germany
http://www.icmu.org/icmu2016/

Kanazawa, Japan
http://www.ieice.org/~icm/apnoms/2016/

Hanoi, Vietnam
http://rev-conf.org

Yangzhou, China
http://ie-wcsp.org

NOVEMBER

MILCOM 2016 — Military Communications Conference, 1–3 Nov.
Baltimore, MD
http://events.afcea.org/milcom16/Public/enter.aspx

IEEE 5G Summit Berlin 2016, 2 Nov.
Berlin, Germany
http://www.5gs ummit.org/berlin/

IEEE SmartGridComm 2016 — IEEE Int’l. Conference on Smart Grid Communications, 6–9 Nov.
Sydney, Australia
http://sgc2016.ieee-smartgridcomm.org/

Bangalore, India

Hanoi, Vietnam
http://rivf2016.tlu.edu.vn/

Palo Alto, CA
http://nfvdsn2016.ieee-nfvdsn.org/

FRUCT19 2016 — 19th Conference of Open Innovations Association FRUCT, 7–11 Nov.
Jyvaskyla, Finland
http://fruct.org/cfp

WPMC 2016 — Int’l. Symposium on Wireless Personal Multimedia Communications
Shenzhen, China
http://www.wpmc2016.org/
IEEE Members Voice and Raise the Importance of Science, Engineering and Technology at The Capital

By Fawzi Behmann, Vice Chair, ComSoc NA and Chair ComSoc/SP/CS Central Texas Austin Chapters

As a senior member of IEEE and active volunteer in the Communications Society, I had the opportunity to participate in the Science, Engineering and Technology Congressional Visits Day (SET CVD), an annual two-day event held April 12-13 to bring scientists, engineers, mathematicians, researchers, educators, and technology executives to Washington to raise visibility and support for science, engineering, and technology.

The overall objective is to improve the innovative climate in the United States and the professional lives of America’s technology professionals. Considering the tight budget environment and discussion of the 2017 Federal budget, it was timely to join other colleagues and express IEEE’s position seeking sustaining research budgets and independence.

One of the objectives is to support maintaining National Science Foundation (NSF) funds to advance research. This improves America’s STEM K-12 science, engineering, and math education program, enabling IEEE to collaborate with education institutions and industry to bring advanced technology to students in a simplified, innovative, compelling and interactive way, while equipping students with hands-on experience to prepare them as future engineers and scientists.

The second objective is to seek support of NIST funds to help American businesses succeed and enable IEEE collaboration with industry so American companies are strengthened to compete in the global market. NIST will contribute to the health of our research infrastructure, e.g. improving America’s electrical generation and distribution systems, including promoting the use of alternative energy and the development of the Smart Grid.

For example, in the area of STEM, IEEE supports STEM initiatives with funding for K-12 and the PACE program for Regions 1-6. However funding is limited. On the other hand, the U.S. is endorsing the STEM program partially funded by the NSF. Such funding will help expand IEEE STEM activities in terms of outreach and coverage.

We are seeking STEM funds (from NSF and others) to implement the program and prepare students in the 50 states for greater challenges with an entrepreneurial mindset. The outcome is to eventually contribute to the economy with jobs and new products and services, and to maintain students’ interest in STEM careers and foster collaboration with academia and industry.

The Alliance for Science & Technology Research in America (ASTRA) recently surveyed six million high school students to determine their interest in STEM-related fields. Nearly 30 percent (more than 1.6 million) would like to pursue STEM careers in the future. Keeping STEM students from dropping out of the STEM talent pipeline is essential in meeting the future demand for US STEM jobs.

SCIENCE-ENGINEERING-TECHNOLOGY WORK GROUP

I had the opportunity to travel to Washington, D.C. as a part of the Science-Engineering-Technology Work Group to express the importance of research and development activities to the nations’ economic growth and stability. On April 12-13 I joined with more than 150 scientists, engineers, and business leaders who made visits on Capital Hill as part of the 20th “Congressional Visits Day,” an annual event sponsored by the Science-Engineering-Technology Work Group.

While visiting congressional offices I discussed the importance of the nation’s broad portfolio of investments in science, engineering, and technology to promote our country’s prosperity and innovation. I also spoke about maintaining the budget for NSF and research for STEM programs to fund initiatives to bring advanced technology such as 3-D printing, Science in a Box, Internet of Things in a Box, and others to students in more simplified, creative, and interactive ways. I also called for support to maintain the budget for NIST budget and research for standards development required for smart alternate energy distribution from the grid to consumers and vice versa.

More than 50 percent of all industrial innovation and growth in the United States since World War II can be attributed to advances pioneered through scientific and technological progress. Achievements from federally funded science, engineering, and technology include global environmental monitoring, lasers, liquid crystal displays, and the Internet, among many other scientific and technical advances.

The federal government supports a unique research and education enterprise that fuels the American economy. This enterprise provides the underpinning of high-technology industries and expands the frontiers of knowledge in every field of science. Much of this research is carried out at academic institutions across the United States.

(Continued on Newsletter page 4)
CHAPTER REPORT

Community Development Opportunity Through IEEE ComSoc Workshops in New Zealand

Nurul I Sarkar, Chair of the IEEE Joint NZ North, South and Central ComSoc Chapter

The IEEE New Zealand (NZ) Communications Society (ComSoc) Chapter is a joint chapter of IEEE NZ North, South, and Central Sections. Last year (2015) was very productive for us in terms of professional development of members in the wider community. We hosted several professional development programs, including an IEEE Distinguished Lecture (DL), invited speakers, and several workshops and seminars. While these programs were very effective for professional development of staff, tertiary students, engineers, and industry practitioners, we have not had a chance to work with local community schools in New Zealand. ComSoc chapter chair, Associate Professor Nurul Sarkar, took the initiative to organize IEEE workshops in the community schools around Auckland. One such event was held at the IQRA (http://www.iqra.org.nz/) Community School on Saturday 12 December 2015. This article highlights the motivation/background and the overall effectiveness of the IEEE workshops on the development of the local community schools.

IQRA is a distinct Auckland-based community school/academy, run by a group of dedicated volunteers/teachers. It aims to provide students ages 6 to 15 with a solid foundation of Islamic knowledge and ethics through technology enabled methodologies supplementing religious education to the traditional primary, intermediate, and high schools. The classes are held once a week mostly on weekends at the Mount Albert community complex (Rocket Park), Auckland. We worked with IQRA Academy and introduced IEEE activities to students and parents through seminars/workshops. The children are very talented here and are highly motivated to learn and find a link between Quran and science/technology. The IEEE workshop was very effective in meeting their requirements.

Another motivation for holding this workshop at IQRA was to meet and greet highly educated teachers and parents. The idea was to introduce IEEE to young minds (in terms of technological development and standardization on science and technology) who will be the future members of IEEE, especially when they enter the tertiary institutions.

The workshop began with a short PowerPoint presentation highlighting the IEEE activities in New Zealand and worldwide. An overview of IEEE was presented along with the role it plays in the

Global Communications Newsletter

July 2016

3

(Continued on Newsletter page 4)

CHAPTER REPORT

17th Congress of Spain IEEE Student Branches: Promoting Engineering and Research to the Next Generation of Scientists

By Oscar M Bonastre, Chair of Technical Activities, IEEE Spain Section

The 17th Congress CNR (http://umh.ieeespain.org/cnr/) for all Spanish IEEE student branches (IEEE SB) was held on 7–9 April 2016 at the Miguel Hernández University, Elche, Spain. The event was supported by the IEEE Spain Section, organized by the IEEE Student Branch located at the same university, and sponsored by industry and institutions located in the regional area. During three days, the congress attracted more than one hundred participants from all IEEE student branches distributed around the country. The event successfully met its main objective, which was to facilitate ample interaction among participants and invited speakers from academia and industry. The program included keynote speeches given by the Chair of the IEEE Spain Section, experts on emerging trends of optical communications, advanced space propulsion technologies, and IEEE distinguished lecturers on brain-machine interfaces.

The program also included one technical session to promote scientific research as one of the significant missions of IEEE for the advancement of engineering and to foster technological innovation and excellence for the benefit of the community. The talk was scheduled for the last day with the aim of integrating all information received through the Congress. The invited speaker was

Global Communications Newsletter

July 2016

2
Technical Colloquium Conducted by the ComSoc Chapter of Hyderabad Section

By N. Venkatesh, Chair, ComSoc/SPS Joint Chapter, IEEE Hyderabad Section

The IEEE Hyderabad Section includes a large and vibrant student community. On 29–30 January 2016, the Student Chapter of CVR College of Engineering, Hyderabad, organized a technical colloquium jointly with the ComSoc/SPS Joint Chapter of the Hyderabad Section. This event had several distinguished speakers from industry, apart from the IEEE Hyderabad Section, who addressed 330 participants, including 240 students and 50 faculty members from the college and 40 participants from other colleges. More than 130 of the participants were IEEE Members or Student Members, with the event providing the others with a good insight into IEEE and the motivation to join.

An objective of the event was to acquaint the students with current technology. This colloquium provided two days of plenary sessions for formal discussions and presentations on current technical topics that made the students aware of many technologies from different streams, such as emerging trends in 5G communications, VLSI technologies, wireless communications for indoor positioning, the Internet of Things, electronic warfare, and radar target detection, among others.

Mr. N. Venkatesh, Chair of the ComSoc/SPS Joint Chapter and a Sr. Vice President at Redpine Signals, provided technical insight into the technologies of indoor positioning. The session covered historic and current navigation methods, and their current move into indoor use. It included details of wireless techniques in locationing using time of flight, received signal strength, and angle of arrival. It provided details of applications of indoor positioning: the locationing of people and assets and the benefits accruing for safety and operational efficiency in various environments. The session also covered the implementation of such systems using RTLS devices and software.

Dr. Abhinav Kumar, a Professor at the Indian Institute of Technology, Hyderabad, conducted a session on 5G communications. The session covered 5G scenarios and their challenges. The speaker compared various mobile communications and emerging trends in 5G communications. This lecture gave the audience an insight into future communication scenarios, the technologies

(Continued on Newsletter page 4)

XII International Siberian Conference on Control and Communications (SIBCON-2016)

By Oleg Stukach, Tomsk ComSoc Chapter Vice-Chair, Russia

The flagship event of the joint COM/AP/ED/MTT/EMC Tomsk Chapter, the Siberian Conference on the latest advances in communications and control systems, was held in Moscow, the Russian Federation, on 12–14 May 2016. Approximately 180 participants from 23 countries presented 232 technical papers in 12 oral sessions. There were five workshops and seven short courses in addition to a number of related events. Discussions included emerging wireless related growth areas such as 5G technologies, networking control, electron devices, wearable electronics, Internet of Things (IoT), and coms/microwave technology in life. The Conference as a whole demonstrated continuing interest in analysis and control methods for communications.

The best activity of a chapter is the organization of conferences. To some extent we are united by a joint passion for organizing events. “Oh, conferences once again,” you exclaim, “there are a lot of conferences worldwide.” I will try to argue that our event is unique. It is completely unpredictable both in the planning and in the background. As a rule, well known flagship events of ComSoc and other societies follow a similar scenario. Also, participants know exactly what they can expect. They can plan their participation in individual sessions as well as in the cultural program.

Probably that model is not for us. In our conditions the planning of a conference 18 months to two years in advance is a great achievement. We had a few examples of needing to change the venue three days before the start of an event, despite preliminary agreements. A positive aspect of this event was more comfortable meeting buildings.

(Continued on Newsletter page 4)
professional development of the members of the wider community. The potential benefits of joining IEEE were highlighted. A short video was presented highlighting the significant contribution made by IEEE in the field of science and technology. The students enjoyed the video presentation and were excited to learn more about IEEE’s contributions worldwide. Finally, a number of quizzes were given to the students to test their knowledge and understanding of various aspects of IEEE. The students did very well with the quizzes, indicating they had learned about IEEE effectively.

Approximately 35 students and parents attended the event. With ample opportunity for discussion, people enjoyed talking and networking during lunch/refreshment break. The event was co-sponsored by IEEE ComSoc and the IEEE NZ North section. Organizing chair Associate Professor Nurul Sarkar received positive feedback from the teachers and parents, indicating that the event was successful.

In conclusion, IEEE NZ ComSoc workshops were very effective last year in terms of community development. We are hoping to offer more workshops of this kind to various community schools this year. We thank IEEE ComSoc and the IEEE NZ North Section for their support. We also thank Mr. Ataur Rahman (principal of IQRA Academy) for providing logistical support for the workshop.

At the Capital/Continued from page 1

the country, ensuring knowledge transfer to future generations of scientists, engineers, mathematicians, physicians, and teachers. Additionally, technology transfer from academic research adds billions of dollars to the economy each year and supports tens of thousands of jobs.

I was pleased to have the opportunity to participate in this event. I feel strongly that making our voices heard to our elected representatives in Washington is critical to ensure ongoing support of federal R&D programs. The most rewarding experience came when I was welcomed in the offices of Senator Ted Cruz, Senator John Cornyn, Rep. John R. Carter, and Rep. Joe Barton. Highlights of the two-day event included a series of briefings and talks by Members of Congress and executive branch officials including Matt Hounihan, Director, AAAS R7D Budget and Policy Program. The George E. Brown Award for outstanding leadership in support of federal R&D was presented to U.S. Senators Chris Coons (DE) and John Thune (SD), and Representatives G. K. Butterfield (NC) and David McKinley (WV). The awards recognized outstanding efforts to promote science, engineering, and technology on Capital Hill.

SIBCON-2016/Continued from page 3

SIBCON has a 20-year history, but for the first time it took place in Moscow, at the Higher School of Economics (HSE). Because of our geographic location, for many people participation in a conference is a rare event. That is why, since 2011, we have located the conference in different cities around the country. Here it is generally difficult to limit the time for presentations, as people want to spend a lot of time communicating and discussing reports. At SIBCON-2015 (see ED-S Newsletter, January 2016), despite planning for additional time, one session continued into the next day, and according to rumors, that was still not enough time. Truly there is nothing more valuable than the luxury of human communication.

The collaboration of industry and academia has great value for us, bringing together and cross-fertilizing ideas from applied areas that have many similarities. Such opportunities arise due to the support of the Conference by National Instruments Company and HSE. A number of papers addressed the problems of engineering design of cyber-physical systems. Significant attention was paid to the analysis and design of network control systems and communication. Also, the papers with applications to problems of measurement based on the National Instruments technologies were presented.

The major goals of the Conference, to bring together researchers from various fields, to advance the state-of-the-art of control theory and technology for communications, and to gain some general and unified perspectives in this interdisciplinary field of advanced research, were achieved.

There are two things that made our work much better. First is a revolutionary reorganization of MCE with additional service to the conference organizers, including cross-check, electronic copyright forms, PDF-express, etc. Second is the change of evaluation criteria for research activity. Many authors, to have stability in the job market, want to write papers indexed in databases. We did not pay attention to scientific writing before. Now we have the opportunity to choose to present the best papers among many submitted works. The percentage of accepted papers is now almost equal to 30 percent, which was almost unattainable previously. It is true that we have additional work due to the higher number of submitted papers submitted. As a result, we keep an eye on the increasing interest of ComSoc: more engineers and professionals are reading in English, and they compare their results with high quality papers. We hope that this sharp move to quality will continue in the following years.
InterDigital’s Creating the Living Network™ Webinar Series

Live It: IoT Use-Cases and Lessons Learned

The Internet of Things (IoT) is driving innovation and development in both connected devices and wireless networks. IoT promises constant connectivity, intelligent devices, and access to more data for better real-time decisions. But, where are we in the race to implement real-life IoT scenarios? Join us on June 23rd to find out!

The 5th webinar of InterDigital’s Creating the Living Network™ webinar series will focus on IoT use-cases such as smart buildings, smart cities, and smart healthcare. During the webinar, IoT experts from InterDigital, Harman and Hitachi will also discuss lessons learned and real-life applications for IoT technologies and services.

Emerging Technologies in IEEE 802.11 WLAN (Wi-Fi)

The IEEE 802.11-based wireless local area network (WLAN) technologies (popularly known as “Wi-Fi”) have grown into one of most ubiquitous wireless access technologies across consumer and enterprise markets. The evolution of IEEE 802.11 standards has significantly increased data rates of wireless access.

Notwithstanding the phenomenal progress made in the past years, the IEEE 802.11 launched a task group to discuss the next generation of the standard after 802.11ac. It is clear now that the days when the Wi-Fi network was considered a mere convenience are behind us. It has become a critical part of our home, enterprise and, even, cellular operator access networks.

In this tutorial, panelists discuss the physical layer (PHY) and medium access control layer (MAC) technologies being considered by 802.11ax such as orthogonal frequency-division multiple access (OFDMA), uplink MU-MIMO, dynamic sensitivity control, dynamic clear channel assessments, and share the latest results on performance improvement provided by these changes. The panelists also discuss the challenges they see in the future regarding high-performance usage of 2.4GHz and 5GHz unlicensed bands by 802.11ax and beyond.

IEEE ComSoc content sponsored by: National Instruments

Limited Time Only at >> www.comsoc.org

For this and other sponsor opportunities contact Mark David // 732-465-6473 // m.david@ieee.org
We live in a world in which there is a great disparity between the lives of the rich and the poor. Information and communication technology (ICT) offers promise in bridging this digital divide through its focus on connecting human capacity with computing and informational content. It is well known that Internet access has the capability of fostering development and growth by enabling access to information, education, and opportunities. Unfortunately, the availability of Internet in worldwide terms is limited, with an estimated 4 billion people — an estimated 60 percent of the human population — lacking Internet access. People in rural areas are particularly hard hit since socio-economic factors preclude the provisioning of Internet access and mobile telephony in these sparsely populated low-income areas.

There is a growing interest in using novel wireless solutions — such as TV white space (TVWS), satellites, drones, and free space optics — to unfetter rural areas from the encumbering constraints of infrastructure (traditionally associated with broadband Internet provisioning). The aim of this Feature Topic (FT) is to highlight the research being done on leveraging wireless technologies for development (W4D), and thereby increase the quality of life for a larger segment of human society by providing opportunities to connect resources and capacity, especially by provisioning affordable universal Internet access. This FT is especially timely since it coincides with the recent push by various companies (e.g., Facebook/Internet.org) and organizations (e.g., the Global Access to the Internet for All [GAIA] research group at the Internet Research Task Force [IRTF]) for the vision of global access to the Internet for all.

Overall, 16 papers were submitted to our FT, out of which we have selected four high-quality articles. These articles, written by researchers from leading groups around the world, have been selected after a rigorous peer review process, and present a broad snapshot of the W4D work going on in the broader ICT for development (ICTD) community. A salient feature of all the selected articles is that they present insights developed from real-world field deployments. A brief description of the accepted articles follows.

The first article of this FT is “Research Advances on Wireless Community Networks with the Community-Lab Testbed,” authored by Leandro Navarro along with contributors from Community-Lab. This article present insights the authors have developed through managing the Community-Lab (Community-Lab.net) wireless community networking testbed, which comprises more than 200 hosts and is built on top of a federation of existing community IP networks constituted by more than 40,000 routers and 60,000 km of links.

The second article accepted for this FT is “Toward Enabling Broadband for a Billion Plus Population with TV White Space” authored by Kumar et al., which proposes TVWS as a robust and affordable backhaul solution to reach a billion plus population of users within India. In particular, the proposed solution uses TVWS for the “middle mile” that connects the last mile (local Wi-Fi clusters) with an optical fiber backbone. In this article, the authors articulate the insights they have developed by commissioning a real-life TVWS testbed that covers more than 13 villages in rural India (spanning an area of 25 km²).

The third article accepted in this FT is “Wireless Technologies for Isolated Rural Communities in Developing Countries Based on 3G small-Cell Deployments.” This article, authored by members of the EU-funded TUCAN3G project, discusses the technical and socio-economic factors related to mobile voice and data service provisioning in isolated rural areas of developing countries. The authors propose a small-cell-based 3G solution that uses 3G femto-cells for the access network along with heterogeneous backhauling. The authors evaluate their solution using a pilot deployment in the Peruvian jungle and demonstrate that the proposed solution is long-term sustainable.

Finally, the fourth article selected for this FT is “SmartCell: Small-Scale Mobile Congestion Awareness” authored by Schmitt et al. In this work, the authors provide an insightful look at the cellular quality divide through an observa-
tional study based in three different settings, including San Cristóbal Verapaz, Guatemala; the Za’atari refugee camp in Jordan; and Santa Barbara, California. Based on their findings, the authors propose SmartCell, an Android-based quality of service application that can detect congestion on a cellular base station and improve mobile connectivity by automatically switching in real time between networks (when multi-SIM handsets are available) or through user-initiated manual switching between networks (when single-SIM handsets are assumed).

We hope that the networking community benefits from the insights presented in this FT and that it provides a balanced snapshot of the range and breadth of W4D research. We sincerely thank all the authors and reviewers for their help and efforts. We would also like to thank the staff members and Editor-in-Chief of IEEE Communications Magazine for their guidance and help.

**Biographies**

Junaid Qadir [SM] (junaid.qadir@itu.edu.pk) is an associate professor at the Information Technology University (ITU) Punjab, Lahore, Pakistan. He served as an assistant professor at the School of Electrical Engineering and Computer Sciences, National University of Sciences and Technology, Pakistan, from 2008 to 2015. He serves as an Associate Editor for IEEE Access, IEEE Communications Magazine, and Big Data Analytics. He is a member of ACM.

Aruna Sathaseelan is a senior research associate at the Computer Laboratory, University of Cambridge, where he leads the Networking for Development (NG4D) Lab. He is Chair of the IRTF Global Access to the Internet for All research group and a member of the Internet Research Steering Group. He previously worked at the University of Aberdeen, where he founded the ICT4D group as an associate with the Center for Sustainable International Development.

Marco Zennaro works at the Abdus Salam International Center for Theoretical Physics on projects involving networking and wireless communications for scientific institutions in developing countries. His research interests are ICT4D and wireless sensor networks. He has given lectures on wireless technologies in more than 20 different countries. When not traveling, he is the Editor of wsnblog.com. He is a co-author of the book Wireless Networking in Developing Countries (www.wndw.net).

Adam Wolisz [SM] is a chaired professor of electrical engineering and computer science at the Technische Universität Berlin, where he founded and leads the Telecommunication Networks Group. Currently he is executive director of the Institute for Telecommunication Systems. In parallel he is also an adjunct professor in the Department of Electrical Engineering and Computer Science, University of California, Berkeley. He is a member of ITG.

Saleem N. Bhatti [M] is a professor at the School of Computer Science, University of St Andrews, United Kingdom. He has been performing collaborative research since 1991 with network operators, equipment manufacturers, and research and academic institutions in Asia, the United States, and Europe. His research interests are in the general area of networked and distributed systems, with a current focus on ICTD, energy-aware systems, and security-related topics. He is a member of ACM SIGCOMM.

Kannan Govindan [SM] is currently working as a senior chief engineer in Samsung Research India. He was the Microsoft Research India Ph.D. Fellow at IIT-Bombay for four years. He was a postdoctoral researcher in the Department of Computer Science, University of California Davis from 2009 to 2011, and also held a visiting position at the University of St. Andrews in the summer of 2007. He is a member of ACM.

---

**2016 IEEE COMSOC STUDENT COMPETITION**

“Communications Technology Changing the World”

**1st Prize**
USD 2000 + plaque

**2nd Prize**
USD 1000 + plaque

Submission Deadline: 15 September 2016
www.comsoc.org/communications-technology-changing-world
ABSTRACT

Beyond traditional telecom providers, citizens and organizations pool their own resources and coordinate in order to build local network infrastructures to address the digital divide in many parts of the world. These crowdsourced network infrastructures can be self-organized and shared by a community for the collective benefit of its members. Several of these networks have developed open, free, and neutral agreements, and are governed as a common-pool resource: community networks. These are built using a variety of commodity wireless hardware (e.g., Wi-Fi long-range point-to-point links, Wi-Fi and GSM access points, and mesh networks), sometimes optical fiber links, heterogeneous nodes, routing protocols, and applications. A group of researchers, developers, and community networks developed the Community-Lab testbed, and for the last five years have worked together to overcome obstacles, improve the technologies, tools, and operational models being used, as well as model best practices for more effective and sustainable community networks. This article presents the challenges for experimentation, the testbed built, results learned, and the impact of that work to place wireless community networks as one sustainable way toward an Internet accessible to all.

INTRODUCTION

Access to the Internet is increasingly seen as an opportunity, allowing individuals to participate in society and benefit from services related to all aspects of our lives. Despite the fact that connectivity is also considered a key component for socio-economic development all over the world, there is a digital divide. The Global Internet Report (Internet Society) estimates that there were 3 billion people online in 2015 out of a population exceeding 7 billion. However, alternatives and choice of connectivity services are considered a necessity for accessibility and sustainability.

Traditionally, this access has been mainly offered by commercial Internet service providers (ISPs) and telecom providers. However, their economic model has left a large fraction of the population behind, such as inhabitants of remote areas, with low population density and high deployment costs, people of low economic capacity, or those simply not satisfied with the local commercial offerings.

Beyond traditional telecom providers, there is an alternative method through which to address the digital divide. Citizens and organizations can pool their own resources and coordinate in order to build local network infrastructures. These crowdsourced network infrastructures can be self-organized and shared by a community for the collective benefit of its members.

These models of participation, organization, and financing of local network infrastructures can vary widely. For example, some networks are freely accessible, others are cooperation-based, and some are run by federations of microISPs. The following examples demonstrate the diversity described:

Broadband for Rural North (B4RN), in Lancashire, United Kingdom, deploys and operates optical fiber in a cooperative way. The Nepal Wireless Networking Project (NWNP) is a social enterprise that, due to the lack of alternatives, provides Internet access, e-commerce, education, telemedicine, environmental, and agricultural services to many remote villages using wireless technologies. The French Data Network Federation (FFDN) is a federation of French do-it-yourself ISPs, which comprises digital subscriber line (DSL) resellers, wireless Internet service providers (WISPs), co-location centers, and other similar organizations. Village Telco is a model of wireless mesh networks replicated in several communities around the world, such as the Zenzeleni rural telecom cooperative, the first legally established in rural South Africa in 2015. Rhizomatica, in Mexico, has deployed open source GSM base stations [12] in more than 10 rural communities under a new national spectrum regulation scheme for rural and indigenous areas. The Internet Society has supported several pilots of Wi-Fi-based networks deployed in remote areas of India by locals [5].

Several crowdsourced networks have developed open, free, and neutral agreements or licenses in a similar way as the free and open source software model [2]. Openness suggests that everyone has the right to know how they are
built. Freedom indicates that the network access is driven by a non-discriminatory principle, and therefore, they are universal. Finally, neutrality implies that any technical solution available could be used to extend the network, and that it should be used without discrimination. This means that the network can be used to transmit data of any kind by any participant, including for commercial purposes. They are so-called community networks (CNs). Representative community network examples are Freifunk (FF) in Germany, the Athens Wireless Metropolitan Network (AWMN) in Greece, FunkFeuer (0xFF) in Austria, guifi.net in Spain, and Ninux.org in Italy.

Community networks such as the ones presented above are built using a variety of commodity wireless hardware (e.g., Wi-Fi and GSM access points, and mesh networks), sometimes optical fiber links, heterogeneous nodes, routing protocols, and applications. They support a large number of end users (guifi.net reaches over 30,000), and follow an innovative model of self-provisioning using unlicensed and public spectrum [4]. These infrastructures, having been developed cooperatively for the community, become a key resource for local communities, a networking community where digital social life takes place. Individual citizens, public institutions, non-governmental organizations (NGOs), businesses, and service providers themselves contribute and take advantage of this common infrastructure. CNs are an emerging and successful model for the future Internet, with hundreds of examples around the world.

Aiming to support CN infrastructures, a group of research and community organizations developed the Community-Lab testbed. This testbed was launched with the support of the CONFINE Future Internet Research and Experimentation project (2011–2015), supported by the European Commission as one of its future Internet testbeds, and federated with other global testbeds (Fed4FIRE and GENI). The project and the testbed include well established CNs, with large end-user bases and diverse application providers, research institutions, non-profit organizations, and small enterprises. The team comes from 24 partner organizations from Europe, America, and Africa.

IN 2016 the Community-Lab testbed has more than 200 hosts embedded in several community networks, and is deployed within the federation of existing CNs comprising more than 40,000 routers and 60,000 km of links. Community-Lab offers unified access to the testbed, with tools that allow researchers to deploy, run, monitor, and experiment with services, protocols, and applications on real-world CNs. Experimenting in production networks is challenging, as previous experiences such as Planetlab [6] demonstrate. For the last five years communities of practitioners have worked together with researchers and developers to overcome obstacles, improve the technologies, tools, and operational models being used, as well as model best practices for more effective and sustainable CNs. This article presents the challenges for experimentation, the testbeds built, the results, lessons learned, and the impact of the research to advance wireless CNs as one sustainable way toward an Internet accessible to all.

CHALLENGES AND ISSUES

Community networks face many challenges [9]. For instance, many of the technologies used were not designed or optimized for this particular purpose. As a result, local infrastructures can be very fragile. Moreover, self-organization and regulation are necessary in order to avoid abuse and collapse, which involves extra effort. These goals should be achieved by creating an environment able to accommodate many people in a sustainable manner.

The active involvement of end users in these networks, usually very skilled, curious, and with an entrepreneurial attitude in trying and adopting new solutions even in prototype form, offers a unique opportunity for discussing and suggesting ideas and improvements. In several situations (e.g., routing issues such as loops or scalability problems), the research community has been successful in characterizing problems (e.g., from logs and benchmarking) and contributing solutions (e.g., deploying experimental routing extensions). However, collaborations require intensive effort from practitioners and researchers. The existence of a testbed has proved to facilitate, automate, and systematize the collection of information (key to understanding and formalizing research problems) and the execution of experiments under realistic conditions and different scenarios.

The aim of the Community-Lab testbed is to advance research and empower society by understanding and removing obstacles to the adoption of CNs and their services. The main objective is to support experimentally driven research, improving software and network development as well as the evaluation of the multiple aspects of community networks. Community-Lab also provides a set of hosts (embedded in several CNs) that are specifically adapted to running experiments and allow the involvement of end users. The main features of the testbed can be summarized using the following adjectives: realistic, heterogeneous, integrated, open, and participatory.

The testbed is realistic because it functions under the operational conditions of several CNs and allows the identification of potential issues in the development and operation of new services.

The heterogeneity of the testbed is the result of providing access to very diverse environments in terms of node and link characteristics, availability, and reliability, among other aspects.

The integration of experimental nodes and services in order to run experiments within a common framework, with a common web portal (community-lab.net), saves time and costs for users. This simplifies the life cycle of experiments and pilot testing, and allows the porting of experiments across other federated experimentation testbeds.

The openness of the testbed facilitates the participation of diverse research groups and CNs for both short-term and long-term experimentation. Moreover, service providers and entrepreneurs can use the testbed to try out innovative services. In addition, openness facilitates verification and repeatability of the experimental conditions and
The facilities in Community-Lab are: the Community Network testbed deployed inside several CNs, the Virtual testbed that can be instantiated within a single computer, and the Campus testbed that can be deployed in specific areas, temporarily or permanently.

![Node/functional architecture of the Community-Lab (C-Lab) testbed](image)

**Figure 1.** The node/functional architecture of the Community-Lab (C-Lab) testbed.

...results since the infrastructure is always available.

The main challenges in the implementation of the testbeds was the design and implementation of the following mechanisms and technologies:

- Resource sharing is critical to allow multiple concurrent experiments, but isolation is an issue.
- Cross-layer solutions are needed in order to optimize services such as routing (with development and experiments of mesh routing algorithms such as BMX6 or OLSR2), content distribution, peer-to-peer live video streaming, information-centric networking, and aggregation and sharing of Internet access capacity [11].
- Self-managing systems running in large and dynamic networks, self-configuration (experiments in adaptive radio channel allocation, power, and rate), self-healing (experiments with services that adapt to node or link failures), self-optimization (adaptation according to available resources and depending on internal or external influences).
- Availability of open data sets such as routing graphs, traffic maps and traces, topology change dynamics, load, failures, number of users, volume of traffic, and spectrum occupation.
- Development of a benchmarking framework, enabling experiments under controlled and reproducible test conditions, offering automated procedures for experiments and performance evaluation and allowing a fair comparison between different networking mechanisms.
- Technical adoption, evaluation of how to incorporate the research results into the hardware and software infrastructure of existing communities and its application to other areas.
- Social mechanisms allowing the assessment of the social impact of the development of these networks on the future Internet and the inclusion of more people in the digital society.

- Economic mechanisms in order to explore the conditions for the sustainability of these infrastructures, influence of cooperative schemes for infrastructure sharing, financing, and the provision of services such as a community cloud to end users and local communities.
- Legal mechanisms that would enable us to explore the feasibility of these models, the role of governments in the regulation of the use of public space, spectrum, and provision of services.

**COMMUNITY-LAB TESTBEDS**

The set of facilities for experimentation developed in the CONFINE project has the collective name of Community-Lab. A key aspect considered in the design of the Community-Lab was the fact that experimentation can have a strong impact on production networks with limited resources, and therefore, these networks would need to be protected. In that sense, we realized that experimentation can benefit from different testbeds specifically optimized to give control to users over some aspects while providing an abstraction level for others.

The facilities in Community-Lab are shown as columns in Fig. 1: the **Community Network testbed** deployed inside several CNs, the **Virtual testbed** that can be instantiated within a single computer, and the **Campus testbed** that can be deployed in specific areas, temporarily or permanently.

Each of these experimentation environments can be divided into three groups of components (rows in Fig. 1): servers and services controlling the testbed, **testbed nodes** (i.e., research devices, RDs), where experiments are performed, and **network routers**, which are part of CNs.

The largest, widest, and most publicly accessible testbed (for registered researchers) in Community-Lab is the Community Network testbed...
(first column from the left in Fig. 1) with testbed nodes deployed as hosts embedded in several production community networks (guifi.net, AWMN, FunkFeuer, FKIE, Wireless België, Ninux, Sarantaporo.gr), with a few more CNs involved in social experiments and development (Fig. 2). This testbed has three main services:

- The Federation service, which interconnects with other testbeds in the Fed4FIRE federation
- The Controller service, which allows the experimenter to define slices and select nodes (RDs) in order to deploy sliver templates, informs each RD about its desired state (slivers requested for a given RD), and collects and aggregates sliver status information
- The Monitor, which collects, aggregates, and presents monitoring information from each RD connected to the underlying CN (this can be seen as an IP network of interconnected routers)

We used Tinc, an IPv6 overlay management network with multiple gateways, to offer a common bidirectional network on top of either IPv4 or IPv6 networks, NAT, or filters, and across Europe using tunnels over the GÉANT academic network and the Internet [1]. This overlay offered a seamless management network able to reach any testbed host, efficient in optimizing routes across gateways and nodes, and protected from the open Internet by a public key mechanism. Following the Planetlab slice-base architecture, an experiment controller allows the selection of which nodes (RDs) to use for an experiment. Each node checks with the controller and creates any required sliver (Linux container).

The coexistence with production networks is a key differentiating factor for Community-Lab, but it depends on the enforcement of limitations on uncontrolled experiments that may otherwise stress the underlying network in terms of traffic and low-level network access, and cause interference, intrusions, or leakage of private data. With that goal in mind, the Virtual testbed can be used to test experiments during development, test the testbed software itself before deployment in production networks, or run a very intrusive experiment using the same set of tools as in the real testbed.

The Virtual testbed is a software package that allows the deployment of a complete testbed in a single computer (or a cluster), using virtual machines and a virtual network. This can also simulate virtual wireless links through the integration of the ns-3 network simulator. It is depicted in the central column in Fig. 1. An instance of a virtual testbed should contain one controller, a set of virtual RDs, a virtual network topology, and a virtual network. Since the software is nearly the same as in the main testbed, and there is also an underlying simulated network, an experiment can easily be ported from the virtual to real testbeds.

The Campus testbed, also known as WiBed [3], was developed in order to avoid the network-protecting limitations set by the community network testbed. It has a specific architecture that allows radio and router-level experiments in a separate network for potentially network-disturbing and radio-interfering experiments, which may not be compatible with a production CN. It is depicted in the first column from the right in Fig. 1. The WiBed testbed nodes can run experimental firmware images that are downloaded from a control server, the WiBed controller, which provides a web interface and application programming interfaces (APIs) to control nodes and experiments. The most typical experiment deployed on the Campus testbed has been the testing and performance characterization of mesh routing protocols.

There are several instances of this testbed; Around 40 router-like devices are deployed at different offices in several buildings at the Universitat Politècnica de Catalunya. These router devices have three radios. Two of them also include an Ethernet connection for control. As WiBed is...

Figure 2. The Community-Lab Community Network testbed (2015).
Table 1. Summary of the main results.

<table>
<thead>
<tr>
<th>Result</th>
<th>Use/impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Testbed software and experimentation services</td>
<td></td>
</tr>
<tr>
<td>C-Lab system</td>
<td>Experiment controller, API, Confinde Node System (CNS)</td>
</tr>
<tr>
<td>C-Lab deployment</td>
<td>200+ worker nodes (research devices) and servers</td>
</tr>
<tr>
<td>C-Lab VCT</td>
<td>Software product, integration</td>
</tr>
<tr>
<td>C-Lab tools</td>
<td>Monitor, federation of testbeds and experimental service clouds</td>
</tr>
<tr>
<td>C-Lab WiBed</td>
<td>Software product, documentation, experiments</td>
</tr>
<tr>
<td>Results from experimental research</td>
<td></td>
</tr>
<tr>
<td>OLSRv2</td>
<td>Algorithms, reference implementation, input to standardization</td>
</tr>
<tr>
<td>DLEP</td>
<td>Reference implementation, input to standardization</td>
</tr>
<tr>
<td>BMX66</td>
<td>Algorithms, software product, research publications</td>
</tr>
<tr>
<td>NodeDB (netJSON)</td>
<td>API, reference implementation</td>
</tr>
<tr>
<td>Cloudy Community Cloud</td>
<td>Software product, integration, experimental use</td>
</tr>
<tr>
<td>Open datasets</td>
<td>Topology, traffic, routing, configuration, participation</td>
</tr>
<tr>
<td>Knowledge on computer networks</td>
<td>Routing algorithms: evaluation and evolution, resource allocation algorithms, cross-layer optimizations</td>
</tr>
<tr>
<td>Knowledge of systems</td>
<td>Scalable systems, interference, resilient systems</td>
</tr>
<tr>
<td>Business models, systems, and services</td>
<td></td>
</tr>
<tr>
<td>Knowledge of communities</td>
<td>Organizational models, common pool resource model, structural and topologic properties</td>
</tr>
</tbody>
</table>

There are several software tools for large-scale experimentation, under an open source/free software model. (Software repository: http://redmine.confine-project.eu and documentation: http://wiki.confine-project.eu). These software tools allow extending a testbed or building a separate testbed elsewhere.

**SERVICES FOR EXPERIMENTATION**

Community-Lab provides three environments in which to perform experiments in a set of nodes and networks (a slice):

- The Community Network testbed, to deploy experiments embedded in several production CNs. It can be accessed through its Community-lab.net web portal or through the Fed4FIRE SFA federation.
- The Virtual testbed, which is deployed on a single computer and can emulate a network for development of testing purposes.
- The Campus testbed (WiBed), deployed in a campus environment.

**OPEN DATASETS**

The participating CNs and research activities have created open datasets (http://opendata.confine-project.eu) in open access mode, already used in several published research experiments. Some datasets are generic, such as network graphs, traffic, and participation. Other datasets are used in specific publications which are made available to enable others to reproduce, verify, or extend the results, or simply use them in their own research activities.

**EXPERIMENTAL RESEARCH**

The main experimental results are presented here.

**DYNAMIC LINK EXCHANGE PROTOCOL**

DLEP is a protocol that standardizes a local connection between a bridging radio and a router, to allow the router to learn about the available link-layer and physical-layer data (e.g., bit rate, frequency, or signal strength). All control data is sent via the same interface where the user data traffic is sent. Our implementation for OpenWRT included additional physical-layer data such as signal strength, frequency, and bit rate, and additional link-layer data, mainly including statistics such as the number of sent and received frames, lost frames, and retransmissions. Research devices and CN routers can use this information for improved routing metrics, network measurements, and application-layer cross-optimizations. [15]

**OLSRv2**

We improved and extended the second generation optimized link state routing version 2 (OLSRv2) implementation for OpenWRT Linux and tested the protocol in the Virtual and Campus testbeds (BattleMesh 2015). We have also developed and evaluated multi-topology extensions and the directional airtime metric for OLSRv2 experimentally. While most CNs do not integrate the concept of priority traffic into their networking mechanisms, to retain fairness among users, there are reasons why some traffic could be prioritized or restricted to a subset of links.

**RESULTS, LESSONS, AND IMPACT**

The main results of this collaboration between communities, researchers, and developers in the last five years can be broadly divided into three areas (Table 1). A first set of results is derived from the testbed software and experimentation services themselves. The second set concerns experimental research. The third set focuses on the development of business models, systems, and services, applied or applicable to community networks in general.

Both a specific deployment and a software package, it can also be deployed in other locations. This has been done during BattleMesh workshops in Leipzig, Germany (2014), and Maribor, Slovenia (2015), and Porto, Portugal (2016).
such cases include certain real-time applications, for example, voice over IP or video communication that need reliable low-delay links, which, in contrast to BitTorrent, are more robust against delay and packet loss. [15]

**ReceIver-drIven routIng In BMx6**

A desirable feature for CN routing protocols is the incorporation of a list of preferred receiver nodes in the routing algorithm, thus enabling routing algorithms to respect the community social contract and not restrict the freedom of community users. In our receiver-driven discretionary routing mechanism, each receiver (the intended destination of the packet) can freely specify delivery objectives and remain compatible with the collaborative approach of CNs. Each node has a unique identifier and can announce the description of its offer and also the description of its routing policy with preferences to deliver traffic to that node. This receiver-driven routing can be applied to express preferences for desirable nodes and paths, or to restrict traffic to trusted nodes, enabling trust- and security-aware routing. A proof-of-concept implementation of key concepts, developed as an extension of the BMX6 routing protocol, validated in our Virtual and Campus testbeds, confirms its feasibility and scalability [10]. This work also allows a routing protocol to perform cryptographically secured negotiation and establishment of concurrent and individually trusted routing topologies for infrastructureless networks without relying on any central management. Finally, this feature can help in protecting a CN from problems that result from the trustworthiness of participants in an increasing amount of users and diversity.

**NodeDB (net4JSON)**

For community wireless networks (CWNs), a node database serves as a central repository of network information. This “registry” functionality is separate from the testbed controller, described above. The common NodeDB manages the network information per se for the CN. The NodeDB is simultaneously a registry, a link planning tool, an IP address assignment tool, and more. The information stored concerns nodes deployed at certain locations, devices installed at these locations, information regarding Internet addresses, and, for networks that use explicit link planning, links among devices. All this information is maintained via a web or RESTful interface by the community members. Therefore, the common NodeDB contains the static as well as dynamic information about each CN as opposed to the experimental testbed network information.

**Community Clouds**

Cloudy is a Linux distribution designed for building community cloud infrastructures in CNs [14]. The distribution contains the platform and application services of the community cloud model. It has been tested in the Community-Lab testbed and then deployed across approximately 70 Atom nodes, most of which are located in community members’ homes. It has a decentralized discovery service (Serf-based) and a small set of services, including social interaction, video streaming, and file storage (Dropbox-like), and the possibility to deploy new services such as a set of Linux containers (Docker-based).

**Routing Protocols**

One important self-management mechanism in a CN is routing. We studied, compared, and characterized the scalability, performance, and stability of three proactive mesh routing protocols: OLSRv1, BMX6, and Babel, three common routing protocols in WCNs, used in the networks involved in our testbed. These routing protocols have been further characterized by studying their control overheads, convergence delay, CPU and memory consumption, and stability. Our experimental evaluation results show the relative merits, costs, and limitations of the three protocols [7]. The results show a very lightweight Babel protocol to be too expensive in large, dense, and changing networks. In such scenarios OLSRv1 with the MPR mechanism and BMX6 seem to perform better, with lower control overheads for BMX6 but higher memory requirements than other protocols.

**Resource Sharing and Isolation**

Resource sharing is a critical mechanism for a testbed that allows multiple concurrent experiments, but isolation is also an important issue. Slicing of computing resources is done using Linux containers with mostly limited resources (CPU, storage). An experiment or experimental service takes a slice of the overall resources, a set of slivers (i.e., containers). The network can be virtualized for privacy reasons to isolate experiments from user traffic, or allow experiments to capture a fraction of anonymous production traffic (partially implemented and disabled by default for privacy reasons). Network virtualization can be useful for software-defined networking experiments where a slice contains its own set of virtualized slivers, its OpenFlow controller, and its virtualized network [11].
There are various self-managing systems necessary in large and dynamic networks: self-configuration (experiments in adaptive radio channel allocation, power, and rate), self-healing (experiments with services that adapt to node or link failures), and self-optimization (adaptation according to the available resources depending on internal or external influences). In these areas we have experimented with potential interference detection. Mitigation among different uses can be assessed through a monitoring system that collects detailed metrics about each node, link, and sliver, and an experimental resource interference detection and protection system for co-located slivers [13], protecting the most resource-sensitive processes in nodes with high utilization.

**BUSINESS MODELS, SYSTEMS, AND SERVICES**

Although CNs have already been studied from many angles, there is still insufficient understanding of the practices and methodologies that have given rise to such complex collaborative systems. We have analyzed guifi.net in detail [2]. It started in 2004 and is currently the largest CN worldwide with more than 30,000 routers (2015). The guifi.net model is an excellent example to make CNs sustainable and scalable. In fact, guifi.net has received the first ever European Broadband Award (2015) in the category of innovative model of financing, business, and investment.

The fundamental principles of guifi.net, in seeking to be fully inclusive, revolve around:
- The openness of access (usage) of the infrastructure
- The openness of participation (construction, operation, governance) in the development of the infrastructure and its community

These fundamental principles applied to an infrastructure result in a network that is a collective good, socially produced, and governed as a common pool resource (CPR).

The network is a collective good in which participants contribute their efforts and various goods (routers, links, and servers) that are shared to build a computer network, which combined with several Internet protocols, results in a peer property, provided that the community rules, in the form of a community license, are respected by all participants.

The development of a CN is a social or peer production because the participants work cooperatively, on a local scale, to deploy an infrastructure by building network islands, and on a global scale to share knowledge and coordinate actions to ensure the interoperability of the infrastructure deployed at the local level. The CPR is the model chosen to hold and govern the network. The participants must accept the rules to join the network and must contribute to the required infrastructure, while always maintaining the ownership of any hardware they contributed and also the right to withdraw.

Nonetheless, as with any other CPR, CNs are fragile. More precisely, they are prone to congestion, as connectivity is subtractable and subject to the free riding problem, because CNs are intentionally non-excludable. Thus, efficient and effective governance tools are needed in order to protect the core resource from depletion, that is to say, to protect it from the “tragedy of the commons.”

To build an effective governance architecture, it is essential to clearly identify the stakeholders based on interests, specific tasks, and potential conflicts of interest. These are volunteers, professionals and their customers, and public administrations. All participants who obtain connectivity must contribute to the infrastructure, directly or indirectly, and can participate in the knowledge creation process.

Figure 3 depicts the architecture of the governance tools. Effective means of communication are essential for any peer production project. The license establishes the participation framework and sets the boundaries of the CPR. The participation framework must be flexible enough to enable mechanisms for self-sustainability such as economic activity, but at the same time must safeguard the essence of the project. The monitoring system is essential for the network operation and accounting for resource usage. A coherent conflict resolution system avoids arbitrariness and minimizes the number of disputes. Through the mechanism of expenditures declaration, the participants indicate the resources they contributed, which must be accounted for. The collaboration agreement system regulates the for-profit usage of the resources. The economic compensation system balances the contributions that have been accounted for and the resource usage of for-profit participants. Finally, an authoritative organization recognized by all the participants is needed to operate the aforementioned tools and mechanisms.

**OPPORTUNITIES FOR AN INTERNET FOR ALL**

Community networks have received growing attention in the last few years as an alternative model used to develop networking infrastructure commons from a local and self-organized perspective, alongside traditional telecom models that cannot serve at least half of the population in the world. We hope to have contributed to this trend. The Internet Research Task Force (IRTF) GAIA Working Group emerged from our efforts, and became an active and global community exploring all models in order to bring connectivity to all. The Internet Society Wireless for Communities program is bringing resources in order to build capacity and help accelerate the development of community networks globally [5]. The Internet Governance Forum is also starting to study CNs for the same purpose. In South Africa, the Zenzeleni network, together with UWC and other organizations, is creating a second layer organization to expand networking cooperatives across rural South Africa [9]. The ongoing NetCommons.eu project in Europe is looking further at CNs from the technical, legal, socio-economic, and political perspectives for the next three years.

**CONCLUSIONS**

Network infrastructures are needed to provide Internet access to every person in the world, enabling them to participate in our global society and benefit from the many opportunities arising from being online. Underserved communities
have found ways to develop their own local network infrastructures. These self-organized infrastructures, developed using commodity network devices and open licensed spectrum, become a key local infrastructure, managed and governed as a common pool resource, with as other local community resources for the common interest. These community networks are an emerging and successful model for the future Internet, with hundreds of examples around the world.

However, these community networks face many challenges. For that reason, a large team of research and community organizations developed Community-Lab.net, which offers unified access to an open testbed with tools that allow researchers to deploy, run, monitor, and experiment with services, protocols, and applications on real-world community networks.

Research and development in the last five years has produced a large set of results (software, algorithms, models, evaluations) and improvements around resource allocation, routing, cross-layer optimized systems and services, benchmarking methods and tools, and detailed evaluations of social, technological, economic, and legal aspects in several developing and developed areas of the world.

ACKNOWLEDGMENT

This work was supported by the CONFINE Integrated Project 288535, the Fed4Fire project 318389, the NetCommons project 688768. We are grateful to a long list of organizations and community members involved in the development of the testbed, the research, and the social and technical experiments, and everyone involved in community networks who we met during at least five years of work.

REFERENCES


BIOGRAPHIES

LEANDRO NAVARRO (leandro@cs.bc.edu) is an associate professor at the Computer Architecture Department of Universidad Politécnica de Catalunya (UPC). He received his Ph.D. from UPC in 1992. Research interests include the design of scalable and cooperative Internet services and applications. He coordinates the Computer Networks and Distributed Systems (CNDS) research group at UPC. He coordinated the CONFINE IP project and currently coordinates the Erasmus Mundus Joint Doctorate in Distributed Computing (EMEDOC). He is a council member of the Association for Progressive Communications (apc.org).

Bart Braem holds a Ph.D. in computer science from the University of Antwerp. He is currently technology lead of the iMinds smart cities project City of Things where he studies large-scale experimentation on IoT in realistic conditions. His research interests include smart cities, large-scale networking, and IoT. He was the research WP leader of the EU FIRE project CONFINE, which developed and operated a community network testbed.

CHRISTOPH BARZ is a senior researcher at the Fraunhofer FKIE in the field of communication systems. He is working on secure and robust IP-based communications solutions and QoS for heterogeneous wireless networks as well as on their integration into core network infrastructures. He received his degree in computer science with a focus on communications systems from the University of Bonn in 2002. Until 2009, he was a research assistant at the University of Bonn.

FELIX FRITJAC is an associate professor at the Computer Architecture Department of UPC, which he joined in 1999. He received his Ph.D. in telecommunication engineering from UPC in 1998 in the area of speech recognition, neural networks, and statistical modelling. His research interests include the performance evaluation of distributed systems, integration of virtualization, and mechanisms for resource allocation in distributed systems. He led the EU funded project Community on Community Clouds.

ROGER BAG was the founding Board of the Neutral Free and Open Network – guifi.net in 2011 as a researcher and currently is actively involved in two EU projects, RIFE and NetCommons. He has been volunteering in the guifi.net community network since 2006. There he has actively contributed in many aspects such as events, projects, and talks. His research interests are in mesh routing protocols, socio-economics, and management of community networks.

JOSEPH BONICIOLI has walked the IT arena for more than 15 years. Since 2011 he has served for the second time as the president of the Non-Profit Association Athens Wireless Metropolitan Network. Since October 2011 he has been responsible for AVMN participation in the EU FP7 funded research project (IP) CONFINE. With a both technical and business minded attitude, he has been involved in projects involving wireless technology, community networks, cloud computing, network security, and high availability, among others.

ANJA VLAATJA-BLÅNGBERG is a computing engineer with a focus on networking, virtualization, and distributed systems based on free software and open standards and technologies, as well as the sociological impact of these in empowering people. He has experience in developing, designing, and administering systems for ISPs, NGOs, and SMEs. He has taken a main role in the design of the CONFINE architecture.
Toward Enabling Broadband for a Billion Plus Population with TV White Spaces

Animesh Kumar, Abhay Karandikar, Gaurang Naik, Meghna Khaturia, Shubham Saha, Mahak Arora, and Jaspreet Singh

ABSTRACT

One of the major impediments to providing broadband connectivity in semi-urban and rural India is the lack of robust and affordable backhaul. Fiber connectivity in terms of backhaul that is being planned (or provided) by the Government of India would reach only up to the rural offices (called Gram Panchayat) in Indian villages. In this exposition, we articulate how TV white space can address the challenge in providing broadband connectivity to a billion plus population within India. The villages can form local Wi-Fi clusters. The problem of connecting the Wi-Fi clusters to the optical fiber points can be addressed using a TV white space based backhaul (middle mile) network. The amount of TV white space present in India is very large when compared to the developed world. Therefore, we discuss a backhaul architecture for rural India that utilizes TV white spaces. We also present results from our TV white space testbed that support the effectiveness of backhaul by using TV white spaces.

INTRODUCTION

In the past decades, India has witnessed ever increasing wireless telecom connectivity. Currently, India has around 1 billion cell phone subscriptions with an overall tele-density\(^1\) of 79, and India is the second largest telecom market in the world. These numbers along with the advent of third generation (3G) and fourth generation (4G) cellular wireless systems would hint that broadband access in India has been solved; however, the reality is far from it. The total number of broadband subscribers, including wireless data plan subscriptions and wireless dongle users, is 120 million. Of this, only 16 million users are subscribed to wired broadband services. The rural India scenario is even worse. Rural teledensity is 48 compared to the urban teledensity of 147. The number of broadband subscriptions in rural India is even lower. According to the Telecom Regulatory Authority of India (TRAI), the current definition of broadband is ≥512 kb/s connection. By 2017, this definition of broadband will be upgraded to a 2 Mb/s connection. The targets of the Government of India are very ambitious: by 2020, the Government of India plans to have a broadband subscriber base of 600 million.

Based on the above mentioned numbers, it appears that broadband access (more so in rural and semi-urban areas) is a largely untapped market. However, there are significant challenges in providing broadband access in the rural areas, including:

- Low average revenue per user
- High capital and operation expenditure (including license fees)
- Affordable backhaul, which is exacerbated due to a very large population
- Energy cost, which is worsened by lack of reliable power supply
- Geographic accessibility issues such as right of way problems

To alleviate the lack of broadband in rural areas, the Government of India has been working with the initiative BharatNet (formerly National Optical Fiber Plan or NOFN). Within BharatNet, which is being implemented in two phases, points of presence (PoPs) with optical connectivity at all village offices, called Gram Panchayat, will be provided.

There are 250,000 Gram Panchayat in India. The total number of villages is 638,619; thus, each Gram Panchayat serves about 2.56 villages on average. Each village has around four hamlets at the periphery on average. As mentioned above, PoPs will be provided at Gram Panchayat with optical fiber backhaul by BharatNet. Since the villages can be at a maximum distance of a few kilometers from the Gram Panchayat, BharatNet will allay but not solve the problem of rural broadband in India. To address last mile access, in each hamlet or village, a wireless cluster can be formed (e.g., by using a Wi-Fi access point, AP), but backhaul of the data from APs remains a challenge. We envisage that TV white spaces (in the UHF band) can be utilized to backhaul data from village Wi-Fi clusters to the PoPs provided by BharatNet. It must be noted that mobility, at the moment, is not a major driver for broadband; instead, primary (fixed) broadband service is the biggest requirement in rural India.

---

\(^1\) Tele-density is measured by number of telephones per 100 people.

The authors are with the Electrical Engineering Department, Indian Institute of Technology Bombay.
This vision raises the following important questions: Can TV UHF band/TV white spaces be used to solve the above-mentioned backhaul problem? How is the TV UHF band/TV white space utilized in the rest of the world? How much TV white space is available in India, and how does it compare with other countries? What network topologies in the TV UHF band can be exploited to solve the backhaul problem? What results are obtained from an actual experimental testbed while performing backhaul in the TV UHF band over sparsely populated rural areas? These questions are subsequently answered in this article.

**TV WHITE SPACE OVERVIEW**

With rising demand for bandwidth by various applications, researchers around the world have measured the occupancy of spectrum in different countries. The observations suggest that except for the spectrum allocated to services like cellular technologies, and the industrial, scientific, and medical (ISM) bands, most of the allocated spectrum is heavily underutilized (c.f. [1–3]). The overall usage of the analyzed spectrum ranges from 4.54 percent in Singapore to 22.57 percent in Barcelona, Spain [1, 2]. Licensed but unutilized TV transmitters, TV white space, is a band of TV white space in the literature [4]. These white spaces in the TV UHF band-IV, 470–590 MHz, henceforth the TV UHF band for brevity, have been of particular interest due to the superior propagation characteristics (from a received signal strength standpoint). Its status in the world is reviewed next.

**TV WHITE SPACE IN VARIOUS COUNTRIES**

The amount of available TV white space varies with location and time. The available TV white space depends on regulations such as the protection margin given to the primary user, height above average terrain, transmission power of secondary users, and separation of unlicensed users from licensed ones. Since the actual availability of TV white spaces varies in both location and time, operators of secondary services are interested in the amount of available white space. TV white space estimation has been done in countries like the United States, United Kingdom, Europe, Japan, and India [5–9]. For instance, in Japan, out of 40 channels, on average, 16.67 channels (41.67 percent) are available in 84.3 percent of the areas [8]. The available TV white space by area in Germany, the United Kingdom, Switzerland, and Denmark on average ranges between 48 and 63 percent of the 40 TV channel bands [7]. It must be noted that in these TV white space studies, the International Mobile Telecommunications-Advanced (IMT-A) 698–806 MHz band is also included.

The Federal Communication Commission (FCC) regulations in the United States and the Office of Communications (Ofcom) regulations in the United Kingdom have allowed for secondary operations in the TV white spaces [4]. For example, FCC regulations declare a band as unutilized if no licensed user (primary) signal is detected above a threshold of –114 dBm [4]. Under this provision, a secondary user can use the unutilized spectrum provided it does not cause harmful interference to the TV receivers and it relinquishes the spectrum when a primary user starts operation.

**STANDARDS AND TECHNOLOGIES TO ADDRESS TV WHITE SPACES**

IEEE 802.11af has been designed by extending IEEE 802.11ac. IEEE 802.11af supports 8 MHz channels and uses a TV white space database to control or inform the use of spectrum by devices [4]. For spectrum usage and interference management, IEEE 802.11af uses carrier sense multiple access at the base station (BS) as well as the clients. The spectral efficiency of 802.11af varies from 0.3 to 4.5 b/s/Hz at the physical layer; as a result, its maximum throughput is 35.6 Mbps over the 8 MHz channel. IEEE 802.11a/b/g-based Wi-Fi can also be made to work in the TV band by appropriate changes in the radio frequency (RF) section. This approach has been adopted in our testbed.

IEEE 802.22 is also designed for enabling broadband wireless access in TV white spaces [4]. At the medium access control (MAC) layer, IEEE 802.22 uses orthogonal frequency-division multiple access (OFDMA). The spectral efficiency of IEEE 802.22 in a single antenna configuration varies from 0.6 to 3.1 b/s/Hz.

IEEE 802.15.4m is geared toward low-rate wireless personal area networks, with applications that include machine-to-machine networks. IEEE 802.19 defines an architecture and protocol for enabling coexistence between different secondary networks operating in TV white space. Finally, 1900.7 has been established for advanced spectrum management and next generation radios [4]. In the future, we expect more technologies to be designed for operation over TV white spaces, including Long Term Evolution (LTE) and IEEE 802.11ah.

**GEOLOCATION DATABASE AND WHITE SPACE DEVICE ACCESS RULES**

To ensure coexistence of TV broadcasters with secondary devices, geolocation databases have been mandated by FCC and Ofcom [4]. All devices should have a location accuracy of ±50 m and must query a certified TV white space database to obtain an allowable channel with associated transmit power. The list of available (unutilized by primary) channels, the channel access schedule for 48 hours, and the transmit power allowed is provided by the TV white space database.

**TV WHITE SPACE AVAILABILITY IN INDIA**

In India, the sole terrestrial TV service provider is Doordarshan. Currently, Doordarshan has 373 TV transmitters operating in the TV UHF band in India. The TV UHF band consists of 15 channels of 8 MHz each. In India, a small number of transmitters operate in the TV UHF band; as a result, apart from the 8–16 MHz band depending on the location, the TV UHF band is not utilized in India. Comprehensive quantitative assessment and estimates for the TV white space in the 470–590 MHz band for four zones of India have been presented in the literature [9]. It has been shown that in almost all cases at least 12 out of the 15 channels (80 percent) are available as TV white space at any location in India [9] (Fig. 1).

**TV UHF BAND UTILIZATION IN INDIA AND ITS POLICY ASPECTS**

Currently, there are no regulations for TV white space usage in India. The National Telecom Policy (NTP) 2012 of India notes that white space...
of rural broadband access.

a solution for solving the problem tower is cheaper. This exactly fits as therefore, the construction cost of power and a 10–15 m high tower. be reached with 4 W of transmit towers. Distances up to 10 km can characteristics of the TV UHF band, deploy technology. Due to the low-maintenance, and easy-to-areas where we seek a low-cost, not suitable for backhaul in rural require more maintenance. These towers are expensive and require more maintenance. Therefore, these technologies are not suitable for backhaul in rural areas where we seek a low-cost, low-maintenance, and easy-to-deploy technology. Due to the relatively superior propagation characteristics of the TV UHF band, we do not need very large height towers. Distances up to 10 km can be reached with a 4 W of transmit tower and a 10–15 m high tower. The tower is light, and therefore, the construction cost of tower is cheaper. This exactly fits as a solution for solving the problem of rural broadband access.

low-power devices, while protecting the primary or licensed user, will be considered in specific frequency bands. TV white space in India is significantly larger than that in the other countries reviewed above. The common approach for TV white space utilization is through the use of white space devices and associated country-specific regulations [4]. Such white space devices and regulations utilize the presence of database lookup, with transmit power limitations on the unlicensed user. While the 470–590 MHz band has been licensed for TV broadcasting, its usage for rural broadband can be fundamentally different in India. This fundamental difference is explained next.

India belongs to Region 3 of the International Telecommunication Union (ITU) terrestrial spectrum allocations. In the 470–590 MHz band, ITU permits fixed, mobile, and broadcasting services in Region 3 [10]. As per the National Frequency Allocation Plan (NFAP) of 2011, India Remarks 36 and 37 [10], fixed services in the 470–590 MHz band are allowed in India. This is in contrast with Region 1 (including Europe) where only broadcasting services are allowed in this band, and Region 2 (including the United States) where fixed services are allowed only in the 470–512 MHz band. This difference accommodates high-power transmissions by any fixed service, such as wireless broadband, in the 470–590 MHz band in India.

We suggest a registered shared access based regulatory approach that can be adopted to use TV white spaces in India. The operators will have to register themselves with a database before using the spectrum, and may have to share a channel or a sub-channel with other users in the vicinity. The operators might have to pay a usage fee for the spectrum, and they must cooperate with each other for achieving high average spectral efficiency. Spectrum sharing between multiple operators can be done by either centralized or cooperative distributed resource allocation. In centralized resource allocation, a central entity (e.g., a database) will allocate orthogonal resources to the operators according to their demand, while possibly considering fairness and bandwidth efficiency before allocation of resources. In the cooperative distributed resource allocation, listen before talk (LBT) can be used along with assistance from the central entity for sharing the channel. The key point in this regime is that the central entity will not do the resource allocation; it will only assist the operators in mitigating the interference between them.

**A BROADBAND ACCESS NETWORK TOPOLOGY FOR RURAL INDIA**

The TV UHF band in India is heavily underutilized [9], and its radio propagation characteristics are much better than those of unlicensed band such as 2.4 and 5 GHz. In fact, its radio propagation characteristics are suitable for non-line-of-sight connectivity. It is envisaged that a broadband access network can be provided by extending Internet coverage from a rural PoP (an optical fiber point such as the one provided by BharatNet) by using TV white space in the TV UHF band. Each village can be served by an unlicensed-band Wi-Fi AP. The end users (for economy of scale and ubiquity) will connect to the 2.4 GHz Wi-Fi AP. The connectivity between the PoP and the Wi-Fi AP can be provided in the TV UHF band, that is, a UHF-BS connected to the PoP and UHF customer premises equipment (UHF-CPE) connected to the Wi-Fi AP. The middle mile network can connect in different topologies:

- **Point-to-point**
- **Multipoint-to-point**
- **Multihop mesh network**

Of these, Fig. 2 illustrates the most general topology of a mesh network. The advantages of this broadband access network are as follows:

- A Wi-Fi access device is affordable for the rural population in India.
- Each link in the middle mile network can cover large distances due to TV UHF band propagation characteristics.
- Non-line-of-sight links can be formed.
- The power consumption of each UHF band radio is low (5–10 W in our testbed), so it can be powered through solar energy.2

Some testbeds have also been deployed in other parts of the world with a focus on rural broadband and the associated backhaul problem. In a very interesting approach named HopScotch [11], the primary focus is to provide wireless broadband access to rural areas using renewable resources and 5 GHz Wi-Fi. The approach also proposes to deploy “WindFi” base stations as an overlay over the TV UHF band. In a mobility-driven setup, a TV white space testbed has been used by Microsoft Corporation, Redmond, Washington, to backhaul data from a Wi-Fi cluster in a moving shuttle [12]. Trials in Africa by Microsoft in Tanzania and Kenya, and Google in the Republic of Liberia and South Africa have
extended broadband using Super Wi-Fi and TV white space [13]. ROSALNet [14] also discusses a middle mile backhaul testbed, but it has not been deployed in a rural setting. The next section highlights the results of the testbed implemented in a point-to-multipoint topology in a cluster of villages near Mumbai, India.

**The First TV White Space Testbed in India**

An instance of the point-to-multipoint network discussed in the previous section has been realized as a TV UHF band testbed in the Palghar district, Maharashtra, India, located at about 107 km distance from the Indian Institute of Technology Bombay. This is the first TV white space testbed of such scale in India. Seven villages in Palghar (as illustrated in Fig. 3) have been selected for the testbed, which extends over an area of 25 km². The Palghar district was selected for deploying the testbed as it represents a typical broadband scenario in a rural setting of India. Also, it is accessible by road from Mumbai, which made the deployment of the testbed less difficult.

**Network Topology of the Testbed**

The testbed comprises 10 UHF-CPEs spread over seven villages and 1 UHF-BS at Khamloli village where the PoP is located (and a 20 Mb/s leased line has been provisioned via an optical fiber link). The UHF-BS at Khamloli has an omnidirectional antenna with a gain of 8 dBi, and is mounted at a height of 30 m. There are four UHF-CPEs, each set up at a height of 30 m with antenna gain of 11 dBi. These result in Khamloli-Maswan, Khamloli-Haloli, Khamloli-Ganje, and Khamloli-Pargaon links (Fig. 3a). There is no line of sight in Khamloli-Maswan. There is heavy vegetation in the Khamloli-Ganje link. All other links have moderate vegetation. Note that there are significant differences in the topographic profiles between the UHF-BS and the UHF-CPEs.

There are seven UHF-CPEs named Khamloli 1, Khamloli 2, Bahadoli 1, Bahadoli 2, Dhuktan 1, Dhuktan 2, and Dhuktan 3 (Fig. 3b). Each of these UHF-CPEs is set up at a height between 4–6 m with an antenna gain of 11 dBi, with small houses between. Khamloli 1, Khamloli 2, Bahadoli 1, and Bahadoli 2 have near line of sight with the Khamloli UHF-BS. However, Dhuktan 1 and Dhuktan 2 do not have line of sight with the Khamloli PoP. Vegetation is large only in Dhuktan 3. Each kiosk is connected to a UHF-CPE via a Wi-Fi point-to-point link (Fig. 3b). At each location, a suitable site was identified for setting up the UHF-BS, UHF-CPEs, and associated accessories including power supply. At every location marked in the testbed layout (Fig. 3b), Wi-Fi APs have been installed to test the Internet connectivity. The end user equipments are Wi-Fi-enabled devices such as tablets or smartphones.

**The Developed Economical TV UHF Band Device Prototype**

Products based on IEEE 802.22 and IEEE 802.11af standards are available off the shelf, but are expensive (US$4000–5000 for a BS and US$1000–2000 for CPE). From an affordable broadband service point of view, we developed a low-cost prototype of a TV UHF band device that costs US$650/device.
The prototype developed at IIT Bombay consists of four major components: baseband processor, modulator (2.4 GHz standard Wi-Fi modulator), frequency down converter, and antenna, as shown in Fig. 4. The prototype comprises a commercially available off-the-shelf IEEE 802.11a/b/g wireless embedded board that can connect with a RF card having a mini PCI interface. The baseband processing is performed on the wireless embedded board, whereas the modulation as well as the frequency conversion from 2.4 GHz to TV UHF band are performed on the RF card. An embedded Linux kernel based operating system, OpenWrt, is ported on the prototype in order to configure the device for operation. The device was configured to a modulation and coding scheme of 16-quadrature amplitude modulation (QAM) and coding rate of 3/4. The receiver sensitivity of the device for this setting varies as –83 dBm to –80 dBm (depending on the bandwidth). A few pictures from our testbed illustrating the prototype box are shown in Fig. 5.

The prototype developed at IIT Bombay consists of four major components: baseband processor, modulator (2.4 GHz standard Wi-Fi modulator), frequency down converter, and antenna, as shown in Fig. 4. The prototype comprises a commercially available off-the-shelf IEEE 802.11a/b/g wireless embedded board that can connect with a RF card having a mini PCI interface. The baseband processing is performed on the wireless embedded board, whereas the modulation as well as the frequency conversion from 2.4 GHz to TV UHF band are performed on the RF card. An embedded Linux kernel based operating system, OpenWrt, is ported on the prototype in order to configure the device for operation. The device was configured to a modulation and coding scheme of 16-quadrature amplitude modulation (QAM) and coding rate of 3/4. The receiver sensitivity of the device for this setting varies as –83 dBm to –80 dBm (depending on the bandwidth). A few pictures from our testbed illustrating the prototype box are shown in Fig. 5.

THE COEXISTENCE HANDLER: A TV UHF BAND DATABASE FOR INDIA

The UHF-BS queries an OpenPAWS database to select the frequency of operation and avoid interference to any terrestrial TV services in other geographical regions of India [15]. The OpenPAWS client was implemented in OpenWrt (port on UHF-BS), while the OpenPAWS server was implemented in Linux. In response to UHF-BS query, OpenPAWS gives a list of available channels with transmit power allocations.

For the Khamloli village (PoP) in the testbed, the OpenPAWS database declared Channel 1 to Channel 15 as available for transmission. The transmit power was restricted to 30 dBm. After assigning the power and channel to the UHF-BS, the database is updated to reflect the same. The database created is open for public access [15] and can be used to view the list of all TV towers operating in India in the TV UHF band along with all operational parameters. It also displays TV towers operating on any particular channel along with the tower’s coverage area calculated in [9].

RESULTS OBTAINED FROM THE TESTBED

The results obtained from the testbed after extensive experimentation over a few months are highlighted in this section. These comprise throughput and latency analysis for various links. An automatic teller machine (ATM) use case is also explained toward the end of this section.

There are many point-to-point links in the testbed of which we exemplify one. This is a long-distance link with line of sight (Khamloli-Pargaon). For a bandwidth of 5 MHz, the obtained uplink and downlink throughputs (while using User Datagram Protocol [UDP] as well as Transmission Control Protocol [TCP]) are illustrated in Fig. 6. In the plots, observe that the received signal-to-noise ratio (SNR) varies with the variation in transmit power (0–27 dBm), keeping the receiver’s distance constant. Consistently, a throughput of 4–5 Mb/s (in TCP) and 5–6 Mb/s (in UDP) were obtained for the uplink as well as downlink. The results for 5 MHz bandwidth have been presented to demonstrate the reconfigurability of our device to 5 MHz rather than the conventional configuration of 20 MHz available in standard Wi-Fi based devices. This feature was introduced in the device to work in scenarios where the available bandwidth is limited. Experiments were also carried out for bandwidths of 10 and 20 MHz. UDP throughput of 10–11 Mb/s and TCP throughput of 7–8 Mb/s were obtained on the Khamloli-Dhuktan1 (2.3 MHz).
km) link for 10 MHz bandwidth. UDP throughput of 18–19 Mb/s and TCP throughput of 14–15 Mb/s were obtained on the Khamloli-Dhuktan1 (2.3 km) link for 20 MHz bandwidth. The results are not illustrated here due to lack of space. The tool used to monitor the network was iPerf.

The variation of latency was also examined. Two extreme ranges of wireless topography were considered: Khamloli-Ganje, which represents a 6.7 km long-distance link at 5 MHz bandwidth, and Khamloli-Dhuktan, which represents a 2.3 km moderate-distance link at 5 MHz bandwidth. The latency for the Khamloli-Ganje link varies from 2 to 15 ms, while its UDP throughput varies from 5.6 to 8 Mb/s. The latency for Khamloli-Dhuktan link varies from 2 to 11 ms, while its UDP throughput varies from 11 to 17 Mb/s.

The implications of the measured throughput in the average case are discussed briefly. The average population in India is about 1000 per village. Considering the income constraint of the rural population, it is assumed that the number of subscribers is limited to 300 per village. Hence, with a contention ratio of 1:50, as prescribed by the Telecom Regulatory Authority of India, about 6 active subscribers per village will be present at any given time. This will necessitate a rate of 3 Mb/s per village (considering 512 kb/s per subscriber). Based on our measurements, on average, one UHF-BS can easily serve two to three villages with 10 MHz bandwidth. The exact network planning will depend on the topography of villages and the availability of white space channels.

High-speed Internet access has been provided to the villagers in Khamloli, Bahadoli, and Dhuktan using Wi-Fi APs deployed at seven locations across three villages with one kiosk in each village. About 60 villagers from three villages and surrounding hamlets (called pada) have been trained as “e-Sevaks” (electronic service-man). An e-Sevak assists other villagers in using Internet services for simple tasks like filling out college forms, paying electricity bills, and booking train tickets. Kiosks set up in the three villages are run by these e-Sevaks on a daily basis for three hours for this purpose. The e-Sevaks have been given tablets to get familiar with these Internet services.

In order to demonstrate the use of e-Finance capabilities of TV white space for rural areas, an ATM provided by TATA Indicash has been set up at the Dhuktan Gram Panchayat (village office). A virtual private network (VPN) circuit is provisioned between Khamloli UHF-BS and the Tata Indicash data center by TATA Teleservices Limited. The ATM, located at Dhuktan Gram Panchayat, is connected to a Wi-Fi AP. This Wi-Fi AP is connected to the Khamloli UHF-BS over a TV UHF band link. At Khamloli, a security gateway appliance statically routes the packets coming from the ATM to the multiprotocol label switching (MPLS) leg of the VPN circuit and vice versa.

**CONCLUSIONS AND FUTURE WORK**

In this article, we have articulated how the TV UHF band can address the challenge in providing broadband connectivity to a billion plus population of India. As outlined in the article, one of the major impediments to providing broadband con-
nectivity in semi-urban and rural India is the lack of robust and affordable backhaul. Even in urban areas, one of the major impediments for widespread deployment of Wi-Fi clusters is the lack of connectivity from Wi-Fi APs to optical fiber gateways. Fiber connectivity in terms of backhaul that is being planned (or provided) by the Government of India would reach only up to the Gram Panchayat in rural areas. In such a scenario, the problem of connecting Wi-Fi clusters to the optical fiber PoP can be addressed using a TV white space backhaul (middle mile) technology.

We believe that a cost-effective solution for backhaul would require a database-assisted approach for TV white space spectrum management. Since the TV UHF band is sparsely utilized in India by the broadcaster, the challenge is not primary-secondary coexistence (as in many countries) but secondary-secondary coexistence. Multiple operators should be able to share the TV spectrum and coexist. While an LBT approach as in IEEE 802.11af is one of the options for coexistence, it will pose challenges in rural regions with large cell radius, due to the superior propagation characteristics of sub-gigahertz spectrum. A combination of database assisted and LBT is a topic for future investigation for providing “primary” broadband connectivity by possibly many local operators.

Dynamic resource allocation algorithms with fair sharing of resources between multiple operators (co-primary) is another area that requires more attention. As the TV band network is being proposed as a middle mile network for backhauling Wi-Fi clusters, both Wi-Fi APs and TV band radios can be controlled through a software defined networking controller. Deployment of middle mile fixed services can set the vision for 5G in India.

ACKNOWLEDGMENT

The authors would like to thank Ms. Rita Teotia, former Additional Secretary, Mr. J. S. Deepak, Secretary, Dr. Ashok Chandra, former Wireless Adviser, Mr. R. J. S. Kushvaha, Wireless Adviser, Mr. A. K. Bhargava, former Member (T), and Mr. A. K. Pattanaik, Senior Deputy Wireless Adviser, all from the Department of Telecom India; Mr. N. Srinath, Mr. Neeraj Dindore, Mr. Durga Prasad, and Mr. Anand Rai of Tata Tele-services Ltd.; and Dr. Anita Patil Deshmukh of PUkar for providing support for setting up this testbed. The authors benefited from extensive discussions and consultations with Dr. Apurva Mody, Chairman, White Space Alliance, and Dr. Ravina Aggarwal. The authors also benefited from discussions with Mr. Sudeep Singhal, Mr. Souvik Ghosh, and Mr. Rajeev Panyari in the planning of the testbed.

REFERENCES


BIOGRAPHIES

Arunesh Kumar (arunesh@iitb.ac.in) is an associate professor in the Department of Electrical Engineering at the Indian Institute of Technology Bombay. He obtained his Ph.D. degree from the Electrical Engineering and Computer Sciences at the University of California, Berkeley. He is an affiliate member of the IEEE Signal Processing Society (SPS) Technical Committee. His current research interests include sampling theory and quantization, statistical and distributed signal processing, and TV white space.

Abhay Karandikar (karandikar@iitb.ac.in) is currently an assistant professor in the Department of Electrical Engineering at the Indian Institute of Technology Bombay. He was the head of the Department from January 2012 to January 2015. He is the coordinator of the National Center of Excellence in Technology for Internal Security and a founding member and Vice-Chairman of the Telecom Standards Development Society (TSDSI), India's standards body for telecom. His research interests include resource allocation in wireless networks, heterogeneous networks, and rural broadband.

Gaurav Nak (gaurangnak@iitb.ac.in) is currently a graduate student at Virginia Tech. He received his B.E. degree from the University of Mumbai, India in 2012, and his M.Tech. degree from the Indian Institute of Technology Bombay in 2015. His research interests include wireless communications, dynamic spectrum sharing, and cognitive radio networking.

Meghna Khatur (meghna@iitb.ac.in) is currently a graduate student in the Department of Electrical Engineering at the Indian Institute of Technology Bombay. She received her Bachelors from Saradar Vallabhbhai National Institute of Technology, Surat, in 2013. Her research interests are TV white spaces and spectrum sharing in wireless networks.

Shubham Saxa (shubhamsaxa@iitb.ac.in) is a graduate student in the Department of Electrical Engineering at the Indian Institute of Technology Bombay. He completed his B.Tech. degree from West Bengal University of Technology. His current research focus is to develop low-cost wireless technologies to provide broadband Internet access to rural parts of India. His other interests include communication networks and machine learning.

Manav Agrawal (manav.a@iitb.ac.in) is a graduate student in the Department of Electrical Engineering at the Indian Institute of Technology Bombay. He completed his B.Tech. from Rajasthan Technical University in 2012. His research interests are wireless communications, TV white spaces, and affordable broadband access for rural India.

Jaipreet Singh (jaipreet.singh@iitb.ac.in) is a research engineer at the Infronet Laboratory, Indian Institute of Technology Bombay. He received his B.Tech. degree in 2010 from Kurukshetra University and his M.Tech. from Jaypee University of Information Technology, India, in the field of electronics and communication engineering with wireless communications as his research area in 2013. His research interests remain in the design and development of the low-cost-based solution in the wireless domain.
CALL FOR PAPERS AND PROPOSALS

IEEE ICC 2017 will be held at Palais des Congrès - Porte Maillot, Paris, France, 21-25 May 2017. Located in the heart of the City of Lights, IEEE ICC 2017 will exhibit an exciting technical program, complete with 13 Symposia highlighting recent progress in all major areas of communications. IEEE ICC 2017 will also feature high-quality Tutorials and Workshops, Industry Panels and Exhibitions, as well as Keynotes from prominent research and industry leaders.

Prospective authors are invited to submit high-quality original technical contributions for presentation at the conference and publication in the IEEE ICC 2017 Proceedings and IEEE Xplore. Proposals for Tutorials, Workshops, and Forums are also invited. Visit http://icc2017.ieee-icc.org for more details.

TECHNICAL SYMPOSIA

Cognitive Radio and Networks
Oliveira, Larkins, King's College London, UK
Wald Saad, Virginia Tech, USA
Communication and Information Systems Security
Peter Mueller, ETH, Switzerland
Cong Wang, University of Hong Kong, China
Communications Software, Services and Multimedia Applications
Maria G. Martinelli, King's College London, UK
Said Hoseini, University of Paris 12, France
Aaida Shamali, Western University, Canada
Communication Theory
Endre Ayanjo, University of California, Irvine, USA
Fahile Babakh, University of Trieste, Italy
Steven Weber, Drexel University, USA
Green Communications Systems and Networks
Cicen Cavdar, Royal Institute of Technology Stockholm, Sweden
Michel Mao, Palackichoro, Greece
Communications QoS, Reliability and Modeling
Dzymitry Kuzovkin, University of Luxembourg
Samir Souhi, UPC, France
Kohel Shimomoto, NTT, Japan
Next Generation Networking and Internet
Shihua Mao, Auburn University, USA
Mehran K. Marina, University of Edinburgh, UK
Siti-Mohammed Senouci, University of Bourgogne, France
Signal Processing for Communications
Michael Buckner, Virginia Tech, USA
Tomohiko Taniguchi, Fujitsu Laboratories, Japan
Optical Networks and Systems
Grazzop Dobrowolski, Poznan University of Technology, Poland
George Roukas, North Carolina State University, Raleigh, USA
Wireless Communications
Mohamad Assaad, CentraleSupelec, France
Azeddine Benslimane, University of Ottawa, Canada
David Darderi, University of Bologna, Italy
Hajm.jpgan Jan, Aalto University, Finland
Yahong Cao, Missouri University of Science and Technology, USA
Mobile and Wireless Networking
Mohamed Aloufouma, University of Alabama, USA
Mendi Bennis, University of Oulu, Finland
Jaket Ben-Othman, University of Paris 13, France
Shaoxi Wang, Nanjing University, China

INDUSTRY FORUMS AND EXHIBITION PROGRAM
IEEE ICC 2017 will feature several prominent keynote speakers, major business and technology forums, and a large number of vendor exhibits. Submit your proposals to the IFEEC Co-Chairs (Luis M. G. Cernea (L.M.G.Cernea@lboro.ac.uk) and Jamshed Khan-Jush (khanjush@l.qslcomm.org)).

TUTORIALS
Proposals are invited for half- or full-day tutorials in all communication and networking topics. For enquiries, please contact the Tutorials Program Co-Chairs (Hanna Bogucka (hanna.bogucka@psl.poznan.pl) and Luc Vandendorpe (luc.vandendorpe@louvainvalenciennes.be)).

WORKSHOPS
Proposals are invited for half- or full-day workshops in all communication and networking topics. For enquiries, please contact the Workshops Program Co-Chairs (Abbas JamaliPour (a.jamali@ieee.org) and Constantinos Papadis (cppap@alt.it)).

IMPORTANT DATES
Paper Submission: 14 October 2016
Tutorial Proposals: 11 November 2016
IFEEC Proposals: 11 November 2016
Workshop Proposals: 24 June 2016
Paper Acceptance Notification: 27 January 2017
Camera-Ready Papers Due: 24 February 2017

Full details of submission procedures are available at icc2017.ieee-icc.org
Recent years have witnessed a massive penetration of cellular systems in developing countries. However, isolated rural areas (sparsely inhabited by low-income population) have been disregarded because classical access and backhaul technologies do not ensure the return on investment. This article presents innovative techno-economical solutions to provide these areas with cellular voice and data services. We first analyze the general characteristics of isolated rural communities, and based on this information, low-cost solutions are designed for both access (using 3G access points) and backhaul networks (using non-carrier grade equipment as WiFi for long distances or WiMAX in non-licensed bands). Subsequently, a study of population-dependent income vs. costs is presented, and a new business model is proposed involving mobile network operators, rural operators, and infrastructure providers. In order to test these solutions, we have built two demonstration platforms in the Peruvian jungle that have allowed validation of the technical feasibility of the solution, verifying the business model assumptions and the scalability of the initiative.

**ABSTRACT**

Recent years have witnessed a massive penetration of cellular systems in developing countries. However, isolated rural areas (sparsely inhabited by low-income population) have been disregarded because classical access and backhaul technologies do not ensure the return on investment. This article presents innovative techno-economical solutions to provide these areas with cellular voice and data services. We first analyze the general characteristics of isolated rural communities, and based on this information, low-cost solutions are designed for both access (using 3G access points) and backhaul networks (using non-carrier grade equipment as WiFi for long distances or WiMAX in non-licensed bands). Subsequently, a study of population-dependent income vs. costs is presented, and a new business model is proposed involving mobile network operators, rural operators, and infrastructure providers. In order to test these solutions, we have built two demonstration platforms in the Peruvian jungle that have allowed validation of the technical feasibility of the solution, verifying the business model assumptions and the scalability of the initiative.

**INTRODUCTION**

According to data provided by the International Telecommunication Union (ITU), the mobile broadband market is the most dynamic in the telecommunications industry. It is estimated that, at the time of writing, 89 percent of the world’s urban population has third generation (3G) or Long Term Evolution (LTE) coverage. However, the situation is different for the rural population: only 29 percent will have 3G coverage on the same date [1].

Focusing on developing countries (DCs), only 34 percent of households have Internet access (in the lowest DC, the percentage is only 7 percent), in contrast to 80 percent in the developed world. This means that over 4 billion people are not yet connected to the Internet. Moreover, a significant number of African countries still do not report mobile broadband coverage in rural areas. As an example, in Guatemala an urban citizen is 12 times more likely to have Internet connectivity than a rural inhabitant. Differences are far from decreasing: in Colombia, the difference between rural and urban dwellers raised from 18 to 35 percent between 2009 and 2012 [2].

The challenge of connecting rural areas is not trivial. Classic infrastructures, requiring costly cell towers and satellite or optical backhaul, are not suitable. Only decisive actions from administrations (promoting or even fully funding rural infrastructure) and regulatory offices (extending the concept of universal service, creating new figures for rural mobile network operators [MNOs], or relaxing the quality of service (QoS) requirements in rural areas) may generate enough motivation to push researchers, equipment manufacturers, and MNOs toward rural inclusion.

Recent works have proposed very interesting bottom-up strategies (based on community initiatives) to expand voice services in rural areas through technologies such as OpenBTS [3], a combination of OpenBTS with WiFi [4], or WiFi mesh networks [5]. However, rural communities do not always have the knowledge, the legal capacity, or the resources to deploy those technologies. This article presents the approach adopted by the project TUCAN3G (funded by the European Commission), which proposes technically feasible but economically sustainable solutions for the progressive introduction of voice and broadband data services in small communities of rural areas of DCs, using conventional 3G cellular terminals.

**TECHNICAL AND SOCIOECONOMIC OBJECTIVES**

TUCAN3G proposes the introduction of 3G access points (APs) in outdoor environments, with heterogeneous backhauling in unlicensed bands using WiFi for Long Distances (WiLD) [6] and WiMAX to provide profitable mobile services to remote rural areas of DCs. The project was structured around three distinct pillars con-
considered as key for the development of a viable solution:

- Study of the technical feasibility of QoS-enabling access and transport solutions, adopting low-cost low-power-consumption technologies (the only source of power is solar panels) that facilitate the scaling up when the demand grows.
- The elaboration of a comprehensive and sustainable business case study based on a market survey that focuses on analyzing three key areas: service demand, cost of technology (split in capital and operational expenditures, CAPEX and OPEX), and financing models (including mixed public-private).
- The demonstrative pilot. The project deployed two networks that provide 3G services in the Department of Loreto, Peru. These networks validate in a real scenario the technology developed, and allow empirical testing of the hypotheses of the proposed business model.

**Rural Scenario and Use Cases**

The TUCAN3G project seeks global solutions that can be applied to most rural areas in DCs. To attain that goal, the Peruvian case is highly representative because three clearly identifiable regions are encountered: the coast (highly populated and well connected), the highland (poor and with some connectivity issues), and the jungle (flat terrain covered by the Amazon rainforest, with a very low population density and very severe problems of connectivity). While almost 50 percent of the urban population uses the Internet, only 11 percent of the rural population does. 30 percent of urban households have Internet access, while only 1 percent of rural households have it [7].

We can observe on the left of Fig. 1 the mobile service cells deployed in the country (small circles). Most of the populated locations (in brown) located in the forest departments (in green on the right of Fig. 1) lack cellular coverage. Similar conditions are found in the highlands. The correlation between cellular coverage and roads in rural areas is almost perfect. It is straightforward to conclude that unconnected locations correspond to the low-income sparse rural population who also lack electricity supply and roads.

The demonstrative pilot focuses on two regions in the Amazon forest as worst cases from the point of view of access to telecommunication. One is along the Napo River, and the other one is in Balsapuerto, along the Parana-pura River, both in the Department of Loreto (Fig. 1).

**Network Design Challenges and Solutions**

**Access Network**

TUCAN3G has departed from a traffic model provided by Telefonica del Peru (TdP) in rural areas in Peru both for day-long voice and data traffic per inhabitant. As for the long-term traffic evolution, TUCAN3G considers a traffic increase of 180 percent in the second year of service, followed by 5 percent in the third year and 2 percent in the fourth and fifth years. Covering these traffic demands, and being low-cost and low-power-consumption, has led to the selection of a 3G radio access network (RAN) low output power APs, which provide adequate coverage and require simpler installation than large base stations. This solution implies low CAPEX and allows a progressive deployment when demand increases. The approach adopted addresses several questions.

**Energy Provision:** The power consumption of micro base stations is around 100 W, while pico and femto APs have around 10 W and 5 W, respectively [8]. The energy dimensioning for the different sites has been calculated assuming commercially batteries (Ritar 12 V 100 Ah) and solar panels (Solar World 85 W/panel) under the worst case assumption that each small cell should work up to 24 h at full power during 3 days without solar radiation. The energy elements required per site appear in Table 1.

**Coverage Area:** Femto APs are suitable for outdoor scenarios if they are placed in high positions and can take advantage of line-of-sight propagation. Additionally, wooden walls create reduced path losses in outdoor-to-indoor channels. In those conditions, the APs deployed in the 800 MHz band licensed in Peru are able to reach a coverage area up to 2 km. The communities are placed 20–70 km apart, so interference is negligible.

**Number of APs Needed:** The communities considered generate low traffic, and the population is fairly concentrated, so one or two 3G APs of 16 (or 24) channelization codes collocated in the same tower at high positions are typically enough. The second AP is required to satisfy peak traffic demands in some scenarios, and operates at a different frequency. The network planning and energy provision dimensioning results are shown in Table 1.

**Adoption of Energy-Aware Self-Optimization Techniques:** Self-organization techniques enable quick unsupervised network management pro-

---

Figure 1. Base stations (left) and road network (right) in Peru (adapted from MTC-Peru). Red circles show the positions of the demonstrative platforms deployed by TUCAN3G.
Table 1. Access network parameters for the two networks deployed in TUCAN3G.

<table>
<thead>
<tr>
<th>Community</th>
<th>APs from ip.access</th>
<th>Antenna configuration</th>
<th>Solar panel units (P_{nom} = 85W)</th>
<th>Battery units (C = 1200 Ah × V)</th>
<th>Backhaul DL in 5 years</th>
<th>Backhaul UL in year 5</th>
<th>Inhabitants including people in itinerancy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Balsapuerto network</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Santa Clotilde</td>
<td>2 S-Class 16</td>
<td>Gain: 7db Dominant: 10° Height: 70m</td>
<td>3</td>
<td>3</td>
<td>1872 kb/s</td>
<td>864 kb/s</td>
<td>3222</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negro Urco</td>
<td>1 S-Class 16</td>
<td>Gain: 7db Dominant: 10° Height: 70m</td>
<td>2</td>
<td>2</td>
<td>1238 kb/s</td>
<td>608 kb/s</td>
<td>316</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Libertad</td>
<td>1 S-Class 16</td>
<td>Gain: 7db Dominant: 10° Height: 15m</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tuta Pishoo</td>
<td>2 S-Class 16</td>
<td>Gain: 7db Dominant: 10° Height: 50m</td>
<td>3</td>
<td>3</td>
<td>1066 kb/s</td>
<td>557 kb/s</td>
<td>287</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Varadero</td>
<td>2 S-Class 16</td>
<td>Gain: 7db Dominant: 10° Height: 30m</td>
<td>3</td>
<td>3</td>
<td>1987 kb/s</td>
<td>892 kb/s</td>
<td>948</td>
</tr>
<tr>
<td>San Juan</td>
<td>1 E-Class 24</td>
<td>Gain: 2db Dominant: 0° Height: 20m</td>
<td>1</td>
<td>1</td>
<td>562 kb/s</td>
<td>285 kb/s</td>
<td>118</td>
</tr>
</tbody>
</table>

Table 1. Access network parameters for the two networks deployed in TUCAN3G.

...cedures, which entails little human intervention and reduction of OPEX. In the rural scenarios self-organization implies reconsidering two paradigms usually overlooked in urban wireless cellular networks:
- Limited access to energy
- Stringent backhaul limitations

First, the integration of the status of the battery and the energy flow from solar panels in self-optimization mechanisms has led to the definition of a switch-off solution as a function of the daily traffic demand, which allows a reduction of 15–20 percent in the size of solar panels and batteries [9]. Second, self-allocation of primary sync codes has been implemented. Third, user association techniques: unlike conventional practice, which is based on the user association with the AP received with the strongest pilot signal, battery powered APs require energy-aware constrained association [10, 11]. Evaluated results indicate significant benefits in terms of bit rate for cell edge users, improving Jain’s fairness index by 30 percent with respect to conventional max-signal-to-interference-plus-noise ratio (SINR) criterion.

To account for the limited capacity backhaul, proper design of the access network is needed. Conventional schedulers address the backhaul limitations by imposing a maximum instantaneous rate as a function of the backhaul bandwidth information, which is an average measure and hence yields an excessively conservative solution. In contrast, TUCAN3G proposes a scheduler that specifically considers the average backhaul state information as a constraint to be satisfied in the long term, in addition to other constraints like the battery status and energy harvesting. This way, the system sum-rate and fairness can be considerably improved [12], and the system is made robust against energy outages and backhaul congestion. For instance, for a backhaul capacity of 500 kb/s, the difference between forcing an instantaneous constraint on the backhaul or taking the constraint in the long term results in an improvement around 200 percent in both the bit rate of the worst user and the system sum-rate [13].

**Transport Network and Interconnection to the Core**

Connection of the access network requires a low-cost transport network that ensures QoS and low energy consumption. The cost can be dramatically lowered by sharing part of the backhaul infrastructure among several locations, which in turn generates traffic management issues. Additionally, the use of low-cost technology operating in unlicensed bands can be adopted here for carrier-class deployments due to low or nonexistent interference. In this context, a combination of WiLD [14] and/or WiMAX systems operating in non-licensed bands is a viable solution for a multihop heterogeneous backhaul provided that the capacity is sufficiently high, and the per-hop delay and packet loss are controlled. The performance of point-to-point links has been theoretically analyzed and experimentally tested, proving that WiLD operating in 20 MHz channels allows capacities over 45 Mb/s at 20 km, and over 20 Mb/s at 55 km, while keeping the average delay under 5 ms and a negligible packet loss. WiMAX allows capacities of approximately 60 percent of those obtained for WiLD with 10 MHz channels, for the same delay and packet loss figures [14]. Higher capacities may be achieved under saturation conditions, but at the cost of QoS degradation.

Geostationary satellite links can be contemplated as a solution: they require simple Earth stations and low cost compared to non-geostationary. However, the downsides are the high propagation delay (a round-trip time in excess of...
...a recurring expensive monthly cost (around US$3000/dedicated Mb/s) compared to alternative solutions.

Ensuring End-to-End QoS in Multihop Backhaul Infrastructures: The QoS management must be designed in such a way that each AP perceives acceptable performance regardless of the traffic generated by other APs connected to the backhaul. In this sense, the main hurdle is adjusting the operation point per link in order to avoid saturation and limit delay and packet loss. Second, end-to-end traffic differentiation is a must (i.e., prioritizing voice and signaling traffic over data traffic). Third, it is fundamental to limit the traffic circulating across the multihop backhaul so that none of the links saturate. Such traffic limits must be imposed in a distributed way, as the traffic flows in shared links are introduced to the backhaul networks at different points. Figure 2 illustrates the concept: three edge routers and one gateway have to prioritize and shape the traffic in order to keep the rightmost backhaul link unsaturated. Inner nodes also have to collaborate in respecting priorities. Once the network elucidates how much traffic should be accepted to/from each AP, different strategies could be used to ensure end-to-end QoS over the shared infrastructure: distributed QoS architectures based on differentiated services (DiffServ) at the IP layer, or multiprotocol label switching (MPLS) at a lower layer. In [14] the backhaul architecture satisfying all these requirements is detailed, justifying the need for a traffic control system in every node, a function that can be efficiently performed by a low-cost embedded router. Results in [14] demonstrate that the proposed backhaul architecture achieves carrier-class QoS for voice and signaling traffic, while ensuring the best possible quality for data traffic.

Dynamic Transport Network Configuration: Any static solution is not adapted to reality. On one hand, the network might temporarily be able to accept more traffic than the nominal value from/to a given AP without any negative impact, provided that the edge nodes have enough information about the network state. On the other hand, the capacity of wireless links may be perturbed by weather or other environmental conditions. Additionally, nodes are likely to be powered with solar photo-voltaic systems, prone to power shortages. Beyond that, the possibility of dynamically changing parameters as a function of the network state must be examined so that either the performance is fostered or costs can be reduced. Hence, TUCAN3G has proposed dynamic mechanisms for resource allocation that take into consideration both the dynamic evolution of the state of the backhaul and the dynamics of traffic demands [13]. The problem of dynamic optimum distribution of the resources can be cast into a convex optimization problem. Experimental results show that the backhaul architecture adopted in the platforms is able to implement a distributed version of the optimization algorithm based on local information in each node.

Interconnection to the MNO Core Network: The traffic coming from/to the rural network is exchanged with the radio network controller (RNC) residing in the MNO core network. Switching that traffic across the distribution network and the core network may create several issues, depending on the technology chosen. The following aspects have to be considered:

- Conditions imposed by the AP manufacturer for the interconnection of the rural AP to the RNC
- Conditions imposed by the MNO on how the traffic has to be switched between the rural network and the RNC in the core network
- Considerations related to the inner architecture of the MNO’s core network, which is subject to strict security protocols, and might have architectural separations, either logical or physical, in circuit-switched (CS) and packet-switched (PS) networks, and/or in the data plane and control plane

To cover these requirements it was necessary to create virtual interfaces on the RNC to fit the architecture of the core network. Also, point-to-point VPN links were implemented between the controller and each edge router, which differed from the usual policies of TdP. Furthermore, in order to achieve communication between the RNC and AP via the transport network of TdP and installed heterogeneous networks, it was necessary to adopt various specific solutions and variations to the standard configurations, such as:

- Increase the maximum number of received signal code power (RSCP) retransmissions in the RNC in order to improve packet loss tolerance
- IPSec tunnels were implemented between each AP and RNC, and enable network address translation (NAT) in the edge router to allow traffic to be routed properly
- Signaling prioritization was implemented as part of a complex process of customizing the equipment configuration of the demonstration platforms

Sustainable Business Model

The previously described access and transport technologies have to be integrated into a sustainable business model. The first step is to analyze the structure of costs and compare it to con-
ventitional strategies. The analysis shows that the main difference is the backhauling deployment: MNOs usually deploy a satellite backhaul for each rural community with total ownership cost (TOC) that in five years doubles the TOC of low-cost wireless terrestrial backhauling [14].

The second step is to analyze how this solution could be applied in a real scenario considering the characteristics, expectations, and potential of key stakeholders: consumers, MNOs, and administration. The market survey sheds the following results:

**Demand Side:** 363 interviews were conducted in six rural populations. Results show that 42 percent of participants earn less than $140/mo. Despite their low and variable income, 69 percent own a conventional mobile phone and 11 percent a smartphone that are used when traveling to the urban areas. They are estimated to complete an average of 20 calls/mo (with cost of $0.20/call), and 60 percent of interviewed inhabitants were willing to pay over $3.6/mo for Internet access. These figures justify assuming an average revenue per user (ARPU) of US$7/mo, and a service penetration of 50 percent.

**Offer Side:** Large MNOs maintain a moderate to low interest in Peruvian rural areas. In addition to the high deployment and operating costs, the service quality established by law in rural and urban areas is similar, a requirement that is difficult to guarantee due to the time required to reach these isolated communities. If we also take into account that urban areas concentrate the largest volume of business, we can easily understand where most MNO investment strategy is focused. As a consequence, they are leaving a niche of opportunity for smaller MNOs or network communities willing to offer services as long as regulation allows it.

**Public Sector:** The Peruvian administration has decided to support the deployment of communications networks as a tool to promote local development and to bring administration services closer to the rural population. To accomplish this objective they play a dual role. On one hand, the government administrates public funds coming from taxes applied to companies with large business volumes (telecom operators, oil, mining, etc.) through FITEL, and devotes them to reducing inequalities between rural and urban areas by deploying new infrastructures through different mechanisms such as subsidies or loans that usually do not cover operation costs. On the other hand, the administration can act through regulation to establish incentives or enforce MNOs work in rural areas, establish new quality requirements, or promote innovative approaches.

Based on this information, several business models were proposed to cover both the demand for 3G services and the revenue expectations of large telecommunications operators. The most relevant are compared in Fig. 3:

- **Satellite:** satellite backhauling per location, which is the traditional approach and is used as reference
- **Dedicated:** dedicated terrestrial backhauling based on the low-cost technologies previously proposed, which would exclusively be used for 3G services
- **Shared:** shared terrestrial backhauling based on the low-cost technologies proposed earlier, which would be shared with other services (3G, fixed Internet access, etc.)
- **Existing towers:** terrestrial backhauling deployed over existing subsidized infrastructure, which would be exclusively used for 3G services

The figure of merit adopted for comparison is the minimum number of inhabitants per location, which would provide a positive margin between revenue and TOC in 5 years (assuming an ARPU of US$7 and a penetration of 50 percent). Results in Fig. 3 show that wireless terrestrial networks can reduce the number of inhabitants required to reach the break even point from 800 (with satellite backhauling) to 400. It is economically feasible to reach even smaller communities (around 200 inhabitants) with alternative business models based on sharing the wireless network with other services or taking advantage of previously mentioned public funds. This analysis is based on actual information from our high-cost deployments, which means that the TUCAN3G solution could reach even smaller communities in more accessible (but still isolated) scenarios, like the Andean region where the deployment cost is significantly lower.

Finally, it is important to note that wireless terrestrial backhauling could be deployed by an MNO but also by a tower-rental company, a local community, or a public administration. Therefore, it would be possible to combine local initiatives with public support.

**Demonstrative Pilots**

In order to implement mobile services on the six target communities, TUCAN3G has deployed an access network (Table 1) and a transport infrastructure. In all communities the coverage field measurements for the access network accurately confirmed the values issued by the access plan-
ning tool. The Napo network is connected to the MNO’s core network through a VSAT terminal, while the second connects to the MNO’s fiber optic distribution network directly in Yurimaguas city. The RNC is an NC200 (by IP.Access) installed in TdP premises in Lima.

The main components of the demonstration platform are displayed in Fig. 4 and described next.

**Napo Network**

This network consists of six APs and four backhaul hops that communicate via point-to-point links operated by WiMAX and WiLD routers at 5.8 GHz placed in 70 m high towers. The traffic supported by the backhaul concentrates on the gateway router located in Negro Urco, which connects to the MNO’s backhaul via a VSAT in Ku-band whose main network operating center (NOC) is located in Lima. The minimum end-to-end capacity in the Napo network backhaul corresponds to the WiMAX link, showing a stable capacity of 3.7 Mb/s for the uplink (UL) and 3.7 Mb/s for the downlink (DL) (the symmetry may be broken to ensure more capacity in the DL, although this has not been implemented because the available capacity exceeds what is needed according to the planning in Table 1). The single-input single-output (SISO) WiLD links have shown a minimum end-to-end capacity of 4.2 Mb/s in each direction and peaks of up to 8.11 Mb/s in each direction. The satellite link has a minimum of 1.2 Mb/s capacity for the upstream and the same for the downstream, although peaks are admitted up to 1.5 Mb/s for the upstream and 2.6 Mb/s for the downstream (capacities have been validated by injecting traffic with the iPerf tool). The round-trip end-to-end delay between APs in the Napo network and the RNC in Lima has been measured while the network was exposed to real traffic, having obtained an average value of 630 ms. In Tuta Pishco (WiMAX link), the delay is 60 ms higher due to the frame duration. Packet losses are below 1.7 percent and jitter below 0.2 ms in all links.

**Balsapuerto Network**

This network consists of three APs and three multiple-input multiple-output (MIMO) WiLD point-to-point links connected to TdP’s Yurimaguas station, and from there to the nation-wide optic fiber network. The capacity measured end-to-end across the three hops is 5 Mb/s UL and 15 Mb/s DL (measurements taken using the bandwidth tester embedded in Mikrotik systems acting as edge routers and gateway). The average round-trip end-to-end delay between femtocells and the RNC in Lima is 77 ms. All capacities and delays obtained meet the expectations and requirements of the access network planning. Packet losses are below 0.25 percent and jitter below 0.56 ms.

**Radio Network Controller**

The NC200 (Fig. 5) is responsible for separating/integrating CS and PS traffic, which are diverted from/to the mobile switching center (MSC) or serving GPRS support node (SGSN) as appropriate. The RNC aggregates traffic from the APs through the Iuh interface and provides standard Iu-CS and Iu-PS interfaces to the core network, carried over IP.

From the experience of deploying the demonstration pilots, some lessons can be drawn.

1. Platform operation in 3G cellular licensed bands. The cost of spectrum licenses is very high, so it is not easy to get a permit of use in a pilot project. In our case, TdP is part of the consortium and is providing service to mobile users. Its full cooperation for spectrum use and import of equipment has been necessary.

2. Costs. The cost of each AP is in the range of US$1000, whereas the cost of the RNC is in the range of hundreds of thousands of U.S. dollars. Clearly, the deployment only makes sense if the RNC is going to provide service to many nationwide APs. In addition, there is also a high cost
The Andean Development Corporation and the Regional Government of Loreto have offered additional funds to extend coverage to the entire basin of the Napo River, adding another 15 locations to the pilot demonstration project. The verification of the business model in this larger sample will reinforce the findings of our study.

Figure 5. A radio network controller and its integration in the access and core networks.

for the integration of the RNC in the core of the MNO, which normally involves interacting with other manufacturers. Also, our pilot demonstration is leveraging existing infrastructure (tower) that the EHAS Foundation had installed in these locations for previous telemedicine projects. The size and cost associated with the towers in the jungle (average height of 70 m and costs above US$20,000) are much higher than in mountain areas (about 12 m and costs below US$3000).

3. Continuity of service and penalties. In many countries, the law enforces maintaining mobile service once it is launched. Moreover, in many cases, penalties for low QoS and blackouts are the same for rural and urban areas, which discourages MNOs from participating in pilot testing or innovative projects. In TUCAN3G, TdP and FITEL were partners of the consortium, which helped to negotiate a consensus position: should the project have proved economically non-viable in the long term, the operator would have maintained the voice service in the communities.

4. The resistance or acceptance of MNOs of these deployments depends largely on the position of the public sector. The approval of legislation to facilitate the introduction of new players in the rural sector (in the form of rural mobile operators) and more flexible conditions of service delivery will help to bridge the coverage gap in these remote areas.

5. It is crucial to build mutual trust relationships with the communities: such projects require the participation and understanding of the population, their acceptance, and formal approval. Usually, the expectation of the people is very high, and they take an active role in proactively reporting incidents during the works and the stabilization process of the platforms.

CONCLUSIONS

The pilot deployment has been operating since October 2015, and has reached an average of 40 daily calls per location. It is believed that 70 daily calls will be reached soon and make the service sustainable.

Several international development agencies have shown interest in the TUCAN3G initiative. The Andean Development Corporation (CAF) and the Regional Government of Loreto have offered additional funds to extend coverage to the entire basin of the Napo River, adding another 15 locations to the pilot demonstration project. The verification of the business model in this larger sample will reinforce the findings of our study.

In conclusion, TUCAN3G has developed technology adapted to the rural reality of DCs. It has also verified the technical, economic, and social viability of a model providing a 3G service alternative to the current one, whereby administration, rural infrastructure MNOs, and large MNOs can jointly provide coverage to remote areas usually neglected in DCs.

ACKNOWLEDGMENTS

This work has been funded by the European Commission FP7 program through project TUCAN3G (IST-601102 STP), and also by Fondo de Inversión de Telecomunicaciones de Perú (FITEL). The authors acknowledge the contribution of all partners: UPC, URJC, PUCP, UCAUCA, TdP, TIWS, IPAccess, EHA, CREPIC, FITEL, and KINNO.

REFERENCES


**Biographies**

**Andrés Martínez-Fernández** received his telecommunications engineer (1994) and Ph.D. (2003) degrees from the Polytechnic University of Madrid (Program of Biomedical Engineering and Health Technology). He is currently an associate professor in the Department of Theory of Signals and Communications, Universidad Rey Juan Carlos, Madrid. He is also a director of the EHAS Foundation. His current research interests include telemedicine and low-cost telecommunication systems for less developed regions.

**Josep Vidal** (josep.vidal@upc.edu) received his Ph.D. in telecommunications engineering (1993) from Universitat Politècnica de Catalunya, where he is currently a professor. He has coordinated five collaborative EC-funded projects in different areas of MIMO relay communications, self-organization, cooperative transmission, and heterogeneous networks, areas where he has authored +170 journal and conference papers. He served as Associate Editor of IEEE Transactions on Signal Processing and is a member of the IEEE ComSoc Signal Processing for Communications and Electronics Technical Committee.

**Francisco Javier Simo-Reigadas** received his telecommunications engineering and Ph.D. degrees from the Polytechnic University of Madrid, Spain, in 1997 and 2007, respectively. Since 1999 he has worked in the field of ICT for development, initially as a practitioner in Africa, later as a researcher with the EHAS Foundation, and since 2005 as an associate professor with Rey Juan Carlos University. His main fields of research are broadband wireless rural networks.

**Ignacio Prieto-Escudo** received his Telecommunication Engineer degree from Carlos III University in Madrid. Nowadays he works as a visiting lecturer at King Juan Carlos University and is the Project Director of the EHAS Foundation. His research is mainly focused on applications of broadband wireless technologies to provide ICT services in remote areas of developing countries, and he has coordinated several projects on this issue.

**Adrián A. Paco Fernández** is the coordinator of the Rural Telecommunications Group of the Pontifical Catholic University of Peru, the leading university in the country. He received his M.S. degrees in telecommunication (2000) and electronic engineering (2002) and his Ph.D. degree (2006) from Universitat Politècnica de Catalunya (UPC), Barcelona, Spain. After working at Indra-Espacio, he joined the Signal Theory and Communications Department at UPC in 2002, becoming a research associate in 2008. He has participated in seven European Commission funded projects, the latest ones being TROPIC and TUCAN5G. His research interests include wireless interference management.

**Juan A. Paco Fernández** is the coordinator of the Rural Telecommunications Group of the Pontifical Catholic University of Peru, the leading university in the country. He received his M.S. degrees in telecommunication (2000) and electronic engineering (2002) and his Ph.D. degree (2006) from Universitat Politècnica de Catalunya (UPC), Barcelona, Spain. After working at Indra-Espacio, he joined the Signal Theory and Communications Department at UPC in 2002, becoming a research associate in 2008. He has participated in seven European Commission funded projects, the latest ones being TROPIC and TUCAN5G. His research interests include wireless interference management.
Despite improvements and expansion of cellular coverage in developing regions, a substantial qualitative divide remains. Maps that display the presence or absence of cellular coverage mask critical differences in infrastructure performance and client load. In order to illuminate challenges faced by users of such mobile networks, we collect and analyze GSM network measurements at the local scale. We discover higher network congestion in developing regions as well as performance differences between available carriers in each location. Based on our findings, we propose an app, called SmartCell, that informs and empowers users in near real time to seek out improved mobile connectivity.

INTRODUCTION

A look at the worldwide mobile cellular subscription rate indicates that by the end of 2015, there will be more than 7 billion mobile cellular subscriptions, corresponding to a penetration rate of 97 percent.\(^1\) While this seems like an immediate cause for celebration, a deeper look is required to more fully understand what this number represents. If we parse the numbers by region, we find that in 2015, 78 percent of residents of Europe and 77 percent in the Americas had mobile broadband subscriptions, compared only 42.3 percent in Asia and the Pacific, and 17.4 percent in Africa. Breaking coverage into geographic region begins to give us an understanding of access differences; however, stopping there masks critical disparities in the quality of available cellular connectivity. In particular, the cellular technology (Long Term Evolution, LTE, vs. third generation, 3G, vs. 2G), the data rate, the number of users per base station, and the cost of the subscription are all critical factors that play into the persisting inequities between cellular users in developed and developing regions. Simply looking at who has access does not give an indication of whether the quality of that access is equivalent for different users in different locations.

To dive deeper into these disparities, we conducted a cellular infrastructure measurement campaign in three locations of diverse population characteristics. Specifically, we passively observed, collected, and analyzed messages broadcast by cellular base stations in order to assess coverage quality and usability. We collected measurements in San Cristóbal Verapaz, Guatemala, a city of roughly 20,000 in one of the poorest countries in Latin America; in the Za’atari refugee camp in Jordan, the oldest and largest refugee camp in Jordan with a population of roughly 80,000 in only 6 km\(^2\); and, for comparison, in Santa Barbara, California, a community of about 90,000 residents in 42 mi\(^2\). Through our measurements, we reveal a number of interesting and important anomalies about coverage in these different locations that begin to paint a clearer picture of cellular quality divides. Most importantly, we discover chronic network congestion in some of the networks, which leads to a consistently poor quality of experience for associated users. Interestingly, we find that in our measurement locations, there are multiple providers available, which have varying levels of traffic load, ranging from minimal load (completely uncongested) to very heavy loads (chronically congested).

Residents in developing countries are acutely aware of coverage disparities between carriers and often use SIM cards from multiple carriers, switching between carriers depending on location. Our interviews with Za’atari residents indicate that they typically carry SIM cards for each available provider, but do not have any specific algorithm for selecting which to use at a given time. Reduced cost calling and texting between users on the same carrier creates a network effect, and may cause users to prefer a given SIM simply because the provider is the most popular. This leads to uneven traffic distribution and load between carriers.

Based on our findings, we propose a system called SmartCell, an Android application that gives users awareness of quality of service by detecting congestion on their cellular base station. SmartCell informs the user when congestion is detected, thereby empowering her with information about the local network. With this information, the user has multiple options for obtaining better quality of service. Most simply, the user could decide to use one of her alternate SIM cards, switch to an underloaded network, and obtain higher quality service. If she has a multi-SIM phone, SmartCell can automatically select a SIM from a less congested carrier to use for voice calls, SMS, and mobile data traffic. Without information about the quality of the

---

given cellular network provided by SmartCell, the user is not informed that a better connectivity option is available, and is likely to continue to suffer with unacceptable performance.

**RELATED WORK**

Two common approaches to understanding cellular performance are application-level studies that look at end-to-end characteristics and radio-level studies that focus on the access link. Our measurement of the cellular infrastructure is a radio-level study that specifically utilizes messages broadcast over the GSM air interface. Recent work has illustrated the potential impact that cellular radio state has on end-user experience [1–3], and how air interface messages can be used to infer cellular user activity [4, 5]. Our work demonstrates how detailed, small-scale analysis can support research on local-level infrastructure. This approach is in contrast to efforts to measure network performance on a global or nationwide scale, such as the FCC’s Measuring Broadband America,[2] and enables us to separate performance attributable to client-facing infrastructure serving the measurement locations from that is related to a carrier’s core networks.

Our system explores advising users to manually switch between cellular providers based on observed quality of service. In April 2015 Google announced Project Fi, which shifts users between the T-Mobile and Sprint networks based on expected data speeds.[3] The goals of Project Fi and SmartCell are thematically related, but the implementations are distinct due to Google’s integration with two existing commercial providers and active user switching. Additionally, while SmartCell infers performance of BTSs through passive and independent observation, we expect Google has access to back-end carrier metrics for determining performance. The always-best-connected concept [6] also touches on the use of multiple networks; however, this work assumes business relationships between providers. We do not require such relationships and simply inform the user of current conditions, leaving the choice of using an alternate carrier to users themselves.

**BACKGROUND**

Our measurement studies are focused on GSM networks as that is the most prevalent cellular technology in our selected communities. In this section we provide an overview of the GSM system and control messages we use for infrastructure characterization. We then briefly describe the locations we use for our case study.

**GSM TECHNOLOGY**

The GSM core network is a complex hierarchy composed of many logical entities. The pertinent objects for our study are mobile stations (MSs) and base station transceivers (BTSs). The MS is the user device, commonly a mobile phone equipped with a SIM card. BTSs are the components that communicate with the user device over the air interface. MSs communicate with only one BTS at a time.

GSM control messages enable the inference of BTS control channel congestion serving the measurement locations. Figure 1 displays the message sequences that take place when an MS requests a private communication channel from the BTS, which is necessary for voice/SMS or data. We capture messages broadcast from the BTS to the MS (solid arrows in Fig. 1) on the common control channel. When an MS needs to use the network for a voice/SMS or data session, it issues a channel request to the base station. If a channel is available, the base station responds with an immediate assignment success message that provides information about the available channel. A BTS operating at full capacity such that it is unable to allocate a channel will issue an immediate assignment reject message, in accordance with the Third Generation Partnership Project (3GPP) 04.08 specification, indicating no channel is available. Because these control messages are broadcast to all MSs connected to the BTS, we can leverage the immediate assignment success rate to approximate standalone dedicated control channel (SDCCH) blocking [7, 8], a GSM industry-standard key performance indicator (KPI) used to measure BTS congestion and inform capacity expansion. We calculate the immediate assignment success rate by dividing the number of observed successful immediate assignment messages by the total number of immediate assignment (success and reject) messages.

**Message Capture:** We capture broadcast messages using a total of eight cellular phones, which consist of Samsung Galaxy Nexus, S2, and Galaxy S4 handsets with radio debug mode enabled, which logs all cellular communications to a computer via USB. We use xgpdmon, an open source tool that converts debug logs into packet capture files using the GSMTAP pseudo-header. Each phone records all of its own uplink traffic as well as all broadcast traffic sent by BTSs. Critically, our message capture is non-invasive and non-intrusive. Messages broadcast over common control channels are received in plaintext and intended for all MSs connected to a BTS. The GSMTAP pseudo-header does not include private user data. In all locations we capture using two phones per carrier, one set to prefer 2G and another set to prefer 3G. We do not capture 4G LTE as LTE is only available in the U.S. location.

---

1 https://www.fcc.gov/measuring-broadband-america
2 https://fi.google.com/about/faq/network-and-coverage-1
3 https://www.3gpp.org/DynaReport/0408.htm
We analyze GSMTAP captures from three locations, each of which represents a different level of development, in order to discover manifestations of cellular infrastructure digital divides. In this section we provide background on each of the locations.

**San Cristóbal Verapaz, Guatemala:** Guatemala is one of the poorest countries in Latin America with a GDP per capita of around $3512. Despite this, mobile carriers within Guatemala have invested heavily in cellular infrastructure, providing connectivity to a large portion of the country. We collected measurements in the rural city of San Cristóbal Verapaz over eight days in January 2015.

Three GSM carriers offer service in Guatemala: Tigo, Claro, and Movistar. All three are available in San Cristóbal Verapaz. Tigo is Guatemala’s most popular carrier and has built the most extensive cellular infrastructure network. Claro is the least popular carrier in San Cristóbal Verapaz, and coverage is more focused on 3G in urban areas. Movistar is the second most popular, and its network consists of a higher percentage of 2G base stations compared to the other carriers. It is difficult to compare prices for the carriers as there is no equivalent package offered on all three. Generally, Tigo is slightly more expensive than the others (e.g., GTQ30 for 500 MB vs. GTQ25 for 500 MB from Movistar). Movistar is the most affordable carrier, and prepaid scratch cards used for credit are widely available.

**Za’atari Refugee Camp, Jordan:** Za’atari is the largest of four refugee camps in Jordan. Located in a previously rural desert region near the border with Syria, the camp was opened in July 2012, and its population quickly rose to 120,000. It currently hosts about 80,000 residents in roughly 6 km². We collected cellular measurements in Za’atari over three days in January 2015.

Mobile service in the camp is offered by three carriers: Zain, Umniyah, and Orange. Much of their infrastructure is located outside the camp and was initially built to serve the surrounding rural community that existed prior to the camp’s development. We discussed carriers with camp residents and were told that Orange is popular in the more urban areas of Amman, while Zain is popular for rural customers. Umniyah was generally not used by people with whom we spoke. Zain is the most popular carrier in the camp as SIM cards for the network are given to residents by camp administration as they first arrive and register at the camp.

**Santa Barbara, California:** For comparison purposes, we capture messages on two cellular carriers in and around Santa Barbara, California (~90,000 residents) including a mix of urban and suburban base stations. We focus on the two major GSM carriers in the United States: AT&T and T-Mobile. We collected traces over roughly 10 days for AT&T and five days for T-Mobile in October 2015.

### Cellular Measurement Analysis

We investigate network performance and congestion via immediate assignment success and rejection messages. As described earlier, an MS issues a request to the BTS for a communication channel at the time when the MS needs to use the network for voice/SMS or data. In response, the BTS broadcasts an immediate assignment message to inform the MS whether or not its request is granted.

<table>
<thead>
<tr>
<th>Capture location</th>
<th>Capture duration (s)</th>
<th>Mean immediate assignments per second</th>
<th>Total immediate assignment success messages</th>
<th>Total immediate assignment rejection messages</th>
<th>Immediate assignment reject percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT&amp;T (U.S.)</td>
<td>840,114</td>
<td>0.04</td>
<td>34,539</td>
<td>9</td>
<td>0.03%</td>
</tr>
<tr>
<td>T-Mobile (U.S.)</td>
<td>379,897</td>
<td>0.05</td>
<td>19,112</td>
<td>2</td>
<td>0.01%</td>
</tr>
<tr>
<td>Claro (Guatemala)</td>
<td>625,051</td>
<td>0.01</td>
<td>7661</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>Movistar (Guatemala)</td>
<td>625,073</td>
<td>0.21</td>
<td>127,996</td>
<td>1055</td>
<td>0.82%</td>
</tr>
<tr>
<td>Tigo (Guatemala)</td>
<td>622,340</td>
<td>0.21</td>
<td>129,837</td>
<td>846</td>
<td>0.65%</td>
</tr>
<tr>
<td>Orange (Jordan)</td>
<td>243,678</td>
<td>0.09</td>
<td>21,675</td>
<td>1355</td>
<td>5.88%</td>
</tr>
<tr>
<td>Umniyah (Jordan)</td>
<td>243,678</td>
<td>0.12</td>
<td>27,385</td>
<td>1550</td>
<td>5.36%</td>
</tr>
<tr>
<td>Zain (Jordan)</td>
<td>243,678</td>
<td>0.12</td>
<td>19,804</td>
<td>8800</td>
<td>30.76%</td>
</tr>
</tbody>
</table>

Table 1. Overall mobile network statistics.

The asynchronous nature of congestion across providers implies that users can potentially achieve improved connectivity and performance by switching points of attachment when the BTS or carrier on which they are connected experiences congestion.
0.04 and 0.05 immediate assignments per second, respectively. These results could indicate networks comprising a larger number of BTSs, where each BTS serves a smaller area and hence a smaller number of users. Claro also exhibited very few messages. We believe this may be caused by Claro’s lack of popularity in our measurement location. Tigo and Movistar exhibit similar message loads and have the highest rates of immediate assignment messages, roughly four times that of the U.S. networks. The three carriers in Jordan have similar messaging rates compared to one another, and fall in between the observed load on the U.S. networks and those in Guatemala.

**BTS CONGESTION**

In terms of congestion, we find qualitative performance divides between locations. Predictably, we capture almost no immediate assignment rejection messages on the U.S. provider networks. This indicates adequately provisioned BTSs for the given client load. Apart from Claro, which has a very low load, the Guatemalan networks experience higher rejection rates (minimum 21 times higher than AT&T, the U.S. carrier with the highest rejection rate). Turning to the refugee camp network, we see an even greater divide. The Jordanian carriers all have vastly higher rejection percentages than in the other locations, with Orange and Umniah both experiencing more than 150 times higher rejection rates than AT&T. Zain, the most popular carrier in the camp, has an extremely high rejection message rate (more than 1000 times higher than AT&T). For perspective, prior work considers a 5 percent rejection rate (more than 1000 times higher than AT&T).

While the rejection percentages may seem low, their impact can be drastic. Each rejection includes a value for a backoff timer that forces the MS to wait up to 255 s before resubmitting a resource request. In terms of user quality of experience, such backoffs result in very poor performance. For instance, the GSM specification classifies call setup times, the time between a user pressing the call button and the call connection, into three categories. Fast” call setup is defined as 1–2 s, “normal” 2–5, and “slow” 5–10. We analyze rejection messages in the traces and find the Jordanian networks frequently send backoffs of up to 128 s during congestion events. Clearly, the additional time attributable to backoffs we observe exceeds call setup times far beyond the classification of “slow.”

**TEMPORAL CONGESTION CHARACTERISTICS**

Because the carriers serving Za’atari experienced the greatest congestion, we take a deeper look into carrier-specific performance in the traces to investigate any cross-network traffic load relationships. Figure 2 shows the percent of immediate assignment messages that were rejections in one-minute bins on January 6, 2015 in Za’atari.

![Figure 2. Percent of immediate assignment messages that were rejections in one-minute bins on January 6, 2015 in Za’atari.](gsmts_0267v050001p.pdf)

We see that the carriers’ rejection patterns do not resemble one another. The Orange network was congested in short, severe bursts. It also appears as though congestion on Orange was higher at particular times of the day that correspond with weekday schedules (i.e., before and after lunch). Umniah, on the other hand, exhibited almost no evidence of congestion throughout the day. The Zain network experienced sustained congestion, occurring throughout the day and frequently reaching rejection percentages above 50 percent.

The asynchronous nature of congestion across providers implies that users can potentially achieve improved connectivity and performance by switching points of attachment when the BTS or carrier to which they are connected experiences congestion.

**SMARTCELL**

Informed by our analysis, we propose SmartCell, an Android application that helps users avoid congested cellular base stations. SmartCell observes GSM control channels, allowing it to passively detect when the BTS serving a user is congested. Essentially, SmartCell exposes the SDCCH blocking KPI in near real time to end users; such metrics are typically only known to the cellular carriers themselves.

The operation of SmartCell depends on a user’s preferences and the features of their mobile device. SmartCell periodically logs a timestamp, mobile network code, cellular base station ID, approximate location, and a BTS availability (BTSA) value based on observed cellular network congestion. If permitted by the user, SmartCell may share these measurements with other SmartCell users by periodi-
SmartCell relies on a diagnostic interface of the phone’s cellular baseband to collect the cellular broadcast messages used to detect congestion. This data collection technique has been used to detect baseband attacks, and is supported on a wide range of Android handsets with popular baseband chipsets.

Figure 3. BTS availability (BTSA) for carriers in Za’atari on the morning of January 6, 2015.

SmartCell relies on the use of multiple SIM cards, which is common in developing regions. Unlike areas where “good” cellular connectivity is ubiquitous, carriers in rural and developing regions often have vastly disparate coverage and quality of service. In contrast to the multi-year cellular contracts common in countries like the United States and Japan, the availability of low-cost and contract-free prepaid SIMs enables users in developing regions to use multiple cellular networks. This leads to users carrying SIM cards from multiple carriers and switching as needed to obtain acceptable connectivity. For example, the number of cellular subscriptions per 100 residents in Guatemala is 106.6, and in Jordan it is 147.8, while in the United States it is 98.4. Respondents to the survey administered in Za’atari use three SIM cards on average; switching SIMs to take advantage of less congested networks, “same network” discounts, and cheaper data-only plans. This level of comfort in switching carriers is promising for a system such as SmartCell.

Switching SIMs traditionally requires powering off the phone. However, phones that support multiple SIM cards are increasingly popular. OpenSignal reports 57 percent of Android users in Guatemala have multi-SIM phones, and more than 50 percent of Android users in several other developing countries own dual-SIM phones. When SmartCell is used on a multi-SIM phone, physically swapping SIM cards is unnecessary. Instead, SmartCell can take the user directly to the SIM settings activity, allowing the user to select which SIM to use for voice calls, SMS messaging, and data traffic. Alternatively, SmartCell can reconfigure the phone automatically, attempting to select the least congested mobile network in the area.

**SMARTCELL OPERATION**

SmartCell observes cellular control messages, specifically immediate assignment and immediate assignment rejection messages, to understand network availability and load on a user’s current BTS. Recall that these messages are BTS-specific and allow an attached MS to infer SDCCH blocking affecting not only itself, but also other MSs connected to the BTS. SmartCell detects these messages and uses them to compute a BTSA metric. When this metric surpasses a threshold, SmartCell notifies the user that the network is congested or automatically switches the priority of a user’s SIM cards to an alternate network. We use 0.9 as the threshold based on doubling the classification of 5 percent blocking as high congestion.

**Availibility Algorithm:** SmartCell estimates BTS availability periodically based on an exponential weighted moving average of immediate assignment success rate, defined earlier, as shown in [Figure 3](#fig3).

**Hardware Platform**

During our visit, our team surveyed 228 residents of the Za’atari refugee camp. We found that Android phones are very popular in Za’atari: 64 percent reported owning an Android device, 22.4 percent owned a Nokia device, while 4 percent owned an iPhone. We believe the universal availability of Android handsets (e.g., 85 percent global market share), powerful features and application programming interfaces (APIs), and a wide range of models at varying prices make Android an ideal platform for SmartCell.

**USE OF MULTIPLE NETWORKS**

SmartCell relies on the use of multiple SIM cards, which is common in developing regions. Unlike areas where “good” cellular connectivity is ubiquitous, carriers in rural and developing regions often have vastly disparate coverage and quality of service. In contrast to the multi-year cellular contracts common in countries like the United States and Japan, the availability of low-cost and contract-free prepaid SIMs enables users in developing regions to use multiple cellular networks. This leads to users carrying SIM cards from multiple carriers and switching as needed to obtain acceptable connectivity. For example, the number of cellular subscriptions per 100 residents in Guatemala is 106.6, and in Jordan it is 147.8, while in the United States it is 98.4. Respondents to the survey administered in Za’atari use three SIM cards on average; switching SIMs to take advantage of less congested networks, “same network” discounts, and cheaper data-only plans. This level of comfort in switching carriers is promising for a system such as SmartCell.

Switching SIMs traditionally requires powering off the phone. However, phones that support multiple SIM cards are increasingly popular. OpenSignal reports 57 percent of Android users in Guatemala have multi-SIM phones, and more than 50 percent of Android users in several other developing countries own dual-SIM phones. When SmartCell is used on a multi-SIM phone, physically swapping SIM cards is unnecessary. Instead, SmartCell can take the user directly to the SIM settings activity, allowing the user to select which SIM to use for voice calls, SMS messaging, and data traffic. Alternatively, SmartCell can reconfigure the phone automatically, attempting to select the least congested mobile network in the area.

**SMARTCELL OPERATION**

SmartCell observes cellular control messages, specifically immediate assignment and immediate assignment rejection messages, to understand network availability and load on a user’s current BTS. Recall that these messages are BTS-specific and allow an attached MS to infer SDCCH blocking affecting not only itself, but also other MSs connected to the BTS. SmartCell detects these messages and uses them to compute a BTSA metric. When this metric surpasses a threshold, SmartCell notifies the user that the network is congested or automatically switches the priority of a user’s SIM cards to an alternate network. We use 0.9 as the threshold based on doubling the classification of 5 percent blocking as high congestion.

**Availibility Algorithm:** SmartCell estimates BTS availability periodically based on an exponential weighted moving average of immediate assignment success rate, defined earlier, as shown in [Figure 3](#fig3).

**Use of Multiple Networks**

SmartCell relies on the use of multiple SIM cards, which is common in developing regions. Unlike areas where “good” cellular connectivity is ubiquitous, carriers in rural and developing regions often have vastly disparate coverage and quality of service. In contrast to the multi-year cellular contracts common in countries like the United States and Japan, the availability of low-cost and contract-free prepaid SIMs enables users in developing regions to use multiple cellular networks. This leads to users carrying SIM cards from multiple carriers and switching as needed to obtain acceptable connectivity. For example, the number of cellular subscriptions per 100 residents in Guatemala is 106.6, and in Jordan it is 147.8, while in the United States it is 98.4. Respondents to the survey administered in Za’atari use three SIM cards on average; switching SIMs to take advantage of less congested networks, “same network” discounts, and cheaper data-only plans. This level of comfort in switching carriers is promising for a system such as SmartCell.

Switching SIMs traditionally requires powering off the phone. However, phones that support multiple SIM cards are increasingly popular. OpenSignal reports 57 percent of Android users in Guatemala have multi-SIM phones, and more than 50 percent of Android users in several other developing countries own dual-SIM phones. When SmartCell is used on a multi-SIM phone, physically swapping SIM cards is unnecessary. Instead, SmartCell can take the user directly to the SIM settings activity, allowing the user to select which SIM to use for voice calls, SMS messaging, and data traffic. Alternatively, SmartCell can reconfigure the phone automatically, attempting to select the least congested mobile network in the area.

**SMARTCELL OPERATION**

SmartCell observes cellular control messages, specifically immediate assignment and immediate assignment rejection messages, to understand network availability and load on a user’s current BTS. Recall that these messages are BTS-specific and allow an attached MS to infer SDCCH blocking affecting not only itself, but also other MSs connected to the BTS. SmartCell detects these messages and uses them to compute a BTSA metric. When this metric surpasses a threshold, SmartCell notifies the user that the network is congested or automatically switches the priority of a user’s SIM cards to an alternate network. We use 0.9 as the threshold based on doubling the classification of 5 percent blocking as high congestion.

**Availibility Algorithm:** SmartCell estimates BTS availability periodically based on an exponential weighted moving average of immediate assignment success rate, defined earlier, as shown in [Figure 3](#fig3).
Eq. 1. Recall that we infer SDCCH blocking from the immediate assignment success rate.

\[ BTSA_t = A_t \times \alpha + (1 - \alpha) \times BTSA_{t-1} \]

Using our algorithm, we plot the BTSA for each of the carriers serving Za’atari on the morning of January 6, 2015 in Fig. 3. We use a period of 1 min and an \( \alpha \) of 0.25. Note that \( \alpha \), the periodicity, and the threshold are tunable values as system responsiveness requirements may differ depending on the environment. We observe that the carriers have distinct availabilities that roughly correspond to the immediate assignment rejection percentage plotted in Fig. 2. Importantly, congestion on one carrier does not necessarily correlate with congestion on another. If the value crosses below the threshold (indicated by the red line in Fig. 3), SmartCell informs the user that their current BTS is congested, prompting the user to switch BTSs or networks, or, in the case of a multi-SIM phone, switches to a different network automatically.

An example scenario of SmartCell’s operation can be observed in Fig. 3. A Zain customer running our application would cross the availability threshold at around 8:30 (shaded region) and be advised they may wish to switch SIMs or move to a different location. At that time, Orange also experienced congestion, although to a lesser degree, while Umniah remained congestion-free. The Zain user could switch to either network and obtain a BTS with higher availability, and hence achieve better quality of service.

**Reachability:** The use of multiple SIM cards negatively impacts user reachability, which is the ability of a correspondent to connect with a user via one of the user’s phone numbers. Issues with reachability manifest as delays in correspondence, as calls and SMSs sent to an inactive SIM will not be received until the user inserts or re-activates the inactive SIM. SmartCell may exacerbate or improve this problem, depending on usage and the cellular networks in the area. In cases like Za’atari, where one carrier has significantly lower availability than others, users of SmartCell may visit the worst performing network less often, decreasing their reachability via their phone number from that network. However, as widespread use of SmartCell would likely result in loose load balancing across networks, this decrease in reachability is likely to be short-lived.

In contrast, in more balanced network environments, SmartCell may encourage users to switch between networks more frequently than they would otherwise, reducing the average time their SIMs remain inactive and therefore reducing the delay before a user receives a missed call notification or SMS. We do not believe that attempting to contact a user of SmartCell will be drastically different than contacting non-users, as users commonly switch between cellular networks currently. In contrast, SmartCell should improve reachability when a user is receiving a call on an active SIM, as SmartCell users prefer base stations with higher availability, which are more likely to have the capacity to provide service for the incoming call.

**Global Congestion:** In the case of global congestion in an area, where all carriers are severely congested, SmartCell may cause users to switch between carriers when the switch will not actually improve connectivity. However, assuming a large population of users, SmartCell should lead to homeostasis where all carriers in an area have roughly the same availability, rather than drastically different levels as witnessed in Za’atari.

In cases of severe congestion on a single carrier, shifting many users onto a different carrier network may lead to simply migrating service traffic from the egress network to the ingress network. As such, SmartCell can be tuned to avoid rapid, drastic changes in estimated availability by increasing the weight assigned to past measurements and/or setting the threshold at a lower value. By increasing the weight assigned to past measurements the availability metric is smoothed.

**DISCUSSION AND CONCLUSION**

SmartCell opens multiple possibilities for further research. A current limitation of SmartCell is that each phone has limited a priori information about whether selecting alternate carriers or BTSs will result in improved service. Each phone with SmartCell relies on its own measurements of the user’s current cell and recent measurements from nearby SmartCell users. This presents a challenge in areas with a low density of SmartCell users. Unless other nearby users have recently reported congestion information for multiple nearby carriers and cells, the selected network could be as congested or worse post-switch. An algorithm to predict availability for carriers and BTSs based on past data submitted to the SmartCell repository is one area for future improvement. The repository could serve tiles to clients in a similar manner to map applications, where each tile contains long-term congestion history of BTSs in an area. This would reveal long-term trends, give users a better intuition of which carriers and cells are likely to be congested at any given time, and enable a more informed choice of networks and cells in the absence of real-time measurements.

The use of a central repository in congested environments can also be a limitation. If poor connectivity precludes the use of a central repository, cooperative peer measurements could be shared using direct channels such as local WiFi networks, WiFi Direct, or Bluetooth. As each measurement is only a few bytes of information, a measurement can be encoded into short strings. For example, the string “1458049642:410:54012:34.4312:-119.7598:8” indicates that at epoch time 1458049642, a cell on ATT (MNC 410) with cell ID 54012 was observed near 34.4312:-119.7598 with availability. These strings can be shared in a connectionless fashion, such as in Wi-Fi SSIDs or in the UUTID field of Bluetooth Low Energy beacons.

Although cellular coverage is rapidly increasing throughout developing regions, it is clear that disparities in the quality of the coverage exist, and will likely continue to exist for the foreseeable future due to a variety of economic and geographic factors. Our work scratches the surface of understanding these inequities.
ic and geographic factors. Our work scratches the surface of understanding these inequities in two specific, very different regions. Clearly, more work remains to be done to fully characterize the problem.

ACKNOWLEDGMENTS

This work was funded through NSF Network Science and Engineering (NetSE) Award CNS-1064821 and NSF Catalyzing New International Collaborations (CNIC) Award IIA-1427873.

REFERENCES


BIOGRAPHIES

Paul Schmitt (pschmitt@cs.ucsb.edu) is a Ph.D. student in the Department of Computer Science at the University of California at Santa Barbara (UCSB). His research focus is on mobile network connectivity in challenging networking environments, and information and communication technologies for development (ICTD).

Daniel iland (iland@cs.ucsb.edu) is a Ph.D. student in the Department of Computer Science at UCSB. He was awarded a Bachelor’s degree in computer science from the Rochester Institute of Technology in 2011. His research focuses on enabling robust and reliable communication in emergency and disaster scenarios, wireless localization, and ICTD.

Elizabeth M. Belding (ebelding@cs.ucsb.edu) is a professor in the Department of Computer Science at UCSB. Her research focuses on mobile networking, including network performance analysis and ICTD. She is the author of over 100 technical papers and has served on over 60 program committees for networking conferences. She is currently an Editor-at-Large for IEEE Transactions on Networking. She is an ACM Distinguished Scientist.
CALL FOR PAPERS
IEEE COMMUNICATIONS MAGAZINE
SUSTAINABLE INCENTIVE MECHANISMS FOR MOBILE CROWDSENSING

BACKGROUND
Mobile devices are explosively growing in our daily lives. It is estimated that the number of smartphones in use globally has reached the 2 billion milestone in 2015. These mobile devices are widely equipped with sophisticated embedded sensors, such as accelerometer, digital compass, gyroscope, GPS, microphone, and camera. The emerging paradigm of crowdsensing allows this large number of mobile devices to measure phenomena of common interest, which provide a new societal fashion of data sensing and sharing. A typical crowdsensing application leverages the ubiquitous mobile devices and the pervasive wireless network infrastructure to collect and analyze sensed data far beyond the scale of what was possible before, without the need to deploy thousands of static sensors.

The incentive mechanism is the most critical concern in the development of mobile crowdsensing because “crowd” participants are the foundations of all crowdsensing applications. Classic incentive mechanisms attract numerous participants by competitive payment designs. However, to achieve a sustainable crowdsensing, advanced incentive mechanisms need to pay attention to not only the payment but also many other features such as energy conservation and secure communications. For example, mobile devices may be compromised by hackers, important data may be stolen by eavesdroppers during wireless communications, and partially sensed data may expose participants’ private information. Obviously, without secure communications, a crowdsensing application will keep losing its participants even with good pay. Although plenty of incentive mechanisms have been developed for mobile crowdsensing, many challenges still remain to be addressed. It is important to explore this timely research topic to support the promising crowdsensing in practice.

This Feature Topic is intended to promote high-quality research in “Sustainable Incentive Mechanisms for Mobile Crowdsensing,” and move the theoretical and practical boundaries forward for a deeper understanding in fundamental algorithms, modeling, and analysis techniques from academic and industrial viewpoints. Authors from both academia and industry are invited to submit unpublished papers to this Feature Topic. The topics suggested can be discussed in terms of concepts, the state of the art, standards, implementations and evaluation, and running experiments and/or applications. Topics of interest include, but are not limited to:
• New platform/architecture/infrastructure for mobile crowdsensing
• Security and privacy in incentive mechanisms for mobile crowdsensing
• Energy-efficient incentive mechanisms for mobile crowdsensing
• Data-driven incentive mechanisms for mobile crowdsensing
• Practical implementations of large-scale mobile crowdsensing
• Sustainable social based mobile crowdsensing
• Reliable communication paradigm for sustainable crowdsensing
• Game theory in incentive mechanisms
• Green wireless communications in sustainable crowdsensing

SUBMISSIONS
Articles should be tutorial in nature and written in a style comprehensible and accessible to readers outside the specialty of the article. Complete guidelines can be found at http://www.comsoc.org/commag/paper-submission-guidelines. It is important to note that IEEE Communications Magazine strongly limits mathematical content, and the number of figures and tables. Mathematical equations should not be used. Article length (Introduction through conclusions, excluding figures, tables, and captions) should not exceed 4500 words. Figures and tables should be limited to a combined total of 6. The number of archival references is limited to 15. All articles must be submitted through the IEEE Manuscript Central site (http://mc.manuscriptcentral.com/commag-ieee) to the “March 2017 / Sustainable Incentive Mechanisms for Mobile Crowdsensing” category by the submission deadline according to the following schedule.

IMPORTANT DATES
• Manuscript Submission Deadline: July 15, 2016
• Decision Notification: October 15, 2016
• Final Manuscript Due Date: December 15, 2016
• Publication Date: March 2017

GUEST EDITORS
Linghe Kong
Shanghai Jiao Tong University, China
linghe.kong@sjtu.edu.cn

Kui Ren
State University of New York at Buffalo, USA
kuiren@buffalo.edu

Muhammad Khurram Khan
King Saud University, Kingdom of Saudi Arabia
mkhurram@ksu.edu.sa

Qi Li
Tsinghua University, China
liqi@csnet1.cs.tsinghua.edu.cn

Ammar Rayes
Cisco Systems, USA
rayes@cisco.com

Mérouane Debbah
Huawei, France
merouane.debbah@huawei.com

Yuichi Nakamura
NEC Corporation, Japan
yuichi@az.jp.nec.com

Mérouane Debbah
Huawei, France
merouane.debbah@huawei.com

Yuichi Nakamura
NEC Corporation, Japan
yuichi@az.jp.nec.com
The mobile Internet that will be enabled by LTE-Advanced Pro and the way in which we interact with it will change significantly within the next few years. From a usage point of view, it is anticipated that by 2020, each person, globally speaking, will consume on average as much as 5 GB of data each month, in addition to the traffic generated by 20–30 billion connected things. More significantly, not only will the mobile network be significantly faster, but even more applications will become possible. The video experience will be improved globally; mobile virtual reality will be available; networked vehicles and perhaps self-driving cars will roam our streets; the Internet of Things will enrich and make our lives more productive; and a mobile cloud will allow access to our data anytime, anywhere.

The requirements to support the anticipated applications can be classified into three main types: higher speed, massive connections, and low latency. Applications such as virtual reality and 4k video will boost the demand for network capacity beyond 1 Gb/s [1]. Massive connectivity on the order of 1000 connections per cell will be required for IoT [2]. Industry 4.0, enabled by cyber physical systems, will need to control physical equipment, and real-time cooperation will require end-to-end latency of less than 10 ms.

2016 will be pivotal in the development of LTE-Advanced Pro networks. Mainstream global operators have unveiled large-scale deployments with over 100 networks expected to be operational by the end of the year [3]. Already by the end of 2015, over 1 Gb/s transmissions rates were demonstrated in numerous commercial networks, and many countries have started pre-commercial deployments of IoT services.

This third and concluding part of the LTE-Advanced Pro Feature Topic presents several articles on potential technologies that will help to meet the above mentioned expected demand for network capacity.

The first article provides a review of the recent results on advanced coordinated beamforming techniques with less overhead than joint transmission CoMP. The discussion focuses on assessing the resilience of these schemes to uncoordinated interference. The article identifies key roadblocks and research directions to address.

The second article elucidates the multi-carrier ultra-dense network for LTE-Advanced Pro. Key advanced carrier aggregation technologies such as carrier switching and dynamic on/off are discussed, and performance benefits are demonstrated based on numerical simulations.

The third article discusses several key techniques for non-orthogonal transmission. Multi-user superposition transmission is proposed for the downlink, while multi-user shared access and sparse coded multiple access schemes are promising for the uplink. MIMO techniques for non-orthogonal transmissions are also considered.

The fourth article addresses the flexible use of spectrum in the future to increase the usage efficiency of radio spectrum. In particular, the business and technology enablers for sharing the ultra high frequency broadcasting spectrum with digital terrestrial TV and mobile broadband are reviewed. The results indicate that the scale, cost synergies, adaptability to regulatory regimes, and differentiation regarding user orientation inherent in the LTE standard are critical for flexible spectrum usage.

In the fifth article, the authors propose a unified traffic steering framework as an approach for orchestration of various traffic steering related issues that impact scheduling decisions for mobile data requests. The approach assumes awareness of the traffic demand and the ability to optimize power consumption in heterogeneous networks.

The last article in this Feature Topic discusses the trends and challenges in wireless channel modeling, which are foundational to system evaluation. Key drivers for new channel models are given, and new trends in channel modeling are reviewed. The article concludes with remarks on the paths taken to resolve current issues.

As is clear from this Feature Topic, LTE-Advanced Pro will be a significant evolution of LTE while maintaining good support for legacy user equipment. The key technologies leading to a more powerful network that provides enhanced user experience on capacity and latency will
still be orthogonal frequency-division multiplexing-based but with evolution in the directions of flexible spectrum utilization, service extension, network optimization, massive multiple-input multiple-output and massive carrier aggregation. With the enhancements discussed within this Feature Topic, users will be able to enjoy a rich experience with ultra high definition voice and video, virtual reality shopping, and maybe even virtual reality social networking. Such a network will allow for a quantum leap in the human-machine interface. Together with artificial intelligence, the promise of seamless and almost magical connections between systems and humans can finally be realized. Moreover, the integration of different systems will bring unprecedented productivity gain to mankind. It will usher in the anywhere, anytime, hyper-connected world of our near future.

Finally, the Guest Editors would like to thank the many reviewers and the IEEE publications staff for their support. Without their hard work, this Feature Topic would not have been possible.

REFERENCES


GUEST EDITORIAL

Robert W. Heath Jr. is a Cullen Trust Endowed Professor in the Department of Electrical and Computer Engineering at the University of Texas at Austin and a member of the Wireless Networking and Communications Group. He received his Ph.D. in electrical engineering from Stanford University. He is a co-author of the book Millimeter-Wave Wireless. His current research interests include millimeter-wave for 5G, cellular system analysis, communication with low-resolution ADCs, and vehicle-to-X systems.

Michael Honig (mnh@eecs.northwestern.edu) is a professor in the Department of Electrical Engineering and Computer Science at Northwestern University. He received his B.S. degree in electrical engineering from Stanford University in 1977, and his Ph.D. degree in electrical engineering from the University of California, Berkeley, in 1981. Prior to joining Northwestern he worked at Bellcore in the Systems Principles Research Division. His recent research has focused on resource allocation for wireless networks and spectrum markets.

Satoshi Nagata 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 working on 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 Chairman of 3GPP TSG-RAN WG1 since 2013.

Stefan Parkvall (S’92, M’96, SM’05) is a principal researcher at Ericsson Research, active in the area of 5G research and 3GPP standardization. He received his Ph.D. degree from the Royal Institute of Technology in 1996, served as an IEEE Distinguished Lecturer 2011–2012, and co-authored several popular books such as 4G/LTE/LTE-Advanced for Mobile Broadband. He received the Ericsson Inventor of the Year award and the Swedish government’s Major Technical Award for contributions to HSPA, and was nominated for the European Inventor Award for contributions to LTE.

Anthony C. K. Soong (S’88, M’91, SM’02, F’14) (anthony.soong@huawei.com) is the chief scientist for wireless research and standards at Huawei Technologies Co. Ltd., in the United States. His research group is active in the research, development, and standardization of the next generation cellular system. He has published numerous scientific papers and has over 90 patents granted or pending. He received his Ph.D. from the University of Alberta, and the 2013 IEEE Signal Processing Society Best Paper Award and 3GPP2 2005 Award of Merit.

THE GLOBAL COMMUNITY OF COMMUNICATIONS PROFESSIONALS

Special Member Rates

50% off - Membership for new members. (Offer valid through 15 August 2016.)

Member Benefits

IEEE Communications Magazine (electronic & digital delivery)
IEEE Communications Surveys and Tutorials (electronic)

Member Discounts

Valuable discounts on conferences, publications, IEEE WCET Certification program, IEEE Training courses and other exclusive member-only products.

Join Now!

http://bit.ly/1sdSno

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

www.comsoc.org
Advanced Coordinated Beamforming for the Downlink of Future LTE Cellular Networks

George C. Alexandropoulos, Paul Ferrand, Jean-Marie Gorce, and Constantinos B. Papadias

Abstract

Modern cellular networks in traditional frequency bands are notoriously interference-limited, especially in urban areas, where base stations are deployed in close proximity to one another. The latest releases of LTE incorporate features for coordinating downlink transmissions as an efficient means of managing interference. Recent field trial results and theoretical studies of the performance of JT CoMP schemes revealed, however, that their gains are not as high as initially expected, despite the large coordination overhead. These schemes are known to be very sensitive to defects in synchronization or information exchange between coordinating base stations as well as uncoordinated interference. In this article, we review recent advanced CB schemes as alternatives, requiring less overhead than JT CoMP while achieving good performance in realistic conditions. By stipulating that in certain LTE scenarios of increasing interest, uncoordinated interference constitutes a major factor in the performance of CoMP techniques at large, we hereby assess the resilience of the state-of-the-art CB to uncoordinated interference. We also describe how these techniques can leverage the latest specifications of current cellular networks, and how they may perform when we consider standardized feedback and coordination. This allows us to identify some key roadblocks and research directions to address as LTE evolves toward the future of mobile communications.

Introduction

Current dense and future super dense mobile broadband networks are and will be subject to various scenarios of simultaneous interfering communication links. In cellular networks, interference from neighboring base stations (BSs) is still one of the most prominent performance degradation factors resulting in outages or performance losses at the cell edges, as well as increasing the need for complex handovers. A classical approach to tackle interference is through medium access control and medium sharing techniques, which in turn severely compromise the performance of each individual user in the network due to explicit time sharing over the common resources. As we move toward denser networks with BSs and access points covering smaller areas to get antennas closer to the users, interference is becoming increasingly challenging [1].

Interference management in cellular networks has been first and foremost implemented through the smart reuse of network resources, mostly through the so-called frequency-division multiple access (FDMA) techniques. Previous generations of cellular network standards employed orthogonal reuse-\(n\) schemes, where neighboring cells do not interfere with each other’s resources. A frequency band used by a cell is not allowed, in this paradigm, to be used by neighboring cells, thereby greatly lowering the inter-cell interference floor. While the previous generation of mobile communications, Universal Mobile Telecommunications System (UMTS), moved from the reuse-\(n\) to a reuse-1 paradigm, today’s Long Term Evolution (LTE) specifications include a more fine-grained approach [2]. In classically deployed networks with large homogeneous cells, a core observation was that interference is mainly an issue for mobile terminals (MTs) lying far from their respective BSs, that is, at the cell edges. According to this approach, LTE BSs separate frequency bands dynamically and ensure that those allocated to the cell edges are non-overlapping. Such fractional frequency reuse (FFR) schemes are a very efficient form of interference management as they requires relatively low coordination on the BSs’ part. On the other hand, it may require more advanced power control in the downlink, and from the network point of view, BSs inefficiently use the time and frequency resources.

Capitalizing on the wide deployment of multiple antennas, especially on the BS side, and the advances in multi-antenna signal processing techniques, a new approach for interference management has made its way into mobile communication standards. Coordinated multipoint (CoMP) [3] is a broad umbrella name for coordination schemes aimed at realizing multi-user communications, that is, sharing the medium among multiple network nodes over space, on top of the possible sharing over time and frequency resources [4]. Focusing on the downlink and considering joint transmission (JT) CoMP, in
the theoretical limit of infinitely many distributed antennas, one could exactly pinpoint each MT and ensure that the signal intended for it adds up at its position, while creating no interference for the other MTs in the network. In this case, interference is not only removed, but is actually harnessed and exploited to increase the received signal power at each MT. However, for the practical implementation of JT CoMP schemes, sharing of channel state information (CSI) and data for the targeted MTs among the coordinated BSs will still be a challenge, especially at the data level among them is necessary. These requirements actually constitute the major downfall of JT CoMP in practical cellular networks, making it hard to achieve its theoretical gains in practice. On top of that, it was shown in [5, 6] that, imperfect and/or outdated CSI and uncoordinated interference have a very large impact on the performance of conventional JT CoMP schemes. Practical RF components, such as oscillators with phase noise, were also shown to have a similar negative effect [1].

As an alternative to JT CoMP for the downlink of cellular networks, coordinated beamforming (CB) uses shared knowledge of the spatial channel between the coordinated BSs and their intended MTs to separate the different data streams without exchanging MTs’ data. As such, CB schemes come with less stringent synchronization and coordination requirements [5], while retaining at least a large part of the JT CoMP performance. With CB, coordinated BSs only share CSI, and as long as the CSI is up to date, synchronization is unneeded, and each BS in the coordination cluster may transmit independently. Recent releases of the LTE specifications by the Third Generation Partnership Project (3GPP) have integrated the necessary elements to estimate the interfering channels on the MT part, with added reference signals and coordination of these signals among the coordinated BSs [7, 8]. 3GPP also included advanced 3D beamforming capabilities and more complex antenna patterns in the latest standards as well as associated simulation tools. Although the standardization of CSI exchange between BSs is still left to the discretion of the vendors, the aforementioned improvements enable the practical implementation of CB schemes, on which we focus in the present article. The theoretical design of CB schemes has lately been the subject of many research papers, representative examples of which are [9–12]. Among these schemes, some [9, 12] target the so-called multiple-input multiple-output (MIMO) interference channel (IFC), where each multi-antenna BS belonging to the coordination cluster wishes to serve exactly one multi-antenna MT, while [10, 11] are intended for the more general MIMO interference broadcast channel (IBC), where each coordinated multi-antenna BS may serve concurrently more than one multi-antenna MTs.

In this article, we present comparative performance evaluation results among the recent CB schemes that constitute future candidates for implementation in practical cellular networks due to their offered theoretical performance gains coming with reduced coordination overhead, and their increased level of compatibility to the latest relevant standards’ specifications [7, 13]. To advocate for the adequacy of interference coordination, only at the beamforming level, as an enabling approach for boosting the performance of dense networks, we consider as example scenarios of interest small-cell network deployments, where high capacity and tightly synchronized (on the signal level) links among the BSs belonging in a coordination cluster are not feasible. In such scenarios, coordination may be fully dynamic as a result of a scheduling mechanism, and hence carried out through dedicated wireless links. We focus on revealing the potential resilience of the CB schemes [9–12] to uncoordinated interference and investigating their performance with standardized feedback [8]. The latter goal may also serve as an indicator of the impact of the quality or latency of CSI to the performance of the considered schemes. To achieve the former goal, we propose a parametric system model where the powers of intra-cluster interference (ICI) and out-of-cluster interference (OCI) are defined relative to the power of the desired signal. The impact of OCI on both the clustered and centralized CB schemes designed for the IFC, and on the decentralized schemes that can be applied to the IBC is assessed. We then discuss how to adapt these schemes in current and future standards, and how practical feedback and quantization may impact their performance. Finally, we conclude with some specific research directions that may be pursued to improve the performance and integration of advanced CB schemes in future LTE networks and beyond.

**Modeling Interference in Cellular Networks**

To investigate the impact of interference in coordinated transmission schemes, we present a simple system model that captures the relative effect of ICI and OCI on the received signal. For the interference experienced by each MT associated with a BS belonging to a coordination cluster, we make the following assumptions:

- The aggregate ICI is of relative power \( a \in [0, 1] \) compared to the power of the desired signal. For example, \( a = 1 \) models cases where MTs are at similar distances from all BSs in the coordination cluster. The latter case is well suited for MTs located at the edges of the separate cells, and thus the center of the cluster, where JT CoMP and CB schemes are expected to perform best.
- The aggregate OCI is of relative power \( \beta \in [0, 1] \) compared to that of the desired signal. This parameter indicates the effectiveness of BS clustering for coordinated transmission. Low values of \( \beta \) indicate that most of the interference for a specific MT has been included within the cluster.

Adopting these two assumptions, the proposed system model is mathematically described as follows. We consider an infinitely large cellular network from which we single out \( B \) BSs, indexed in the set \( B = \{1, 2, \ldots, B\} \), to form a coordination cluster. On some time-frequency resource unit, the BS cluster aims at providing service to \( U \) MTs indexed in the set \( U = \{1, 2, \ldots, U\} \). A set of MTs associated to BS \( b \in B \) is
denoted by $U_b$, such that all sets $U_b$ for all $b$ form a partition of the set $U$. Without loss of generality, we assume that each BS is equipped with an $N$-element antenna array, whereas each MT has $M$ antennas. Let also $x_{b,u}$ represent the $N$-dimension vector with the information bearing signal transmitted from BS $b$ and intended for MT $u$. Then the baseband received $M$-dimension vector at MT $u$ can be expressed as

$$y_u = H_{u,b}x_{b,u} + \sum_{k \in U_b \setminus U_u} H_{u,k}x_{b,k} + \sum_{g \in B_l} \sum_{n \in U_b} H_{u,n}x_{n,b} + g_u + n_u,$$

where $H_{u,b}$ denotes the $M \times N$ channel matrix between MT $u$ and BS $b$, and the $M$-dimension vector $g_u$ is the OCI, for which we model the amplitude of its elements as independent and identically distributed Nakagami-$m$ random variables. It can be shown that this modeling of OCI includes that of [6, 14]. In addition, the $M$-dimension vector $n_u$ represents the noise modeled as additive white Gaussian such that $\mathbb{E}(\|H_{u,b}x_{b,u}\|^2)/\mathbb{E}(\|n_u\|^2) = \text{SNR}$. We further normalize the channel matrices in order to have, on average, ICI power at the signal level as $\mathbb{E}(\|H_{u,b}x_{b,u}\|^2) = \alpha (B - 1) - \beta (1 - \|H_{u,b}x_{b,u}\|^2)$ and OCI power at the signal level as $\mathbb{E}(\|g_u\|^2) = \beta \mathbb{E}(\|H_{u,b}x_{b,u}\|^2)$, where we have assumed that $\mathbb{E}(\|H_{u,b}x_{b,u}\|^2) = \mathbb{E}(\|H_{u,b}x_{b,u}\|^2)$ for $k \in U_b$ with $k \neq u$. The system model of Eq. 1 is capable of describing a wide range of interference scenarios by varying the parameters $\alpha$ and $\beta$ as well as the distribution of $g_u$, thereby capturing how interference coordination might perform for MTs in different network setups. An example illustration of this model is depicted in Fig. 1. The three BSs in the center of the figure are assumed to form a coordination cluster. The MTs falling into the regions covered by these BSs are subject to relative interference $\alpha$ from intra-cluster BSs, and aggregate interference $\beta$ from each of the out-of-cluster BSs.

**Advanced Coordinated Beamforming Schemes**

In this section, the system model of the previous section is first employed to a simplistic cellular network in order to demonstrate the theoretical gains of JT CoMP and CB schemes over representative non-coordinated ones as well as to compare JT CoMP with CB. Then we present performance comparisons among CB schemes requiring full CSI exchange among coordinating BSs as well as schemes that operate with limited coordination overhead. The compared schemes differ on the considered design objective and the level of taking network interference under consideration.

**Theoretical Gains through an Example**

We consider a cluster of $B = 2$ BSs as a part of a large cellular network, which aims to serve two MTs in every time-frequency resource unit; one MT is associated with one BS and the other MT to the other BS. Focusing on the presented system model and using the classical bounds for the individual MT rates in multiple-input single-output (MISO) IFCs [4], the following holds:

- With full reuse of time-frequency resources, each MT is subject to interference from every BS with which it is not associated, and its rate is upper bounded by $\log_2(1 + \text{SNR}/(\alpha \text{SNR} + j \text{SNR} + 1))$.
- With orthogonal allocation of the resources, ICI is absent, but the prelog factor 0.5 appears on each individual MT rate, yielding $0.5 \log_2(1 + \text{SNR}/(j \text{SNR} + 1))$.
- With the CB scheme based on interference alignment (IA) [9], ICI can be completely nulled, and the individual MT rate becomes $\log_2(1 + \text{SNR}/(j \text{SNR} + 1))$.
- With ideal JT CoMP, the interference power actually boosts the intended signals, and the individual MT rate is given by $\log_2(1 + (1 + \alpha) \text{SNR}/(j \text{SNR} + 1))$.

The latter rates for each individual MT are sketched in Fig. 2 with OCI being 6 dB lower than that of the power of the intended signal.
(i.e., $\beta = 0.25$) and for two different values of $a$, which reveals the relative power of ICI. As expected, both coordinated transmission schemes provide substantial gains compared to full reuse and orthogonal transmission when the network operates in the interference-limited regime (i.e., when $\text{SNR}$ increases). As $a$ approaches 0, the gain of JT CoMP over IA decreases. For example, for $\text{SNR} = 15$ dB and $a = 1$ in Fig. 2, IA results in a nearly 100 percent gain over orthogonal transmission, while this gain becomes nearly 180 percent for JT CoMP. When $a$ decreases, the latter gain of IA remains the same, whereas that of JT CoMP decreases to nearly 110 percent. This example illustrates that in many cases of interest, a large part of the coordination gain comes from the removal of interference from the signal of interest rather than from stacking the powers of multiple BSs. It is also noted that when considering practical implementation issues in achieving JT CoMP, the bonus of full coordination becomes even lower, since JT CoMP is more afflicted by degraded CSI and “dirty” RF than CB [1].

**CB Schemes with Full CSI Exchange**

We hereinafter focus on the $B$-user $N \times M$ IFC, which constitutes a special case of the system model above where each $I_{b}\kappa$ comprises exactly one MT. In Figs. 3 and 4, we consider a coordination cluster of $B = 3$ BSs with $N = 4$ and $M = 2$, and compare the ergonomic performance with optimum receivers for different values of $a$ and $\beta$, and spatially independent Rayleigh fading of the following CB schemes:

- IA [9], which aims at aligning and then nulling interference at each MT belonging in the BS cluster.

- Maximum signal-to-interference-plus-noise ratio (SINR) [9], which targets maximizing the received SINR of each transmitted information data stream in the cluster.

- Weighted minimum mean squared error (WMMSE) [10], that minimizes a metric for the whole network that is based on the MMSE.

- Reconfigurable [12]. The latter scheme combines a network-wide MMSE criterion with the single-user MIMO waterfilling solution in order to maximize the rate of each MT associated with the coordination cluster according to the condition of its desired channel and the whole network’s interference level.

Although for all aforementioned CB schemes we consider here a centralized implementation with full CSI exchange among coordinating BSs, it is noted that for the maximum SINR, WMMSE, and reconfigurable schemes, distributed versions are also available, where explicit CSI exchange among BSs is avoided, and thus, coordination overhead can potentially be reduced.

The IA, maximum SINR, WMMSE, and reconfigurable CB schemes are linear schemes, which means that each BS transmits its signal using precoded symbols as $x_{a,b} = v_{a,b} s_{b}$, where $v_{a,b}$ represents the $N \times d_{b}$ precoding matrix and $s_{b}$ is the $d_{b}$-dimension information stream vector. Upon signal reception, each MT estimates the desired transmitted symbols using a $d_{b} \times M$ decoding matrix $U_{b}$, forming $\hat{s}_{b} = U_{b} y_{b}$. For the IA and maximum SINR schemes in Figs. 3 and 4, each $d_{b}$ was set to $0.5 \min(N, M) = 1$ according to the IA feasibility conditions [9], and $v_{a,b}$ for all $b$ was obtained in closed form for IA and iteratively for maximum SINR. For both the WMMSE and reconfigurable iterative schemes, each $d_{b}$ was initialized as $d_{b} = \min(N, M) = 2$ and obtained at the end of the algorithmic iterations or upon

**Figure 3.** Achievable sum rates for different CB schemes with full CSI exchange and $a = 1$ and $\beta = 0$. The coordination cluster comprises three 4-antenna BSs and one 2-antenna MT associated with each BS. A maximum of 10 iterations was used for each of the iterative schemes, maximum SINR, WMMSE, and reconfigurable. The performance of full reuse and orthogonal $4 \times 2$ MIMO transmission is also depicted.

**Figure 4.** Achievable sum rates for different CB schemes with full CSI exchange and $a = 0.25$ and $\beta = 0.25$. The coordination cluster comprises three 4-antenna BSs and one 2-antenna MT associated with each BS. A maximum of 10 iterations was used for each of the iterative schemes, maximum SINR, WMMSE, and reconfigurable. The performance of full reuse and orthogonal $4 \times 2$ MIMO transmission is also depicted.
Figure 5. Achievable rates per BS for different values of \( \alpha \) and \( \beta = 0.25 \) for an IBC with 4-antenna nodes and a main interferer for each coordinated BS. For both the eigenbeams and downlink IA schemes, IRC receivers have been used. The performance of the WMMSE scheme is also illustrated for comparison purposes.

convergence, explicitly for the reconfigurable scheme and implicitly for WMMSE together with all \( V_i \) matrices. More specifically, the reconfigurable scheme outputs \( d_b \) to be sent by each coordinated BS \( b \) together with their beamforming directions, whereas WMMSE only generates the transmit covariances matrices with possibly some streams set to zero power — thus rendered unusable. This means that for the latter scheme the optimum \( d_b \) needs to be searched in some way, a fact that will cause extra overhead in practical networks and possibly decrease performance. As can be concluded from Figs. 3 and 4, for a maximum of 10 iterations per iterative scheme, the performance of all considered CB schemes is susceptible to ICI and OCI. This behavior for OCI was also observed in [15] for IA. For example, for \( SNR = 15 \) dB, it is shown that the performance of all CB schemes drops approximately 45 percent between the two interference scenarios, according to which \( \alpha \) decreases from 1 to 0.25 and \( \beta \) increases from 0 to 0.25. Interestingly, for the considered interference cases in both figures and \( SNR < 17.5 \) dB, the maximum SINR, WMMSE, and reconfigurable schemes, which take OCI interference under consideration, provide equal to or slightly more than 100 percent improvement compared to IA. This behavior shows that maximum SINR, WMMSE, and reconfigurable schemes are highly resilient to the \( \alpha \) values; however, their resilience to \( \beta \) values is low, especially for WMMSE and high \( SNR \) values. This result tends to reinforce the necessity of considering OCI when designing CB schemes, and justifies their study under practical network conditions. As also demonstrated in Fig. 4, the majority of the CB schemes perform very close to or slightly better than full reuse, and for \( SNR < 17.5 \) dB, all CB schemes outperform orthogonal MIMO transmissions. However, for \( SNR > 17.5 \) dB, orthogonal transmissions are the best option, a fact indicating that to achieve the best performance for general values of \( \beta \), the coordination cluster needs to adopt a dual-mode operation, which switches depending on the \( \beta \) values between the reconfigurable CB scheme and orthogonal MIMO transmissions, for example.

In the CB schemes discussed before, MTs served by the clustered BSs are assumed to be clustered so as to create a separate group. This transpires in the current LTE standard; in particular, LTE Release 11 describes a CoMP cluster in which BSs may coordinate their transmissions [13]. This CoMP cluster forms the basis on which the techniques [9–12] may be implemented, although as we discuss in the following section, information exchange between the coordinated BSs is still not standardized. Inside a CoMP cluster, an MT may estimate the channels of its interferers through specific CSI structures and commands. This CSI may then be used to compute interference-aware receive filters [7], or it can be fed back to their associated BS for further processing. The BSs inside a CoMP cluster may also be able to exchange CSI when operating in time-division duplexing mode [5], by making use of the channel reciprocity property, whereas WMMSE is lacking in the more common frequency-division duplexing (FDD) mode. Notwithstanding, CSI exchange among coordinated BSs is still a complex operation; it weighs heavily on the backbone network [1], and as of today, there is no specific standardized mechanism on how or when to transfer this information. Therefore, the straightforward implementation of the described CB schemes outside of a vendor-locked configuration is still out of reach.

**CB Schemes with Limited Coordination Overhead**

One example of a CB scheme with limited coordination overhead is the downlink IA presented in [11], which is suitable for the more general IBC. This schemes capitalizes on the standards’ specifications [13] to allow each MT to estimate its strongest interferer and feed back good precoder candidates to its associated BS. Consider a cellular network with coordination clusters of \( B = 2 \) BSs, where each coordinated BS aims at sending a single information stream to its \( K = N - 1 \) associated MTs. The basic idea of downlink IA is to force the received signal at each MT associated with the cluster from the non-intended coordinated BS in a signal subspace of rank \( N - 1 \), thus freeing up one decoding dimension from interference for the desired signal. These decoding directions create equivalent MISO channels for the MTs associated with the cluster, which can then feed them back to their associated BS. Each BS can then employ any multi-user MIMO technique [4] to multiplex the \( N - 1 \) information streams toward its respective MTs, such as zero-forcing beamforming. The benefit of the downlink IA scheme is that although each BS only frees a single dimension, the interference-free direction is different for each MT, thereby enabling multi-user diversity. One can contrast downlink IA with FFR schemes [2], where the dimension freed in frequency is the same for all MTs. In our performance evaluation, we further assume that each MT can learn
the precoder chosen by the BS for their stream at the end. Since they have already estimated the channel from the non-intended clustered BSs, they know the necessary information to update their receivers to interference rejection combining (IRC) receivers, as described in [7].

The performance of downlink IA with IRC receivers over spatially independent Rayleigh fading channels is illustrated in Fig. 5 as a function of the SNR for different values of, $\alpha = 0.25$, $B = 2$, $M = N = 4$, and $K = 3$. Within this figure, we also plot the performance of a more classical multi-user MIMO scheme where interference is not exploited, and for which the decoding direction of each MT assigned to the cluster is chosen as the strongest eigenvector of its intended channel; this scheme is denoted as the eigenbeams scheme. Note that with the eigenbeams scheme, each BS can support four MTs, whereas with downlink IA each BS serves three MTs. As seen from the figure, and as expected, for cell edge MTs there is a potential very large SINR gain coming from the removal of the interfering coordinated BS. In that case, downlink IA shows a 50 percent gain from the eigenbeams scheme in the average sum rate per coordinated BS. On the other hand, the gain of downlink IA for MTs that are not at the cell edge, and thus do not experience a strong interferer, is reduced. As highlighted by the theoretical example above, the performance of downlink IA depends on the values of both $\alpha$ and $\beta$; if $\alpha$ is much larger than $\beta$, we have a strong interest in removing the interference even at a somewhat misaligned with our own channel. On the other hand, if the remaining OCI is on the level of the ICI, downlink IA provides less gain than a straightforward multi-user MIMO scheme like eigenbeams. This is in line with recent analyses, for example, in [15], where it was shown that blindly applying IA in a clustered cellular network is altogether detrimental. We can also conclude from Fig. 5 that the performance of the WMMSE scheme is poor in this context, since it targets minimizing the interference from the unintended clustered BS even if it has to shut down transmission to its MTs. At convergence of this iterative algorithm, a subset of the MTs will experience a very high SINR, while some streams will be unused, the overall performance is lower than that of downlink IA or eigenbeams.

**COORDINATED TRANSMISSION WITH THE LTE FEEDBACK SPECIFICATION**

The CB schemes presented in the previous section require CSI exchange among the BSs belonging to the coordination cluster. However, there is still no standardized mechanism in the current LTE specifications for full CSI exchange in cellular networks. This means that non-proprietary attempts to achieve CB are not truly possible as of today. As such, CB is not feasible outside of a vendor-locked set of BSs, or within a single BS with remote radio heads. This precludes many of the presented advanced CB schemes, which require CSI exchange and possibly joint computation of the transmission parameters among the coordinated BSs.

Focusing on LTE Release 12 for feedback specifications, we henceforth compare the performance of the downlink IA and eigenbeams schemes for the IBC scenario of Fig. 5 under standardized feedback, and compare it to the ideal feedback case. In particular, the CSI feedback needed in these schemes is limited to the feedback of a channel quality indicator (CQI) and a precoding matrix indicator (PMI) for each frequency subband. The physical layer procedures related to this feedback and the PMI codebooks are described in [8]. In Fig. 6, we evaluate the performance of downlink IA and eigenbeams with practical feedback using 4-antenna network nodes and an 8-bit codebook that creates a family of 256 possible precoders. To apply this codebook to the considered IBC scenario, we feedback the equivalent channel by using the PMI to the closest precoder in the family. As depicted in Fig. 6, this procedure results in a net performance loss of about 10 percent for the downlink IA scheme and 20 percent for the eigenbeams scheme. It can be shown that this loss is not entirely linked to the somewhat coarse feedback quantization, but rather to the way the codebook is constructed in [8]. In fact, increasing the number of bits in the feedback scheme while keeping the same codebook construction does not improve the performance substantially. This indicates that the sheer number of bits for the feedback channel is not itself the strongest indicator of feedback quality, and that codebook construction is in fact a fundamental question. Higher precision in the feedback process as well as accurate CSI estimation are thus still two of the key questions to answer today for coordinated transmission schemes as well as for many other channel-dependent signal processing techniques. In addition, practical CSI exchange between BSs participating in a coordination cluster is undefined as of the latest LTE release.
There is actually no standardized way to encode CSI in FDD systems. The specifications of the backbone communications in a CoMP set are also left to vendor implementations, precluding any inter-vendor CoMP set to be set up in practice.

**Conclusion and Research Directions**

As network deployments become denser, interference arises as a dominant performance degradation factor, regardless of the underlying physical-layer technology. The feature of coordinating BS transmission to manage interference in cellular networks is already a part of the latest LTE release, offering significant potential for performance improvement, especially at the cell edges. Among the recently proposed coordination schemes, there are CB schemes that require coordination overhead that is more or less compatible with the current standard’s specifications, and adapt satisfactorily to ICI while showing some resilience to OCI. However, to maximize the benefit from CB in future communication networks, certain advances need to take place. One of these is BS clustering, which needs to be both dynamic and scalable. Efficient clustering methods, based, for example, on network connectivity or received SINR, that keep OCI levels to a minimum can be combined with CB schemes to boost network performance. Another necessary step in coordination schemes is the design of techniques for information exchange with low overhead among the coordinated network nodes. The coordination overhead of the latest CB schemes is still far from what can be supported in the current LTE release. This necessity becomes even more prominent in fully distributed CB schemes, where information needs to be exchanged iteratively between transmitters and receivers. In fact, it is still unclear how to practically implement iterative CB schemes and cope with their required information exchange overhead. There are issues in the actual form the information messages will take, the structure of the message-passing shells, and, most importantly, the quantization that has to be applied on the message content. Up to this day, there is little research on designing CB schemes where the iterative computation supports noisy or quantized messages. This is also related to the accuracy of CSI as measured by members of the coordination cluster. Last but not least, coordination schemes need to be designed to account for the characteristics of technologies intended for next generation networks, such as full duplex radios and massive MIMO with possibly hybrid analog and digital processing.

**References**


Advanced Carrier Aggregation Techniques for Multi-Carrier Ultra-Dense Networks

Jialing Liu and Weimin Xiao

ABSTRACT

An emerging scenario to be considered for further evolution of Long Term Evolution and beyond is a multi-carrier ultra-dense network (MC-UDN). Advanced carrier aggregation (ACA) techniques, such as carrier switching, lightweight carriers, opportunistic transmissions, and adaptive resource optimization schemes, provide important means for the MC-UDN to allocate carrier-level resources properly and efficiently according to traffic load, channel, and interference conditions across the network. In this article, the general background, motivations, high-level design, and performance benefits of ACA techniques for the MC-UDN are described.

INTRODUCTION

Wireless communication has been one of the most active and rapidly growing areas in the past decade. Fueled by the global popularization of mobile devices, wireless communication has demonstrated a sharp turn toward more explosive traffic demands. This general trend may accelerate even further in the future. A large portion of wireless traffic has gone through cellular networks and will continue to do so, which poses a significant challenge to the cellular industry. Moreover, as more and more applications are integrated or considered to be integrated with cellular networks, including massive Internet of Things (IoT), connected vehicles, and so on, cellular networks are becoming the critical infrastructure on which the entire society relies. Therefore, any inefficiency or failure of future cellular networks can cause high social costs or even catastrophes in our society.

To meet the challenges, many candidate solutions for cellular networks have been considered, for near-term Long Term Evolution (LTE)-Advanced Pro (aka 4.5 G, comprising LTE Releases 13 and 14) and longer-term fifth generation (5G). These candidate solutions aim to increase the capacity many times, support massive connections with large coverage, provide ultra-reliable and low-latency communications, enhance user experience, and improve various efficiency metrics (e.g., spectrum efficiency, energy efficiency). To help bring clarity, we sort the solutions into roughly two main categories:

• Category 1: More communication resources for the networks and user equipment (UE):

This includes more network nodes, wider bandwidth, additional resources (e.g., WiFi) to offload, more antennas, carrier aggregation (CA), coordinated multipoint (CoMP), dual connectivity (DC), and so on.

• Category 2: Higher operational efficiency and spectral efficiency; This includes advanced transmitters and receivers; advanced coding, modulation, and multiplexing schemes; enhanced interference management; overhead reduction; coordination; advanced antennas techniques; improved measurements and feedback; and so on.

The above considerations lead to an important scenario gaining increasing attention: the multi-carrier ultra-dense network (MC-UDN). The MC-UDN possesses a unique capability of achieving multi-fold throughput gains without intractable challenges in design, deployment, or operations in the foreseeable future, and is generally accepted as an indispensable component of the forthcoming wireless systems. The MC-UDN is the main scenario focused on in this article.

However, simply increasing the network resources to obtain an MC-UDN and operating the MC-UDN in a (brute force) way similar to current networks have been found quite inefficient (see [1, 2] and later sections). Therefore, to efficiently deploy and operate an MC-UDN and make the most of the network resources including all available network nodes and carriers, advanced CA (ACA) techniques are introduced. These include, but are not limited to, carrier selection and switching, lightweight carriers, opportunistic communications, and adaptive resource optimization schemes. The main rationale is to introduce higher adaptation capabilities and support efficient adaptations that are found to be powerful in addressing issues such as interference, channel, and traffic load variations. Thus, the ACA techniques advocated in this article lie in the intersection of both categories 1 and 2 as they aim to efficiently (and even optimally, if feasible in practice) utilize the increased resources in an MC-UDN.

HISTORICAL REVIEW:

FROM RELEASE 8 TO RELEASE 12

There have been many features standardized in the Third Generation Partnership Project (3GPP) LTE for cellular networks. In this sec-

The authors are with Huawei Wireless Research and Standards.
tion, the general trends behind the evolution of the cellular networks are identified (mainly focusing on radio access network, RAN, aspects), which helps motivate the study of the MC-UDN and ACA. See Fig. 1 for an illustration and more descriptions below.

**Trend 1: Toward More Resources**

One main thread for 3GPP standardization is to support more resources on the network side and UE side (corresponding to category 1 above). Densification of networks, leading to smaller and smaller cell coverage areas, has been a major focus in 3GPP. Releases 8/9 were focused on macro networks of high-powered enhanced NodeBs (eNBs) at regularly located locations. Releases 10/11 support heterogeneous networks (HetNets) with pico or relay cells overlaid with macrocells. Release 12 is concerned with clustered small cells. Densification has been shown to deliver multi-fold capacity gains [2].

Cellular networks have been incorporating more and more new spectrum resources. Release 8/9 user equipment (UE) supports only one carrier of up to 20 MHz bandwidth. Release 10/11 CA-capable UE supports up to 5 carriers, each with up to 20 MHz bandwidth, within the same band (i.e., intra-band CA) or in different bands (i.e., inter-band CA) [3]. Release 12 considers CA with time-division duplex (TDD) and frequency-division duplex (FDD) joint operations. CA allows UE to receive simultaneous downlink transmissions from multiple aggregated component carriers (CCs) and hence significantly increases user peak data rate and shortens latency.

In addition, to provide more resources for UE, CoMP (in Release 11), DC (in Release 12), and various offloading methods were supported, among other motivations.

**Trend 2: Toward Higher Efficiency**

To achieve higher operational efficiency and spectral efficiency (corresponding to category 2 in the previous section), one of the most important ways is to perform interference avoidance and coordination, mainly including:

- Static frequency reuse, effective for Release 8 macro networks with regularly located interferers
- Inter-cell interference coordination (ICIC) for power- and frequency-domain coordination over a slow timescale (Release-8)
- Enhanced ICIC (eICIC) for time-domain avoidance of one dominant interferer over a slow timescale, suitable for co-channel HetNets (Release 10)
- CoMP for interference coordination and turning interferers into serving resources (Release 11)
- Cell on/off for more flexible time-domain reuse, effective, for example, for small cell clusters (Release 12) [4]. This may also help improve energy efficiency.

Furthermore, interference may also be mitigated at a UE receiver as done in network-assisted interference cancellation and suppression (NAICS) in Release 12. Other methods for higher efficiency include improved measurements for channel and interference, dynamic TDD, higher order modulation, multiple-input multiple-output (MIMO), and so on.

It is clear now that there are two major threads for cellular network evolution: more resources (especially densification and more spectrum resources) and higher efficiency (especially interference management), which may be seen as the underlying principles for many features. These threads continue into 4.5G and 5G, to which we now turn.
ADVANCES IN 4.5G RELEASE 13

To further alleviate capacity limitations and increase user peak data rate, 3GPP considered drastically increasing the spectrum resources that may be used for the network and UE. This included mainly two aspects: massive or enhanced CA (MCA or eCA) and incorporating the unlicensed band in licensed-assisted access (LAA) LTE.

In MCA, CA enhancements beyond five carriers were considered. Up to 32 CCs for a UE are supported. The use cases for MCA include CA within licensed spectrum or CA in licensed and unlicensed spectrum. The resulting supported downlink peak data rate is increased to about 25 Gb/s.

Unlicensed spectrum is abundant. The 5 GHz unlicensed band contains 14 non-overlapping 20 MHz carriers (or channels). In LAA, the licensed spectrum can be used as the primary carrier for cellular services, whereas the unlicensed carriers can be used as secondary resources for downlink offloading and best effort transmissions.

On the unlicensed spectrum, some regulations are imposed to avoid interference among different radio access technologies (RATs) without coordination. LAA standardized listen-before-talk (LBT) for eNB to comply with coexistence regulations. An LAA eNB needs to perform carrier sensing and then determines if a transmission can follow, all done on small timescales of about a few milliseconds. This results in discontinuous opportunistic communications. This can be seen as a further evolution of the semi-static cell on/off developed in Release 12 toward more dynamic and opportunistic cell on/off.

ACA FOR 4.5G RELEASE 14 AND BEYOND

Moving forward, it is expected that a cellular network will continue the trend of densification, and operate with many carriers with diversified spectrum resources in the near future, resulting in an MC-UDN. To efficiently support the MC-UDN, ACA techniques are proposed for Release 14 and beyond.

SCENARIO CONSIDERATIONS: MC-UDN

We now consider the two aspects of future deployment scenarios: further densification and many carriers.

Further Densification: Releases 12/13 mainly address small cell clusters, that is, there are several disconnected clusters, and each cluster contains a few small cells. There can be several ways to further densify:

- The clusters may get much closer to each other, and each cluster may contain more small cells.
- Continued densification leads to the clusters being connected with each other, forming a continuously dense area.
- A massive number of various types of new links (e.g., D2D, massive IoT, wireless inband backhaul) may be integrated into the network, considerably increasing the network density of links.
- 3D dense networks may be formed in high-rise buildings, with each floor in each building deployed with a set of network nodes.
- Dense macro networks are being deployed in some dense urban regions.

However, significant densification also poses challenges, among them interference becomes a more severe problem.

First, there may be even more strong interferers but none dominant, and the interference fluctuations in the time, frequency, and spatial domains are more severe and random. This leads to interference statistics different from those of a sparse network, which generally has one (or a very small number of) dominant interferer(s). Hence, static (or semi-static) interference avoidance and cancellation, as specified above, are generally not effective; even if UE avoids or cancels interference from several top interferers, the residual interference may still be rather strong.

Second, existing LTE design may not scale well toward a UDN. For example, cell-specific reference signals (CRSs) need to be transmitted in almost every transmission time interval (TTI, i.e., 1 ms for LTE) except for cases with cell on/off and LAA, which may account for about 10 percent (or even higher) of total network resources and causes (almost) persistent interference. The impact of CRS interference can be more severe for a UDN due to the extreme densification.

Some other issues may also become exacerbated with further densification, such as traffic load imbalance across the network, availability of channel and interference measurements, and so on.

Many Carriers: More spectrum resources may be integrated when Release 14 is commercialized. A UE may support up to 32 carriers, and the network may support even more. However, there are also issues and limitations to be addressed.

eNB-UE Carrier Number Asymmetry Due to UE Capability Limitations: In typical situations, an eNB supports many more carriers than UE does. How to enable such UE to efficiently access more carriers (e.g., for better interference avoidance semi-statically or even dynamically) available at the eNB without significantly increasing UE capability should be considered.

UE Uplink-Downlink Carrier Number Asymmetry Due to UE Capability Limitations: It is common that UE can transmit on only one or two uplink carriers at the same time, but it can receive simultaneously from many more carriers in downlink. This not only limits the network’s ability to perform efficient interference management and load balancing/shifting in uplink, but also impacts TDD downlink beamforming performance. The latter is because TDD downlink beamforming based on sounding reference signal (SRS) effectively exploits channel reciprocity and outperforms beamforming based on precoding matrix index (PMI) feedback. Hence, if UE has more downlink carriers than uplink carriers, some downlink-only TDD carriers do not have corresponding SRS in uplink and may lead to downlink performance degradation.

High UE Monitoring Complexity and Power Consumption: Even though UE may be capable of aggregating, say, 32 carriers, to monitor all 32 carriers at all times is not efficient. Here “monitoring” includes data buffering, control channel blind detection, and so on, which are needed in almost every TTI in current systems. This prob-
Carrier selection and switching can enable UE to potentially access over time as many carriers as the network can provide without drastically increasing UE capability requirements, thus overcoming the eNB-UE carrier number asymmetry issue due to UE capability limitations.

**Motivations for ACA**

ACA is proposed for Release 14 and beyond to better support the MC-UDN scenario and address the issues and limitations described above.

**Considerations for Interference:** To alleviate the impact of interference, more adaptive resource utilization is needed; that is, the set of resources experiencing lower interference and/or causing less interference to other transmissions needs to be selected more dynamically. As resource utilization and interference are largely driven by traffic loads, higher capability of resource adaptation is also crucial for networks with traffic loads varying significantly in time, frequency, and spatial domains.

In addition, to reduce interference, overhead such as (almost) persistently transmitted CRS in current networks needs to be minimized.

**Considerations of UE Capability Limitations and Others:** To address the issues and practical limitations above, it is desirable to allow UE to utilize more carriers available at the eNB in uplink and downlink without significantly increasing UE capability, which may be done via fast carrier selection and switching. For instance, for UE capable of simultaneously receiving data from 5 downlink carriers, the eNB may select a set of 5 carriers for the UE for a certain time out of the 30 carriers available at the eNB, change to another set of 5 carriers for the UE for the subsequent time duration, and so on. The UE switches its carriers per eNB instructions as quickly as possible, and hence effectively and efficiently utilizes possibly all 30 carriers at the eNB. Moreover, to reduce UE monitoring complexity and UE and eNB power consumption, lightweight carriers and monitoring on demand can be considered. More details are presented below.

**Fast Carrier Selection and Switching**

Fast carrier selection and switching are important for more efficient carrier-level resource adaptation, fast load balancing and shifting across carriers, and fast interference coordination and avoidance across carriers, while considering various UE limitations. In particular, carrier selection refers to the selection of a subset of carriers available at the network, performed by the eNB; carrier switching refers to switching from one carrier to another carrier, performed by the UE as instructed by the eNB, preferably via a fast indication. Carrier selection and switching can be done semi-statically (generally on a timescale of tens of milliseconds and longer, via high-layer signaling and procedures) or dynamically (generally on a timescale of at most a few milliseconds, via physical-layer signaling and procedures).

For carrier switching and selection, shorter transition times associated with switching can yield better performance gains. Therefore, faster switching is especially important. For this purpose, it is beneficial to pre-configure more carriers to UE than the UE’s CA capability and allow the UE to switch among them via a fast carrier switching mechanism. This can be considerably more efficient than eNB pre-configuring only five carriers if the UE capability is only five carriers, allowing no carrier switching beyond the five carriers, or allowing only slow carrier switching beyond the five carriers via slow carrier reconfiguration procedures. Hence, carrier selection and switching can enable UE to potentially access over time as many carriers as the network can provide without drastically increasing UE capability requirements, thus overcoming the eNB-UE carrier number asymmetry issue due to UE capability limitations.

Furthermore, carrier switching can help reduce UE monitoring complexity and power consumption in CA, MCA, and LAA. With carrier switching, UE monitors only a subset of carriers, and based on the indication received on these carriers, it quickly switches to other carriers for data reception whenever needed. This can be especially useful in LAA as the carrier availability is subject to LBT and not known a priori. For example, suppose UE can support all 14 carriers on the unlicensed band. Without fast carrier switching support, while the eNB performs LBT on these carriers and is uncertain about which carriers may become available, it is necessary that the UE monitors all 14 carriers all the time. Then when one or more of the carriers become(s) available, the eNB transmits data on them for a few milliseconds, and the UE detects the transmissions and receives data. This is not efficient. In contrast, with fast carrier switching, the UE does not need to monitor any of the 14 carriers until it receives a fast indication. The fast indication, telling the UE to monitor some carriers upon receiving the indication, may be sent from the eNB on a licensed carrier based on its decision on the availability of some channels. Therefore, fast carrier switching can lead to substantial reduction of UE monitoring activities and power consumption.

**Fast Carrier Switching in Uplink:** The above description of fast carrier switching is mainly focused in downlink, but the approach nevertheless applies to uplink as well, so UE can efficiently utilize more carriers over time for uplink transmissions.

In addition, fast carrier switching in uplink is important for efficient TDD operations. It helps overcome the limitations due to UE uplink-downlink carrier number asymmetry via SRS switching among multiple carriers. To enable fast carrier switching to and between TDD carriers, the network needs to first configure a UE with SRS on possibly all TDD carriers, even if the UE uplink CA capability is much less. Then the UE can switch among those TDD carriers and transmit SRS. The switching may be according to the network configuration or network indication. For example, the network may indicate to a UE to suspend its transmission on carrier 1 (TDD or FDD), switch to TDD carrier 2, transmit SRS on carrier 2 on the indicated...
time/frequency resources, and finally switch back to carrier 1. The switching may be triggered by dynamic signaling to facilitate fast load shifting among the downlink carriers with improved spectrum efficiency. Evaluations showed the potential of about 10 to 30 percent downlink spectrum efficiency improvement via SRS switching [5].

**Lightweight Carriers and Opportunistic Communications**

As opposed to existing carriers with (almost) persistently transmitted overhead (mainly CRS), a lightweight carrier sends such overhead channels only when needed. Omitting overhead in the absence of data transmission not only reduces the interference to neighbor cells, but also enables the eNB and UE to save power if some operations can be suspended and some components are turned off. Large throughput gains can be attained with lightweight carriers [4].

To support lightweight carriers, UE behavior has to be considerably modified from the current one in which UE (almost) always expects and relies on CRS. UE has to assume that CRS may not always be transmitted, and relies on alternative ways (e.g., discovery signals with much longer duty cycles, introduced in Release 12 for semi-static on/off) for long-term measurements and coarse synchronization. As data transmission starts, UE needs to be instantly “tuned” for the carrier upon receiving a fast indication from the network. To facilitate this, data transmission may be accompanied by or preceded by CRS or some enhanced RS (which may be termed “preamble”) for faster “tuning” (including automatic gain control, AGC adjustment, synchronization, channel estimation, link adaptation, etc.).

The above design leads to opportunistic communications. In opportunistic communications, both eNB and UE reduce transmission, measurement, and monitoring activities, relying on long duty cycle signals for maintaining connections on a carrier. Then they rapidly turn on (or off) the transmission, measurement, and monitoring as soon as data arrives (or completes, respectively).

It is preferred to provide a unified design to support the indications of fast carrier switching, opportunistic communications, and LBT-based discontinuous transmissions (for LAA), due to the inherent connections among the operations.

It should be pointed out that the proposed fast adaptation is meant to be complete within a very short amount of time such as at most a couple of milliseconds, but not necessarily meant to be performed every couple of milliseconds.

**Adaptive Carrier-Level Resource Optimization Schemes**

Some aspects of the lightweight carrier and opportunistic communications are closely related to fast carrier switching. For instance, the data arrival and completion described in opportunistic communications may be the result of adaptive cross-carrier and cross-eNB interference coordination and avoidance as well as load balancing and shifting. Thus, carrier-level resource adaptation algorithms are needed for fast carrier selection and switching and opportunistic communications to perform sensible operations.

**Other ACA Techniques**

A potential enhancement for fast carrier selection and switching is cross-carrier hybrid automatic repeat request (HARQ). With cross-carrier HARQ, one HARQ process is not restricted to be on only one carrier. Instead, the same HARQ process may switch among multiple carriers depending on instantaneous availability of the carriers. If the current carrier becomes unfavorable, the HARQ process can resume on another carrier rather than being terminated as is done currently. Another potential enhancement comprises CRS reduction in the frequency domain, which restricts CRS transmission in the central region of a carrier, and thus less interference to neighbor cells. If the current carrier becomes unfavorable, the HARQ process can resume on another carrier rather than being terminated as is done currently. Another potential enhancement comprises CRS reduction in the frequency domain, which restricts CRS transmission in the central region of a carrier, and thus less interference to neighbor cells, but also enables the eNB and UE to save power if some operations can be suspended and some components are turned off. Large throughput gains can be attained with lightweight carriers [4].

**Solutions and Performance**

It should be pointed out that ACA techniques do not necessarily require standards support, but standards support (fast indication design, etc.) helps improve ACA operations and performance. On the other hand, ACA techniques also heavily rely on network implementation solutions, such as adaptive carrier-level resource optimization schemes.

An adaptive carrier-level resource optimization scheme formulates optimization problems to maximize a chosen utility by determining:

- Which cells and carriers should be selected for which UE’s data transmission, and at which power level
- The on/off status of cells and carriers

There may be a few ways to implement the scheme.
Figure 3. Example of carrier-level resource optimization to assign portions of carrier resources to UEs (see Fig. 2 for the setting with 3 carriers). UEs have the same amount of average traffic load and are capable of receiving data from all 3 carriers simultaneously, but each UE is assigned to at most 2 carriers according to the optimal solution. UE1 is assigned the 2 blue portions of site A on carriers 1 and 2; UE2 is assigned with the 1 green portion of site A on carrier 3, and so on. UE4 and UE6 are on different carriers due to higher interference, UE3 and UE7 are on the same carrier due to lower interference. Site A serves more UEs; site B blanks on carriers due to higher interference. UE3 and UE7 are on the same carrier.

- Cell-level adaptation without coordination: Each cell (or each carrier of an eNB) decides the on/off status and power level independently. For example, when the cell serves no users, it may turn off.
- Cross-carrier coordinated adaptation by each eNB: An eNB controls multiple carriers, and it decides the on/off status, power level, and UE-carrier association for each carrier jointly, for purposes such as fast load shifting across all the carriers.
- Cross-carrier cross-eNB coordinated adaptation by each cell cluster: A cell cluster controls multiple eNBs and all their carriers. It decides the on/off status, power level, and association of UE to cells and carriers for eNBs and carriers jointly, for purposes such as fast interference management and fast load shifting among all radio resources within the cluster.
- A centralized coordinated adaptation by a network controller: The network controller controls many clusters and eNBs with a large number of carriers, and possibly a massive number of UEs. It decides the carrier-level resource allocation for the network, but may not be necessary for each UE due to the high computation complexity. Instead, UEs aggregated into UE groups may be considered.

Moreover, as the traffic loads, traffic distributions, and interference conditions change, the optimal allocation will adapt accordingly, generally on the timescale of at least tens of milliseconds. Interested readers may refer to [7–11, references therein] for more detailed discussions.

Figure 2 shows an example of a deployment scenario with 3 eNBs at 3 sites and 12 UEs. A centralized optimization problem is formulated to minimize the weighted sum of UE-experienced latencies [8]. Figure 3 shows an example of carrier resource allocation results based on solving the optimization problem. In this example, each UE is assigned a fraction of the time/frequency resource of one or more carriers, based on traffic load, channel, and interference conditions in the network. Some general “rules” of reuse patterns can be observed: some UEs are assigned to orthogonal carrier resources to avoid significant interference among them, while some other UEs are assigned to reuse the same carrier resource if the interference is determined to have less impact on them. A certain carrier may be turned off completely in a site as the site has lighter traffic load and/or it may cause higher interference to other sites. However, although such rules of reuse patterns appear straightforward to interpret, they are almost impossible to obtain heuristically for a rather general deployment setting, and solving an optimization problem may be the only way to obtain proper complicated reuse patterns in practical networks. The involved optimization problems are generally non-convex and intractable for small- and medium-scale networks, but several simplifications, convex relaxations, and convex approximations [8] have been proposed or are under current study, which may yield suboptimal solutions but still be effective in practice with much reduced computation complexity.

Figure 4 illustrates the performance benefits of adopting the ACA techniques. A considerable reduction in user experienced latency is observed, as well as a substantial increase of network capacity. Here the capacity corresponds to the maximum packet arrival rate (i.e., traffic load) that the network can support without “exploding” user experienced latency. The results in Table 1 show that carrier resource allocation can improve throughput performance, especially the average perceived throughput (UPT) performance for edge users without trading off cell center user performance. Note that the UPT for a packet is the ratio between the number of bits in the packet and the sojourn time for the packet, wherein the sojourn time is from the time of packet arrival to the time of complete delivery to the UE, including both the queuing time and transmission time.

The adaptive carrier-level resource optimization schemes for MC-UDNs have been applied.
to more complicated situations. For example, much denser networks with 57 cells and hundreds of UEs have been evaluated as well with similar conclusions. For another example, the average traffic loads for the UEs may not be the same across the network, which leads to more complicated reuse patterns obtained by the optimization problem. Moreover, the reuse of time/frequency resources can be based on frequency-division multiplexing (FDM) for the UEs, which may be associated with a simple $M/M/1$ queueing model, but the reuse can be based on time-division multiplexing (TDM) or more complicated rules, which may be associated with more complex models such as $M/G/1$ queues. Ongoing research has revealed further increased performance gains with the latter. Finally, wireless inband backhaul can be incorporated in this framework with appealing performance benefits.

**CONCLUSIONS**

The MC-UDN is an important scenario moving forward to 4.5G and 5G. It seems to be a necessary component for meeting the various challenges faced by the cellular industry and for fulfilling IMT-2020 requirements for 5G with very high area traffic capacity of up to 10 Mb/s/m$^2$. To address interference issues and fully utilize the increased radio resources, a number of ACA techniques are proposed, with a common theme of all of them leading to a more adaptive and opportunistic system for coordinated interference management and resource allocation. These technologies are shown to be effective in improving throughput performance and end-user experience in a dense network with multiple carriers. It is anticipated that these research studies will have profound impacts on 4.5G in Release 14, and the impacts may also extend to 5G in the near future.

**REFERENCES**


**BIographies**

**Jialing Liu** received his B.S. degree in mechanical engineering and B.S. degree in automatic control, both from Tsinghua University, Beijing, China, in 1998, and his M.S. and Ph.D. degrees in electrical engineering from Iowa State University, Ames, in 2002 and 2006, respectively. He joined Garmin International in 2009, Motorola Inc. in 2007, and Huawei Technologies in 2010. He is interested in communications, information theory, dynamic systems, control, and the interconnections among them.

**Weimin Xiao** received his B.S. degree from Huazhong University of Science and Technology, Wuhan, China, in 1992, his M.S. degree from Tsinghua University in 1994, and his Ph.D. degree from Northwestern University, Evanston, Illinois, in 2001, all in electrical engineering. He joined Motorola Inc. in 2001 and Huawei Technologies in 2009. He is interested in communications and information theory, signal processing, and network optimization. He received the Communications Society & Information Theory Society Joint Paper Award in 2002.
Non-Orthogonal Transmission Technology in LTE Evolution

Yifei Yuan, Zhifeng Yuan, Guanghui Yu, Chien-Hwa Hwang, Pei-Kai Liao, Anxin Li, and Kazuaki Takeda

ABSTRACT

Non-orthogonal transmission, although not entirely new to the wireless industry, is gaining more attention due to its promised throughput gain and unique capability to support a large number of simultaneous transmissions within limited resources. In this article, several key techniques for non-orthogonal transmission are discussed. The downlink technique is featured by MUST, which is being specified in 3GPP for mobile broadband services. In the uplink, grant-free schemes such as multi-user shared access and sparse code multiple access, are promising in supporting massive machine-type communication services. The multi-antenna aspect is also addressed in the context of MUST, showing that MIMO technology and non-orthogonal transmission can be used jointly to provide combined gain.

INTRODUCTION

Multiple access technologies of previous generations of cellular systems are mostly orthogonal: frequency-division multiplexed (FDM) in the first generation, time-division multiplexed (TDM) in the second generation, code-division multiplexed (CDM) in the third generation, and orthogonal frequency-division multiplexed (OFDM) in the fourth generation. Orthogonal transmission simplifies receiver design and scheduling implementation.

There are a few instances where signal transmissions or receptions are non-orthogonal. In code-division multiple access (CDMA) uplink, different users are distinguished by spreading codes as well as pseudorandom number (PN) codes. PN codes are not strictly orthogonal, and the uplink transmission timings are deliberately unsynchronized. Hence, signals of different users would interfere with each other at the base station (BS) receiver. In CDMA downlink, although the spreading sequences are orthogonal, multipath-induced delay spread would destroy the orthogonality at the user equipment (UE) receiver. When the spreading factor is large and the data rate is low, such cross-user interference would cause little degradation in performance. However, in the case of high data rate with small spreading factor, the interference would be a prominent issue.

Assuming that interference can be properly handled by optimal transmission schemes such as dirty paper coding and/or advanced receivers, it is proved that non-orthogonal transmission can more closely approach the capacity bound of multi-user systems compared to orthogonal transmission [1]. Figure 1 shows an example where UE1 and UE2 represent a far user and a near user, respectively, from the base station. Their transmit powers are $P_1$ and $P_2$. The channel gains are $|h_1|^2$ and $|h_2|^2$ for each link. In orthogonal multiple access (OMA), UE1 occupies $\beta$ of total time-frequency resources, with the rest of the resources, that is, $(1 - \beta)$, to UE2. The rate of each user can be written as

$$R_1 < \frac{\beta}{\beta N_0} \log \left(1 + \frac{P_1 |h_1|^2}{\beta N_0} \right)$$

$$R_2 < (1 - \beta) \log \left(1 + \frac{P_2 |h_2|^2}{(1 - \beta) N_0} \right)$$

(1)

In the non-orthogonal case, resources are shared by UE1 and UE2. The rate of each user becomes

$$R_1 < \frac{1 + \frac{P_1 |h_1|^2}{P_2 |h_2|^2 + N_0}}{1 + \frac{P_2 |h_2|^2}{N_0}}$$

$$R_2 < \frac{1 + \frac{P_2 |h_2|^2}{P_1 |h_1|^2 + N_0}}{1 + \frac{P_1 |h_1|^2}{N_0}}$$

(2)

Note that in Eq. 2, UE2’s rate is computed assuming that the interference due to UE1 is completely cancelled. This seems feasible since the modulation and coding scheme (MCS) of UE1 would be low, while the signal-to-noise ratio (SNR) at UE2 tends to be high. The likelihood of correct detection of UE1’s signal can be further improved by using the transmission schemes to be discussed later. When the channel gain difference is 20 dB between UE1 and UE2, the rate region can be plotted in the right part of Fig. 1. It is observed that the sum capacity of non-orthogonal is higher than that of OMA.

With the fast advancement in receiver implementation, interference cancellation becomes

---

Yifei Yuan, Zhifeng Yuan, and Guanghui Yu are with ZTE Corporation; Chien-Hwa Hwang and Pei-Kai Liao are with MediaTek; Anxin Li and Kazuaki Takeda are with NTT DoCoMo.
more affordable, as seen in the study on network assisted interference cancellation (NAIC) in the Third Generation Partnership Project (3GPP) [2]. This makes non-orthogonal transmission more feasible.

There are two major applications for non-orthogonal transmission: mobile broadband (MBB) and massive machine-type communication (mMTC). For MBB services, the performance target is system capacity and data rate, as illustrated in Fig. 1. The recent study on multi-user superposition transmission (MUST) [3] in 3GPP is for downlink MBB services. For mMTC, the aim is to simultaneously serve a large number of low data rate devices. Technologies such as multi-user shared access (MUSA) [4] and sparse code multiple access (SCMA) [5] can allow many users to share the same time and frequency resources.

Several techniques of non-orthogonal transmission are discussed in this article. More emphasis is on downlink since its specification work is already started in 3GPP. The article is organized as follows. The following section is devoted to multi-user superposition transmission (MUST) for downlink where we describe major schemes for transmitter assumption, control signaling aspects, and multi-antenna aspects, followed by the performance evaluation. The next section begins with the principle of grant-free transmission for uplink, and then two schemes are discussed: MUSA and SCMA. The conclusions are provided in the final section.

**DOWNLINK MULTI-USER SUPERPOSITION TRANSMISSION**

**MAJOR SCHEMES FOR TRANSMISSION**

In general, more than two users can be superposed. However, considering the diminishing returns when multiplexing more users, as well as the signaling complexity, two-user superposition per spatial layer is the focus of the study in [3].

MUST categories are featured by the transmitting schemes since they define the key mechanisms of superposition. Three MUST categories have been identified:

- **MUST Category 1:** direct linear superposition of component constellations (NOMA)
- **MUST Category 2:** linear superposition of component constellation, to ensure Gray mapping
- **MUST Category 3:** superposition with label-bit assignment on Gray-mapped composite constellation

Figure 2 shows the examples of these categories. Category 1 (NOMA) [6] can be seen as a special case of Category 2. Both support flexible power partition between near and far UEs, denoted as \( P_1 \) and \( P_2 \) in Fig. 2a. \( P_1/P_2 \) is chosen to maximize the sum-rate under certain fairness criterion, while taking into account channel gains \( h_1, h_2 \), and so on. MUST Category 2 can be generally represented as Fig. 2b, where the receiver side processing is carried out per modulation symbol [7] where \( r \) and \( m \) coded bits of near and far UEs, denoted as \( a_1, ..., a_n \) and \( b_1, ..., b_m \), respectively, are input to a bit converter denoted as \( G(.) \). The output bits are \( c_1, ..., c_r \) and \( d_1, ..., d_m \), which are separately modulated with legacy quadrature phase shift keying (QPSK)/quadrature amplitude modulation (QAM) mapper, each having \( 2^n \) and \( 2^m \) constellation points. After being scaled by appropriate powers, the two modulation symbols, \( s_1 \) and \( s_2 \), are superposed. The composite symbol \( s \) has a constellation with \( 2^{n+m} \) points. The bit conversion table ensures Gray property in the composite constellation. In MUST Category 1, coded bits are modulated directly by the legacy QPSK/QAM mapper, without going through \( G(.) \); thus, its composite symbol is non-Gray mapped.

The 16-point constellation in Fig. 2a is an example when QPSK is used for both near and far UEs, and when \( P_1 < 0.5 \) (near UE is allocated with smaller portion of the power, so its symbol has smaller amplitude, as opposed to far UE, with a symbol determining in which quadrant the composite symbol would lie). The bit conversion table \( G(.) \) in this case is simplified as

\[
\begin{align*}
c_1 &= a_1 \oplus b_1 \\
c_2 &= a_2 \oplus b_2 \\
d_i &= b_i, \text{ for } i = 1, 2
\end{align*}
\]

Four colors in Fig. 2a represent four clusters of constellations in four quadrants. Each cluster corresponds to a state of the bit-pair of far UE. These two bits form the most significant bits of the composite symbol. Due to the flexible power allocation, constellation of the composite symbol of MUST Category 1 and Category 2 is generally not uniformly spaced as in legacy QAM.

Category 3 follows an opposite approach. The constellation is legacy QAM mapped: a uniformly spaced rectangular grid with Gray property. Coded bits of the composite constellation are partitioned between near UE and far UE. These two bits form the most significant bits of the composite symbol. Due to the flexible power allocation, constellation of the composite symbol of MUST Category 1 and Category 2 is generally not uniformly spaced as in legacy QAM.

Figure 1. Comparison of sum rate region of a two-user system using orthogonal and non-orthogonal transmissions.
ally speaking, MUST Category 3 is more suitable for broadcast services where the transmitter does not distinguish between far UE or near UE. In that case, bit partition can be based on the quality of service (QoS), for instance, higher data rate with smaller coverage, or lower data rate with wider coverage.

**Receiver Assumption**

In order for near UE to achieve the rate $R_2$ in Eq. 2, the user needs to perform interference cancellation (IC) to subtract the signal intended for far UE. In the network assisted interference cancellation system (NAICS) [2], extensive study was on advanced receivers with various capabilities of IC. While the interference scenarios are not the same between NAICS and MUST, it is elaborated below that many of these receiver types can be reused for MUST. NAICS and MUST receivers are employed by different groups of users. Specifically, the NAICS receiver targets the inter-cell interference and is typically used when the user is at the cell edge, while a user close to the BS is likely to be scheduled as a MUST near user. Therefore, the same receiver can be shared by MUST and NAICS when the user is at different locations of the network.

For far UEs, receivers can be less advanced, since the interference due to the transmission to near UE would be much weaker than far UE’s own signal as well as the neighboring cells’ interference. Hence, a linear minimum mean squared error with interference rejection combining (LMMSE-IRC) receiver, which suppresses the inter-cell interference (together with the interference due to the signal intended to the near UE when the power allocated to the near UE is known), may be enough.

For near UEs, more advanced receivers are desirable as the interference due to far UE would be stronger than near UE’s own signal. Depending on the transmission scheme, different receivers would be needed. For example, for Category 1 (NOMA), codeword-level successive interference cancellation (CWIC) [1], that is, a receiver utilizing successive application of detection, decoding, re-encoding, and cancellation, is preferred since it can effectively overcome the inferior property of non-Gray mapped constellation. For Categories 2 and 3, a reduced complexity maximum likelihood (R-ML) receiver [2], that is, reduced complexity joint detection of useful and interference modulation symbols in accordance with the ML criterion, would be enough.

Apart from the interference between near UE and far UE, the receiver would also suffer from inter-cell interference, and inter-spatial-layer interference if single-user multiple-input multiple-output (SU-MIMO) is supported (e.g., in 2Tx/2Rx, 4Tx/2Rx antenna configurations). For inter-cell interference, LMMSE-IRC is assumed. For inter-spatial-layer interference, either LMMSE-IRC or R-ML can be used. When R-ML is used, modulation symbols of the inter-spatial-layer interference and the MUST near UE’s signal are detected jointly.

**Control Signaling Aspects**

Control signaling aspects include:

- Configuring a UE for MUST operation
- Providing UEs with assistance information for interference cancellation

As mentioned previously, near UEs can use advanced receivers such as CWIC or R-ML to cancel the strong interference due to the transmission to far UE. To perform interference cancellation, some parameters about the far UE’s signal are needed. The required parameters or assistance information can be obtained by blind detection or the network control signaling, or tied to specific UE assumptions.

The assistance information required by a near UE depends on the receiver type. For example, in the case of an R-ML receiver, the network assistance information should help near UE to obtain the following parameters of far UE: the modulation order, the transmission power ratio relative to near UE’s signal, and the effective channel if different precoders are applied to near and far UEs. For MUST Category 3, additional...
information about the bit allocation of the composite constellation is also needed. If a CWIC receiver is used, in addition to the parameters for R-ML, listed above, knowledge about the transport block size, hybrid automatic repeat request (HARQ) information, limited buffer rate matching assumption, and parameters for descrambling and cyclic redundancy check of the paired user are needed as well.

However, for far UEs using QPSK, the necessity of MUST control signaling would be significantly reduced because far UEs can use an LMMSE-IRC receiver with negligible performance loss. Demodulation of QPSK does not require transmission power ratio. Thus, far UEs can be legacy UEs if their modulation is limited to QPSK. According to [3], QPSK accounts for most (i.e., about 90 percent or more) of far UE’s modulation order, especially if the system traffic load is high and therefore the inter-cell interference is strong. This allows the network to enjoy the performance gain of MUST even when there are both MUST and legacy UEs in the system. In Long Term Evolution (LTE), dynamic control signaling is supported by the downlink control information (DCI) carried in the physical downlink control channel. We discuss the DCI format of MUST in the context of multi-antennas.

**Multi-Antenna Aspects**

Two aspects on multi-antennas are addressed. One is in regard to the optimality of the MUST scheme in the downlink MIMO broadcast channel (in the terminology in antenna theory, not the primary broadcast channel in LTE), and the other is about the support for various TMs.

It is known that superposition coding at the transmitter in conjunction with successive interference cancellation at the receiver achieves the capacity of a single-input single-output (SISO) broadcast channel [10]. The throughput gap between the superposition coding and the orthogonal transmission is larger when the asymmetry of channel gains of the two users gets widened. However, such a scheme is not applicable to a MIMO channel since, in the multiple transmit antenna case, there is no guarantee that there is a single user who would have sufficient signal-to-interference-plus-noise ratio (SINR) to decode everyone else’s data. In MUST, a restriction on precoder assignment is imposed to resolve the problem. Specifically, when the same precoder is used for users whose signals are superposed, a natural ordering of the users in the interference cancellation still exists even in a MIMO channel.

MUST is a universal physical layer technology to take advantage of the near-far effect that often exists in the macrocell environment. In this sense, MUST can be considered as a common feature that may generally be applied to various TMs thus far specified in LTE, and provide extra gain on top of that from antenna technologies. Currently, there are 10 different TMs defined for various multi-antenna transmission schemes, and each TM is associated with a specific DCI format. At the time of writing this article, it has not been decided yet how the DCI of MUST is designed to support many TMs. There are two possible ways. One is to consolidate the DCIs of different TMs into a new format in the context of MUST. This new DCI format would either need to carry heavy payload or only support limited choices of TMs. The other method is to redefine some of the fields in legacy DCI formats. This not only reduces the effort in defining a totally new DCI, but also makes it easy to support the legacy UEs when they are scheduled as far UEs.

**Performance Evaluation**

In order to verify the gains of MUST categories, system-level evaluations have been carried out with consideration of real deployment scenarios and practical assumptions in [3] for a fair comparison with OMA. In the following evaluations, we assume homogeneous network layout and 2 × 2 antenna configuration with closed-loop SU-MIMO. When MUST is applied, up to two UEs can be superposed in the power domain, and each UE can further have one or two data layers in the spatial domain. Four power sets are assumed for power allocation of MUST category 2, that is, (0, 1 – α), where α = 0.14, 0.17, 0.23, 0.36, which would be determined by imperfect interference cancellation due to channel estimation errors, modulation scheme, and rank. In this article, the values of α that are found to provide the highest user perceived throughput (UPT) are used. In MUST scheduling, dynamic switching among OMA and MUST is applied where the multiple access scheme that maximizes proportional fairness (PF) metric criteria is dynamically selected for each subband [6]. SINRs of scheduled user pairs are approximated from channel quality indicator (CQI) feedback [6]. Open-loop link adaptation (OLLA) is further applied to compensate the CQI imperfections. The R-ML receiver is assumed for near UE to deal with the inter-user interference. Channel estimation is realistic, based on a common reference signal (CRS). In the evaluations, FTP traffic model 1 is assumed, with packet size of 0.1 MB and resource utilization (RU) of around 60 and 80 percent.
Figure 3 shows the cumulative distribution function (CDF) of the UPT for MUST Categories 2 and 3. The UPT for MUST Category 1 is not shown here since MUST Category 2 performs better than Category 1. Transmission power alignments and the same precoder for paired UEs are assumed. It can be observed that the overall ranges of UPT are improved by MUST. In MUST, although the received SINRs become worse due to the power split between paired UEs, more time and frequency resources can be allocated to each UE, thus improving the UPT. In Table 1, MUST performance gains at 5, 50, and 95 percent and mean UPT are summarized. It is also shown that MUST category 2 provides a slightly better balance between UE UPT gain and UE fairness than Category 3, due to more flexible power allocation between paired UEs.

In Fig. 3, it is assumed that different UEs are paired for non-orthogonal transmission only when they feed back the same precoder matrix indicator (PMI) to the base station. This restriction on the UE pairing can reduce the complexity of a UE receiver. On the other hand, if each UE is allowed to report the best and second best PMIs to the base station, the scheduling and UE pairing can be more flexible, and the UPT would be further improved. This can be observed in Fig. 4 and Table 2. From single-PMI feedback to multi-PMI feedback, the MUST superposition probability increases from 20 to 50 percent.

**Non-Orthogonal Schemes for Uplink Grant-Free Transmission**

In LTE uplink, each UE is scheduled in a UE-specific manner, normally with orthogonal radio resources. Such scheduling and control leads to heavy signaling. On the other hand, in massive connection scenarios, the payload is very small, and the number of connections is huge. Large overhead will increase the energy consumption of devices. The tight control mechanism makes terminal design more complex and expensive.

Non-orthogonal transmission has been used in the uplink of IS-95, cdma2000, and Universal Mobile Telecommunications System (UMTS), which primarily serve circuit-switched voice with small packet size. A commonality of these systems is spread-spectrum, which allows multiple users to share a resource pool, hence reducing the need to indicate the time-frequency resources of an individual user. The concept of spreading can be used in uplink for massive connectivity with more advanced techniques.

**Principle of Grant-Free Transmission**

Uplink grant-free transmission is composed of the following key ingredients.

**Code Spreading:** As evidenced in the third generation (3G), spreading increases the systems’ robustness to interference and resource collision. The idea of spreading can be refined to further improve the performance.

A factor graph is an effective tool for design optimization. In a factor graph [11] a number of variable nodes are connected to a number of factor nodes. The connections between variable nodes and factor nodes define the key property of non-orthogonal access schemes. In addition, the map of connections provides guidance for receiver implementation, for example, low density spreading (LDS) [12], which can reduce the detection/demodulation complexity by avoiding metric calculation of full connections of variable nodes and factor nodes. Non-binary complex sequences have lower cross-correlation compared to binary sequences, even when they are very short, that is, 8 or 4 [4]. Such characteristics can significantly accommodate more active UEs in shared resources when those UEs randomly choose sequences.

**Mode of Operation:** In grant-free mode, link adaptation is done in a long-term fashion. Long-term means that the MCS depends only on the path loss/shadowing and/or the open loop power control. MCS is not dynamically adjusted. Since fast fading is short-term, such grant-free scheduling is not to maximize the system capacity.

Long-term MCS would be configured via radio resource control (RRC) or semi-persistent

<table>
<thead>
<tr>
<th>Throughput (Mb/s)</th>
<th>OMA</th>
<th>Category 2</th>
<th>Category 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NOMA</td>
<td>Gain</td>
<td>NOMA</td>
</tr>
<tr>
<td>Average</td>
<td>7.00</td>
<td>7.73</td>
<td>10.43%</td>
</tr>
<tr>
<td>95%</td>
<td>24.24</td>
<td>25.81</td>
<td>6.45%</td>
</tr>
<tr>
<td>50%</td>
<td>3.98</td>
<td>4.55</td>
<td>14.20%</td>
</tr>
<tr>
<td>5%</td>
<td>0.76</td>
<td>0.86</td>
<td>13.61%</td>
</tr>
<tr>
<td>Average of last 5%</td>
<td>0.42</td>
<td>0.55</td>
<td>29.35%</td>
</tr>
<tr>
<td>RU (%)</td>
<td>88.23%</td>
<td>86.99%</td>
<td>—</td>
</tr>
</tbody>
</table>

Table 1. UPT gains of MUST Categories 2 and 3.
scheduling. This is often accompanied by the resource pool configuration for multiple devices.

**Receiver Type:** Two types of receivers have gained attention:

- Successive interference cancellation (SIC) at the bit level
- Message passing algorithm (MPA)-based [5]

The SIC receiver here is similar to CWIC discussed above. For uplink, bit level SIC seems more affordable than the downlink since the base station receiver anyway needs to decode the bits for all active devices. MPA is a sub-optimal algorithm of ML detection of factor graph. The detection/demodulation process is iterative, similar to that of low density parity check (LDPC) where belief metric or extrinsic information flows back and forth between the variable nodes and factor nodes.

**Grant-Free Transmission Schemes**

In the following, two spreading code based techniques are presented.

**Multi-User Shared Access:** The basic idea of MUSA is illustrated in Fig. 5. In this example, the codeword length is 4 and there are six users. Each element of the codeword can take one of nine complex values (marked in different colors). The codebook contains \( n \) specially selected codewords. Each user can randomly pick a codeword to spread its data symbol. For simplicity, without losing generality, data symbols here are assumed to be the same (e.g., all being 1). Since the six users share four orthogonal resources, their spread sequences (channel gain scaled) are added up at the BS receiver. Codeword-level SIC is used to separate data between different users.

The design of spreading sequence is crucial to MUSA since it determines the interference between different users and system performance. Moreover, the impact on the complexity of SIC implementation also needs to be considered. A family of non-binary complex spreading sequence is being studied to achieve low cross-correlation at very short length.

**Sparse Code Multiple Access:** The recently proposed SCMA [5] is an enhanced version of low density spreading (LDS). The basis of LDS and SCMA is the same: to use low density (sparse non-zero element sequence) to reduce the complexity of MPA processing at the receiver. However, in SCMA, bitstreams are directly mapped to different sparse codewords, as illustrated in Fig. 6. Each user has a codebook, and there are six users in Fig. 6. All codewords of a particular codebook contain zeros (marked in light blue) in the same two dimensions, and the positions of zeros in different codebooks are distinct to facilitate the collision avoidance of any two users. Non-zero values in the codebooks can take various complex values (marked by different colors). For each user, bit pair (e.g., 00 for user 1, 10 for user 2, etc.) is mapped to a complex codeword in each codebook. Codewords for all users are multiplexed over four shared orthogonal resources (e.g., OFDM subcarriers).

<table>
<thead>
<tr>
<th>Throughput (Mb/s)</th>
<th>OMA</th>
<th>Category 2</th>
<th>Category 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Gain</td>
<td>Gain</td>
</tr>
<tr>
<td>Average</td>
<td>7.00</td>
<td>7.73</td>
<td>10.43%</td>
</tr>
<tr>
<td>95%</td>
<td>24.24</td>
<td>25.81</td>
<td>6.45%</td>
</tr>
<tr>
<td>50%</td>
<td>3.98</td>
<td>4.55</td>
<td>14.20%</td>
</tr>
<tr>
<td>5%</td>
<td>0.76</td>
<td>0.86</td>
<td>13.61%</td>
</tr>
<tr>
<td>Average of last 5%</td>
<td>0.42</td>
<td>0.55</td>
<td>29.35%</td>
</tr>
<tr>
<td>RU (%)</td>
<td>88.23%</td>
<td>86.99%</td>
<td>–</td>
</tr>
</tbody>
</table>

Table 2. UPT gains of multiple PMI feedbacks.

![Figure 5. Block diagram of MUSA.](image)

![Figure 6. Example of codebook of SCMA.](image)
The key difference between LDS and SCMA is that a multi-dimensional constellation is designed to generate codebooks for SCMA. The constellation takes into account the probability of each occurrence, and thus can bring a certain “shaping” gain that is not possible for LDS.

**Conclusions**

In this article, non-orthogonal transmission technology was discussed for both downlink and the uplink. In downlink, the focus was on mobile broadband services where three multi-user superposition transmission categories were described from the aspects of receiver algorithms, control signaling, and multi-antennas, together with performance evaluations. Non-orthogonal transmission can provide significant gain in downlink system capacity compared to orthogonal transmission. In uplink, two grant-free non-orthogonal schemes were briefly described, both of which are spreading-code-based and can potentially be used for massive MTC services.

**References**


**Biographies**

Yifei Yuan (yuan.yifei@zte.com.cn) received Bachelor’s and Master’s degrees from Tsinghua University, China, and a Ph.D. from Carnegie Mellon University. He was with Alcatel-Lucent from 2000 to 2006, working on 3G/4G key technologies. Since 2008, he has been with ZTE, responsible for standards research on LTE-Advanced physical layer and 5G technologies. His research interests include MIMO, iterative codes, resource scheduling, and multiple access. He was admitted to the Thousand Talent Plan Program of China in 2010. He has extensive publications, including two books on LTE-A.

Zhifeng Yuan (yuan.zhifeng@zte.com.cn) received his M.S. degree in signal and information processing from Nanyang University of Post and Telecommunications in 2005. He has been a member of the Wireless Technology Advanced Research Department at ZTE since 2006 and a leader of the team of new multi-access (NMA) for 5G wireless systems since 2012. His research interests include wireless communication, MIMO systems, information theory, multiple access, error control coding, adaptive algorithms, and high-speed VLSI design.

Chen-Max Yowucv (chienhwa.hwang@mediatek.com) received his Ph.D. in electrical engineering from the University of Southern California, in 2003. He is currently with MediaTek working on LTE and 5G advanced receivers and system design. His research interests include statistical signal processing for wireless communications, MIMO communication, and interference management of cellular systems.

Pei-Ku Liao (pk.liao@mediatek.com) received his Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, in 2007. In July 2007, he joined MediaTek Inc. in Hsinchu. His research interests focus on 4G and 5G physical-layer technologies. He participated in the specification projects of IEEE 802.16m in IEEE 802.16WG (July 2007–January 2011), and LTE Release 10/11/12/13 in 3GPP RAN (August 2009–present). He currently serves as 3GPP Rapporteur of LTE Release 14 MultiUser Superposition Transmission (MUST).

Anxin Li (anxin.li@docomolabs-beijing.com.cn) (M’05) received his Ph.D. degree from Tohoku University, Sendai, Japan, in 2003, 2004, and 2007, respectively. Since 2005, he has been with DOCOMO Beijing Labs. He is currently a vice principal researcher and research manager in the DOCOMO Beijing Wireless System Research and Standardization Division. His research interests include new radio access techniques, waveform design, space time processing, and V2X communications.

Keisuke Takokō (keisuke.takeda.ts@nttdocomo.com) received his B.E., M.S., and Dr.Eng. degrees in communications engineering from Tohoku University, Sendai, Japan, in 2003, 2004, and 2007, respectively. Since 2005, he has been a Japan Society for the Promotion of Science (JSPS) research fellow. In 2008, he joined NTT DOCOMO INC. His research interests include multiple access techniques for new radio access technologies.
CALL FOR PAPERS
IEEE COMMUNICATIONS MAGAZINE
Fog Computing and Networking

BACKGROUND
Pushing computing, control, data storage and processing into the cloud has been a key trend in the past decade. However, cloud alone is encountering growing limitations in meeting the computing and intelligent networking demands of many new systems and applications. Local computing at the network edge is often necessary to, for example, meet stringent latency requirements, integrate local multimedia contextual information in real time, reduce processing load and conserve battery power on the endpoints, improve network reliability and resiliency, and overcome the bandwidth and cost constraints for long-haul communications.

To meet the growing local and distributed computing needs, the cloud is now “descending” to the network edge and sometimes diffused onto end user devices, which forms the “fog.” Fog computing distributes computing, data processing, and networking services closer to the end users. Instead of concentrating data and computation in a small number of large clouds, fog computing envisions many fog systems deployed close to the end users or where computing and intelligent networking can best meet user needs. Fog computing and networking present a new architecture vision where distributed edge and user devices collaborate with each other and with the clouds to carry out computing, control, networking, and data management tasks.

Fog computing and networking see rapidly increasing applications in, and demands from, many industries such as manufacturing, smart cities, connected transportation, smart grids, e-health, and oil and gas. Fog computing will also be a key enabler for the Internet of Things (IoT) and 5G mobile networks. For example, fog-based services can prove effective ways to address a wide range of unique IoT challenges such as help securing resource-constrained endpoints or supporting local analytics. Fog-enabled 5G radio access networks can improve network performance, enable direct device-to-device wireless communications, and support the growing trend of network function virtualization and separation of network control intelligence from radio network hardware.

Realizing fog computing and networking imposes many new challenges. For example, how to compose, deploy, and manage distributed fog services, how to enable highly scalable and manageable fog networking and computing, how to secure fog computing systems, how should the fog interact with the cloud, and how to enable users to control their fog services provided by fog operators. Addressing these challenges necessitates rethinking of the end-to-end network and computing architecture.

This Feature Topic (FT) is designed to attract papers that will address key challenges such as those mentioned above. Authors are invited to submit complete unpublished papers that are not under review in any other conference or journal in any of, but not limited to, the following or related topic areas:

• Fog computing and networking architectures, including fog-based radio access networks
• Fog system and service management
• Fog-cloud interactions and enabling protocols
• Fog-based data services, including distributed data centers, edge data analytics, edge caching
• Edge resource pooling
• Security and privacy in fog computing environment
• Fog-enabled applications
• Trials and experimentation on fog computing and networking

SUBMISSIONS
Papers should be tutorial in nature, and authors must follow the IEEE Communications Magazine guidelines for preparation of their manuscripts. For further details, please refer to ‘Information for Authors’ on the IEEE Communications Magazine web site at http://www.comsoc.org/pubs/commag/sub_guidelines.html. Manuscripts should be submitted through Manuscript Central at http://commag-ieee.manuscriptcentral.com/. Please select “April 2017 / Fog Computing and Networking” in the drop-down menu.

IMPORTANT DATES
• Manuscript Submission: September 1, 2016
• Decision Notification: November 15, 2016
• Final Manuscript Due: January 15, 2017
• Publication Date: April 2017

GUEST EDITORS
Mung Chiang Sangtae Ha Chih-Lin I
Princeton University, USA University of Colorado at Boulder, USA China Mobile
chiangm@princeton.edu Sangtae.ha@colorado.edu icl@chinamobile.com

Fulvio Risso Tao Zhang
Politecnico di Torino, Italy Cisco Systems, Inc., USA
Fulvio.risso@polito.it tazhang2@cisco.com
**INTRODUCTION**

Profound transitions in policy, business and technology affecting broadcast (BC). The BC industry is being reshaped due to significant developments in consumer use habits, new service innovations, changing regulatory environments, tightening competition from over-the-top (OTT) media, and the scale and growth of mobile broadband (MBB) technologies and the business ecosystem. The variety of media content and services has extended from traditional linear over-the-air TV and radio to a long tail of time-shifted, on-demand, consumer-produced, hybrid content, and broadband services. Consumers have access to an increasing range of devices from large screen high-definition (HD) TV receivers to laptops, tablets, and smartphones, which are increasingly used to access and consume media services at any time and at any location, at home and on the move as a complementary second screen or, increasingly, as the primary media device. In addition to merely consuming media services, user generated content (UGC) platforms account for a growing share of consumers’ TV and video viewing, which has an impact on traffic flows [1].

The viability of digital terrestrial television (DTT) in particular has been challenged by the continuous compression and reduction of the BC spectrum, higher market-driven fees, and the evolution of competing and complementary fixed and wireless platforms for media content delivery [1]. At the same time, the momentous development of the Internet, MBB, and the associated digitalization of services are having a deep impact on the media industry. The exponential growth of MBB usage and demand for capacity from, to an increasing degree, downstream video places growing pressure for LTE to evolve further and calls for prompt availability of additional spectrum. This transformation of the operating environment affects BC spectrum holders and opens up new business opportunities as well as risks that are associated with increasing pressure to be innovatively flexible and engage in cross-domain collaboration and sharing in spectrum usage.

With these sights in mind for the future, the BC and MBB stakeholder communities have recently started to study more flexible distribution approaches including traditional cable and satellite broadcast platforms as well as fixed IP and MBB networks from a technological, business, and regulatory perspective [2]. National regulatory authorities (NRAs) — the third key stakeholder group with a double role as a key enabler and framework provider — have also shown growing interest in flexible and shared spectrum usage models. The traditional spectrum auctioning process is becoming increasingly complex due to the difficulty of finding unused exclusive spectrum as well as the high costs and time needed for re-allocation. The European Commission (EC) recently participated in this platform and spectrum policy debate by setting up
a High-Level Group (HLG) for the mobile and broadcasting sectors to deliver strategic advice on the use of the UHF spectrum [1]. Furthermore, the European Conference of Postal and Telecommunications Administrations (CEPT) Task Group 6 (TG6) report “Long Term Vision for the UHF Broadcasting Band” [3] spurred the Radio Spectrum Policy Group (RSPG) to produce an opinion on a long-term strategy on the future use of the UHF band in 2015 [4]. On one hand, NRA around the world are thinking about gradually compressing and withdrawing some lower-demand DTT licenses so that these frequencies could be repurposed for MBB use. On the other hand, the most used licences, especially those held by national public service broadcasters, can, for the foreseeable future, continue to fulfill national public service media (PSM) obligations. This necessitates regulatory flexibility, which takes into account differences in pace, and the amount of licenses across states and within specific DTT regions and border areas. In this article, we focus on analyzing flexible usage of the UHF BC spectrum by both DTT and MBB, particularly when utilizing Long Term Evolution (LTE) technology. Generally speaking, it is important to consider future services with the underlying enablers that make it possible to create business models which scale and are attractive for all involved stakeholders when developing new flexible spectrum usage concepts. Earlier business analysis regarding flexible use of the UHF BC spectrum has mainly focused on the mobile TV concept [5]. Business opportunities associated with the flexible use concept, including their framing conditions and related strategic choices for a mobile network operator (MNO), have been described in [6]. Additionally, some consideration has been given to business opportunities related to the European licensed shared access (LSA) concept for spectrum sharing [7]. Standardization of evolved multimedia broadcast multicast services (eMBMS) by the Third Generation Partnership Project (3GPP) standardization group “Enhancements for TV Video Services” (FSenTV) has studied use cases and potential requirements for TV services like linear TV as well as live, video on demand, and OTT content [8]. However, research that focuses on both BC and MBB services, and their business models and potential areas of convergence has been scarce to date. There is no analysis of the scalability, feasibility, and attractiveness of novel services and enablers related to the flexible use of UHF in earlier research, as the primary focus has been on technological capabilities and requirements. This article seeks to answer the following research questions:

- Which features of LTE technology could enable flexible usage of the UHF band?
- What service opportunities could flexible usage concepts and the evolution of LTE open?
- How scalable are these services and the associated business enablers?

The article is organized as follows. First, we present the flexible UHF spectrum usage concept and related technology enablers. After this, we introduce the theoretical background for business model scalability analysis. Next, we derive the applied research methodology as well as the scalability of service and business models for using the flexible spectrum concept, and finally, we draw our conclusions.

**Flexible Use of the UHF Broadcasting Spectrum Concept**

**System Concept Scenario**

Based on the TG6 vision, CEPT holds the view that the following four classes of long-term system scenarios could cover developments in the 470–694 MHz band while accommodating both the delivery of TV content and additional capacity for MBB in terrestrial networks:

- **Class A:** primary usage of the band by existing and future DVB
- **Class B:** hybrid usage of the band by digital video broadcasting (DVB) and/or downlink LTE
- **Class C:** hybrid usage of the band by DVB and/or LTE including uplink
- **Class D:** future communication technologies

Classes B and C represent collaborative and complementary hybrid usage scenarios, which introduce a flexible way to transfer TV channels for mobile use on smartphones and tablets while maintaining the ability to deliver TV content for conventional large screen use in the sitting room. System scenarios are positioned in the developed scenario map summarized in Fig. 1. The vertical delivery axis has been divided into unicast and BC delivery mechanisms, and the horizontal consumption axis positions technologies, with traditional BC technologies such as DVB on the left and cellular technologies like LTE on the right. Traditional BC technologies are suitable for fixed TV content provision, whereas LTE targets more mobile media. Hybrid usage concepts combine...
both technologies in a flexible manner. If the majority of content is live BC, more resources can be used for BC technology like DVB, and if the content is largely personal unicast, resources can be used mostly for LTE supplemental downlink (SDL). The BC technology used could be either traditional DVB or LTE eMBMS sharing the common LTE platform with SDL depending on the amount of mobile data.

**Spectrum Scenarios**

During the past decade, the DTT broadcasting community has experienced growing pressure from regulators to further compress UHF bands to make room for new MBB spectrum without being offered real incentives for changing their spectrum usage. As illustrated in the top bar of Fig. 2, broadcasting services were originally allocated spectrum from 470 to 862 MHz on a primary basis. The 800 MHz band (790–862 MHz) has now been deployed for MBB use throughout Europe. Furthermore, the World Radiocommunication Conference (WRC) in 2012 made a decision that the 700 MHz band will be used for MBB after the 2015 WRC. The 2015 WRC also addressed co-primary allocation of the lower UHF band (470–694 MHz) to mobile and decided to consider it as a provisional agenda item for the 2023 WRC. Some individual administrations showed strong interest in awarding co-primary allocation to mobile in their countries, however.

In the United States, for example, the Federal Communications Commission (FCC) decided to implement 600 MHz incentive auctions [9]. In the meanwhile, the key issue for regulators is to determine how the current regulatory framework associated with broadcast allocation on the 470–694 MHz band should be interpreted when considering the innovative services that could be provided over SDL in scenario class B [3]. In support of this scenario, the European Commission (EC) released a proposal in February 2016 to limit non-BC terrestrial use of this band to downlink only [10].

Recent studies [1] predict decreasing demand and relative value for DTT as the primary TV content delivery vehicle. Based on this, it could be further assumed that some “underutilized” TV spectrum might be reassigned and/or shared on a co-primary basis with MBB. As the utilization level varies between different areas and across national borders, one solution would be to re-assign channels first for MBB downlink use only, employing the 3GPP LTE SDL concept [11]. As both DTT and MBB LTE downlink-only transmitters are at fixed known locations, interference between systems could be controlled by careful network planning and, in special cases, appropriate mitigation systems. Available TV channel resources could be taken flexibly into MBB use across different regions and countries over a transitional period by leveraging functionalities developed recently for shared spectrum access such as the LSA concept [7], which maintains quality of service (QoS) and incumbents’ rights. SDL deployment enables both MBB unicast downlink and LTE broadcast usage in a flexible demand-based way. eMBMS already provides tools for cell capacity optimization, for example, to cope with growing mobile data asymmetry. Several evolution scenarios for the UHF broadcasting spectrum are being discussed in Europe to follow the market, as shown in Fig. 2.

Initially, the SDL concept could speed up deployment of the 700 MHz band for MBB through better coexistence characteristics with potential cross-border TV transmitters. The hybrid scenario increases the flexibility of band use and the efficiency of spectrum utilization by taking advantage of regional differences in DTT spectrum use. It also allows for future changes in DTT spectrum use due to new technology (e.g., HDTV). The key results of the compatibility simulations performed for the German case — where interference to LTE SDL was caused by neighboring countries’ DVB-T on band 734–790 MHz (channels 54–60) — are illustrated in Fig. 3. Simulations were based on the principles and propagation prediction methods agreed in the Regional GEO6 agreement [12] and real BC network site information. On the left, the area where LTE uplink, for example, in traditional frequency-division duplex (FDD) use, would be free from co-channel interference caused by neighboring countries’ BC transmitters is colored green. On the right, the area in which LTE terminals could receive in the SDL mode without interference in similar circumstances is likewise colored green.

Next, in a flexible scenario, DTT use could be flexibly moved to the lower UHF band when...
additional spectrum is freed from DTT. Local SDL-enabled base stations could start utilizing DTT frequencies one by one, while there is no change in the availability of interleaved spectrum used by, for example, program making and special events equipment. In the long-term integrated UHF multimedia network vision, and depending on the regulatory situation and market demand, it should also be possible to fully migrate to a converged LTE platform to deliver TV media content and thus completely replace current DTT technologies.

**Enabling LTE/LTE-Advanced Features**

Today’s fast-growing MBB traffic is primarily caused by increasing media consumption, video content in particular, and this has resulted in significant network asymmetry. This has led to growing requirements for the downlink spectrum. The SDL concept [11] shown in Fig. 4 is a form of carrier aggregation (CA), which enhances the downlink capacity of MBB networks via bonding a paired downlink channel with a supplemental unpaired downlink channel(s) in a different band into a single downlink channel with wider bandwidth. In this flexible usage concept, media is delivered over a large SDL channel operated as a secondary component carrier (SCC) while a smaller FDD band provides the primary component carrier (PCC) with supporting authentication and management functionalities. In this way, operations are not dependent on the availability of the SDL carrier. Availability of an uplink channel can enhance BC services by enabling an interactive path for BC in addition to normal broadband services. SDL and CA are enabled in 3GPP LTE-Advanced (LTE-A) standardization, but have been deployed only recently due to the unavailability of the required chipsets.

Additionally, changing consumer habits have been addressed in LTE via developing the eMBMS standard [13], which enables point-to-multipoint distribution of data on a multicast or BC basis in MBB networks. Time multiplexing of MBMS and unicast enables dynamic BC/unicasting as shown in Fig. 5. The user could receive linear broadcast content while simultaneously using unicast services, which enables novel interactive and hybrid broadcast/unicast services. A future area of interest for eMBMS standardization in 3GPP is to study use cases, assumptions, and potential requirements to enhance 3GPP systems for TV service support, including PSM requirements [8].

**Business Model Scalability Factors**

When developing any new spectrum regulatory framework and related technology enablers, it is fundamental to consider services and underlying business enablers for all the key stakeholders at an early stage. This brings our attention to value co-creation and co-capture as well as to the scalability of business model designs. The BC and MBB industries are confronting strategic environmental changes caused by novel flexible UHF spectrum usage scenarios. These are associated with, for example, emerging competitive market structures, policy and regulatory changes, as well as technological progress, convergence, and complexity. These changes all require companies to adapt or reinvent aspects of their dynamic capabilities and business models. A business model generally consists of the following elements: partners, activities, resources, value proposition, customer relationships, customer segments, channels, cost structure, and revenue streams. In analyzing business model designs and their elements, scalability has been identified as the prime factor for new business growth. Stampfl defined and categorized the antecedents of business model scalability as five mutually exclusive factors and created an exploratory model of business model scalability [14]. Proposed scalability factors for assessing service opportunities and business model design elements are:

- Technology: scalability of technical infrastructure, automation of processes
- Cost and revenue structure: superior value proposition that generates sustainable revenue early while keeping initial fixed costs low
- Adaptability to different legal, regulatory, and policy regimes
- Network externalities: network effect “lock-ins” and critical mass
- User-orientation: uniqueness and “need pull,” offer to solve real problems simply by exploiting existing user knowledge

**Analysis of Services and Their Business Enablers**

This section summarizes the research methodology used in the study and analyzes potential services utilizing hybrid use of the UHF band and the scalability of the associated business models.
Scenarios, user stories, and service business opportunities with related business model design elements were created utilizing the future-orientated anticipatory action learning (AAL) methodology [15] in a series of workshops in April–June 2015 organized in the context of the Finnish Future of UHF (FUHF) project. The AAL methodology is used to develop alternative foresights, thus representing a reflexive and iterative process of questioning and creating the future from a transformational point of view. Dialogue between cross-disciplinary participants from multiple domains is essential for this interactive and collaborative approach. Dialogue allows a range of different world views to be shared and negotiated as a basis for studying, theorizing, and otherwise engaging the future, thus helping to create it.

**Methodology**

Scenarios, user stories, and service business opportunities with related business model design elements were created utilizing the future-orientated anticipatory action learning (AAL) methodology [15] in a series of workshops in April–June 2015 organized in the context of the Finnish Future of UHF (FUHF) project. The AAL methodology is used to develop alternative foresights, thus representing a reflexive and iterative process of questioning and creating the future from a transformational point of view. Dialogue between cross-disciplinary participants from multiple domains is essential for this interactive and collaborative approach. Dialogue allows a range of different world views to be shared and negotiated as a basis for studying, theorizing, and otherwise engaging the future, thus helping to create it.

**Service Opportunity Scenarios**

The AAL method was employed to create end-user-oriented user stories for the flexible UHF concept. Results were incorporated into the developed service business opportunities. Key business model design elements were analyzed for each service opportunity. Mobile Broadband: Mobile broadband enables extra capacity to cope with asymmetric data traffic utilizing spectrum where and when not used by DTT. MNOs benefit from faster access to below-1-GHz spectrum with superior rural and indoor coverage characteristics without mandatory coverage obligations. This service opportunity allows operators to cope with the rapid growth of on-demand streamed video and UGC services with increasing user requirements on flexibility, quality experience, and content. Enhanced capacity and coverage offer differentiation through improved flexibility, QoS, and quality of experience (QoE), and enable service level pricing/data packages as well as rural MBB and interactive TV. Furthermore, extra capacity could open up nationwide wholesale options to sell BC and unicast capacity to MNOs and BC content providers to support both their own managed TV services and OTT providers’ TV services over the same 3GPP network. Flexibility also allows service capability in certain time slots in the area to be leased to broadcasters, venue owners, or content providers. New revenue stream opportunities could be fixed or shared from services, or novel time-slot auction innovations provided.

A Public Service Media: Can broadcast aggregated TV channels flexibly via broadcast or unicast, whichever way demand can be met most efficiently. It is essential that services contribute to the general national objectives framed by the regulatory, operational, and business models. In addition to statutory universal “must carry” directive obligations, PSM sets specific service requirements related to, for example, content integrity, robustness, ease of use, and emergency service availability. These are addressed in the ongoing eMBMS standardization process. Free-to-air is kept available without subscription with various government funding models. The perceived value of scheduled viewing has remained high due to ease of use, premium content, and social aspects.

Live TV/Radio Broadcast: Is very similar to PSM, although it may have different coverage, content, content protection, and funding models. It delivers linear BC of popular TV and audio channels or other curated high-quality content to smart devices and large screens either by built-in LTE or using mobile devices as sort of “set-top boxes” with screen cast. Live broadcast service offers customers user- and content-personalized interactive mobile video and audio media service innovations while reducing direct unicast capacity demand, network resource requirements, and cost for an MNO. Revenue models include subscription/data packages, pay per use, free-to-air with ads, free for premium, and revenue share from content and ad partners. New collaborative opportunities could be built on, for example, shared content ownership by multiple MNOs and enabling content providers to collaborate with multiple MNOs unbundling content access from transport subscription. Event and venue casting delivers premium content services at key events, high-density locations like sports stadiums, or local service businesses. A radio edge server solution could enhance the service and performance, for example, local recording, orchestration, and production of video streams with smaller delays and a true real-time experience. The MNO benefits via traffic offloading gains, improved QoS and QoE, and service differentiation based on good customer data, for example, purchase of in-app services like camera angles, replays, statistics, and special offers as well as pushed video ads with reach and quality of delivery. Revenue streams are generated by subscriptions, sharing or distribution fees from local entities (pushing ads), and venue or event organizers.

Media on Demand: Allows numerous users to subscribe to relevant content like news, sports, stock info, weather, educational and instructional videos, and a variety of user-generated content through either live broadcasts or device caching to view at their convenience. Users benefit from access to the latest content with or without an Internet connection (pushed and stored). MNOs benefit from cost savings (as content could be regularly updated to devices), ease of implementation, and improved QoS. Service breadth, improved QoS, and seamless service experience could, in turn, improve customer satisfaction and help in new customer acquisitions. Revenue models include subscription, bundle with unlimited data, premium to data package; ad supported, purchased premium content, and pushed by a service provider with a transmission fee.

Off-Peak Media and Software: Delivers high-demand pre-recorded content (e.g., TV shows, movies, YouTube), subscription content (e.g., eNewspapers, eMagazines, and music), applications, and firmware updates at off-peak times. This enables more efficient use of network resources, offloading benefits, increased service breadth, and improved user experience and churn reduction. Revenue models include subscription, share from content partners, pay per view, free for premium, and transmission fees.

The Internet of Things (IoT): In the flexible UHF concept, UHF could connect with the cloud to provide ease of management, location-based media services, updates, and content deliveries.
Table 1. Service and business model scalability factors.

<table>
<thead>
<tr>
<th>Service/factor</th>
<th>Technology platform</th>
<th>Cost and revenue structure</th>
<th>Policy adaptation</th>
<th>Network externalities</th>
<th>User orientation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mobile broadband</td>
<td>+ LTE scale</td>
<td>+ Existing infra</td>
<td>+ 700 MHz now</td>
<td>+ Known channels and</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ SDL CA</td>
<td>+ QoS, QoE</td>
<td>– Co-primary for lower</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ Intra sharing</td>
<td>+ Coverage</td>
<td>UHF</td>
<td>+ Billing</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>+ Wholesale</td>
<td></td>
<td></td>
<td>+ Need for video</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+ User generated content</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+ Rural digital inclusion</td>
</tr>
<tr>
<td>Public service media</td>
<td>– Adaption to PSM</td>
<td>– Existing infra + content –</td>
<td>– PSM obligations</td>
<td>– Media partnership</td>
<td></td>
</tr>
<tr>
<td></td>
<td>specific requirements</td>
<td>– densification</td>
<td>– Spectrum policy support</td>
<td>– BC prominence concerns</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ SFN, eMBMS</td>
<td>– Rural coverage</td>
<td>– standardization enhancement</td>
<td>– Timing uncertainty</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>– Business model</td>
<td></td>
<td></td>
<td>+ Reach digital natives</td>
</tr>
<tr>
<td></td>
<td></td>
<td>– Revenue share</td>
<td></td>
<td></td>
<td>+ Interactivity</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+ Multi screen</td>
</tr>
<tr>
<td>Event/venue casting</td>
<td>+ LTE scale</td>
<td>+ Sharing with venue/</td>
<td>+ Existing higher spectrum first</td>
<td>+ Local uniform user</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ eMBMS</td>
<td>organizer</td>
<td>– Availability of UHF</td>
<td>community</td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ Mobile edge</td>
<td>+ QoS, QoE differentiation</td>
<td></td>
<td>+ Participation</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+ Existing problem</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+ Good customer data</td>
</tr>
<tr>
<td>Live TV/radio broadcasting</td>
<td>+ LTE SDL/eMBMS on</td>
<td>– Net neutrality</td>
<td>+ Existing groups</td>
<td>+ Converging MBB and</td>
<td></td>
</tr>
<tr>
<td></td>
<td>demand</td>
<td>+ Coexistence</td>
<td>+ OTT</td>
<td>media needs</td>
<td></td>
</tr>
<tr>
<td>Media on demand</td>
<td>+ LTE SDL CA/eMBMS</td>
<td>– Flexible uni/broadcast</td>
<td>+ Existing groups</td>
<td>+ Converging MBB and</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>+ Efficiency</td>
<td>+ OTT</td>
<td>media needs</td>
<td></td>
</tr>
<tr>
<td>Off-peak media and software</td>
<td>+ LTE scale</td>
<td>+ No change</td>
<td>+ Apps group</td>
<td>+ Multi screen</td>
<td></td>
</tr>
<tr>
<td></td>
<td>– Local terminal storage</td>
<td>+ Revenue share</td>
<td>+ IoT segments</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IoT</td>
<td>– LTE scale for some services</td>
<td>– Cost of LTE –</td>
<td>– Uncertainty with timing</td>
<td>– Scattered IoT platforms</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>immature business</td>
<td></td>
<td></td>
<td>+ Known “things”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>models</td>
<td></td>
<td></td>
<td>+ Momentum</td>
</tr>
</tbody>
</table>

(e.g., smart meters, public TV terminals, connected cars). IoT services for vertical segments benefit from scalability, improved QoE, and extensive coverage that avoids cell edge issues. Revenues could be gathered from IoT operators that want to ensure their devices are always updated, subscription data services, and revenue from local advertisements.

**Business Model Scalability Discussion**

Developed service opportunities and business enablers were analyzed and assessed using the business model scalability factors discussed earlier. A summary of the scalability analysis is presented in Table 1.

Additional mobile broadband capacity could be first flexibly introduced with LTE SDL to address an existing problem to meet growing demand for downstream video while protecting DTT and its vital national interests. Using mobile networks on the lower UHF spectrum for broadcasting content delivery could also help promote affordable rural digital inclusion. At a later stage, when combined with scalable and flexible dynamic switching between unicast and MBMS operation on demand, or driven by the event or venue, SDL provides a scalable, efficient global platform for the various above-mentioned novel media services. In addition, sharing infrastructure between mobile and BC operators could significantly reduce the required initial fixed costs. Personalized converged MBB and media BC services help the MNO to better retain and grow a customer base with rapidly changing demands and habits, while enabling opportunities for differentiation and new revenues through sharing with media partners and IoT service companies from vertical segments. As a collaborative benefit for broadcasters, a flexible use concept extends their reach to smart devices, builds intimacy with digital natives, and enables the flexible combination of linear and nonlinear offerings and interactive services. Value will be generated via deepening audience engagement and captured through advertising and monetizing the gathering and metering of user data. Furthermore, as a future option, a global converged platform could help in addressing the high cost and long life cycle of dedicated infrastructure.

The key enabling factors for PSM and eMBMS TV services to be taken into account in 3GPP LTE BC evolution are capacity/coverage, service layer enhancements, and the enabling of novel services. The following technology platform antecedents were identified as essential for the enabling and scaling of new business opportunities: decoupled transport, service, and content; enabling of shared BC services with multiple MNOs; and simultaneous use of any LTE services and TV services from different networks. Various architecture and business model design options could be built around different current and future spectrum allocation models, spectrum framework, licensing, and operating options. In MNO scenarios, a “must carry” channel obligation could also be leveraged in shared networks through receiving incentives from PSM taxes, contributing positively to investment in the network and unbundling regulatory processes for media service and network operations.

Lack of harmonization and adaptability to different regulatory regimes could create a strong barrier in terms of scalability, in particular, for politically sensitive PSM service cases, and this has the potential to limit others through late availability of spectrum decisions. A fragmented national and global market structure deprives economies of scale and scope, raises costs, and
hampers innovation in the business ecosystem. Furthermore, from a strategic perspective, the introduction of flexible use and sharing models may impact the current exclusive spectrum licensing model of MNOs and broadcast network operators, and affect future availability.

While none of the discussed services may on their own amount to a fully scalable and compelling business case, various service opportunities built on common technology enablers and platforms could, as a whole, add significant new value for an MNO and the whole ecosystem.

**CONCLUDING REMARKS**

Co-primary allocation and its well-timed implementation using LTE carrier aggregation supplemental downlink and broadcasting concepts allows MNOs to access new unpaired downlink UHF-licensed QoS spectrum bands. This is vital in allowing MNOs to cope with growing asymmetric downlink traffic driven by video, user generated content, and new media services, and enables consumers to enjoy a wide variety of content and interactive services anywhere. As a collaborative benefit for broadcasters, the model extends reach to smart devices, builds intimacy with digital natives, enables flexible combinations of linear and nonlinear offerings and interactive services, and provides real potential for globally converged platforms for DTT to address the high cost and long life cycle of their dedicated infrastructure. For regulators, this offers additional opportunities to free up long-range spectrum for use in, for example, digital inclusion, and thereby ensures affordable access to broadband services for everyone and enables more efficient use of spectrum resources while providing flexibility to adapt to the national broadcasting context and regulations.

As in the case of any novel concept, it is of utmost importance to thoroughly investigate the associated business potential for all stakeholders. Therefore, in this article, we have developed user stories, service scenarios, and key business model elements for mobile broadband network operators utilizing the UHF spectrum with flexible hybrid usage concepts for the broadcasting operator. The results were analyzed using a business model scalability framework in order to identify value co-creation and co-capture business enablers for the ecosystem and, in particular, for BC and MBB stakeholders.

MNOs are optimally positioned to explore new business model opportunities in parallel with their traditional business models. The flexible hybrid use concept with new service co-creation models in the media domain could open up new opportunities for MNOs through access to content and improved efficiency in monetizing infrastructure. All this could help operators retain and win over new customers by offering personalized mobile broadband data and media delivery services first locally and with a focus on selected customer segments, and, at later stage and as the regulatory environment permits, on a wider national scale to provide a converged mobile broadband and media distribution platform.

The results indicate that the LTE technology platform’s scale and cost synergies, adaptability to regulatory regimes, and differentiation regarding user orientation are becoming of critical importance in the context of flexible UHF broadcasting spectrum usage for the evolution of LTE and novel 5G architectures. This can significantly re-shape the business ecosystem around MBB and media, and open up new converging and cooperative business opportunities for the media and TV industries while they undergo transformation. Interim hybrid DTT-MBB-scenario-enabled services address all the key business model scalability antecedents, leveraging the existing strong asset and capability base of established MNOs, and could be taken flexibly into use today to meet user-driven “need pull.” Meanwhile, the concept could be utilized as a test platform for integrated multimedia network services, in particular public service media, while we wait for regulatory enablement, technological enhancements, and cooperative business models within the ecosystem. While none of the discussed services may on its own amount to a compelling business case, various service opportunities could add significant new revenue when taken as a whole. In the future, business modeling studies of flexible UHF usage concepts will need to be expanded to cover cooperative business scenarios in the broadband, media, and broadcast domains. These should utilize the most prominent use cases and service scenarios in regulation, identify requirements for the evolution of LTE, and establish standardization steps, particularly for LTE broadcasting.

**ACKNOWLEDGMENT**

The authors would like to acknowledge the Future of UHF (FUHF) project consortium and particularly the partners from Nokia, Digital Networks, Elisa, the Finnish Communications Regulatory Authority, RF-Tuote, Schneider Finland, Telia Sonera, Turku University of Applied Sciences, University of Turku, VTT Technical Research Centre of Finland, YLE Finnish public service broadcasting company, Åbo Akademi University, and the Finnish Funding Agency for Innovation Tekes for their active participation in workshops.

**REFERENCES**


Biographies

Seppo Yrjölä (seppo.yrjola@nokia.com) is currently a principal innovator at Nokia Innovation Steering. Prior to joining Innovations in 2007, he headed wireless technology research at Nokia. He holds an M.Sc. degree in electrical engineering from the University of Oulu and did his postgraduate studies in the area of telecommunication and radio technology. His current focus is innovating in the area of future radios with cognitive radio businesses opportu-nities and potential disruptions as an important aspect to scout.

Miia Mustonen (miia.mustonen@vtt.fi) received her M.Sc. degree in electrical engineering from the University of Oulu in 2005. She has been working as a research scientist at the VTT Technical Research Centre of Finland in Oulu since 2006. Her current research interests include various spectrum sharing models, for example, licensed shared access (LSA). Since joining VTT, she has been actively involved in European and international regulatory activities on various topics, for example, IMT-Advanced, cognitive radio systems, and LSA.

Marja Mätinmikko (Marja.Matinmikko@ee.oulu.fi) received her M.Sc. degree in industrial engineering and management and D.Sc. degree in telecommunications from the University of Oulu. She is a university researcher and project manager at the Centre for Wireless Communications (CWC) at the University of Oulu. Previously she was with VTT Technical Research Centre of Finland, where she led Cognitive Radio Trial Environment (CORE) projects in 2011–2015. She conducts multi-disciplinary research on 5G from business, regulatory, and technology perspectives.

Pekka Talmonla (Pekka.hk.talmola@nokia.com) received his M.Sc. degree in electronics engineering and his Lic. Tech, both from Helsinki University of Technology, Espoo, Finland, in 1986 and 1994, respectively. He has worked at Nokia since 1985 in various R&D areas including TV, satellite TV, digital TV-systems, and development of mobile TV technology. Since 2000 he has worked on standardization of mobile TV, cognitive radio, local connectivity, and energy in organizations like DVB, DE, IEC, ETSI, WPC, Continua, ITU-T, and MHL. He has been involved in spectrum and regulatory work, especially related to the UHF spectrum in CEPT and ITU-R. He has also been taking an active part in various international and national research projects like Moti-vate, Wing-TV, B21C, Engines, Wise, and FUHF.
This article presents an overview of recent 3GPP developments in the area of traffic steering that are shaping the road toward the approaching 5G systems. Due to the increasing availability of RAN features addressing mobile data delivery over a variety of spectrum access techniques, a new traffic steering design is required for the evolving LTE-Advanced Pro standard. LTE evolution toward 5G brings an opportunity to introduce a native and unified approach to the coordination of radio access mechanisms in multi-radio access technology networks for efficient data delivery in mobile networks. This leads to a design of a unified traffic steering framework, aiming at the orchestration of traffic steering related features for optimal radio resource utilization. Load-based radio access network coordination is presented and accompanied by illustrative examples to visualize how the multitude of LTE-Advanced Pro features can be handled in a holistic manner. Based on the proposed solution, potential evolution directions of mobile networks are discussed as an initial step toward the standardization of 5G.

The authors present an overview of recent 3GPP developments in the area of traffic steering that are shaping the road toward the approaching 5G systems. Due to the increasing availability of RAN features addressing mobile data delivery over a variety of spectrum access techniques, a new traffic steering design is required for the evolving LTE-Advanced Pro standard. LTE evolution toward 5G brings an opportunity to introduce a native and unified approach to the coordination of radio access mechanisms in multi-radio access technology networks for efficient data delivery in mobile networks. This leads to a design of a unified traffic steering framework, aiming at the orchestration of traffic steering related features for optimal radio resource utilization. Load-based radio access network coordination is presented and accompanied by illustrative examples to visualize how the multitude of LTE-Advanced Pro features can be handled in a holistic manner. Based on the proposed solution, potential evolution directions of mobile networks are discussed as an initial step toward the standardization of 5G.

**INTRODUCTION**

Network densification techniques, new spectrum access schemes, and the increasing availability of radio access solutions geared toward spectral efficiency improvements are all enlarging the flexibility of the radio resource management (RRM) landscape. However, RRM flexibility enhancements yield an increase in complexity.

One example scenario addressing the spectrum access mechanism selection problem is presented to illustrate this problem: To cope with mobile data traffic growth, one approach may be to exploit carrier aggregation (CA) using macro sites acting as mobility anchors for moving user equipments (UEs), supported by dual connectivity (DC). Alternatively, small cell (SC) layers can offload time- and location-specific traffic, exploiting licensed or unlicensed spectrum bands. In the latter case, a licensed assisted access (LAA) [1] scheme may be more efficient for outdoor users due to the availability of interference management mechanisms. For the indoor environment however, an LTE-WLAN aggregation (LWA) [2] scheme might serve better due to the ease of WiFi access points deployment in the case of SCs using the time-division duplex (TDD) spectrum, enhanced interference management and traffic adaptation (eIMTA) allows LTE frame adaptation to instantaneous uplink/downlink (UL/DL) traffic changes. Furthermore, multi-radio access technology (MRAT) joint coordination [3] includes mechanisms to manage the shared spectrum in a dynamic manner, for example, by opportunistically refarming the underutilized 2G spectrum and assigning it to LTE. Taking into account the selected alternatives mentioned above, it is far from obvious how to coordinate these solutions. This leads to the conclusion that to steer the traffic efficiently, a proper orchestration mechanism for radio access network (RAN) features is required.

To handle non-uniform data demand, traffic steering (TS) is considered one of the most important functions in current mobile networks, being responsible for routing user data traffic via the most suitable radio resources (RRs). In heterogeneous networks (HetNets) with a multilayer, multi-RAT, multicarrier landscape, candidate cell(s) assignment is not obvious [4], and the selection of serving cell(s) based on the best signal-to-interference-plus-noise ratio (SINR) is not always optimal. Additionally, mobile network operators (MNOs) may set different policies for TS operation. For example, TS may perform load balancing across carriers during the day and push the traffic to the coverage frequency layer for energy saving purposes during the night. In general, UE-specific cell selection is a multidimensional optimization problem. To properly assign radio links to users, a TS engine needs to consider a set of inputs (radio conditions, UE capabilities, available RATs and RAT-specific features, frequency bands and layers, cell load, quality of service (QoS) requirements, network and UE power consumption, etc.).

Selected aspects of this problem have already been addressed in prior work. The METIS II project has identified the need for a holistic approach in the “agile RRM” proposal, covering a multitude of use cases [5], while the COHERENT project addressed the Central RRM Coordinator framework [6]. These concepts, however, focus solely on 5G, and do not cover aspects of current 4G networks and the already existing problems. A coordination framework for HetNets was proposed in [7], but addressed only a selected number of functionalities, including SC
sleeping and cooperative transmission mechanisms. There are more references in the literature on this topic; however, due to the character of this article, we limit ourselves to provide the selected ones.

To address the complexity of the described problem of feature coordination and to address the limitations of prior work, this article proposes a holistic, scalable, and extendable unified traffic steering (UTS) framework. It covers Third Generation Partnership Project (3GPP) features, evolution and provides forward compatibility toward 5G. The design of the UTS focuses on the available spectrum access mechanisms and their orchestration. It also considers scenario-specific feature prioritization and ranking, aiming at optimal radio resources utilization. UTS logic serves as an engine, coordinating TS policies. The goal of the UTS framework is to manage the available spectrum resources with the available radio level for real-time channel and load-aware common RRM. The aim is to provide capacity and quality of experience (QoE) improvements in a HetNet environment. Scenarios cover non-co-located (DC-based) or co-located (CA-based) utilization of carrier WiFi serving as a secondary link.

To enable a common LTE-WLAN RRM, the WLAN management work was triggered covering carrier WiFi performance monitoring and WLAN key performance indicators (KPIs) exchange with 3GPP RAN.

MRAT Joint Coordination: All of the above schemes fall under the MRAT joint coordination framework, which considers 3GPP RATs and carrier WiFi access. It aims at radio-resource-efficient MRAT management, reducing the configuration burden and signaling overhead caused by mobility management or load information exchange.

Licensed Shared Access (LSA): A new spectrum licensing method allows spectrum owners to share their resources with other MNOs [9]. An LSA advantage over the regular spectrum sharing scheme is QoS support, for example, by the use of a spectrum owner’s defined protection or exclusion zones for shared bands. LSA work analyzes the support of static and semi-static sharing in 2.3–2.4 GHz band.

RAN Sharing Enhancements (RSE): Dynamic on-demand capacity negotiations and load balancing schemes for sharing scenarios are considered in [10]. The allocation of shared RAN resources is based on the proportion of assigned RAN usage for each of the participating MNOs.

SON for Adaptive Antenna System (AAS): Release 13 introduces a network densification technique, called self-organizing networks (SONs) for AAS. It allows automated cell shaping by splitting/merging in vertical or horizontal dimensions, as well as by beamforming. Cell shaping is triggered based on the cell load or interference conditions.

Table 1 covers details of the above features, gathering their advantages, limitations, and usage requirements.

**UNIFIED TRAFFIC STEERING FRAMEWORK**

Based on the above analysis of TS related features and their relations, a UTS framework is proposed (Fig. 1). It aims to manage the available spectrum resources with the available features taking into account varying (time-, location-specific) load conditions and MNO-specific TS policies. The goal of the UTS framework is to efficiently deliver capacity when and where needed to fulfill QoS requirements and to save energy in the network when and where possible.

A **mobile devices** entity collects the radio mea-

---

1 Detailed descriptions of all remaining Release 13 features are out of the scope of this article. A Release 13 features overview is covered in [11].
<table>
<thead>
<tr>
<th>Feature</th>
<th>Input to/output from TS</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Usage requirements</th>
<th>Use case</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carrier aggre-</td>
<td>Input: Available CCs, load</td>
<td>• Improved UE throughput</td>
<td>• Increased scheduling complexity</td>
<td>• Assign all CA-capable UEs with multiple</td>
<td>Assign all DC-capable UEs with DC</td>
</tr>
<tr>
<td>gation with enhancements beyond five component carriers</td>
<td>per CC</td>
<td>• Scheduling flexibility (cross-carrier)</td>
<td>• Increased RF complexity</td>
<td>CCs</td>
<td>Switch stationary users to be served by small</td>
</tr>
<tr>
<td></td>
<td>Output: SCeLL assign/addr</td>
<td>• No need for load balancing (UL) – CA scheduling</td>
<td></td>
<td>Use cross-carrier scheduling for HetNet</td>
<td>Assign DC to moving UEs or to cell-edge</td>
</tr>
<tr>
<td></td>
<td>release, CC management, PCell assign/release</td>
<td>distributes load among CCs</td>
<td>• Multiple carriers</td>
<td>Assign random PCell to each UE</td>
<td>UEs (CTU)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Increased scheduling complexity</td>
<td>• UE support</td>
<td>• Alternatives: DC, LAA, LWA</td>
<td>• Alternatives: CA, LWA, LAA, CoMP</td>
</tr>
<tr>
<td>Dual connectivity</td>
<td>Input: User context, cell load, UE measurements</td>
<td>• Increased UE throughput</td>
<td>• Increased mobility for HetNet</td>
<td>• Multiple carriers</td>
<td>• Assign all DC-capable UEs with DC</td>
</tr>
<tr>
<td></td>
<td>Output: Secondary cell group (SCG) management</td>
<td>• Improved cell edge</td>
<td>• UE support for handling two</td>
<td>• UE support</td>
<td>• Switch stationary users to be served by</td>
</tr>
<tr>
<td></td>
<td></td>
<td>performance</td>
<td>independent links</td>
<td>• Cell clustering</td>
<td>small cells</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Decreased HO burden and LB</td>
<td>• &quot;Split bearer&quot; scheduler</td>
<td></td>
<td>Assign DC to moving UEs or to cell-edge</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Improved ROI for MNO</td>
<td>• Improved ROI for MNO</td>
<td>UEs (CTU)</td>
</tr>
<tr>
<td>etTITA</td>
<td>Input: Cell load for DL/UL balance, co-channel interference measurements</td>
<td>• Improved RU utilization</td>
<td>• UE support</td>
<td>• Assign all LAA-capable UEs with</td>
<td>• Assign all LAA-capable UEs with</td>
</tr>
<tr>
<td></td>
<td>Output: DL/UL TDD frame configuration</td>
<td>• Dynamic local traffic adaptation</td>
<td>• Cell clustering mechanism</td>
<td>unlocked carrier if in coverage</td>
<td>unlocked carrier if licensed is high</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Extra signaling</td>
<td>• More traffic in UL/more subframes for DL</td>
<td>• UE support</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Cell clustering required</td>
<td>• More traffic in DL/more subframes for DL</td>
<td>• Cell clustering</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Need for accurate interference measurements</td>
<td>• Alternatives: MBL, SON for AAS, eICIC</td>
<td>• Measurement collection and processing among RATS</td>
</tr>
<tr>
<td>Licensed</td>
<td>Input: Available CCs, load per CC</td>
<td>• Unlicensed spectrum usage enabled</td>
<td>• UE support (link aggregation)</td>
<td>• Measure traffic load, resource availability,</td>
<td>• UE class steering policy: steer video</td>
</tr>
<tr>
<td>assisted access</td>
<td>Output: CC management messages, unlicensed SCeLL assign/release</td>
<td>• Improved ROI for MNO</td>
<td>• Carrier WiFi only</td>
<td>QoS with certain granularity and provide to XHR</td>
<td>users to LTE, distribute among available layers (e.g., CA); steer voice users to CDMA;</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• &quot;Split bearer&quot; scheduler</td>
<td>• &quot;Split bearer&quot; scheduler</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• No QoS support</td>
<td>• No interface</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Lower performance than licensed carrier</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LTE-WLAN radio level integration</td>
<td>Input: Available SSIDs, cell loads, WiFi measurements</td>
<td>• Enables usage of WiFi in a network-controlled manner</td>
<td>• New UE measurements required</td>
<td>• Distribute UEs among carriers (equally or proportionally)</td>
<td>• Assign all LWA-capable UEs with LWA</td>
</tr>
<tr>
<td></td>
<td>Output: WiFi cell add/release, LWA management</td>
<td>• Improvised RU for MNOs</td>
<td></td>
<td></td>
<td>• Switch stationary indoor UE to the available WiFi</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Increased scheduling complexity (split bearer scheduler or CA scheduler)</td>
<td></td>
<td>• Apply LAA for outdoor scenarios</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• No QoS support</td>
<td></td>
<td>• Alternatives: LWA, DC, CA</td>
</tr>
<tr>
<td>Multi-carrier</td>
<td>Input: cell load, available carriers</td>
<td>• Proactive UE distribution control among layers</td>
<td>• New signaling</td>
<td>• UE support (link aggregation)</td>
<td>• Assign all CA-capable UEs with unlocked carrier if in coverage</td>
</tr>
<tr>
<td>load distribution</td>
<td>UE DLE mode camping info/reselection parameters</td>
<td>• Adds new, accurate measurements for reselection/cell camping/HO</td>
<td>• UE support</td>
<td>• UE support</td>
<td>Assign LAA if load on licensed is high</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• New UE measurements required</td>
<td>• New measurements and procedures</td>
<td>• Apply LWA, if load in licensed carrier is high.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• Need alignment with CONNECTED mode LB</td>
<td>• Alternatives: LAA, DC, CA</td>
</tr>
<tr>
<td>Multi-RAT joint coordination</td>
<td>Input: traffic load per RAT, UE measurements (MRAT)</td>
<td>• Improved RU efficiency and spectral efficiency by joint RR coordination</td>
<td>• Interfaces from different RATS</td>
<td>• Dl connectivity among carriers, (equally or proportionally)</td>
<td>• Distribute UEs among carriers (equally or proportionally)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Guaranteed QoS across RATS</td>
<td>• Measurement collection and processing among RATS</td>
<td>• Measure traffic load, resource availability,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Improved network capacity by steering UEs to appropriate RAT, including carrier WiFi</td>
<td>• Possibly new KPIs</td>
<td>QoS with certain granularity and provide to XHR</td>
</tr>
<tr>
<td></td>
<td>Output: user steering decisions (HO and reselction between RATS)</td>
<td></td>
<td>• Improved coordination with load optimization and interference management features (i.e., need updates on the actual RR availability)</td>
<td>• Under high load, request more on-demand RR from the shared spectrum pool</td>
<td>• Measure traffic load, resource availability, QoS with certain granularity and provide to XHR</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Large change in network management implementation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAN sharing</td>
<td>Input: requested RR, required on-demand capacity, MNO priorities, RR usage monitoring</td>
<td>• Dynamic RR usage optimization among MNOs in shared scenarios</td>
<td>• Required coordination with load optimization and interference management features (e.g., need updates on the actual RR availability)</td>
<td>• If spectrum not used, release RR to the shared pool</td>
<td>• If load is high in the inner part of cell, trigger split cell</td>
</tr>
<tr>
<td>enhancements</td>
<td>Output: resource availability</td>
<td>• Required coordination with load triggered network densification (macro scenario)</td>
<td>• Shared RAN scenario</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SON for AAS</td>
<td>Input: cell load info (CDR/CCD ratio) interference measurements</td>
<td>• Automatic mechanism for load triggered network densification (macro scenario)</td>
<td>• AAS antennas</td>
<td>• Under high load, request more on-demand RR from the shared spectrum pool</td>
<td>• Under high load, request more on-demand RR from the shared spectrum pool</td>
</tr>
<tr>
<td></td>
<td>Output: cell split/merge</td>
<td></td>
<td>• Proper measurements needed</td>
<td></td>
<td>If spectrum not used, release RR to the shared pool</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Coordination with other features</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WLAN management</td>
<td>Input: WLAN measurement configuration</td>
<td>• Tight interworking enabled (3GPP/WIFI)</td>
<td>• Required coordination with participating MNOs</td>
<td>• If interference is too high to edge users, trigger merge cell</td>
<td>• IeU class steering policy: steer video to LTE, distribute among available layers (e.g., CA); steer voice users to CDMA;</td>
</tr>
<tr>
<td></td>
<td>Output: QoS load info, WLAN measurements</td>
<td>• Optimized TS decisions (enabler for LWA and offloading to standalone WiFi)</td>
<td>• Inter-MNO signaling with respect to shared resources and required capacity</td>
<td>• Alternatives: MB, MBL, ICIC</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Additional signaling</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• UE measurements needed</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Interfaces between WLAN and 3GPP (XHR)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Spectrum broker entity and new interfaces</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Coordination with cell reselction, HO, DC, CA, MBL to dynamically utilize LWA band</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Measure traffic load, resource availability, QoS with certain granularity and provide to XHR</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Useful available features (e.g., CA, DC, MBL) using additional spectrum bands (e.g., add new SCeLL from LSA band to the CA)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Alternatives: LAA, LWA (antilicensed schemes)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Overview of selected Release 13 traffic steering features [11].
A unified traffic steering framework presenting the orchestration of a complex features landscape.
capacity is upgraded (e.g., more carriers are available)
• Feature ranking updates promoting more aggressive offloading when a network evolves (e.g., more SCs are deployed)
• Suggesting network capacity upgrades in case long-term congestion is identified in a certain area

SON features include:
• Standard self-optimization features (e.g., automatic neighbor relationship, mobility robustness optimization, and capacity and coverage optimization). These features are always enabled to optimize radio resource usage and to improve network KPIs.
• Standard self-configuration and self-healing mechanisms not directly related to TS.
• Measurements and QoS information supported by minimization of drive tests (MDT) and radio environment maps (REMs) presenting radio/interference conditions and user perception for proper traffic steering policies updates.

UTS FRAMEWORK

IMPLEMENTATION CONSIDERATIONS

Based on the above description, the key differentiators characterizing the proposed UTS framework are as follows.

Implementation aspects of a UTS framework:
• UTS logic is distributed among cooperating instances that are exchanging information (e.g., load, interference coordination messages, neighbor relation updates).
• UTS logic instances implemented per macro eNB nodes and X2/Xw interfaces are used for inter-site coordination.
• SCs act as slave nodes under master macro eNB management, controlling SCs within their coverage layers (to enable fast ON/OFF mechanisms and control plane/user plane split).
• UTS logic instances with a limited set of functionalities are located in SCs for the purpose of low load detection and switch-off triggering.
• A logically centralized RAN management entity orchestrates the implementation of distributed UTS logic instances.

The timescale of the UTS operations has to address dynamic load variations in HetNet scenarios.

• Traffic pattern collections and big data predictions of load patterns in the network performed within the RAN management entity are used for the dynamic parameterization of the UTS action triggers in the time domain (directly impacting the amount of the signaling overhead).
• The UTS framework does not imply per transmission time interval (TTI) granularity, in order to avoid excessive signaling overhead and potential instability of the system. The measurements, based on which actions are taken, should
be low-pass filtered, that is, ranging from several tens of milliseconds up to seconds.

**Signaling overhead of the proposed framework.**

- Radio interface signaling includes typical LTE compliant radio resource control (RRC) reconfiguration messages to assign radio links to UEs. Thus, it does not introduce additional signaling overhead compared to a system without a UTS framework. The difference comes from the fact that the UTS-based operation relies on the overall coordinated picture of radio conditions within the cluster, obtained from the UE measurements, which are further heavily low-pass filtered by the *UTS logic* to limit the overhead. The signaling overhead results from the actual algorithm running within the framework to make decisions on radio link management (e.g., change/add/delete link).

- Coordination signaling results from the interworking between *UTS logic* entities, as well as from required features interaction. The interaction includes inputs and action triggers from the *UTS logic* (e.g., cell ON/OFF messages, eICIC updates for interference management among Small Cells, etc.). As the UTS framework operates on Layer-3 level, non-real time signaling is required and thus does not provide excessive RAN overhead.

*UTS logic* operates using layer 3 reconfigurations on the radio interface; however, it interacts with the physical (PHY) and medium access control (MAC) layers in the following ways.

- **PHY layer:** In order to efficiently utilize sleep mode in the case of SCs used as capacity boosters serving the user plane layer only, a “light” radio frame (e.g., lean carrier) with DRSs is required. Additionally, the interference conditions are changed by adaptive operation using eIMTA or eICIC coordination mechanisms.

- **MAC layer:** UTS impacts scheduling decisions with the use of eIMTA, eICIC, or CoMP features. On the other hand, *UTS logic* decisions triggering SC OFF or eIMTA frame adaptations require measurements provided by the MAC layer (e.g., buffer reports, throughput, interference conditions).

**Load-Based Feature Coordination within the UTS Framework**

Individual *UTS logic* controls feature operation via actions taken upon the load changes within an access node or within a cluster of nodes. Load changes are mapped to load-based state machine diagrams (Fig. 2). Load-state definitions, state transitions, and related triggers depend on the available feature set, UTS strategies, as well as traffic load characteristics. The feature coordination approach is as follows:

- **Upon each load state transition:** Features are switched on/off.

- **Within each predefined load state:** All enabled features are operational with internal updates/optimizations (including conflict resolution among features).

UTS strategies are defined per load state and selected by the MNO; for example, in case of medium load state, macro network densification is triggered. Proposed UTS strategies\(^2\) for the load states (Fig. 2) are:

- **Low load:** *save energy:* Utilize the macro layer only; switch OFF certain carriers/cells to save energy during low traffic periods.

- **Medium load:** *densify network, increase macro layer capacity:* Add carriers, split cells, and enable CA. Improve resources utilization and interference across layers: macro IM, macro LOpt, and AAS.

- **High load:** *densify network, enable SC layer:* Switch ON SCs in hotspots; enable LAA, LWA, and DC. Improve resources utilization and interference across layers: enable SC interference management and SC load optimization.

- **Overload:** *decrease congestion:* Utilize congestion management features upon overload detection.

An individual instance of the *UTS logic* coordinates usage of the spectrum resources, RATs, and SCs under a single *macrocell coverage area* (MCCA), which defines the geographical area within the macrocell coverage (for nominal antenna configuration), serving as a coverage layer. The following UTS metrics are defined per MCCA.

- **Area capacity** (*ACap*): is defined as 100 percent of the network capacity available over the MCCA coverage layer per macrocell. It includes the capacity of all licensed frequency bands\(^3\) and all layers (including SCs). *ACap* is a scalar value defined per MCCA (e.g., *ACap* = 5 Gb/s/MCCA) that can be obtained in various ways:
  - Channel-quality-indicator-based: estimation using spectral efficiency statistics within an MCCA (i.e., weighted average of individual MCS usage over the area)
  - MDT-based: UE-report-based estimation; reports from high-load periods
  - REM-based: estimation using the location-specific dynamic interference maps

- **Area load** (*AL*): is defined as the utilization of available *ACap*. *AL* = 100 percent is the *ACap* for a particular MCCA.

- **Area load threshold:** is defined as a percentage of the *ACap*, used to trigger load-state transition, which in turn enables proper feature groups. Area load thresholds are subject to optimization per MCCA by SON mechanisms within the *RAN management* entity. Statistics of the RAN usage per MCCA are collected and stored within *RAN management* for thresholds tuning purposes. Based on potential long-term overload-state duration, network capacity extensions and upgrades could be decided by the MNO.

**Features Coordination Example**

UTS strategies are defined, or selected from the predefined set, by the MNO. These strategies should reflect the scenario-specific traffic characteristics. To visualize the range of scenarios and features covered by the UTS framework an example use case is presented (Fig. 3).

The proposed scenario consists of two areas: a dense urban (DU) area and a suburban (SUB) area. The traffic density and daily traffic pattern differs between these areas. In the case of DU areas, traffic is in bursts and condensed within a limited area, while within SUB areas traffic variations are smoother and distributed over a wider area.
The number of load states for DU areas and SUB areas also differs due to different traffic pattern characteristics, different UTS strategies, and different features used in those areas (Table 2). The network in a DU area consists of multi-band macro layer and SC layer, including various RATs (i.e., LTE-FDD, LTE-TDD, LTE-U, and WiFi). A suburban area is served by a multi-band macro layer with AAS antennas for densification purposes. Each area is covered by a separate MCCA. There are two UTS logic entities involved, handling and coordinating features and radio resource usage within individual MCCAs.

The load state transition is analyzed, in which the traffic shifts from DU areas toward a SUB area, modeling a late afternoon situation when most of the traffic is moving from the city center (business area) to residential areas. Area load (AL) for MCCA_DU peaks during daytime, whereas the AL for MCCA_SUB peaks during the evening/night. It is assumed that the AL in MCCA_DU is much higher than the AL in MCCA_SUB, justified by the fact that the traffic concentrated over a relatively small DU area during working hours (covered by a single MCCA) spreads across multiple surrounding residential areas (covered by multiple MCCAs).

**Table 2. Scenario specification.**

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Dense urban (DU) – City center</th>
<th>Suburban (SUB) – residential area</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Network configuration</strong></td>
<td>MCCA structure: Dense macro and SCs in hotspot areas; Macrocell carriers: F1 (FDD), F2 (FDD), F3 (FDD) SC carriers: F4 (TDD), F5 (unlicensed)</td>
<td>MCCA structure: regular macro Macro cells carriers: F1 (FDD), F2 (FDD) SCs carriers: NO</td>
</tr>
<tr>
<td><strong>Load distribution and load variations</strong></td>
<td>• Very low during night, increasing in morning, rush hours during daytime, decreasing in the evening • Hotspots, highly non-uniform, bursty traffic</td>
<td>• Constant low traffic during the day, more traffic during the evening and early night • Regular, uniform</td>
</tr>
<tr>
<td><strong>Load states</strong></td>
<td>1) LOW, 2) MEDIUM, 3) HIGH, 4) OVERLOAD</td>
<td>1) LOW, 2) MEDIUM/HIGH, 3) OVERLOAD</td>
</tr>
<tr>
<td><strong>Carrier WiFi/open access WiFi</strong></td>
<td>Yes (e.g., shopping malls, municipality deployed, public transportation with onboard WiFi nodes)</td>
<td>No</td>
</tr>
<tr>
<td><strong>UE mobility</strong></td>
<td>Stationary, pedestrian, vehicular: connected cars/Mi-Fi (30km/h), commuters</td>
<td>Stationary, pedestrian, vehicular: connected cars/Mi-Fi (50 km/h+)</td>
</tr>
<tr>
<td><strong>Main TS features</strong></td>
<td>LAA, LWA, CA, DC, eIMTA, RSE (4G capacity sharing)</td>
<td>AAS, CA</td>
</tr>
<tr>
<td><strong>Scenario challenges</strong></td>
<td>• Challenging interference due to uncoordinated customer premises SC deployment (HeNB) • High capacity demand • Complex coordination among features/RATs • Site acquisition problems in mature markets</td>
<td>• Coverage holes (especially for 4G) • Backhaul availability</td>
</tr>
</tbody>
</table>
energy saving and interference management purposes. The following processing is performed by the UTS logic per individual SC node (Fig. 4a):

- UEs are moved from SCs to macro carriers utilizing handover/redirection commands, and CONNECTED and IDLE mode parameters changes are updated.
- Interference Management, Load Optimization and Offload features are switched-off.
- ESM is activated for those SCs, triggering sleep mode (each SC is individually switched OFF in adaptive fashion when the traffic is low within that particular SC).
- The neighbor relation table (NRT) and PCIs are adjusted according to the new neighbor situation. RRC protocol updates measurements and cell configurations with respect to changes from ANR, PCI optimization, and ESM functions.
- When all SC are in sleep mode, the whole SC layer is switched off (i.e., SC switch to deep sleep mode, thus no longer sending DRS signals).

As the traffic shifts away from the DU area, the UTS logic in the SUB area detects the load increase from low load to medium load. The UTS strategy here is to densify the macro layer and increase the available capacity (Fig. 4b):

- Carrier aggregation with a second component carrier and SON for AAS enabling cell splitting is activated.
- Interference management and load optimization mechanisms are switched on and neighboring relations are updated accordingly:
  - MLB is adjusted to properly distribute traffic among new cells.
  - MRO settings are updated for cell splitting (note: MRO parameters can be optimized separately for split and non-split case, to avoid a constant need for adjustments each time the cell is split/merged).
  - MLB and MRO are coordinated, to avoid conflicts between them (e.g., MLB may want to move traffic toward one cell due to load imbalance, while at the same time, MRO may trigger an opposite action due to mobility performance issues).
  - NRT is updated by ANR, followed by PCI adjustment.
  - ICIC is updated to handle new interference scenario.
- RRC invokes CONNECTED/IDLE mobility procedures and parameter changes to reflect feature updates.

**SUMMARY AND FUTURE EVOLUTION**

Recent developments in the area of LTE evolution have been briefly covered in this article. Based on the observed “RRM landscape” complexity, the authors claim that traffic steering requires a unified and future-proof approach to achieve radio-resource-efficient networks. The UTS framework is proposed as a solution to orchestrate a multitude of TS-related RAN functionalities and mapping instantaneous mobile data demand to the available spectrum resources. Non-uniform traffic demand, stemming from HetNet scenarios, are evolving toward more complex cases along with the introduction of novel services like vehicular communication or a wide range of IoT applications. To cope with such variable mobile data demands in future net-
works, it is claimed that the access node capabilities, their locations, and the resulting density should also be deployed in a non-uniform manner. This requires an adaptive unified traffic steering framework, posing requirements on software-defined network-based deployments, to address the limitations of static network planning. UTS assumes an awareness of the traffic demand and the ability to optimize its power consumption within HetNet networks. This requires radio resource coordination on multiple levels: inter-RAT, inter-band, inter-site, inter-layer. UTS logic, the central node in the proposed framework, is considered to serve as an engine to coordinate TS mechanisms with respect to scenarios, feature groups, MNO policies, traffic steering strategies, and load conditions. In conclusion, the expected UTS framework benefits are:

- Unified orchestration of traffic steering features to increase the potential instability of network operation
- Optimized usage of small cells, enabling energy efficiency by the introduction of modified RRM states with two-stage sleep mechanisms
- Forward compatibility and easy integration of new RAT and related features within the UTS framework
- Consistent and unified decisions on radio links selection to serve user traffic, and to enable mobility support for multi-connectivity in future networks.

The future-proof formulation of the UTS framework was one of the working assumptions during its design, to allow its utilization in multi-RAT scenarios moving toward 5G standardization. 5G is expected to address a mobile communications paradigm shift, shifting from the current network-centric approach toward user-centric concepts, such as novel performance metrics handling in the form of user focused key quality indicators or a cell-less network design approach, supported by native SON.

To address these developments beyond LTE-Advanced Pro, the 3GPP System Architecture and RAN groups have already started discussions on Release 14 requirements. From a UTS framework point of view, the relevant architectural aspects were discussed in the Study on Architecture for the Next Generation System, which is expected to define the requirements for a next generation (NG) system [12]. According to this study, NG RAN architecture should support:

- Tight interworking between new RAT and LTE, covering inter-RAT mobility as well as multi-RAT aggregation, for collocated and non-collocated deployments
- Multi-connectivity via multiple transmission points, for collocated and non-collocated deployments
- Separation of control plane signaling and user plane data from different sites
- Interfaces for inter-site scheduling coordination
- Network slicing, allowing different logical networks to utilize the same network infrastructure
- Harmonization of MAC and higher layers
- Network functions virtualization concepts in the context of RAN architecture

The currently discussed set of tools and mechanisms within the spectrum access domain is expected to further evolve in coming 3GPP releases. For example, the addition of millimeter-wave spectrum bands to the RRM toolset opens new opportunities for future mobile broadband applications. It also creates challenges due to demanding channels and network architecture implications requiring coordination with the legacy RATs for fallback solutions.

In parallel to the initial Release 14 studies, 3GPP RAN recently held a 5G RAN Workshop collecting contributions to the 5G vision proposals, requirements, and potential solutions. The discussion indicated that the 5G network service requirements will have to cover a much wider scope of applications than is being handled in current 4G networks, such as automotive, health, energy, or manufacturing applications. This leads to a new, non-backward-compatible RAT requirement to be introduced by 3GPP under the 5G umbrella. LTE evolution will be continued in parallel to NR/5G development to address service continuity and fallback solutions. In that context, the UTS framework is continuously evolving to cover both the new aspects of 5G as well as the support of interworking between evolved LTE and NG RAT.

**REFERENCES**


**Biographies**

**Marcin Dryjanski** (marcin.dryjanski@huawei.com) received his M.S. degree in telecommunication from the Poznan University of Technology in 2008. He served as senior R&D engineer and lead expert at iFS-Wireless responsible for architecting 5G-Wireless software solutions and providing expert level training on LTE/LTE-Advanced. He was a Work Package Leader in FP7-SCINOW project targeting 5G radio interface design. Currently he serves as a RAN system specialist at Huawei Technologies designing algorithms and architecture toward 5G. He is an expert in PHY/MAC/RRM design and co-author of numerous research papers targeting 5G RAN design.

**Michał Szoboszoi** (michal.szoboszoi@huawei.com) received his M.S. degree in mobile telecommunication from Wrocław University of Technology in 2005. He was involved in 3GPP RAN work as a Nokia Siemens Networks delegate, taking responsibility for BS demodulation requirements for WCDMA and LTE/LTE-A. He was involved in the FP7-ICT SAPHIRE research project on spectrum sharing performance evaluation. He co-authored a number of research papers in the area of spectrum sharing and mobile networks evolution toward 5G. Currently he serves as a radio research engineer at Huawei Technologies Sweden AB, working on system architecture and radio networks evolution toward 5G and 6G networks.
**Trends and Challenges in Wireless Channel Modeling for Evolving Radio Access**

Paul Ferrand, Mustapha Amara, Stefan Valentin, and Maxime Guillaud

**ABSTRACT**

With the advent of 5G, standardization and research are currently defining the next generation of radio access. Considering the high constraints imposed by future standards, disruptive technologies such as massive MIMO are being proposed. At the heart of this process are wireless channel models that now need to cover a massive increase in design parameters, a large variety of frequency bands, and heterogeneous deployments. This tutorial presents the key drivers behind the evolution of channel modeling and simulation tools as a platform to evaluate future radio access technologies. This evolution comes with new challenges for the research community; this tutorial surveys these challenges, and the approaches to overcome them.

**INTRODUCTION**

The fifth generation (5G) of mobile wireless standards will be drafted by 2020. Recently, industrial and academic teams are focusing on one of its prerequisites: agreeing on common evaluation scenarios and methodologies. This step is crucial as it will be the foundation to evaluate and compare all proposals to the 5G standards. One key aspect in defining evaluation methodologies lies in the considered channel model, and its analysis and simulation tools.

Modeling how the physical channel impacts the transmitted signals has two main objectives for communication engineers. The first one is to enhance our understanding of the physics of communication systems. Knowledge of the propagation channel’s behavior is critical in order to efficiently design future communication schemes. This is even more prominent in a wireless context where the propagation medium shows high variance in time, frequency, and space. To handle this, the preferred approach is to use stochastic approximations of the behavior of carrier waves. How good these approximations are will determine how well the designed technology will work in practice. This leads us to the second objective of channel modeling: to evaluate the performance of a given technological solution in a realistic environment. Environmental complexity requires some simplifications in order to obtain a tractable tool that is used to simulate a very large number of links simultaneously. Models fulfilling the first goal, design models, will usually abstract away some of the parameters considered to focus on specific properties of the propagation medium. On the other hand, the simulation models need to cover many propagation environments and technologies. Thus, they are more complex than their design counterparts, and can be tweaked and configured to handle a larger number of scenarios.

In this article, we highlight the new challenges and requirements that 5G imposes on channel modeling and channel simulation tools [1]. In particular, we analyze how the massive increase of antennas, the move to higher bands, and the increasing heterogeneity of cellular deployments are moving the frontier of current modeling technology [2, 3]. We provide a unified view of previous analysis and tutorials, in mobile radio access channels [3] and more specific vehicular and machine-to-machine (M2M) channels [4, 5]. We also address current topics and trends such as mmWave [6], massive multiple-input multiple-output (M-MIMO) channels [7, 8], 3D channels [9], map-based modeling [3, 10], and non-stationary channels [8, 11]. We review classical models and background information in the following section, and analyze the key drivers in wireless technology that will shape the future of channel modeling after that. This allows us to then identify the broad research trends to pursue in the standardization process. We conclude with a discussion of the challenges ahead.

**SIMULATING WIRELESS CHANNELS**

Building new models for propagation channels must start with the definition of the specific scenarios of interest. A scenario is defined by a typical usage of the channel in a typical environment, with the goal of answering a specific question. As a bridge between the models used and reality, great care is usually placed on defining the scenarios of interest in the standardization process; their definition is subject to much debate among the actors in standard bodies.

Simulation models then strive to reproduce the behavior of the chosen scenario as realistically as possible. However, there is an inherent trade-off made in practical channel simulation. For example, one could always model the environment exactly and solve Maxwell’s equations. Notwithstanding the computational complexity...
of this approach, it would give a very accurate answer at the cost of parametric complexity. It would also lack generality, in the sense that multiple instances of the environment would have to be averaged to obtain the behavior of the channel in the scenario of interest. We can analyze most models under these terms, and consider how they trade off in terms of:

- Accuracy, that is, *how close is the realization to a specific instance of my scenario?*
- Generality, that is, *how typical is the realization with respect to my scenario?*
- Simplicity, that is, *how complex is it to parametrize and run my simulation?*

These characteristics are somewhat antagonistic when considering abstract scenarios. One intuitively expects models with a large number of parameters to be less general. In essence, solving Maxwell’s equation exactly or in a simplified manner through ray-tracing would score perfectly if the scenario of interest was limited to a specific environment. The map-based model of [3] is an example of such a ray tracer; an even more extreme case is using measurements as a simulation platform (e.g., as in [12]). Overall, most models will favor some objective rather than others. Design models have a high emphasis on simplicity, preferring lower numbers of parameters capturing the essence of the model, while ray-tracer types tend to value accuracy rather than generality.

Most simulation models stem from simplifications of the ray launching paradigm [2]. One way to do so is to concentrate on the endpoints, or *drops*, which are an abstraction for the users’ locations in a large network. The interacting objects in the environment, also called *scatterers*, are then randomly generated, parametrized, and associated with each drop to form a statistically representative channel. This *user-centric* approach is the one chosen by the Third Generation Partnership Project (3GPP) model in its first iteration, the spatial channel model (SCM), and its derivatives. Another approach is to consider a *scatterer-centric* approach, where scatterers are generated globally so as to form a virtual environment. Every user then shares the scatterers, mimicking ray tracing with a finite number of known interacting objects. This is the preferred approach of the directional channel model (DCM), which culminated in the COST2100 channel. All these models and related approaches are described in more detail in [2, 3, references therein].

We illustrate the relative strengths of each model family according to the three aforementioned dimensions of accuracy, simplicity, and generality in Fig. 1. This allows quick assessment of how they trade off the different objectives and thereby how one may choose a model depending on specific objectives. While the latest iteration of these models cover the current standardization needs, further developments are required to account for the requirements of future communication systems [1]. As the use cases of the network evolve, the scenarios of interest evolve with them, and simulation models need to cater to these new simulation needs. Identifying the key drivers behind future technology is therefore of paramount importance in order to efficiently work toward the next iteration of channel models.

**Key Drivers for New Channel Models**

There still is much debate over the specifications of future mobile networks and the associated technologies [1]. While detailed in [3], the state of the art is beyond the scope of the article, at their core is a need for higher spectral efficiency and a move toward more varied usage scenarios of interest in an increasingly heterogeneous network (HetNet) infrastructure. This, in turn, leads to a number of specific technological paths that we discuss now.

**Large Antenna Arrays**

One way to dramatically increase spectral efficiency is by using multiple antennas simultaneously at the transmitter and/or receiver side in order to exploit multi-path propagation and the inherent *multi-user* diversity of the wireless channel [2]. During the last decade, multiple antennas have been integrated on devices and base stations (BSs) in order to achieve the theoretical promises of increased capacity. MIMO has become an essential element of wireless communication and is present in many standards including but not limited to IEEE 802.11/ac, IEEE 802.16d, and the 3GPP standard family related to Long Term Evolution (LTE). The technology has also been extended to improve both the robustness and performance of single links, as well as a multiple access method to simultaneously serve users separated in space on the same frequency band.

The trend now goes in the direction of massively increasing the number of antennas, which opens a number of new technological possibilities [7]. In a way, M-MIMO gives the transmitter much more freedom in designing its output energy pattern. The large number of antennas enables very precise *beamforming*, targeting users without creating interference or selectively removing interference from some points in space. The increase of the number of antennas

![Figure 1. Categorization of the trade-offs made by different modern simulation and design models used for cellular network research and standardization. The SCM and its family (SCME, WINNER, Quadriga, etc.) are derived from the standardized 3GPP model. The COST model family, and ring models at large are discussed in [2, references therein], as well as in [3]. The METIS map-based ray tracing model specifications can be found in [3].](image-url)
can either take the form of co-located antennas placed on wider panels such as walls or be achieved by aggregating distributed antennas from various sites, leading to so-called distributed MIMO (D-MIMO).

A very specific need for MIMO is related to beamforming and the considered antenna patterns. While many cellular simulations have been limited to a 2D plane with linear antenna arrays, advanced MIMO communications require a complete description of the space. Adding the elevation dimension was an intricate task, which required model adaptation as well as new measurements. Most of the current simulation models have integrated this 3D modeling for both transmitting and receiving antennas, thereby enabling the evaluation of more complex antenna structure and scenarios (Fig. 2). Full-dimensional assumptions are now the norm, and are integrated into the latest iterations of the major simulation models [9]. However, the parametrization and calibration of these models for different scenarios are still ongoing matters.

NEW FREQUENCY BANDS

Another approach for increasing the wireless capacity is by reaching toward new frequency bands with a lot of available bandwidth. Advances in hardware have enabled communication scientists to consider alternative bands out of the overloaded conventional licensed ones. Frequencies below 6 GHz are more and more flooded by communication systems with low quality of service (QoS) guarantees such as WiFi and Bluetooth. They also handle most of the land-to-land mobile communications today. As such, the behavior of carrier waves in this frequency range is now well understood.

Beyond 6 GHz, multiple bands are starting to be considered for future communication systems such as the 28 GHz and 73 GHz bands, avoiding the high absorption effect of oxygen at 60 GHz [6]. These frequencies are in the process of being decommissioned from their earlier uses, if any, and re-licensed for mobile communication worldwide. Since they were never used for ground-to-ground communication, their propagation characteristics in radio access scenarios are not yet well known. Being free of regulations and used in the IEEE 802.11ad standard, the 60 GHz band is an exception and has been extensively evaluated in indoor scenarios. A large body of work has thus improved our understanding of these bands, although multiplexing capabilities and multipath behavior are still under study (e.g., [6, 12, references therein]). All these elements, in addition to the standardization efforts, have motivated researchers to launch measurement campaigns aimed at better understanding how carrier waves behave at millimeter-wave (mmWave) frequencies, and how different this behavior is compared to lower bands.

NEW DEPLOYMENTS

In addition to the above requirements related to new technologies that came as a response to the promised 1000-fold throughput improvement [1], there has been a lot of effort in proposing new network architectures. The idea of using smaller and denser cells in particular has emerged to the forefront, after being discarded in the early days of cellular networks due to the complexity and cost of their deployment. Access points getting closer to the user means that the network is more dynamic in nature, with frequent switches between serving BSs from the user point of view. A user could also be served jointly by multiple BSs, a possibility that is already enabled in the more recent LTE standard through the so-called coordinated multipoint (CoMP) schemes. Moreover, the emergence and exponential increase of connected objects ranging from vehicles to wearables bring in a completely new kind of propagation channel to be investigated and modeled. This trend is commonly referred to as the Internet of Things (IoT). In the future, we also expect that direct communication between users and objects, in a device-to-device (D2D) manner, might be used to either offload some part of the traffic in the network or improve coverage. Such extensions are already studied in current 3GPP standards as proximity services (ProSe). Future 5G networks, such as IMT-2020, will thus most certainly use a wide variety of link types and technologies in order to achieve the performance and flexibility that are expected from them [1].

All these new techniques and network deployments differ wildly from the classical architecture from a channel modeling point of view. For example, in a macrocellular network, the BSs are at elevations of several tens of meters. On the other hand, micro and femtocells would be at heights of a couple of meters. This difference of perspective dramatically changes the nature of the propagation channel, the obstacles and their nature, as well as their density and distribution. This implies that the conventional channel models defined for a classical cellular system are no longer valid and must be updated or replaced if necessary. In addition to these propagation constraints, we notice that for some new scenarios of interest, both endpoints are possibly mov-
Figure 3. Propagation between two antennas arrays, comparing the true circular wavefronts (in blue) and the classical planar wavefront approximation (in red). For the planar approximation, the distance between the antennas — and thus the phase offset — depends only on the antenna angle with respect to the impinging angle of the wave and the receiving antenna separation. When the wavefront is not planar, however, the phase offset has to be computed from the distance between antenna pairs. This increases the diversity in the channel and thus improves the overall performance.

Consider a practical case of such a small cell on the side of a building in a street canyon, as exemplified in Fig. 2. Using a comparatively low number of antennas in a vertical array, the smaller cell can then be expected to discriminate its users in space, and achieve large multi-user MIMO gains. While this conjecture has been validated in practice (e.g., in [9]), one remaining question is how to model the channel seen through these larger antenna arrays. In current models, there is a basic assumption that the antenna aperture is so small with respect to the distance between the transmitter, the receiver, and its scatterers. Consequently, they may be considered as points, and the wavefronts on all antennas can be modeled as parallel planes. This has a tremendous impact on the complexity of simulations, as all antennas will see similar paths from the transmitter to the receiver. The path for one antenna pair can be computed and subsequently replicated over the array, accounting for the small spatial displacement between neighboring antennas. This hypothesis is realistic for macrocellular BSs with co-located antennas and distant users. However, in novel uses of the channel, this breaks down in a number of ways.

A key difference lies in the stationarity of the channel model over the antenna array, an issue we discuss below. Another difference is that as the array size increases, cell coverage gets smaller, and BSs move closer to their users, thereby crossing the Rayleigh distance. Consequently, propagation characteristics change from the far field to the radiative near field, and the wavefront can no longer reliably be assumed to be planar [14]. Note that one can still approximate the general solution of Maxwell’s equation by the ray launching paradigm; in this case, however, the phase offset of the wave between adjacent antennas depends on the exact distance between transmit and receive antenna pairs, rather than the receiver antenna separation and angle of arrival of the wave (Fig. 3). While this will increase the computational complexity of the model, this also comes with a positive aspect. Line-of-sight MIMO communications with a low number of multi-path components are not expected to provide much gain in the far field. However, the authors of [14] have shown that in the radiative near field, this planar approximation can actually cause the capacity to be severely underestimated for close range communications.

**NEW TRENDS IN CHANNEL MODELING**

These new scenarios highlight both deficiencies and opportunities in the channel models and simulation tools currently in use. On one hand, they make a number of simplifying assumptions, which have been validated with respect to their original application scenarios. When the application scenarios change or evolve, those simplifications may not hold anymore, and the fundamental model needs to be adapted or deeply revised. On the other hand, analyzing the physical behavior of wireless propagation channels may lead to new opportunities for system engineers. Specific behaviors can be harnessed through advanced algorithms as enablers for new communication algorithms. One common example is the two-level precoding method described in [12].

**HIGHER SPATIAL RESOLUTION**

The consideration of elevation — the third dimension — in channel models has already prepared the technical tools to evaluate advanced beamforming [9]. A very promising aspect of elevation beamforming is for smaller cells with large antenna arrays with respect to the user distance.
large scale, for example, the presence of buildings or massive objects between transceivers. Receivers that are geographically co-located will have a global path loss that is correlated. The treatment of correlation at the scatterer level, on the other hand, varies between the approaches and model families. In scatterer-centric modeling, scatterers can naturally be shared between different users or linked together as twin clusters [2]. The evolution of the channel state on short time and spatial scales is thus handled by the model. Meanwhile, user-centric models have limited this joint modeling to large-scale parameters such as shadowing [3]. This approach has a clear advantage from a complexity point of view. However, it fails to capture the finer aspects in the joint modeling of propagation links, which future technologies are expected to exploit. Remote antennas, and D-MIMO in general, are in part reliant on this joint evolution of the link to provide capacity gains. However, one risks overestimating the gains of cooperation if the joint distribution of the links is not taken into account. Few simulation tools are able to properly assess this, as even ray tracers usually do not correlate the phases of similar paths to the destination [3].}


data-centric propagation models

A common assumption in most channel models is that their distributions are stable in time, frequency, and space. In particular, it is often expected that the autocorrelation of the channel impulse response is wide-sense stationary (WSS) over time, and that fading due to the scatterers is uncorrelated and independent between different drops in space [2, 13]. For example, in the 3GPP SCM, correlation in time is considered only through a Doppler component [2]. Each user is associated with a speed vector, which translates into Doppler effects for each sub-path from each scatterer. The underlying approximation here is that a moving user will see a similar environment during its displacement so that large- and small-scale parameters are fixed during the simulation of a specific drop. How large this displacement can be in time or frequency can be analyzed and estimated through measurements [13]. It may also be analyzed for specific channel models [8]. However, in some scenarios of interest, the channel statistics are not constant over the transmission period, and scatterers around the transceivers can appear and disappear rapidly. For example, this is the case for D2D in very cluttered areas [5], or in communication with or between high-speed vehicles such as trains and cars [4, 11]. As we discuss below, parameter maps can handle a part of this problem from the channel simulation side by correlating large-scale parameters over space and time. Specific simulation models have also been proposed to simulate fast-varying non-stationary channels with adequate parametrization [11].

For larger M-MIMO arrays, there is strong evidence that all antennas will not see the same interacting objects, or even the same users in extreme cases [7]. This relates to stationarity and homogeneity of scattering effects over an antenna array in the spatial domain [2]. While this effect was negligibly small in classical MIMO, it is expected to be of notable importance for M-MIMO. It opens up new options for user separation and grouping in the spatial dimension, as illustrated, for example, in [12] for an M-MIMO system at 28 GHz. In essence, one can use the second order statistic of the spatial channel to extract and group users with common scatterers or similar spatial signatures. How to address such a model this form of non-stationarity in space is still open to discussion, while the preferred models are birth and death processes of scatterers in time and space [8]. A similar phenomenon appears in D-MIMO scenarios, where neighboring users will share scatterers that evolve jointly over time and space.

**Data-Centric Propagation Models**

Heterogeneous cell sizes and deviations from the cellular paradigm, such as D2D, need to be characterized by a plethora of new radio propagation models, each of which comes with a large parameter set. It is a dominating trend that these propagation environments are characterized more and more accurately with the help of environmental data. The 3GPP SCM already defined maps of so-called large-scale parameters in order to capture the joint spatial evolution of characteristics such as shadowing. New models define very specific scenarios based on building geometry and attenuation material, with the best example being the proposed map-based channel model of the METIS project [3]. This solution seems to be one way to answer the problems discussed up to now with tractable complexity, and at the same time support new modeling constraints, new deployments with dual mobility, mesh networks, and non-stationarity in addition to the conventional parameters supported by stochastic modeling. The model uses multi-dimensional maps of attenuators based on the frequency and positions of the transmitter and receiver, and their relative velocities and bandwidth, which are then complemented by randomly placed scatterers and reflectors.
A completed radio map of Berlin, Germany, reconstructed from online measurements [10]: The heatmap shows radio propagation loss between the corresponding position and the strongest serving sector over a geographical area of 56 km², while the red lines indicate the main roads of the downtown area. Further overlays illustrate building structures in light gray and the stations as blue triangles.

Figure 4. A completed radio map of Berlin, Germany, reconstructed from online measurements [10]: The heatmap shows radio propagation loss between the corresponding position and the strongest serving sector over a geographical area of 56 km², while the red lines indicate the main roads of the downtown area. Further overlays illustrate building structures in light gray and the stations as blue triangles.

The resulting propagation models are highly diverse, use a large set of data, and often sacrifice generality for accuracy. But an even more practical approach is to directly use radio propagation measurements for the analysis and design of the radio access. Such propagation data is widely available, and can be constructed via different means. One solution would be through extensive simulations by means of ray-tracers. Since modern handsets provide cost-efficient access to measurements of radio signal strength and geographical position, data collection in measurement campaigns during normal operation could provide another solution. Operators can access this information via the standardized interfaces, for example, in the latest LTE specifications for the minimization of drive tests (MDT), or through over-the-top service providers such as http://opensignal.com/ and http://www.rootmetrics.com/, who collect and offer such information.

These new, cost-efficient data sources are appealing for modeling radio propagation based on data alone. However, accurate channel modeling is still necessary, albeit in an indirect hidden manner. The varying and partly unknown accuracy of the above data sources requires careful processing to detect outliers and to interpolate incomplete information. As it is often impractical or impossible to measure the channel at every geographical position in a given area, radio propagation maps are often incomplete in terms of area, altitude, and sample space. To fill in the gaps, the community has successfully applied general radio propagation models to parametrize generic methods such as Kriging, matrix completion, and support vector machines [10]. As illustrated in Fig. 4, radio propagation measures may be augmented by further integrating information such as building geometry, street maps, and base station positions, to improve the parametrization and to reduce the search space for machine learning methods. Such combination of generic models with specific scenario data provides an interesting trade-off between generality and accuracy, and is thus a promising field of future research.

OUTLOOK AND CHALLENGES

Overall, there is a global need for more advanced modeling of radio links. Advanced models need to be more precise, cope with the fact that common approximations do not hold anymore for new radio technologies, and account for coupled links. This translates into higher computational complexity for simulation tools, and larger sets of parameters to consider for the models. As computational power increases, the former can be handled through better computing equipment and algorithms, improving over time. The latter, on the other hand, is a burden that has yet to be tackled. Radio maps provide an immediate but incomplete and highly complex answer considering the amount of data that needs to be collected, stored, and processed. As models get more complex simulations they lose generality and consequently have to be replicated over varying scenarios to derive general findings.

Beyond the large amount of measurements to parametrize the models, the models' dominating factors are still unclear in new scenarios such as mmWave and high-speed ground transportation. This requires further measurement campaigns before the actual model validations can start. From the technological drivers described here, we identify the key environmental parameters that we expect will influence the performance of future networks. This definition process is already taking place, as we can see in Fig. 5. Large partnerships are aiming toward the 5G standardization process. 2016 will see a number of channel measurement efforts in 3GPP, led by all the partners, as discussed at the last meeting of the RAN work group. In parallel, both the new METIS-II project and the IMT-2020 partnership will produce their standard evaluation scenarios, which will precisely define the needs and expectations of future RANs with respect to their simulation models.

Finally, let us reconsider the primary purpose of channel modeling for wireless telecommunications: providing a foundation for the design of disruptive technologies. In parallel with improved simulation tools, there is still a need to capture these essential channel behaviors into simpler design models used by more theoretical researchers. These models need to be ahead of the standardization curve. In general, the remaining open problems are similar to those expressed in this article, with a small set of parameters and greater generality. For example, as of today, there are no usable design models for D-MIMO that accurately reproduce the joint evolution of links between different access points and specific users. The community also lacks more precise measurements and analysis of the stationarity intervals of channels in time, frequency, and space, a discussion that is related to the joint behavior of links over these dimensions. An accurate estimation of the channel statistics is a requirement for advanced communication algorithms, stationarity is an important operational factor.

As large measurement campaigns are performed to move toward 5G, this data may be used to extract new design models and opera-
Figure 5. Timeline of recent and upcoming milestones in channel modeling, with links to the standardization process, as planned in the last 3GPP Radio Access Network (RAN) group meetings. As of this publication, most partners in the 3GPP will have engaged in large measurement and validation campaigns related to the issues and challenges identified in this communication. As one can see, 2016 is pivotal in the number of measurement efforts, as well as the scenarios and evaluation framework definition for the future 5G standards.

REFERENCES


BIographies

PAUL FERRAND received his M.Sc. degree (with honors) from the University of Lyon, France, in 2009 and his Ph.D. degree from INSA Lyon, France in 2013. In 2014, he joined the Mathematical and Algorithmic Sciences Lab, France Research Center, Huawei Technologies Co. Ltd. as a junior research engineer and member of its Signal and Information Processing team. He is interested in signal processing for transceiver design, cooperative networking, and smart resource management, in both body area networks and cellular networks.

MULTAPASA AGUS received his M.Sc. in 2007 from Telecom ParisTech, and defended his Ph.D. in 2011 at EURCOM. From 2011 to 2013 he occupied a research engineer/teacher position at Conservatoire National des Arts et Métiers (CNAM), Paris, France. He joined the Mathematical and Algorithmic Sciences Lab, France Research Center, Huawei Technologies Co. Ltd. in 2014. His research interests are in future wireless-communication systems, that is, multi-user wireless systems, cognitive radio, and device-to-device communications.

STEFAN VALENTINI received an M.A. in electrical engineering from the Technical University of Berlin, Germany, in 2004 and a Dr.rer.nat. in computer science from the University of Paderborn, Germany, in 2010. In the same year, he joined Bell Labs, Germany, as a member of technical staff, where he worked on wireless resource allocation algorithms for 4G and 5G. Recently, he moved to Paris to lead the Context-Aware Optimization team at Huawei’s new Math and Algorithms Lab. His research interests are channel prediction, context-aware resource allocation, and anticipatory adaptation.

MAXIME GUILLAUD received his M.Sc. degree in electrical engineering from Ecole Nationale Supérieure de l’télécommunication et de ses Applications, Cergy, France, in 2000, and his Ph.D. in electrical engineering and communications from ENST (now Telecom ParisTech), France, in 2005. From 2006 to 2010, he was a senior researcher at the FTW in Vienna, Austria. From 2010 to 2014, he was a researcher with Vienna University of Technology, Austria. Since 2014, he is the head of the Signal and Information Processing team at the Huawei French Research Center. His recent interests revolve around the physical layer of fifth-generation cellular networks.
Recent advances in mobile and wireless technology have affected many areas of our lives, including the way we communicate, and how we interact and collaborate. Along with this, we have also witnessed a rapid increase in the mobility of people, thanks to faster transportation methods, simplified connections, and shorter commute times. These developments mean that studying the technological impact of mobile ad hoc and sensor networking now also requires studying aspects of social and human mobility. Mobile and wireless networking now involves people with networked sensors, mobile devices, and sensing context. The performance of each system then also depends on how people interact with these systems.

The sensing context is often identified with the Internet of Things (IoT) [1]. This term was coined in 1999 but has only turned from a concept to reality in recent years, due to advancements in the enabling technologies. For example, ZigBee and Z-Wave have been used in home automation. Bluetooth has designed a low-energy mode that is tailored for smartphones. IEEE 802.11ah is in the final stages of standardization, and products are expected to hit the market soon, making whole-house coverage for an IoT a reality. The next generation WiFi 802.11ax community has established a Long Range Low Power (LRLP) Topic Interest Group (TIG) to address its interest in expanding into the IoT domain. The fifth generation (5G) mobile networks show ambition to integrate cellular and other technologies into a cloud network, resulting in reduced energy consumption, increased capacity, and scalability.

Various companies have been pushing forward additional devices to the increasingly heterogeneous IoT network; other companies have been working to improve the efficiency of medium use. More and more automation devices, both industrial and commercial, are being deployed and distributed. Some devices have high mobility, such as cameras, smartphones, drones, and wearables (biometric, fitness, wellness, or infotainment devices). Other devices, once deployed, are stationary: these include utility sensors and parking meters. It is expected that the IoT will play a central role in the information ecosystem by 2020, covering the majority of the world’s population. For example, China UNICOM plans to promote 20 million mobile IoT subscribers in 2016. China Telecom is also developing solutions for connected vehicles.

Several results in the 1990s, such as the small world phenomenon [2] and the Dunbar number [3], gave us a surprising view of how our social world is “small” (i.e., connected with a small number of hops) and cannot go over certain physical limits (i.e., we have a limited number of strong connections) [4]. These fascinating results have had an impact on both the research and the use of mobile and wireless technology. In some cases, these results even influenced public policy making. For example, China and the United States have both begun big data initiatives that underscore the importance of massive in-the-field sensing. Quite a few major Chinese cities have built, or are in the process of building, city-wide infrastructures to enable mobile and ad hoc sensor networking. The city of Shanghai has gone beyond this by making a lot of the resulting data available to researchers and practitioners as a means of exploring new applications.

An area of mobile ad hoc and sensor networks where the human and social aspects are fundamental is crowdsourcing or crowdsensing, which relies on contributions from a large number of people to accomplish certain tasks. The success of crowdsourcing therefore depends on the participation of personal mobile devices. One of the challenges in long-running crowdsourcing systems lies in the capacity to attract new volunteers. More importantly, the challenge is to leverage existing social ties between volunteers to keep them involved so as to build long-lasting crowdsourcing communities. In addition, the advent of high-performance devices and ad hoc communication technologies can help to further amplify the effect of sensing actions in the proximity of the volunteer devices.

The article “Spatial Crowdsourcing: Current State and Future Directions” discusses the unique challenges of spatial crowdsourcing, where tasks are location-specific and thus require people to be physically at specific locations in order to complete them. In this article, Yongjian Zhao and Qi Han provide a comprehensive view of this new paradigm.
by introducing a taxonomy, and suggest several interesting areas for future research.

On mobile crowdsensing, the article “Empowering Mobile Crowdsensing through Social and Ad Hoc Networking” describes how to exploit socio-technical networking aspects to increase the performance of mobile crowdsensing campaigns in the ParticipAct living lab, an ongoing real-world experiment that over two years involved about 170 students of the University of Bologna. Luca Foschini, Antonio Corradi, Stefano Chessa, and Michele Girolami also report some significant experimental results to quantify the effectiveness of the proposed techniques.

The concepts of wisdom of the crowd and collective intelligence have been utilized by mobile application developers to achieve large-scale distributed computations known as crowd computing. The effectiveness of this method depends heavily on users’ social interactions and their willingness to share resources. So, similar to the crowdsensing systems described previously, different crowd computing applications need to adopt mechanisms that motivate peers to collaborate and defray participants who share their resources. The article “OpenNRP: A Reputation Middleware for Opportunistic Crowd Computing” proposes a novel, lightweight, and scalable system middleware that provides a unified interface to crowd computing and opportunistic networking applications. Dimitris Chatzopoulos, Mahdieh Ahmadi, Sokol Kosta, and Pan Hui illustrate the benefits of using such middleware by simulating the behavior of message forwarding and task offloading.

The article from Franca Delmastro, Valerio Arnaboldi, and Marco Conti, “People-Centric Computing and Communications in Smart Cities,” presents a mobile app that exploits the participatory sensing, opportunistic sensing, and mobile social network paradigms for people-centric computing. The article illustrates, through real experiments in a smart city scenario, the advantages and drawbacks of the technological solutions adopted by the authors. These results define the technical guidelines for the development of heterogeneous people-centric mobile applications, as both independent solutions and new integrated services developed on top of a middleware platform that integrates the main features of people-centric computing.

The articles in this issue of the Series provide ideas on how the interleaving of technological and social aspects can produce results that are closer to real life, and help related technologies from other fields support the mission of improving the quality of life of citizens. These articles also raise interesting open questions, such as the cultural and policy effects of the technologies and applications, the potential for standardization, and security issues. We hope you enjoy reading these articles as much as we did!

We also thank all the reviewers and the editorial team for their work and their invaluable support.

References


Biographies

Edoardo Biagioni (esb@hawaii.edu) is an associate professor in the Department of Information and Computer Sciences at the University of Hawaii at Manoa. His research interests focus on networking, with an emphasis on ubiquitous wireless networking, but they have over time ranged widely from security to high-performance computing, programming languages, and human-computer interfaces. He received his Ph.D. degree from the University of North Carolina at Chapel Hill, and has been a series co-editor for IEEE Communications Magazine since 2006.

Silvia Giordano (silvia.giordano@supsi.ch) is a professor at SUPSI, a Distinguished Thousand Talents Plan Professor at Tianjin University of Technology, China, and associate researcher at CNR. She directs the Netlab, and is responsible for the Social Area. Her work in the field of ad-hoc networks has produced several seminal papers in key research directions, with hundreds of citations. Her H-index is 33. Currently she is working in the area of human mobility with very innovative results in the field of localization, and with seminal work that calls for a new understanding of the concept of social distance. She is an Editor of several journals, and on the Steering and Organizing Committees of major conferences. She is an ACM Distinguished Scientist and ACM Distinguished Speaker.

Xia Luo [SM’03, M’07, SM’09] received B.S. and M.S. degrees from the University of Electronic Science and Technology, China, in 1999 and 2003 respectively, and a Ph.D. from the University of Illinois at Chicago in 2008. He is currently a Thousand Talents Plan Professor at Tianjin University of Technology. From 2007 to 2008 he worked at Motorola Labs in Illinois, and from 2008 to 2015 he was with Qualcomm Research in California. His research interests include pervasive computing, wireless communications, and visual data analysis. He served as the Tutorial Co-Chair of IEEE GLOBECOM 2012 and a member of the IEEE Communication Society Tutorials Committee in 2015. He is the inventor of 17 international patents.

Tracy Camp [F] received a B.A. degree from Kalamazoo College, an M.S. degree from Michigan State University, and a Ph.D. degree from the College of William and Mary. She is currently a full professor of computer science in the Department of Electrical Engineering and Computer Science at the Colorado School of Mines with research interests in the credibility of ad hoc network simulation studies and the use of wireless sensor networks in geosystems. She is an ACM Fellow and an ACM Distinguished Lecturer. She shares her life with Max (born in 2000), Emma (born in 2003), her husband (Glen), and two cats.

Tao Tuo received his B.S. degree from Tsinghua University, Beijing, China, in 1999, and his M.S. and Ph.D. degrees from the University of California at Los Angeles in 2000 and 2003, respectively, all in electrical engineering. From 2003 to 2004, he was with MediaWorks Integrated Systems Inc., Irvine, California. From 2004 to 2016, he was with Qualcomm Inc., San Diego, California. Since March 2016, he has been with Spreadtrum Communications USA Inc., San Diego. His research interests include wireless communications and signal processing.
Spatial Crowdsourcing: Current State and Future Directions

Yongjian Zhao and Qi Han

ABSTRACT

Crowdsourcing relies on the contributions of a large number of workers to accomplish spatial tasks, and it has drawn more attention in recent years. Many crowdsourcing tasks are completed online due to its convenience and efficiency. However, sometimes this traditional method may not work due to special requirements involving actual physical locations. Thus, a new paradigm of data collection, called spatial crowdsourcing, has emerged in the past few years. Spatial crowdsourcing consists of location-specific tasks that require people to physically be at specific locations to complete them. In this article we discuss unique challenges of spatial crowdsourcing, provide a comprehensive view of this new paradigm by introducing the taxonomy, and give future directions.

INTRODUCTION

The term “crowdsourcing” was first introduced by Jeff Howe in a Wired magazine article titled “The Rise of Crowd-Sourcing” in June 2006 as follows: “Crowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call.” Crowdsourcing tasks can be completed either by individuals or through cooperation. The large group of people and the open call are two key components of crowdsourcing according to this definition. Various terms have been used for crowdsourcing, including social computing, collective intelligence, human computation [1], crowdsensing, crowd computing, peer-production user-powered systems, crowd wisdom, smart mobs, and mass collaboration.

Traditionally, crowdsourcing has been utilized in data mining to ensure getting enough data, excluding bad data, or receiving more precise answers. It has also been used in software engineering to support various activities. Typically, crowdsourcing happens online since it can easily engage plenty of people. For example, Amazon Mechanical Turk (AMT) is a web-based market that engages people to complete tasks published by individuals or corporations. Online crowdsourcing can be utilized for various purposes such as search of missing persons and social science experiments (e.g., demography). Usually, the workers who receive or request the tasks answer the questions based on their own knowledge and do not need to be present at a specific physical location. The increasing popularity of mobile devices has enabled mobile crowdsourcing, that is, mobile users can complete various tasks that cannot be accomplished in one single location. In particular, since mobile devices often have a variety of built-in sensors, crowdsensing — a subclass of crowdsourcing — is now possible where a number of users complete tasks by sensing and reporting certain information using sensors built into their mobile devices.

In this article, we specifically focus on one type of mobile crowdsourcing: spatial crowdsourcing (SC). SC requires workers to physically be at specific locations to complete the tasks. In this type of crowd wisdom, people gather, analyze, and disseminate geographical and/or social information in the real world. SC is sometimes referred to as location-aware or geo-crowdsourcing. There has been a lot of research on crowdsourcing; however, much less attention has been paid to SC, from which many applications could greatly benefit. For instance, the government wants to collect air quality information from different locations in the city at different times (i.e., hour, day, month, year), and each location requires several workers to contribute the data. Since the workers are geographically distributed in the city, a traditional crowdsourcing system, where workers just upload the data in their vicinity at their chosen time, may not provide a complete picture of the air quality status in the city. This is because not enough data may be reported from some locations or times since not enough workers are in those regions at those times. In contrast, an SC system that proactively dispatches workers will resolve this issue. In recent years, the collection of geographical data or geospatial data has drawn attention from several governments. For instance, the U.S. and U.K. governments have granted projects on collecting, producing, and licensing geospatial data to be used as part of national strategies. Grassroots Mapping is a project that utilizes a balloon mapping platform to collect aerial imagery. The gathered imagery is then geo-registered and shared to the public. In addition, Open Street Map (OSM), Google MapMaker, and Wikimapia are all outstanding...
geographical projects that use SC. Therefore, it is important to develop sound SC techniques to support these applications. This article provides a comprehensive review of SC. We compare SC with traditional crowdsourcing and identify the uniqueness of SC. Subsequently, we present a taxonomy of SC to help better understand the key issues in SC and the current state of research in this area. Finally, we discuss several future directions.

**UNIQUE CHALLENGES OF SC**

Typically, a crowdsourcing system is concerned about the following issues.

**Task formulation.** It is important to properly formulate tasks. For instance, if a task requires people to answer a questionnaire, it is crucial to design good questions in order to collect detailed responses; if a task is too big, it is better to divide it into multiple subtasks.

**Task assignment or worker selection.** This issue targets which tasks should be assigned to how many and which volunteers. From a task publisher's perspective, a task should be assigned to as many volunteers as possible with minimum total budget and maximum reliability. Some tasks have stringent time constraints; then the goal would be to maximize the number of tasks completed within the deadline.

**Incentive mechanism.** In SC, workers are often not unconditionally willing to accept tasks, but are usually reward-driven; thus, it is of great importance to design an incentive mechanism to draw people's attention. Designing an incentive mechanism involves taking into consideration several aspects including pay, altruism, enjoyment, reputation, and implicit work [1]. Pay involves rewards like money, the most common incentive method. Reward-based incentives can be formulated as a game and can be solved using game theory algorithms. For certain areas such as politics and sports, reputation is often used as an incentive.

**Scalability.** It is crucial to design a system that can perform well regardless of the number of users and the number of tasks. Crowdsourcing engages a large number of workers. For instance, Amazon Mechanical Turk involves millions of people to perform tasks. Hence, it is important to design a platform that can support numerous task requests and worker responses.

**Quality of user contributed data.** Data quality can be defined as attribute accuracy, semantic accuracy, and temporal quality, and in SC, it requires the system to ensure the quality of user contributed data. These are broken down into user quality, feature quality, and interdependence [2]. Some user-contributed data may be of poor quality or even malicious; thus, it is important to exclude the bad data and ensure the quality of data. Another intriguing issue here is that the measured quality of collected data needs to compare against the ground truth, which a system does not have. How the system estimates the truth from streaming data (some of which may be misleading or unreliable) is an interesting topic.

**Privacy protection.** Crowdsourcing requires workers to report their data. However, some data maybe sensitive and should not be revealed to others; thus, it is important to protect workers’ privacy. All of these issues are not independent of each other; instead, they often interact. For instance, worker selection often needs to consider workers’ reliability, the quality of their contributed data, and incentives; an incentive mechanism cannot ignore workers’ reliability and data quality either.

In addition to these issues, SC also brings in new challenges due to the fact that it requires people to be physically present at a certain place to complete the task. SC can easily reveal workers’ locations, which raises serious privacy concerns. PriGeoCrowd [3] was developed with an interactive visualization and tuning toolbox for privacy-preserving spatial crowdsourcing. System designers can investigate the impact of parameters such as allocation strategy, task assignment heuristics, privacy budget, and data set density of private task matching through PriGeoCrowd. Only when privacy is protected can people be encouraged to participate in SC applications. In addition, several other issues should be tackled.

**Location awareness.** Location awareness is the most unique characteristic of SC. The distinction of location awareness in SC is that users’ movement is involved. This differs from volunteered geographic information (VGI) [4], the goal of which is to collect geographic information provided by individuals and users voluntarily participating by randomly contributing data. In VGI, although user contributed data is location-based, users are not requested to go to particular geographical places to perform the tasks; they may just respond with the information they already have. Similar to VGI, crowdsourcing location-based queries over Twitter [5] involves geographic information. It employs a location-based service to find the appropriate people to answer a given query. Based on the history of users’ Foursquare check-ins, the system selects the most relevant users. However, the users who are selected to answer the question do not have to intentionally move to the location, and they only answer the question based on their own existing knowledge.

Different from the above two projects, gMission [6] is a general platform for crowdsourcing various location-dependent tasks where users can post, answer, and verify different questions. Due to location-based verification, the system checks whether the worker has really reached the specific place. It accepts the user’s answer only if the user is near that location when posting the answer; thus, the workers have to travel to the event place.

**Workers’ path selection.** Since the users should travel to the event place and perform the tasks, it is critical for the users to select the best route and wisely schedule the task sequence. The classic traveling salesman problem (TSP) and the vehicle routing problem (VRP) are commonly used to formulate the problem of finding a path of minimum cost. In VRP, all workers start from the same location, and the number of workers is a fixed number. However, in SC, finding the right paths is much more challenging because:

- The workers start from different locations.
- The number of volunteers may differ from task to task and also vary over time.
Workers seeking Reward

A taxonomy of spatial crowdsourcing. Workers model Voluntary workers Point task Region task Worker numbers Task assignment Multiple worker assigning task Worker selecting task Task required tasks Worker required tasks Task model Categorical data Continuous data Response model Multimedia data System perspective Maximize net reward Minimize system cost Maximize task coverage Maximize data quality Minimize number of tasks with missed deadlines

Figure 1. Taxonomy of spatial crowdsourcing.

Data sets. Only a few real-world datasets can be utilized under certain assumptions. For instance, Gowalla and Brightkite are two location-based social networks in which users are able to check into different locations in their vicinity. The check-ins include the location and the time when the users enter the spots, and users share their locations through check-ins. If we consider the users as workers, and the check-in spots as tasks, both data sets can be used for SC. The Yelp data set was captured in the greater Phoenix, Arizona, area, including the locations of 11,537 businesses (hotels, etc.), 43,873 users, and 229,907 reviews. For this data set, we may consider Yelp users as workers, businesses as spatial tasks, sets of businesses with the same category as spatial complex tasks, and reviewing a business as accepting a spatial task. For a worker $i$, we know his location at different times: $(loc_i, t_i)$ from the data set. We generate his trajectory according to these locations using a certain mobility model. This can be viewed as the worker’s routine trajectory. When a spatial task is assigned to the worker, we let the worker deviate from his trajectory to perform the task if the deviation is acceptable by the worker. If the worker can perform the task before the deadline, we consider the task completed by the worker. Other data sets that contain places of interest and trajectories can also be used for SC. A place of interest is the location of a task, and the user trajectory can be viewed as the task scheduling route.

There are no real-world datasets that can be directly utilized for SC. Due to this lack of data sets, SC algorithms are usually evaluated using synthetic data sets: distribution function is applied to generate user and task location, and then a user mobility model is used to simulate the user behavior. Designing a general-purpose platform for researchers to collect data for SC can have great potential to accelerate research in this area.

Taxonomy of Spatial Crowdsourcing

To assist in identifying the needs of future application needs, we have developed a taxonomy of potential application classes (Fig. 1). The first division relates to how SC workers are modeled, the second division relates to the requirements of the crowdsourcing tasks, the third is about various types of responses provided by workers, and the fourth division is on what an application would like to optimize and what constraints it faces. In Fig. 1, shaded boxes show the issues that need to be modified to suit SC, while other boxes show the issues common in SC and general crowdsourcing.

Worker Model

An SC system comprises a group of human workers. Each worker $w_i$ is associated with a set of attributes, denoted as $(d_i, lat_i, long_i, d_i, exp_i, rel_i, punct_i, traj_i)$, where $id_i$ represents the worker’s unique identifier, $lat_i$ and $long_i$ are used to represent the user’s geographical coordinates, and $d_i$ indicates the user’s proficiency level in expertise $exp_i$. For instance, professional photographers can provide higher quality pictures than an amateur. User reliability is denoted as $rel_i$, which reflects the probability that the worker can provide an accurate answer, and $punct_i$ stands for user punctuality, which measures how likely the user can arrive at the task location within the deadline. $traj_i$ represents the past trajectories of the worker, which may be used to predict workers’ future locations.

According to the motivations of the workers, we can classify workers into two classes: reward-seeking and voluntary.

Reward-seeking workers. Reward-seeking workers complete tasks in order to receive rewards. In most cases, the reward is a certain amount of money or commodities. Workers receive a certain amount of reward if they complete a task correctly and punctually [7].

Voluntary workers. Voluntary workers are willing to complete location-dependent tasks without any compensation. These people are self-motivated to participate for various reasons such as promoting cultural and political reputation or simply enjoying appreciation. For example, in a participatory sensing campaign, a group of people are willing to report traffic information like accidents or documenting geographical information using sensor-equipped devices. Most existing SC systems rely on voluntary participants.

Task Model

Each task $j$ is associated with a set of attributes: $(id_j, deadline_j, loc_j, que_j, num_j, incentive_j, budget_j)$. $id_j$ represents the task’s unique identifier, $deadline_j$ stands for its deadline, which means...
the task has real-time constraints, and quej is the question the worker needs to answer. It can be a yes/no question to verify certain information; it can also ask the worker to take a picture of a certain place, and so on. locj stands for the physical location of the task. If the location is a two-dimensional coordinate, we consider the task as a point task; otherwise, the location may be represented by an area such as a campus or a city, in which case the task is viewed as a region task. For instance, VGI [4] is crowdsourcing a region task where workers collect geo-spatial information voluntarily in a certain area to contribute to the data quality. numj refers to the number of workers required to complete the task. In some systems, workers are assumed trustworthy and accurate; thus, a task is only assigned to a single worker. However, in the real world, users may not always be reliable, so it is better to assign a task to multiple workers to ensure collective quality of received responses. incentivej denotes the reward to worker i if she completes task j and budgetj stands for the total budget for the task.

Based on whether single or multiple workers are needed for a task, we can classify the tasks into the following two classes.

**Single worker required tasks.** These kinds of tasks assume workers are trustable and reliable, so the workers can finish the spatial tasks accurately without any malicious intentions. It is sufficient to assign the task to only one worker, for instance, the nearest, the most reliable, or the least expensive worker [8].

**Multiple workers required tasks.** The success of crowdsourcing relies on the wisdom of the crowd, so it is more reasonable to assign a task to multiple workers. Amazon’s Mechanical Turk is an example that uses multiple workers for one task. In practice, not all workers are trustworthy or punctual all the time. However, it is reasonable to assume the majority of workers can be trusted. Thus, majority voting is often one common way to extract ground truth from all the received responses. Although it is more expensive to involve multiple workers for the same task, bad data can more easily be excluded.

Once the task model and worker model are set up, we can consider task assignment or worker selection. Generally, there are two main modes: one is to have the server assign tasks, and the other is the worker select tasks.

**Server assigning tasks.** After collecting all the locations of the workers, the server assigns each task to workers according to the system optimization goals such as minimizing the total travel distance. This often leads to global optimization for the system. However, the workers’ locations may be revealed. Most of the existing SC systems assign tasks to workers [8].

**Worker selecting tasks.** In this mode, the server publishes various spatial tasks online, and workers may select any tasks based on their own preferences without notifying the server in advance [9]. The advantage is that workers do not need to reveal their current locations and have more freedom to choose tasks they like. However, one drawback of this mode is that the server does not have any control over the allocation of spatial tasks. This may result in a load imbalance problem; in other words, some tasks have no participants, while others may have too many volunteers. Furthermore, since workers choose tasks based on their own objectives such as choosing the closest, easiest, or higher-paid tasks, it does not always result in globally optimized assignment from the system’s perspective.

In both modes, workers’ traveling routes need to be part of the consideration. This problem becomes even more prominent when multiple tasks are either assigned or being selected by a worker. This is because these tasks may not be in the same location, so we need to avoid unnecessary traveling between locations by carefully selecting paths and also scheduling tasks. For example, a coordinated task assignment approach [10] can be used, where the system assigns a sequence of tasks to each worker, taking into account the expected trajectory predicted from the historical movement of individuals.

**Response Model**

Different SC applications may pose different tasks. In each task, workers may contribute different types of data such as categorical data, continuous data, or multimedia data. Some may ask workers to verify certain information at a particular location, so the response is simply yes/no; others may ask questions that require more elaborate answers; still others may ask workers to respond with multimedia data such as pictures or videos. Pictures can be taken at different times, from different angles, from different distances, and so on. An interesting way to consider the responses is the concept of “diversity” introduced in reliable diversity-based spatial crowdsourcing (RDB-SC) [11]. It assigns time-constrained spatial tasks to dynamically moving workers so that tasks can be accomplished with high reliability and spatial/temporal diversity where the diversity value is formulated as entropy of provided responses. Since the problem is NP-hard, three approximation solutions are provided (greedy, sampling, and divide-and-conquer).

**Optimization Goals and Constraints**

Different SC systems have different foci that can vary from the worker’s or system’s perspective. From a worker’s perspective, his goal is often to maximize his total net reward, the difference between the reward he gets from the system and the cost (e.g., travel cost). To achieve this goal, a worker may seek to take as many tasks as possible on his traveling path; then the workers may compete with each other. This can be formulated using different game theory models to achieve Pareto optimality. To reduce the cost, the worker may choose the best route to accomplish all the tasks, so task scheduling and path selection need to be jointly considered in selecting tasks.

From a system’s perspective, the goals are often to maximize the task coverage with least cost and obtain maximum quality.

**Maximize task coverage.** This is to maximize the number of assigned tasks. To achieve this goal, the server first collects all the locations of the workers and then devises a strategy to maximize the overall number of assigned tasks. Some systems do not consider different expertise levels in workers, so they treat the workers equally. The task assignment problem is then formulated as a
matching problem [12], where workers and tasks form a bipartite graph; the edge between a worker and a task has a weight of 1. Other systems take into consideration workers’ skill levels by defining expertise match as an assignment of a task to an expert [13]. Higher scores are given to expertise matches than non-expertise matches. The problem can be formulated as a weighted b-matching problem.

Due to the nature of inherent traveling involved in SC, the task assignment problem often needs to consider task scheduling. Once a set of tasks are assigned to a worker, the worker has to decide the optimal way to complete the tasks. For instance, in order to maximize the number of tasks assigned to each worker, two exact algorithms are developed using dynamic programming and branch-and-bound strategies to solve the task scheduling problem [9]. Also, a bisection-based LALS framework [14] is developed that performs top-down recursive bisection and a bottom-up merge procedure iteratively so that assignment and scheduling can be performed locally in a much smaller promising space.

**Minimize system cost.** The cost can be defined as the total incentives paid to selected workers or total traveling distances of all selected workers. In addition to trying to minimize the total cost, it is also possible to use budget as a constraint to maximize task coverage.

**Maximize data quality.** Depending on how data quality is defined, different strategies can be applied to maximize data quality. Some aim to maximize the total number of people providing correct responses by applying majority voting to get the ground truth [2], others aim to maximize the aggregated probability of getting correct responses [15], and still others target maximizing expected spatio-temporal diversity [11].

**Minimize the number of tasks with missed deadlines.** Spatial tasks may have time constraints, so workers need to complete the tasks before the deadlines. In this case, the system may want to minimize the number of tasks that are not completed within the deadlines. Again, task scheduling and workers path selection became a critical part of task assignment.

Figure 2 shows the flow of SC. Initially, each worker downloads the SC app and is able to see all the tasks published by task publishers. If the server will be assigning tasks, the server gathers all the workers’ information and recruits the appropriate workers; if a worker will be selecting tasks, the worker selects the proper task set to maximize her reward. Both worker selection and task selection utilize the information from the worker model and the task model. An incentive mechanism is designed to motivate workers to perform the tasks. Once the task is completed, the task publisher collects the data from workers. The system measures the contribution from users in terms of quality of contributed and user reliability, and so on. The results are further used to adjust the incentives. For example, once the task is published, a worker’s reliability and punctuality may change, and the incentive for this worker has to be updated. This process iterates as long as there are available tasks.

**Future Research Directions**

Existing work on SC has studied various aspects of SC. We next highlight several directions for future research.

**Consider incentives and budgets.** While incentive mechanism design is a well studied topic in typical crowdsourcing, most existing SC systems have relied on voluntary workers or simple incentive mechanisms where the same reward applies to every participant. However, in real-world scenarios, people tend to seek reward when participating in SC. And, the quality of contributed data varies from person to person; thus, modeling SC with certain reward and budget is closer to reality. For task publishers, how to distribute limited budget to obtain the most workers should be studied. For volunteers, how to gain as much reward as they can is important.

**Integrate location privacy into SC.** Most existing approaches use the server assigning tasks model where the server collects the location information of all potential volunteers, which may reveal workers’ locations. Although privacy protection in SC has been explored, this issue has not yet been fully studied. For example, the task model, worker model, and response model vary under different circumstances; thus, it is challenging to provide the most effective SC while protecting workers location privacy.

**Combine server assignment and worker selection models.** Most existing task assignment models focus on server assigning tasks, which assumes that workers should perform the tasks once they receive the assignments. However, in reality, it is possible that workers may refuse to complete the tasks due to various reasons (e.g., length of travel, laboriousness). Hence, it is more efficient and effective to combine the server assignment model with a worker selection model so that workers may select the tasks they are good at and then compete with others to obtain the reward.

**Jointly optimize the system.** The optimization goals for existing SC research are mainly on a single objective. However, real applications often need to jointly optimize several factors. For example, maximizing the task assignments and also the data quality at the same time may be required by an SC app.

**Integrate SC with social networking.** Existing SC work only considers the geographical information or expertise characteristics of the
worker. However, in a real-world scenarios, workers may interact with each other and belong to social groups. For instance, if the spatial task is to take pictures of a place, it is more reason-
able to assign the task to the social group tagged “Photographers.” Hence, community detection in combination with SC task assignment is an inter-
esting idea to explore.

The issues above address the technical problems that may be targeted. Besides, in reality, there is no available data set that can be utilized directly in SC tasks. Thus, from an engineering perspective, it is a necessity to collect real-world data to help validate different research ideas and systems instead of running simulations over synthetic data.

CONCLUSIONS

In this article, we survey a new branch of crowdsourcing — spatial crowdsourcing — which requires workers to be at specific physical locations in order to complete tasks. We discuss the unique characteristics of SC. We then classify the state-of-the-art research into different categories. In the end, we suggest several promising problems that have not been studied in the existing literature.

REFERENCES

ing, 2015, pp. 1404–07.
tions (PerCom), 2011, pp. 513–18.

BIOGRAPHIES

YONGJIAN ZHAO (qzhao2@mines.edu) is currently a Ph.D. student in the Department of Electrical Engineering and Computer Science, Colorado School of Mines. He received his B.S and M.E degrees in computer science from Sun Yat-Sen University, Guangzhou, China, in 2012 and 2015, respectively. His research interests include mobile and pervasive computing, mobile crowdsourcing, wearable computing, and data mining.

Qi Han [M] (qhan@mines.edu) received her Ph.D. degree in computer science from the University of California, Irvine in 2005. Currently, she is an associate professor in the Department of Electrical Engineering and Computer Science, Colorado School of Mines. Her research interests include mobile and pervasive computing, mobile sensing, network-aware data management, and wireless sensor systems. She is a member of the ACM.
Empowering Mobile Crowdsensing through Social and Ad Hoc Networking

Stefano Chessa, Antonio Corradi, Luca Foschini, and Michele Girolami

ABSTRACT

Mobile crowdsensing (MCS) enables collective data harvesting actions by coordinating citizens willing to contribute data collected via their sensor-rich smartphones that represent sources of valuable sensing information in urban environments nowadays. One of the biggest challenges in a real long-running MCS system lies in the capacity not only to attract new volunteers, but also, and most importantly, to leverage existing social ties between volunteers to keep them involved to build long-lasting MCS communities. In addition, the advent of high-performing devices and ad hoc communication technologies can help to further amplify the effect of sensing actions in proximity of the volunteer devices. This article originally describes how to exploit these socio-technical networking aspects to increase the performance of MCS campaigns in the ParticipAct living laboratory, an ongoing MCS real-world experiment that involved about 170 students of the University of Bologna for more than two years. The article also reports some significant experimental results to quantify the effectiveness of the proposed techniques.

INTRODUCTION

Mobile crowdsensing (MCS) is a recent sensing paradigm that leverages the worldwide availability of smartphones. By installing a crowdsensing application, any smartphone can become part of a (large-scale) mobile sensor network, partially operated by the owners of the phones themselves. A crowdsensing application transforms a smartphone into a data sensing, collection, and sharing terminal to exploit the embedded sensors (cameras, microphones, accelerometers, barometers, etc.) and the mobility of the user carrying the phone to gather information about some events of common interest to users. Considering the widespread diffusion of smartphones and their density (typically in urban areas), the information from which a crowdsensing application draws can be rather dense and fine-grained [1]. This fact, along with the limited investments required to develop and maintain crowdsensing platforms and applications, made this paradigm particularly attractive to smart cities managers to improve the quality of their cities and citizens. Despite its potential, MCS still faces some barriers to user acceptance [2]. Some of them are technological and will likely fade away with the progress of smartphones and crowdsensing technologies and standards: an example is the heterogeneity of hardware/software platforms that raises the costs for development and maintenance. Other barriers are more related to user acceptance, such as the concern of users for privacy, battery consumption, and communication costs. However, other obstacles will not disappear so easily: persistent examples are the difficulty of involving users and the high rate of dropoff due to users who lose interest in contributing to MCS campaigns. With increasing expansion of MCS, these last two obstacles are quite critical since an insufficient number of involved people can compromise the effectiveness of an MCS campaign. In fact, realistic scenarios may involve only a small fraction of the whole citizenship in MCS campaigns, while most citizens do not take part in it.

To overcome these limitations, we propose novel technical solutions to empower and enrich the whole MCS cycle by exploiting both user sociality and physical proximity to emphasize the effect of MCS campaigns. First, we gain higher user participation by leveraging communities of users, that is groups of users with high mutual attendance [3]. We introduce a new type of cooperative task, the community-based task, with multiple stages, and we ask people in the same community to participate to its completion. Second, we explore the possibility of increasing the number of data gathered during an MCS campaign by exploiting other users who are not part of the MCS loop. This capability, which we define as the sensing amplification factor [4], enables the smartphone of an MCS user to cooperate with other devices detected in proximity (owned by users not already involved), as happens already in mobile social networks (MSNs) [5, 6]. Finally, we report a selection of interesting experimental results that assess the proposed facilities in the ParticipAct MCS living laboratory, by discussing and evaluating their strengths and weaknesses.

MOBILITY AND SOCIALITY IN MCS

MCS services typically cross-cut and work at the overlapping of various different research areas that span from management of large socio-technical smart city systems for optimizing the MCS...
Increasing the number of data gathered by crowdsourcing techniques that enlarge the number of people involved and aim to gain higher participation. Without claiming completeness, this section first proposes a general model for MCS, then briefly overviews the current state of the art in these very active research fields.

While there are already some good surveys about MCS systems (e.g., [7]), this article proposes a simplified MCS reference architecture by identifying its basic components: we claim that any MCS architecture must consist of the following building blocks. First of all, the MCS crowd is a set of $M$ people, representing a fraction of the whole population, and we refer to them as volunteers (drawn as light orange full points in Fig. 1), that deliberately express their availability to be involved in crowdsensing campaigns over the smart city.

Volunteers move within the smart city and establish meaningful social interactions (in the real world) that typically present two properties: their mobility and interaction patterns. In addition, volunteers are expected to carry their smartphones provisioned with the MCS app (i.e., MCS client) in charge of receiving MCS requests, also referred to as tasks. A task represents a data collection activity to be accepted and completed by volunteers involved in an MCS campaign. Each task may include multiple data gathering actions to complete, typically in a given time span and within an area of interest; these actions can occur either automatically by the MCS app without any other human interaction (apart the initial acceptance) or they may require some active participation of volunteers that should, for instance, provide their feedback about a focused night event. Finally, the MCS server is the central backend in charge of storing gathered data creating new MCS campaigns and sending tasks to MCS clients, and it is composed of a dashboard for configuring tasks and a database for the storage of sensed data.

In addition to the above basic MCS issues, we argue that the spatial and social dimensions of tasks can greatly enhance the effectiveness of MCS campaigns, and that potentially powerful aspect is often underutilized. Along that line, we claim the need for two extensions to the MCS reference architecture, in two directions:

- **Enhancing the involvement of volunteers and supporting more complex tasks via the introduction of the concept of communities of volunteers**
- **Increasing the number of data gathered by MCS campaigns by exploiting devices of other users that lie in proximity of volunteers and can offer their data in a crowdsensing task**

On one hand, the extended architecture includes the discovery of communities of volunteers with a meaningful and durable social interaction among themselves (in Fig. 1 these social ties among volunteers are depicted as solid lines) and the introduction of the new concept of a community-based task. Those tasks require that volunteers interact closely with one another toward the entire completion by contributing some different subtasks in their turn; the smaller tasks are called community-based task stages or simply stages, and volunteers can offer their work for specific stages, depending on their capacities and availability.

On the other hand, we add discovery of devices close to volunteer devices (referred to as discovery of colocated device) where the connection between devices in proximity typically rely on ad hoc communication protocols (devices in proximity are shown in Fig. 1 as pink blank spots, with the wireless transmission range corolla that overlaps with that of a volunteer device). Focusing now on the literature, the availability of rich smartphone sensing platforms have enabled the creation of several MCS vertical services and apps for different domains including environment monitoring, intelligent transportation, urban dynamics sensing, healthcare, and so forth [7]. Some MCS systems, such as Vita and Medusa [8, 9], take a step further by introducing support functions to ease definition and automatic assignment of tasks to volunteers by also including complex (monetary and non-monetary) incentive mechanisms, which are out of the scope of this article and complementary, to increase participation [10].

More recently, various authors have agreed that one of the key open challenges in MCS is the possibility of exploiting socio-technical network effects to consolidate and extend the crowd by leveraging communities of volunteers [7, 11]. Accordingly, new research trends at the crossroads of MCS, MSNs, and ad hoc sensor networks share with our proposal the goal of facilitating community identification and formation to enable informed scheduling of MCS tasks not only to single volunteers, but also to communities of volunteers [12]. Moreover, some proposals explore the use of opportunistic interactions with nearby devices to further boost the performance of MCS data harvesting [4, 13].

What makes our research efforts unique is the fact that the ParticipAct living laboratory, with long duration (more than two years) and geographical width (the whole Emilia Romagna region in Italy), closely mimics a very real...
AMplIfyIng Mcs tHrougH coMMunIty-bAsed tAsks

This section details the solutions used to enhance the MCS process by leveraging existing communities of volunteers and enabling opportunistic communication with devices in proximity.

AMplifying MCS through Community-Based Tasks

The social interactions among humans are often referred to as social ties to imply the involved relationship among people. Moreover, as a matter of fact, people tend to move according to objectives and activities stemming from their social interactions and derived from specific opportunities and events [15]. Focusing on individual social relationships, they can have different strengths: they can be tight ones, as among friends, or weaker ones, as occur among strangers. The strength of social ties also depends on the time spent together by involved people: typically decide the best relationship category to involve, also depending on the importance of distinguishing different relationship categories. Once statistics on past community-based task completion rates are collected, it is possible, given a target community-based task completion ratio, to probabilistically decide the best relationship category to be involved. In other words, although “friends” communities will typically perform better (as also demonstrated by experimental results collected in the field in our living laboratory), a fair task management strategy suggests avoiding always involving the same category of community. Hence, a manager can more freely decide the community to be involved, also depending on the importance of the task itself. Along the same line, to decrease the MCS burden on single users, a volunteer who belongs to multiple communities is asked only for one community. We schedule a volunteer who belongs to multiple communities along that line, we determine different mean values of DURATION and CONTACTS (MDVALUE and CMVALUE, respectively) to identify and distinguish these four social categories. Indeed, the configuration of MDVALUE and CMVALUE is an important decision, because it affects the cardinality of node communities and has to be tuned for the specific MCS datasets and mobility traces.

Finally, we use communities detected via k-CLIQUE to schedule tasks to communities; as already stated, we introduced community-based tasks because they favor wider participation and social involvement of volunteers, but let us also stress the importance of distinguishing different relationship categories. Among all the available ones, we decided to use k-CLIQUE [3] to evaluate the social relationship between individuals. We choose the k-CLIQUE for some main reasons:

- As recognized in the literature, it is a general-purpose algorithm very easy to apply to a wide range of application scenarios.
- It does not require any previous profiling knowledge of volunteer mobility metrics.
- It has a relatively low computation complexity, \(O(n^2)\) in the size of the explored network, and hence it is suitable for a periodical re-computation.

Going deep in the technical aspects, k-CLIQUE is a distributed spatio-temporal detection algorithm coming in many versions; the basic version of k-CLIQUE requires two parameters to decide whether to add a device to the community (called Familiar Set): the parameters are the cumulative contact duration (DURATION in Fig. 2) and the number of contacts (#CONTACTS in Fig. 2) [3]. In this article, we use k-CLIQUE to identify the four categories of relationship (friends, community members, familiar strangers, and strangers) typically used to quantify the strength of human relationship that can be measured by the time spent together and the frequency of encounters. As recognized in the literature, it is a general-purpose algorithm very easy to apply to a wide range of application scenarios.

defining the communities of volunteers and enabling opportunistic communication with devices in proximity.

AMplifying and MAintaining Users in MCS Platforms

This section details the solutions used to enhance the MCS process by leveraging existing communities of volunteers and enabling opportunistic communication with devices in proximity.

AMplifying MCS through Community-Based Tasks

The social interactions among humans are often referred to as social ties to imply the involved relationship among people. Moreover, as a matter of fact, people tend to move according to objectives and activities stemming from their social interactions and derived from specific opportunities and events [15]. Focusing on individual social relationships, they can have different strengths: they can be tight ones, as among friends, or weaker ones, as occur among strangers. The strength of social ties also depends on the time spent together by involved people: typically, friends are people who usually share lots of interests, and tend to meet often and spend much time together; differently, strangers tend to have fewer overlapping interests and stay in touch only for short and sporadic periods. According to the human tendency to spend time with people with similar interests, the MSN literature has classified the MSN topology to characterize how people move together in a smart city. In particular, it has defined four main categories of relationships to distinguish communities of people presenting different levels of sociality among themselves; from the most to the least social ones, these are friends, community members, familiar strangers, and strangers. Community detection algorithms aim to identify and extract from (volunteers’ mobility traces the communities (namely, pinpointing all communities per category and the people in any one of them) belonging to each of those relationship categories.

In particular, the literature proposes many community detection algorithms motivated by different possible expected outcomes [5, 6]. Among all the available ones, we decided to use k-CLIQUE [3] to evaluate the social relationship between individuals. We choose the k-CLIQUE for some main reasons:

- As recognized in the literature, it is a general-purpose algorithm very easy to apply to a wide range of application scenarios.
- It does not require any previous profiling knowledge of volunteer mobility metrics.
- It has a relatively low computation complexity, \(O(n^2)\) in the size of the explored network, and hence it is suitable for a periodical re-computation.

Going deep in the technical aspects, k-CLIQUE is a distributed spatio-temporal detection algorithm coming in many versions; the basic version of k-CLIQUE requires two parameters to decide whether to add a device to the community (called Familiar Set): the parameters are the cumulative contact duration (DURATION in Fig. 2) and the number of contacts (#CONTACTS in Fig. 2) [3]. In this article, we use k-CLIQUE to identify the four categories of relationship (friends, community members, familiar strangers, and strangers) typically used to quantify the strength of human relationship that can be measured by the time spent together and the frequency of encounters. Along that line, we determine different mean values of DURATION and #CONTACTS (MDVALUE and CMVALUE, respectively) to identify and distinguish these four social categories. Indeed, the configuration of MDVALUE and CMVALUE is an important decision, because it affects the cardinality of node communities and has to be tuned for the specific MCS datasets and mobility traces.

Finally, we use communities detected via k-CLIQUE to schedule tasks to communities; as already stated, we introduced community-based tasks because they favor wider participation and social involvement of volunteers, but let us also stress the importance of distinguishing different relationship categories. Once statistics on past community-based task completion rates are collected, it is possible, given a target community-based task completion ratio, to probabilistically decide the best relationship category to be involved. In other words, although “friends” communities will typically perform better (as also demonstrated by experimental results collected in the field in our living laboratory), a fair task management strategy suggests avoiding always involving the same category of community. Hence, a manager can more freely decide the community to be involved, also depending on the importance of the task itself. Along the same line, to decrease the MCS burden on single users, a volunteer who belongs to multiple communities is asked only for one community. We schedule the community-based task starting with the community with a stronger relationship, friends, first, then community members, and so forth, and we avoid sending the task multiple times, excluding all other communities of the same volunteer. To take a step further, it would be possible to profile single users, and eventually entire communities, according to their interests, and use these profiles to drive and further refine community-based task scheduling decisions. While that is part of
our ongoing efforts on further evolving community-based task scheduling, we already explored those advanced scheduling strategies applied to single-user tasks considering both spatial and interest dimensions; for more detail, we refer interested readers to our previous works [1, 4].

AMPLIFYING SENSING THROUGH OPPORTUNISTIC COMMUNICATIONS

Our amplification factor aims to exploit the sensing services that other (non-MCS) devices may offer to enlarge the number of sensed data. In particular, when an MCS client detects the presence of another device that can offer and publish data of interest for its present task, it can opportunistically pick up the foreign data to shorten its completion time.

This amplification approach requires cooperation among devices that is enabled by two current technologies available in the wearable device market: short-range radio interfaces and applications for content sharing. Short-range interfaces are Bluetooth or WiFi, for instance, now available on most smartphones, tablets, wristbands, and smart watches; they allow detection of nearby devices (i.e., 0 to 10 m) and interact with them without any broadband connection. Along the other line, the diffusion of applications for sharing contents has pushed a deeply novel way of device (and hence people) collaboration. Currently, all the most popular application markets offer apps for novel resource sharing techniques, for sharing either Internet connection (tethering), an instant messaging service, or sensing services.

The amplification factor $f$ quantifies the benefit of exploiting the opportunistic communications with other devices in an MCS campaign. Given a task of an MCS campaign, let $\beta$ be the amount of data that can be gathered by only considering devices of those volunteers who decide to accept the task, say $V \subseteq M$ ($M$ is the total number of volunteers). We define $f$ as the ratio between the amount of data that the task can gather by opportunistically exploiting other devices and $\beta$:

$$ f = \eta \beta. \quad (1) $$

Thus, $\beta$ is the lower bound of the number of retrievable data, while $\eta$ represents the upper bound of it. Moreover, let us observe that data collected by different devices in the same place at the same time might overlap (unless some measurement error occurs); in that case they would account only for one data chunk. Hence, assuming that the sensor readings are requested at discrete time slots, we can provide an estimation of $\eta$ by defining the amplification set $G_i$ of device $i \in V$ at time slot $t$ as the subset of the neighbor set $N_i$ of $i$ that provide $i$ with additional data useful for the task. Letting $q$ be the probability that a neighbor of $i$ provides information useful for the task, we have $G_i = q|N_i|$. Hence, $\eta$ can be written as

$$ \eta = \sum_{i \in V} |G_i|^*p^*|G_i|. \quad (2) $$

Equation 2 provides an estimation (shown in Fig. 3) of the contribution given by one device in $V$ to $\eta$, by varying the number of devices in $V$ and the probability $p$ that in a time slot $t$, a device belongs to any pair of $G_i$ and $G_j$ for some $i, j \in V$. In particular, we observe that the value of $\eta$ decreases for higher values of the cardinality of $V$. This because with a larger number of devices that receive the task, the chance that any two subsets $G_i$ and $G_j$ have a larger intersection is also higher. The same effect is also due to high values of $p$, which also result in small values of $\eta$. Therefore, in order to improve the amplification factor it is not sufficient to send the task to additional devices, but it is also important to choose the additional devices in such a way that their amplification sets do not overlap much with that of the other devices running the task in order to ensure that $p$ is kept low, for example, by selecting users belonging to different communities such as strangers or familiar strangers.

EXPERIMENTAL CAMPAIGN

The experimental assessment of a large-scale crowdsensing system in a realistic scenario poses complex problems and opens up social, technical, and logistic challenges. The reported results have been collected within the long-running and still active ParticipAct living laboratory that we have maintained at the University of Bologna since January 2014. ParticipAct is a large crowdsensing deployment that now involves 178 volunteers, all of them students of our university from different courses, years, and campuses: the Bologna campus (123 students) and the Cesena campus (55 students). Although, as in any similar experiment, it is an open question whether obtained results and implications could be extended to a more general scenario, the ParticipAct dataset is large enough (in both time, almost two years, and space, with different cities involved) to draw some first important observations, rather realistic for urban setting scenarios. As another observation, our university campuses are not self-contained, but spread over some metropolitan areas, and the same is true for most universities in Italy. For this reason, volunteer paths and behaviors are not limited to a specific area, but relate to the whole urban territory that coincides with the same area of all citizens living in the same smart city. In the following, we present a selection of experimental results aimed at quantitatively assessing the functions originally presented in this article.
Figure 4. Evaluation of a) communities; b) community-based tasks success rate.

Our first experimental result shows the effectiveness of the newly introduced community-based task. In particular, we have compared volunteer involvement before and after the introduction of this feature to evaluate the impact of the employed technique in consolidating participation of volunteers for community-based tasks of different complexity. The analysis focused on periods of one month: ParticipAct mobility traces for one month all involved volunteers and typically contain about 7 million data location points that made it possible to extract an average of 41,570 contacts per month between our volunteers. First of all, we run some preliminary configuration tests to tune the $k$, MDVALUE, and #CMVALUE parameters; the goal was to find the best values to distinguish different relationship categories and balance the number of communities identified for each category. After some empirical tuning, we found that putting $k = 4$, MDVALUE = 144 h (20 percent of the total number of hours in one month of 30 days), and #CMVALUE = 5 contacts per day, it is possible to distinguish well an average number of 35 communities for each month with the distribution shown in Fig. 4a. In particular, communities are divided as follows: 52 percent friends, 35 percent community members, and the remainder distributed among 4 percent familiar strangers and 9 percent strangers. Let us stress that volunteers in strangers communities present very low sociality. Due to their very low responsiveness, after some first tentative community-task assignments to them, we decided not to use communities of this category for scheduling community-based tasks, but to simply consider single independent users as the volunteers in this category.

Then our second experimental result allowed us to automatically identify communities in September 2015, that is, 37 communities, and to use them in the scheduling of 16 community-based tasks in October 2015; these 16 community-based tasks present increasing complexities and a number of stages ranging from 1 to 5 (points in Fig. 4b are average values over all these tasks, and all presented measurements have exhibited a limited variance, always under 8 percent). All these tasks are very easy and fast to complete, such as ordering the first three soccer teams classified in the last World Cup championship and ordering the first $N$ lyrics of a song. The first important result is that the introduction of community-based tasks produced an increase in the acceptance rate of the proposed tasks themselves: before the acceptance of a new task requests settled at 38 percent, the rate of volunteers ignoring new tasks settled at 55 percent, and those explicitly refusing at 7 percent; after the introduction of community-based tasks, instead, the acceptance rate improved to 51 percent, while the number of volunteers ignoring and refusing dropped to 45 and 4 percent, respectively. For the sake of fair comparison, we considered simple routine tasks that always required very simple action, that is, multiple choice questions to answer and usually taking less than one minute.

Afterward, we focused our analysis on community-based task success rate (expressed as percentage of success in Fig. 4b), and also on completion of all stages by involved volunteers; in particular, at each stage, we asked all participants in the group to answer part of a simple question by excluding those who had already replied at previous stages. As in Fig. 4b, there is a rather clear distinction between the performances obtainable with the different relationship categories: the higher the sociality, the better the completion ratio; that also allows, given a target completion ratio, (statistically) deciding on the category of communities to involve and those to exclude, thus lowering the load of requests sent to potential volunteers. Let us also add that an important guideline to follow in the design of community-based tasks is to avoid a too complex task with too many stages; in fact, over 4 stages the completion ratio drops below 20 percent for any category of community.

Finally, our third experimental test works on the amplification factor to evaluate the improvements in terms of additional data collected. For the sake of simplicity, respecting the general theoretical model of Fig. 3, we focused on the amplification factor for one single device receiving a task (hence $\beta = 1$), and we set probability $q = 1$ to represent the most optimistic case in which all encountered neighbors provide information suitable for the task. With such a setting, the presented results can be considered as the upper bound of the amplification factor $f$, as shown in Fig. 5, which reports the amplification factor for the whole crowdsensing campaign. On the $x$ axis, we report the hours elapsed since the start of the experimentation for a total of approximately 9 months; on the $y$ axis we report the amplification factor. The experimentation starts in December right before the Christmas holidays, during which very few students meet, resulting in very low amplification factor. From
January to June, the amplification increases and stays in the range [0 to 1.2]. The reason for such an trend is the scheduling of many university lectures and examination breaks that attract students and force them to stay in touch for longer periods. As a result, the average cardinality of the neighborhood also increases. Summer is also significant as in Fig. 5, when few students meet each other, resulting in a low amplification factor. In fact, the amplification factor decreases during the works week (typically Monday to Friday), see weekly pattern in Fig. 5) when people are more likely to meet and stay in contact for longer periods. Conversely, the amplification factor increases during late weekends giving rise to a weekly pattern of encounters.

From the previous analysis, we observe that the number of expected results from a task is deeply affected by social activities of people. The ParticipAct dataset highlights that routinely cyclic aspects, and also other events forcing people to meet and to stay in contact, can increase the average neighborhood relationship and hence the amplification factor. Furthermore, the knowledge of routine patterns like weekly patterns can be exploited in order to synchronize the submission of a task with the daily rhythm of a crowd.

Let us conclude by summarizing some lessons we have learned toward new campaigns of the whole MCS process. First, even with a small number of volunteers, such as our 178 ParticipAct ones, a simple community detection solution is possible for community-based task scheduling. Second, the introduction of community-based tasks might be used as an indirect incentive [7] to increase volunteers’ sense of belonging, so it induces an increase in acceptance rate. Third, MCS tasks should be as simple as possible, avoiding too many stages. Fourth, our experiments show a strong relationship between the amplification factor and the routine behavior of volunteers. Hence, fifth and finally, MCS systems should exploit that additional awareness to refine and synchronize task scheduling strategies with the rhythms of a crowd.

CONCLUSIONS

Mobile crowdsensing is a powerful tool for performing sensing campaigns with citizens; however, one of the most difficult barriers to the spread of MCS campaigns is recruiting volunteers. To overcome this barrier, the social context of people involved in the MCS can effectively increase performance. In particular, we propose two complementary solutions. The first one aims to keep MCS systems involved in the MCS can effectively increase the results of an MCS campaign. That suggested three further lines of investigation. The first one deals with the detection of communities more accurately reflecting the social events shared by volunteers. The solution of this article exploits only spatial-temporal properties for detecting groups of people who are members of the same community. However, a step further can be taken by combining orthogonal sociological markers such as the physical activity of people, speech intensity, and similarity among visited places. All such markers can be pushed together in order to identify strong and durable ties among volunteers. The second line of investigation relies on a deeper knowledge of volunteer profiles; information such as interests, social habits, and preferences may be exploited to decide more accurately the target volunteers for a specific task. Finally, along the third direction and complementary to the work presented in this article, we are also considering the possibility to exploit more direct incentive mechanisms to extend the core volunteer base. On one hand, we are considering the possibility of providing different kinds of benefits and monetary micro-payments to motivate new users entering the MCS system; on the other hand, we are including in ParticipAct novel gamification and entertainment strategies to evolve MCS and sensing tasks into a game that offers virtual rewards to more active users and communities.

ACKNOWLEDGMENTS

This research was supported in part by CIRI, Center for ICT Technology Transfer of the University of Bologna, funded by POR FESR Emilia-Romagna 2007-2013.

REFERENCES


Biographies

Stefano Chessa (stefano.chessa@unipi.it) graduated from the University of Pisa, Italy, where he received a Ph.D. degree in computer science in 1999. Since 2015 he is an associate professor at the Computer Science Department of the University of Pisa. He has also been a research associate of the ISTI-CNR Institute since 2003. He has co-authored more than 100 papers published in international journals and conference proceedings. His current research interests are in the areas of Internet of Things, smart environments, ambient assisted living, activity recognition, and indoor localization.

Antonio Corradi (M) (antonio.corradi@unibo.it) graduated from the University of Bologna, Italy, and received an M.S. in electrical engineering from Cornell University, Ithaca, New York. He is a full professor of computer engineering at the University of Bologna. His research interests include distributed and parallel systems and solutions, middleware for pervasive and heterogeneous computing, infrastructure support for context-aware multimodal services, network management, and mobile agent platforms. He is a member of the Italian Association for Computing (NICA).

Luca Foschini (M) (luca.foschini@unibo.it) graduated from the University of Bologna, Italy, where he received a Ph.D. degree in computer engineering in 2007. He is now an assistant professor of computer engineering at the University of Bologna. His research interests include distributed systems and solutions for system and service management, management of cloud computing, context data distribution platforms for smart city scenarios, context-aware session control and adaptive mobile services, and mobile crowdsensing and crowdsourcing.

Michele Girolami (michele.girolami@isti.cnr.it) graduated from the University of Pisa, where he received a Ph.D. degree in computer science in 2015. Currently, he is a member of research staff at CNR-ISTI with the Wireless Network Laboratory. His research interests include data forwarding and service discovery in mobile social networks, crowdsensing techniques, and context-aware middleware for smart environments.
OPENRP: A Reputation Middleware for Opportunistic Crowd Computing

Dimitris Chatzopoulos, Mahdieh Ahmadi, Sokol Kosta, and Pan Hui

Abstract

The concepts of wisdom of crowd and collective intelligence have been utilized by mobile application developers to achieve large-scale distributed computation, known as crowd computing. The profitability of this method heavily depends on users’ social interactions and their willingness to share resources. Thus, different crowd computing applications need to adopt mechanisms that motivate peers to collaborate and defray the costs of participating ones who share their resources. In this article, we propose OPENRP, a novel, lightweight, and scalable system middleware that provides a unified interface to crowd computing and opportunistic networking applications. When an application wants to perform a device-to-device task, it delegates the task to the middleware, which takes care of choosing the best peers with whom to collaborate and sending the task to these peers. OPENRP evaluates and updates the reputation of participating peers based on their mutual opportunistic interactions. To show the benefits of the middleware, we simulated the behavior of two representative crowdsourcing applications: message forwarding and task offloading. Through extensive simulations on real human mobility traces, we show that the traffic generated by the applications is lower compared to two benchmark strategies. As a consequence, we show that when using our middleware, the energy consumed by the nodes is reduced. Finally, we show that when dividing the nodes into selfish and altruistic, the reputation scores of the altruistic peers increase with time, while those of the selfish ones decrease.

Introduction

The pervasive presence of smartphones, combined with their ever increasing capabilities in terms of CPU, memory, wireless connectivity, and a multitude of sensors, is making the case for mobile crowd computing more realistic. The idea is simple: users exploit their surrounding devices to accomplish some tasks without the help of a central entity. Many distributed device-to-device (D2D) applications have already been deployed in smartphones’ market stores, and many others are being extensively studied in literature [1–3].

Even though different in nature, all crowd computing applications share some important common requirements, which mainly are:

• The need for nearby devices discovery
• The need for evaluating the goodness of collaborating peers

A device that utilizes more than one crowd computing application would have to run multiple modules for device discovery — one for each application — and would be overloaded by scanning requests. The significant extra energy consumed due to redundant wireless scanning would make users limit their utilization of crowd computing applications, which of course would impede the diffusion of these collaboration-based applications. Furthermore, each application would have its own module for evaluating the goodness of peers. They would keep a private interaction history between the current device and peers, and a private evaluation of their willingness to help with collaborative tasks.

In this article we introduce OPENRP, a middleware that runs in the background on mobile devices, offering unified services to all crowd computing and opportunistic networking applications installed on the device. The middleware serves as a bridge for all applications that want to collaborate with other peers. The middleware constantly scans for nearby devices, exposes this information to the interested applications through a well defined application programming interface (API), and exchanges information with nearby devices related to her abilities and the installed crowd computing applications. This way, every time different devices want to collaborate with each other, they can do so through OPENRP. Applications can delegate tasks to the middleware, which takes care of transmitting them to suitable remote counterparts. Apart from handling the communication process, OPENRP also keeps track of the goodness of peers by calculating a reputation score for each of them. Devices exchange the calculated reputation values with each other, updating their local view of the other peers based on the weighted opinions of their trusted collaborators. This reputation exchange process is considered as one of the installed crowd computing applications since it is possible for a few users to not be interest-
OPENRP implements a unified and heterogeneous collaborating platform for all types of applications and integrates a reputation scheme that works based on the combined information obtained by all the applications. The middleware exposes a rich API that hides the technical difficulties related to the opportunistic network, making it transparent for applications to exploit the power of D2D collaboration.

ed in exposing their evaluations. An interesting feature introduced by the use of OPENRP is the circumvention of the cold-start problem. In a classic scenario, where each application maintains its private reputation values, a newly installed application would start with an empty database of reputation scores about other peers. In an OPENRP scenario this is partially solved. The newly installed application can use the collective knowledge of the existing applications maintained by the middleware, dealing in this way with the cold-start problem. OPENRP had to be lightweight, so its energy consumption is minimized, and multifaceted, so it can serve multiple types of applications.

Some of the most prominent opportunistic D2D applications proposed in the literature that can benefit from using the reputation middleware are:

**Context-Aware Applications**: Applications that allow users to connect and share information can benefit from OPENRP since the applications can evaluate the shared data and rate of the user that shared them.

**Opportunistic Networking**: Mobile users send messages to each other without the help of a centralized infrastructure, making use of opportunistic networking. OPENRP can discern the contribution of each mobile user and provide meaningful information to opportunistic networking applications, via its API, in order to select the best devices to forward application messages.

**Remote Sensing**: Users participating in remote sensing are required, usually by a cloud server, to perform local measurements using their devices’ sensors and report the measured values to a central entity. Via OPENRP, the central server will be able to select the best nodes based on their reputation.

**Computation Offloading**: This refers to the process of task migration from one device to more powerful devices. Traditionally, the cloud has been seen as the perfect candidate to offload tasks, but in more recent works proximal resource-rich devices are preferred.

**Cooperative Streaming**: A set of mobile users within proximity of each other are interested in watching the same video from the Internet at the same time in their smartphones (Keller et al. [4]).

**P2P-Based k-Anonymity Location Privacy Service**: Another application case is a peer-to-peer (P2P)-based k-anonymity location privacy service [5], in which users borrow their neighbors’ identity information to hide themselves in a crowd. Given that a user may not be interested in exchanging her collected reputation scores about others, we consider reputation exchange as a separate application.

**RELATED WORK**

The contribution of our middleware is twofold: to serve as a unified communication platform for all crowd computing applications, and to calculate the reputation of each device using a novel distributed reputation algorithm. In this section we present prominent works on crowd computing and popular reputation mechanisms for P2P systems.

The authors of [6] propose a side-payment scheme to make it optimal for software-based market agents to share their reputation information. Their mechanism assumes a set of centralized R-agents, which possess more accurate and qualitative information than others. It is therefore fundamental, for the proper functioning of their system, that the other agents learn to recognize the R-agents and develop a trust model for the agents themselves. The need for centralized agents makes this mechanism not suitable for today’s participatory P2P applications. In the same way, [7] make use of virtual currency in order to use credit-based incentives to motivate collaboration. The authors of SecuredTrust [10] analyze the different factors related to evaluating the trust in a multi-agent system and propose a comprehensive quantitative model for measuring such trust. All these works imply either a centralized server that keeps track of every currency exchange or, in the case of [7], require a tamper-resistant security module or make use of large number of messages for synchronization and distributed consensus, which contradicts the lightweight requirement of crowd computing applications.

SORT [11] is the closest work to the reputation mechanism integrated in our middleware. As in OPENRP, SORT proposes distributed algorithms used by peers to reason about trustworthiness of other peers based on past interactions and mutual recommendations. Different from SORT, our middleware uses information collected from multiple applications, which makes it converge faster to correct reputation scores. Furthermore, in OPENRP we only consider recommendations from trusted peers, thus reducing the number of exchanged control messages.

Mobile crowdsensing applications and frameworks, like [1], make use of devices’ sensors to perform local measurements and share their data with each other. Opportunistic networking applications and opportunistic computation offloading applications [2] make use of intermittent wireless connections between devices to accomplish distributed message forwarding or perform D2D computation offloading, respectively. The work presented in [12] proposes a mathematical model that analyzes the problem of data traffic offloading from cellular networks to the D2D distributed platform. The authors of [3] analyze different aspects of opportunistic applications, arguing that device heterogeneity is one of the key challenges to deal with for their large-scale adoption. OPENRP addresses the needs of the above works by implementing a unified and heterogeneous collaborating platform for all types of applications and integrating a reputation scheme that works based on the combined information obtained by all the applications. The middleware exposes a rich API that hides the technical difficulties related to the opportunistic network, making it transparent for applications to exploit the power of D2D collaboration.

System-wise, OPENRP is close to [13]. The authors of [13] propose the Group Context Framework, a programming toolkit that allows mobile devices to form groups and share contextual information. They mainly focus on the connectivity between the devices and the collaboration/cooperation between them. The core difference of our work with [13] is that we focus on
how to collect data about mobile users in order to build a reputation score for them.

**System Architecture**

As depicted in Fig. 1, OPENRP is modular and is composed of independent daemon components that communicate with each other using inter-process messages. The main components of the system, presented in detail in Fig. 2 and described later, are responsible for accepting requests from applications, collecting information about the surrounding neighbor devices, collaborating with nearby peers, and evaluating the reputation of the collaborators. As we can see from Fig. 1, applications are totally unaware of the underlying details of the system and interact with the middleware through a properly specified API layer, which is presented later. Moreover, OPENRP provides a configuration interface that the final user can use to select the resources she is willing to share when participating on the crowd computing collaboration system.

**Middleware Components**

In this section we give a more detailed description of the role of the main components of the middleware. The *decision engine* deals with the requests of the applications by delegating them to the proper OPENRP components. The *collaboration engine* is responsible for communicating with the outside world through the *communication layer*. At the heart of the middleware we can find four modules that are invisible to the developers and to the applications. The *data collector* collects information about the surrounding environment and the outcome of a collaboration between peers. The collected data are then encoded and stored in a *reputation table*. The *reputation calculator* processes the data of the reputation table and calculates a reputation score about each known device. Finally, the *appraiser* is the component that estimates the costs (e.g., energy) of running a collaborating operation.

**API to Crowd Computing Services**

OPENRP provides an extensive API to frameworks and applications to be used for P2P collaboration. The main purpose of the middleware is to receive tasks from the above applications and delegate them to nearby devices.

In the context of crowd computing, a task can be any remotely executable application part belonging to the set of applications such as the ones that are described in the Introduction.

In Fig. 3 we show the sequence diagram of a classical scenario of collaboration between two devices, where application i running on device dev1 wants to send a task to another device. To keep the diagram clean, we omit the arguments of the API’s methods, leaving only their names. Application i on dev1 uses the request() command to delegate a task to OPENRP. When OPENRP finds an available nearby device, dev2 in our example, it sends a request for collaboration, encoded as REQ, and waits for the response RESP from the other device. Table 1 shows the format and details of the REQ and RESP messages exchanged by the devices. If dev2 decides to collaborate, represented by the “Yes” branch in the diagram of Fig. 3, application i in dev1 is notified by OPENRP through the ack() method. At this point, application i on dev1 registers with OPENRP to be notified when the eventual response of the task is ready. Then OPENRP on dev1 sends the task to the OPENRP on dev2 using the message TASK. The middleware on dev2 receives the task and uses the method process() to delegate it to the appropriate application, which is the same application i as the one running on dev1 that knows how to handle the task. After the application processes the task, it uses the method answer() to inform OPENRP on dev2, which then sends a message.

![Diagram](image-url)
ANS (Table 1) to the requesting device dev1 that the task was correctly processed. OPENRP on dev1 then uses the method response() to pass the eventual result to application i or to simply inform it that the task was processed by dev2, depending on the task type. Application i uses the method feedback() to advertise its experience of the collaboration with dev2. OPENRP collects this information and uses it to update the reputation of dev2.

If, for some reason, dev2 refuses to process the task (the reason can be included in the response message RESP, Table 1), OPENRP in dev1 updates the reputation score of dev2, informs application i using the method reject() that the other device did not accept the task, and keeps scanning for other devices.

Apart from the methods described in the previous example, OPENRP provides the applications with more API commands, as listed here:
- **getMostTrusted(N,S|A)** returns the N most trusted neighbors. This method has both synchronous and asynchronous versions, returning the currently nearby devices or all the known ones, respectively.
- **sync()** command forces OPENRP to contact the neighbors and ask for updates.
- **getAvgContactDuration(dev d)** gives the average contact duration between the current device and device d.
- **clearCache()** is used by an application to clear the cache.
- **removeFromCache(task)** is used by an application to tell the middleware that the task is not needed anymore.
- **cacheResult(task, timestamp)** is used by an application to ask the middleware to cache a result so that for future calls of the same task there will be no need to process the task.
- **getUserRep(id)** returns the reputation of a specific user.
- **getListOfSensors(dev d)** gives the list of the available sensors for use at device d.
- **getMeetingFrequency(dev d)** returns the meeting frequency between the current device and device d.
- **getNmostTrusted(N,S|A)** gives the list of the most trusted neighbors.
- **getNmostTrusted(N,S|A)** returns the N most trusted neighbors. This method has both synchronous and asynchronous versions, returning the currently nearby devices or all the known ones, respectively.
- **notifyTheMiddleware** is used by an application to notify the middleware about new incoming data.
- **getCurrentConditions(dev d)** returns the current battery level, CPU utilization, and other available resources of device d.

**OPENRP: COLLECTED DATA AND COLLECTIVE INTELLIGENCE**

In the context of crowd computing there is no centralized server, which makes things more difficult when it comes to evaluating the reputation of the nodes in a fully distributed way. Inspired by the Oxford Dictionary definition, we define reputation in the context of crowd computing as: "The belief in a peer’s capability and willingness to process a task for others."

To collect the needed data for calculating the reputation scores, OPENRP uses the applications’ feedback; the willingness of the peers to collaborate (which is included in the message RESP, Table 1); and the recommendations of

---

**Table 1.** Format, definition, and size in bytes of the messages exchanged between the OPENRP middleware running on different devices.
other peers. OPENRP on device $i$ calculates the reputation score of device $j$ as a weighted combination of the direct interactions between $i$ and $j$, $r_{ij}^{dir}$, and the recommendations about $j$ received by other devices, $r_{ij}^{rec}$:

$$r_j = \alpha \cdot r_{ij}^{dir} + (1 - \alpha) \cdot r_{ij}^{rec} \quad (1)$$

where the value of $\alpha \in (0, 1)$ is used to give more weight to the direct interaction or to the received recommendations and depends on the number of direct interactions. It is increasing and converging to 1 while the number of direct interactions is increasing. More sophisticated update algorithms are part of our future work.

**Reputation Update Algorithm**

The reputation score of a node changes whenever it interacts with other nodes. There are mainly two events that cause an entry to enter the table:

1. An application that has received a response from a device gives explicit feedback about that device using the `feedback()` method of the OPENRP API.
2. OPENRP on a device tries to send a task to another device, which refuses the task by sending a message `RESP` with the “NO” parameter set (Table 1).

In the first case, the satisfaction of the interaction is determined explicitly by the application, while in the second case the value varies depending on the reason the node refused to collaborate. In our system, a device can refuse a request for collaboration if it is busy executing another task or because the device simply does not have enough free resources (e.g., the battery is below a certain threshold set by the device owner). If the device is busy, OPENRP records this interaction as neutral and does not alter the reputation score of the device. If the device does not have free resources and the thresholds set by the user through the user configuration interface are quite high, OPENRP decreases the reputation score of the device, flagging it as selfish.

One important feature of OPENRP is its approach to the cold-start problem, which is presented when an application is freshly installed and has no information about the surrounding neighbors. OPENRP partially solves this drawback by combining the feedback received by all the applications installed on the device on a unique reputation table. If there are no installed applications, the new application can use the provided functionality to ask for recommendations scored by her neighbors. This way, newly installed applications will have a starting point and will contribute to further enrichment of the table.

**Implementation and Evaluation**

We implement and evaluate the performance of OPENRP on top of real traces that include mobile users and battery levels. We use the Mobile Data Challenge (MDC) dataset [14, 15], which is 30 days long and includes the mobility traces and battery usage of 64 users. We divided the users into selfish, who reject all requests for task processing but are eager to ask for help from other users, and altruistic, who collaborate whenever they have enough resources. Through extensive simulations, we measured the traffic incurred in the system and the energy consumed by the altruistic and selfish nodes, and the evolution of their reputation scores. In order to produce each plot we continue the simulation until we get to an accuracy of at least 0.05 for all the metrics with 0.95 probability. We selected two representative crowd computing applications that can benefit from utilizing our middleware, computation offloading and information spreading. We compare our strategy with two benchmark algorithms that we call lazy and greedy. The lazy algorithm tries to offload the code or forward the message to every nearby device available at the time of task creation. In the case of computation offloading, the code is executed locally if there are no nearby devices, while in the case of information spreading the message is simply stored until a device becomes available. The greedy
algorithm offloads the code or forwards the message whenever it meets another node, not only at the moment of the task creation.

We generate requests using the Poisson process, and each user, after joining the ecosystem, generates 1 request around every 40 min. The duration of the code to be executed in the case of computation offloading is uniformly selected between 250 and 500 s. We assume that each piece of code has a maximum allowed delay, that is, execution deadline, which is set to 30 times the execution duration of the task. The time to live (TTL) of the messages in case of the information spreading application is set to one day.

Figure 4 shows the daily evolution of the reputation scores of the nodes during a one-month simulation for two experiments:
- Only computation offloading requests
- Both computation offloading and information spreading requests

The reputation is calculated based on all the generated requests. Each plot includes three experiments that differ from each other in the percentage of selfish and altruistic nodes. The reputation scores of the altruistic nodes increase with time, while the reputation scores of the selfish nodes decrease. Moreover, in the second experiment we can see that the score values change faster because the users are running two applications at the same time, so the chances for collaboration increase, and the reputation scores are updated more frequently.

In Figs. 5a and 5b we show the traffic generated and the energy consumed by the nodes, respectively, when using the three different strategies on nodes running the computation offloading application. To measure the traffic, we simply count the number of messages exchanged between the nodes. We fixed the bandwidth of the communication channel between the devices to 10 Mb/s. According to the specifications of WiFi-Direct, its bandwidth is 200 Mb/s, while from our experiments we can show that it is at least 20 Mb/s. To estimate the energy consumption, we first measured the energy consumption of the WiFi interface using the highly adopted Monsoon Power Monitor2 when sending and receiving messages between two real devices. From these measurements we fixed the energy consumption of transmitting/receiving one single byte to 0.00000048 J and 0.0000004 J, respectively. We then used these values to estimate the energy consumption of each exchanged message between the devices, also referring to Table 1, by simply multiplying the size of a message by the energy values per byte. As we can see, OPENRP generates down to two times less traffic than the lazy and greedy approaches.

As expected, the lower traffic overhead results in reduced energy consumption, as can be seen from Fig. 5b. OPENRP manages to reduce the traffic and energy consumption due to its selfish behavior detection, since these nodes are correctly identified by the middleware (due to their low reputation scores) and are not considered by the well behaving nodes.

**DISCUSSION**

In OPENRP, every running application has a unique application id, and whenever OPENRP handles executing part of it in nearby devices, it first broadcasts the id of the application, and all the nearby devices that support the same application will initially update their knowledge about this device and then reply with their availability to help. Moreover, underutilized mobile devices broadcast heartbeat messages where they include the ids of their supporting applications. If one device has installed one application, it can receive and execute a task of the same type.

The high research interest in D2D-based technologies like WiFi-Direct and the protocol itself make frameworks like OPENRP feasible. It is worth mentioning that the two most popular problems on WiFi-Direct-related publications are device interconnectivity and group formation, which are the core components of OPENRP.

**CONCLUSION AND FUTURE WORK**

We have proposed and evaluated OPENRP, a middleware and reputation mechanism for mobile crowd computing applications. OPENRP serves as a bridge for all applications that make use of opportunistic encounters, handling all their requests and responses. The middleware makes use of the collaborative device-to-device approach by offering a simple and extensible API to developers, hiding all the complex underlying details. The middleware includes a novel distributed reputation algorithm, which uses the collected intelligence of interactions between devices to estimate peers’ reputations. By extensive evaluations on real mobility traces, we show that OPENRP significantly reduces the
traffic and energy consumed by nodes running two representative crowd computing applications compared to two benchmark algorithms. Furthermore, we show that the reputation scores evolve with time, increasing for altruistic users and decreasing for selfish ones. As future work, we are implementing the middleware on Android phones, and will test its performance on a setting with real devices and real users, and analyze the energy requirements of OPENRP and its basic components.

Acknowledgments
This research has been supported, in part, by General Research Fund 26211515 from the Research Grants Council of Hong Kong, Innovation and Technology Fund ITS/369/14FP from the Hong Kong Innovation and Technology Commission, and the European Commission under the Horizon 2020 Program through the RAPID project (H2020-ICT-644312).

References

Biographies
DIMITRIS CHATZOPOULOS (dcdb@cse.ust.hk) received his Diploma and Msc in computer engineering and communications from the Department of Electrical and Computer Engineering of the University of Thessaly, Volos Greece. He is currently a Ph.D. student at the Department of Computer Science and Engineering of the Hong Kong University of Science and Technology (HKUST) and a member of the HKUST-DT System and Media Lab. His main research interests are in the areas of mobile computing and device-to-device ecosystems.

MAHDI AHMADI (mahmadi@ce.sharif.edu) received her B.S. degree in computer engineering from the University of Tehran in 2013, and her M.S. degree in computer engineering from Sharif University of Technology in 2015. She is currently working toward her Ph.D. degree at Sharif University of Technology. She was a research assistant from July 2015 to September 2015 at the HKUST-DT System and Media Lab. Her main research interests include networking, distributed systems, and mobile cloud computing.

SOKOL KOSTA (kosta@di.uniroma1.it) received his Bachelor’s, Master’s, and Ph.D. degrees in computer science (summa cum Laude) from Sapienza University of Rome, Italy, in 2006, 2009, and 2013, respectively. He was a postdoctoral researcher at Sapienza University and a visiting researcher at HKUST in 2015. He is currently an assistant professor at Aalborg University Copenhagen. His research interests include networking, distributed systems, and mobile cloud computing.

PAN HUI (panhui@cse.ust.hk) received his Ph.D. degree from the Computer Laboratory, University of Cambridge, and earned both his M.Phil. and B.Eng. degrees from the Department of Electrical and Electronic Engineering, University of Hong Kong. He is currently a faculty member of the Department of Computer Science and Engineering, HKUST, where he directs the HKUST-DT System and Media Lab. He also serves as a distinguished scientist of T-labs Germany, and an adjunct professor at Aalto University, Finland.
The extremely pervasive nature of mobile technologies, together with the user’s need to continuously interact with her personal devices and to be always connected, strengthen the user-centric approach to design and develop new communication and computing solutions. Nowadays users not only represent the final utilizers of the technology, but they actively contribute to its evolution by assuming different roles: they act as *humans*, by sharing contents and experiences through social networks, and as *virtual sensors*, by moving freely in the environment with their sensing devices. Smart cities represent an important reference scenario for the active participation of users through mobile technologies. It involves multiple application domains and defines different levels of user engagement. Participatory sensing, opportunistic sensing, and mobile social networks (MSNs) currently represent some of the most promising people-centric paradigms. In addition, their integration can further improve the user involvement through new services and applications. In this article we present SmartCitizen, an MSN application designed in the framework of a smart city project to stimulate the active participation of citizens in generating and sharing useful contents related to the quality of life in their city. The app has been developed on top of a context- and social-aware middleware platform (CAMEO) able to integrate the main features of people-centric computing paradigms, lightening the app developer’s effort. Existing middleware platforms generally focus on a single people-centric paradigm, exporting a limited set of features to mobile applications. CAMEO overcomes these limitations and, through SmartCitizen, we highlight the advantages of implementing this type of mobile application in a smart city scenario. Experimental results shown in this article can also represent the technical guidelines for the development of heterogeneous people-centric mobile applications, embracing different application domains.

**ABSTRACT**

The extremely pervasive nature of mobile technologies, together with the user’s need to continuously interact with her personal devices and to be always connected, strengthen the user-centric approach to design and develop new communication and computing solutions. Nowadays users not only represent the final utilizers of the technology, but they actively contribute to its evolution by assuming different roles: they act as *humans*, by sharing contents and experiences through social networks, and as *virtual sensors*, by moving freely in the environment with their sensing devices. Smart cities represent an important reference scenario for the active participation of users through mobile technologies. It involves multiple application domains and defines different levels of user engagement. Participatory sensing, opportunistic sensing, and mobile social networks (MSNs) currently represent some of the most promising people-centric paradigms. In addition, their integration can further improve the user involvement through new services and applications. In this article we present SmartCitizen, an MSN application designed in the framework of a smart city project to stimulate the active participation of citizens in generating and sharing useful contents related to the quality of life in their city. The app has been developed on top of a context- and social-aware middleware platform (CAMEO) able to integrate the main features of people-centric computing paradigms, lightening the app developer’s effort. Existing middleware platforms generally focus on a single people-centric paradigm, exporting a limited set of features to mobile applications. CAMEO overcomes these limitations and, through SmartCitizen, we highlight the advantages of implementing this type of mobile application in a smart city scenario. Experimental results shown in this article can also represent the technical guidelines for the development of heterogeneous people-centric mobile applications, embracing different application domains.

**INTRODUCTION**

The penetration of smartphones and wearable sensing devices in everyday life is driving the definition and development of people-centric computing and communications paradigms [1, 2]. These are aimed at developing smart services able to increase citizens’ participation and empowerment. Smart cities represent a relevant scenario for people-centric solutions in which, on one hand, sensors and smart objects are used to monitor the city infrastructures (electric grid, lighting, transport, etc.), and on the other hand, users with their personal devices become sensors able to monitor the human behavior in the city and contribute to the definition of smart policies. Without concrete user involvement, new policies for efficient energy consumption, smart mobility, traffic, environmental monitoring, and other aspects are simply ineffective.

From a technological point of view, user involvement can be achieved by leveraging three user-centric paradigms:

- Participatory sensing [3], in which users exploit their sensing devices to collect information during their daily living activities and share them (with other users), mainly through web servers
- Opportunistic sensing [5], in which mobile applications opportunistically exploit all the sensing technologies available in the environment (not requiring direct user interaction)
- Opportunistic mobile social networks (MSNs) [4], in which users directly generate and share heterogeneous types of contents with nearby users in real time by exploiting the physical interactions of their personal mobile devices (i.e., opportunistic communications)

The emergence of these three paradigms indicates a substantial change in the management of legacy ad hoc and sensor networks, creating new research challenges related to sensing devices, wireless communications, data management, and dissemination.

By integrating the features of these user-centric paradigms in a single mobile system, we can design efficient and personalized mobile applications in several domains (e.g., health and well being, social inclusion, smart mobility, environmental and urban monitoring), and integrate them toward the real development of a smart city.

To this aim, we designed and developed CAMEO [6], a context- and social-aware middleware platform for mobile devices and MSN applications. It provides app developers with the functionalities to automatically discover...
users and mobile devices nearby, and to identify users’ common interests, habits, available services, and resources. All this data is collected and exchanged through opportunistic communications, and represent the context used by CAMEO to implement optimized networking protocols, resource management mechanisms, and the exportation of context-aware features to MSN applications. CAMEO is able to support concurrent MSN apps on the same device through dedicated inter-process communications, and it is deployed as an Android app running in the background.

As a real example of an MSN app running on top of CAMEO, we developed the SmartCitizen app. It has been designed in the framework on top of CAMEO, we developed the SmartCitizen, and it is deployed as an Android app running through dedicated inter-process communications, and it is deployed as an Android app running in the background.

As a real example of an MSN app running on top of CAMEO, we developed the SmartCitizen app. It has been designed in the framework on top of CAMEO, we developed the SmartCitizen, and it is deployed as an Android app running in the background.

![SmartCitizen](image)

**SmartCitizen App**

SmartCitizen is aimed at stimulating the active participation of citizens in collecting and sharing data related to the environmental conditions of the city of Pisa. The first approach to the app is represented by an intuitive visualization of the current environmental conditions of some city areas, used as a testbed for the monitoring infrastructure developed inside the SmartHealthyENV project. Then the app stimulates the creation of discussions between citizens on topics related to air quality by borrowing functions from social networking applications, such as posting, commenting, and chatting. Of course, discussions are not limited to environment-related topics, and citizens can use the app as a traditional mobile social network. In Fig. 1 we present some screenshots of the graphical user interface of the app showing a real use case scenario. In the first screenshot, the app presents, on a map, the air quality index measured on each sensing station deployed in the city and presented as a circle of a range of colors from red (polluted area) to green (healthy area) for intuitive comprehension by common citizens. However, the app is designed for expert users as well, such as employees of the local municipalities, able to interpret the detailed information about measured chemical parameters. In this case, an authorized user, clicking on the circle, will visualize the detailed data of the related station.

All the sensing data is downloaded from the SmartHealthyENV server if the user device has Internet connectivity, or it can be downloaded through device-to-device communication if available on other user devices in proximity. Regarding the generation of user contents, Fig. 1 shows the different views provided to the user: the form for the creation of a new post, the list of active discussions on different topics, and the list of generated content for each discussion with appropriate notification of new available contents. In the use case shown in Fig. 1, users commented on the weather in a specific area generally used as a pedestrian path, supported by the green air quality index in the same area and some multimedia content. To further customize SmartCitizen for specific user categories, we integrated some features for sports users (Fig. 2), who are generally interested in environmental conditions for their outdoor activities: they prefer doing activities in healthy areas, and most of them are used to comparing their experiences and performances through dedicated or online social networks. In this case, SmartCitizen allows users to directly record activity paths in the city and associates the air quality information derived from heterogeneous people-centric paradigms in a single mobile app.

The article is organized as follows. First, we present in detail SmartCitizen features. We show the experimental evaluation divided into three different subsections related to heterogeneous context management, user experience, and opportunistic communication issues, respectively. Finally, we present the advantages of the proposed solution and of the entire framework with respect to existing work and future research directions.

---

1. SmartHealthyENV — Smart Monitoring Integrated System for a Healthy Urban Environment in Smart Cities, a project funded by the Tuscany region, Italy, under the program POR-CREO FESR 2007–2013, http://www.progettoshe.it
2. OGC Sensor Web Enablement DWG (SWE), http://www.opengeospatial.org/projects/groups/sensorwebdg
from the monitoring network with each path, and the user can further customize it with her own information, comments, and suggestions. These features allow users to share immediate information about the specific paths and areas in the city.

All the contents generated though the app are disseminated on the network through the context-aware content dissemination protocol provided by CAMEO. To this end, the app developer must only define the application user profile: a set of information characterizing the user in the framework of the current application (e.g., user category, interests related to generated contents and/or specific application features). This profile is automatically shared among mobile devices in proximity running CAMEO, and it is used to evaluate the potential utility of the available contents provided by other users in the network for the local user. To this aim, each user should characterize the generated content (generally multimedia contents like audio/video messages, often associated with text posts) by selecting or generating one or more tags, which are used as input for the matching “user-content” algorithm. Since users’ interests can change over time, CAMEO automatically adapts the dissemination protocol to those changes, maintaining all the procedure transparent to the app developer.

Users’ profiles and contents are distributed on the network through peer-to-peer communications. Nodes’ communications happen through WLAN access or a WiFi Direct self-organizing network, based on the current availability. The app is completely transparent to the type of communication protocol used, which is automatically selected by CAMEO.

In order to enrich the centralized storage of environmental data with user-generated content, SmartCitizen also allows users to upload their contents to a SmartHealthyENV server, maintaining compliance with the SWE data format. This is not automated to limit the connections toward the server and to cause the user to upload only significant contents.

In terms of privacy, users can decide what type of information to share in proximity and what can be stored remotely. They must explicitly agree to share it before starting the app through a specific privacy statement. However, no personal and sensitive data are exchanged with others.

In the next section we present the experimental evaluation of SmartCitizen in terms of:
• Efficient management of heterogeneous sensing data based on the SWE standard
• User experience in a real setting
• Open issues related to opportunistic communications based on WiFi Direct

**EXPERIMENTAL EVALUATION**

**HETEROGENEOUS SENSOR DATA MANAGEMENT**

To perform efficient data exchange with the SmartHealthyENV server, the app uses SME encoding and decoding procedures integrated in CAMEO. SME has two main targets:
• To define the descriptions of phone-embedded sensors and related measurements compliant with the SWE standard
• To efficiently process SWE sensor data on mobile devices

The interaction with an SWE server requires a series of steps as defined by the standard. The application must first query the server to obtain a description of the monitoring service in terms of available sensors and their properties (e.g., vendor name, sampling frequency, range of valid measurements). Then it can ask for the measurements of single or multiple sensors (called sensor observations). In the case of citizen users, the app asks only for the indexes as the elaboration of multiple raw parameters on the same sensing
station. Instead, for an expert user, it can request detailed data in a variable time window in order to also visualize temporal variations of the single parameters. Therefore, the size of the SWE files downloaded from the server can vary depending on the application request, and it can have an impact on the mobile device processing time for data serialization and deserialization procedures.

For this reason, to evaluate the efficacy of SmartCitizen in managing heterogeneous sensor data through SME, we performed several experiments in a real environment. Specifically, we measured serialization and deserialization times of an SWE file containing a variable number of raw sensor measurements. The file contains an increasing number of values related to independent measurements (from 0 up to 50,000 in steps of 10,000). This represents a complex and large SWE file that can be received by a mobile device (for which SME deserialization is required) and subsequently forwarded from a mobile device to another (for which SME serialization is required before transmission).

The XML file size of this observation ranges between 1.9 kB (i.e., containing only SWE XML header and 0 values) and 2.3 MB (i.e., containing 50,000 values). Results shown in Table 1 highlight the efficiency of the SME library to manage this operation in a few seconds. Specifically, deserialization and serialization times remain reasonably small even in the case of 50,000 sensor observations, which represents a high number of observations for a limited area such as the neighborhood of a single user in an urban scenario.

USER EXPERIENCE

Before releasing the app on the store and opening the test to citizens, we performed an experiment to assess the user experience in real settings. We recruited 19 participants in a one-day session to test the app, analyze reactions, and collect suggestions. The participants were 17 students and 2 professors of a high school class who planned a visit to our campus to learn about new technological solutions for smart cities. We briefly presented the research project and the features of SmartCitizen, and we asked them to install CAMEO and the app on their Android smartphones. Then they could use the app while moving around the campus in order to generate contents and communicate with each other. The experiment lasted three hours (excluding the initial training session). At the end of the visit, we asked the users to fill in a questionnaire designed to identify their technological skills and to evaluate their experience while using the app.

The majority of the users declared medium-high skills in the use of new technologies (i.e., all of them spend at least one hour per day on the Internet, and they all use online social network applications, like Facebook and Twitter). They also understood very quickly how to use the paradigm of MSN inside the app, as demonstrated by the amount of generated and shared content, detailed in Table 2.

During the experiment, we contributed by generating a couple of posts just to show the participants how to use the app. Then they actively involved us as additional users in the experiment, adding comments to our posts. Therefore, the effective number of people involved in the experiment was 22, creating in total 39 posts, 151 comments, and 22 photos attached to posts or comments. These also contained 22 additional tags (i.e., other than those already predefined by the application), customizing the content dissemination protocol on new user-defined interests. Out of the 22 participants, 13 created new posts during the experiment, 20 generated at least one comment to an existing post, 13 generated both posts and comments, and no one generated only comments. Moreover, eight users created multimedia content (photos or videos) in their posts or comments.

The statistics of the generated contents in...
As previously described, SmartCitizen has no control over the type of communication protocol used to disseminate data over the network, depending on the availability of a WLAN access. Actually, the deployment of real opportunistic networks based exclusively on device-to-device (D2D) communications is still an open research challenge.

Actually, the deployment of real opportunistic networks based exclusively on device-to-device (D2D) communications is still an open research challenge.

Actually, the deployment of real opportunistic networks based exclusively on device-to-device (D2D) communications is still an open research challenge.

Actually, the deployment of real opportunistic networks based exclusively on device-to-device (D2D) communications is still an open research challenge.

Actually, the deployment of real opportunistic networks based exclusively on device-to-device (D2D) communications is still an open research challenge.

Actually, the deployment of real opportunistic networks based exclusively on device-to-device (D2D) communications is still an open research challenge.

Actually, the deployment of real opportunistic networks based exclusively on device-to-device (D2D) communications is still an open research challenge.

Actually, the deployment of real opportunistic networks based exclusively on device-to-device (D2D) communications is still an open research challenge.

Actually, the deployment of real opportunistic networks based exclusively on device-to-device (D2D) communications is still an open research challenge.

Actually, the deployment of real opportunistic networks based exclusively on device-to-device (D2D) communications is still an open research challenge.

Actually, the deployment of real opportunistic networks based exclusively on device-to-device (D2D) communications is still an open research challenge.

Actually, the deployment of real opportunistic networks based exclusively on device-to-device (D2D) communications is still an open research challenge.

Actually, the deployment of real opportunistic networks based exclusively on device-to-device (D2D) communications is still an open research challenge.

Actually, the deployment of real opportunistic networks based exclusively on device-to-device (D2D) communications is still an open research challenge.

Actually, the deployment of real opportunistic networks based exclusively on device-to-device (D2D) communications is still an open research challenge.

Actually, the deployment of real opportunistic networks based exclusively on device-to-device (D2D) communications is still an open research challenge.

Actually, the deployment of real opportunistic networks based exclusively on device-to-device (D2D) communications is still an open research challenge.

Actually, the deployment of real opportunistic networks based exclusively on device-to-device (D2D) communications is still an open research challenge.

Actually, the deployment of real opportunistic networks based exclusively on device-to-device (D2D) communications is still an open research challenge.

Actually, the deployment of real opportunistic networks based exclusively on device-to-device (D2D) communications is still an open research challenge.

Actually, the deployment of real opportunistic networks based exclusively on device-to-device (D2D) communications is still an open research challenge.

Actually, the deployment of real opportunistic networks based exclusively on device-to-device (D2D) communications is still an open research challenge.

Actually, the deployment of real opportunistic networks based exclusively on device-to-device (D2D) communications is still an open research challenge.

Actually, the deployment of real opportunistic networks based exclusively on device-to-device (D2D) communications is still an open research challenge.

Actually, the deployment of real opportunistic networks based exclusively on device-to-device (D2D) communications is still an open research challenge.

Actually, the deployment of real opportunistic networks based exclusively on device-to-device (D2D) communications is still an open research challenge.

Actually, the deployment of real opportunistic networks based exclusively on device-to-device (D2D) communications is still an open research challenge.

Actually, the deployment of real opportunistic networks based exclusively on device-to-device (D2D) communications is still an open research challenge.

Actually, the deployment of real opportunistic networks based exclusively on device-to-device (D2D) communications is still an open research challenge.
work performance with a high number of nodes. In fact, even with a few nodes, we experimentally found that the throughput drastically decreases when the number of client-to-client communications within a group increases. Specifically, we measured the throughput in a WiFi Direct group with a variable number of peers that communicate with each other for 180 minutes. In Table 3, we present different traffic configurations (GO-client, client-client, etc.) for different group sizes (2, 3, and 4 nodes). In the first configuration, we created a group of two peers, a GO and a client, and generated TCP traffic from the former to the latter (and vice versa). The average throughput after an experiment of 180 minutes of continuous traffic was 34.4 Mb/s. This represents a rather high throughput, also supporting the exchange of large files in short time periods. This makes this technology suitable for opportunistic scenarios in which the high mobility of nodes limits the duration of physical encounters and, consequently, the time opportunity to exchange data, like the urban scenario of SmartCitizen.

In a scenario with three peers per group, considering the communication between the GO and one client, and vice versa, we obtained broadly the same results of the previous case in terms of throughput. Despite this, communications from client to client lead to a throughput value that is nearly half the previous case. This is due to the fact that client-to-client communications require two steps before reaching the destination, forcing the communication to pass through the GO. To better investigate the limitation of the star topology in an opportunistic network, we set up an experiment with multiple client-to-client communications within groups of four peers. In this case, considering the worst case results (i.e., the GO and two clients simultaneously send data to a single client), the throughput related to the GO-client link is around 37 Mb/s, while that obtained for the client-to-client links is around 3 Mb/s. By considering these results, it is clear that client-to-client communications are penalized by the group size, and this aspect has to be taken into account while designing forwarding and content dissemination protocols in opportunistic networks. Specifically, the framework could favor small groups with high reconfigurability based on users’ mobility.

In addition, WiFi Direct works on a dedicated WiFi interface, which allows mobile devices to use both WiFi and WiFi Direct interfaces in specific configurations. For example, two devices belonging to two different WiFi Direct groups can set up an inter-group communication, thus allowing the creation of a multi-group. In this scenario, one of the clients of a group is connected to the GO of a second group through its standard WiFi interface. This peer acts as a bridge that connects the two groups and enables the exchange of data between them. However, only the bridge is able to receive and send data from/to the other group. In fact, in order to forward contents from one node to all the others belonging to the two groups, we should implement a customized routing protocol (in charge of also maintaining information about the physical configuration of each group) as described in [9, 10].

Despite these limitations, we decided to investigate the feasibility and advantages of using multi-groups in our scenario by conducting some experiments with particular attention to throughput analysis. Specifically, we measured the average throughput considering two concurrent communications from two clients of the first group to the client of the second group (connected through a dedicated interface to the GO of the first group). In this scenario the average throughput is 6.8 Mb/s, comparable with the results found in simple groups with the same cardinality. This means that the use of an additional interface and a separate channel between groups does not impact the data traffic. However, Android OS does not support the automatic creation of multi-groups, since each group has a pri-

<table>
<thead>
<tr>
<th>Group size</th>
<th>GO→client</th>
<th>Clients→client</th>
<th>1) GO→client 2) client→client</th>
<th>1) Client(B)→client(A) 2) Client(C)→client(A)</th>
<th>1) GO→client(A) 2) Client(B)→client(A) 3) Client(C)→client(A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>54.4 Mb/s</td>
<td>—</td>
<td>1) 44.3 Mb/s 2) 4.24 Mb/s</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>3</td>
<td>52.6 Mb/s</td>
<td>22.3 Mb/s</td>
<td>1) 40 Mb/s 2) 5.41 Mb/s</td>
<td>1) 12.7 Mb/s 2) 9.07 Mb/s</td>
<td>1) 17.4 Mb/s 2) 2.8 Mb/s 3) 5.22 Mb/s</td>
</tr>
<tr>
<td>4</td>
<td>52.75 Mb/s</td>
<td>17 Mb/s</td>
<td>1) 40 Mb/s 2) 5.41 Mb/s</td>
<td>1) 12.7 Mb/s 2) 9.07 Mb/s</td>
<td>1) 17.4 Mb/s 2) 2.8 Mb/s 3) 5.22 Mb/s</td>
</tr>
</tbody>
</table>

Table 3. Throughput in WiFi Direct groups with 2, 3, and 4 peers. A→B indicates that the throughput is the average for unidirectional communication from A to B and from B to A.
In order to extensively evaluate the entire framework in a real environment, we are planning in the next future to introduce additional features able to support the deployment of large-scale experiments. In this way, we will be able to directly involve citizens in a real smart city scenario.

**RELATED AND FUTURE WORK**

In the last few years, different middleware solutions have been proposed in the literature to support participatory sensing, opportunistic sensing, and mobile social networks. Despite this, to the best of our knowledge, all these solutions operate independently, thus generating several independent mobile applications. The integration of these paradigms in a single mobile platform is still to be achieved and, consequently, the definition of a single mobile application model that exploits all these features. Through CAMEO we intend to fill this gap in order to support a much larger range of people-centric services than existing middleware solutions, and SmartCitizen is an example of how to make it real.

We can list several works on the three single paradigms cited above, each focused on the improvement of a specific feature. As far as participatory sensing is concerned, Phokas et al. [11] have recently proposed a middleware aimed at simplifying applications’ access to devices’ sensors. Mobile [13] is designed to monitor the sensing capabilities of mobile devices over time, trying to limit the amount of redundant data, temporarily deactivating redundant sensors and saving resources. Usense exploits sensors on mobile devices distributed in the environment to perform community-driven sensing tasks.

In the literature there is also middleware implementing device-to-device communications to support opportunistic data exchange between mobile devices, some based on the definition of context information to improve data dissemination, as demonstrated in the Haggle project [15]. Finally, moving toward the application layers, we are planning in the near future to extensively evaluate the entire framework in a real environment, we are planning in the near future to introduce additional features able to support the deployment of large-scale experiments. In this way, we will be able to directly involve citizens in a real smart city scenario.

**REFERENCES**


**BIographies**

**FRANCA DELMASTRO** (Franca.Delmastro@iit.cnr.it) is a researcher at the IIT Institute of the Italian National Research Council. Her main research interests are in the design, development and deployment of innovative ICT solutions for smart cities, with particular attention to the health and well being domains. These solutions range from network protocols for wireless and body area network to context- and social-aware middleware and mobile applications. She has published several papers in international journals and conferences. She is an Editorial Board member of the International Journal of Communications, Network and System Sciences, and has participated in the organization of several international conferences.

**VALERIO ARNABOLDI** (Valerio.Arnaboldi@iit.cnr.it) is a researcher at the IIT Institute of the Italian National Research Council. He works on the analysis and characterization of social structures and information dissemination in online social networks, and on context- and social-aware middleware solutions for mobile social networking applications. He received a Ph.D. in computer engineering from the University of Pisa in 2014. He co-authored the book Online Social Networks: Human Cognitive Constraints in Facebook and Twitter Personal Graphs.

**MARCO CONTI** (Marco.Conti@iit.cnr.it) is a research director of the Italian National Research Council. He has published 300+ research papers and five books related to computer networks, pervasive systems, and social networks. He received the Best Paper Award at ITIP TGC Networking 2011, IEEE ISCC 2012, and IEEE WkhpM 2013. He is Editor-in-Chief of Computer Communications and Associate Editor-in-Chief of Pervasive and Mobile Computing. He has served as TPC/General Chair for several major conferences, including Networking 2002, IEEE WoWMoM 2005 and 2006, IEEE PerCom 2006 and 2010, ACM MobiHoc 2006, and IEEE MASS 2007.
CALL FOR PAPERS
IEEE COMMUNICATIONS MAGAZINE

NETWORK SLICING IN 5G SYSTEMS

BACKGROUND

Network slicing has evolved from a simple network overlay concept to a fundamental feature of the emerging 5G systems enabling dynamic multi-service support, multi-tenancy and the integration means for vertical market players. Network slicing can drastically transform the networking perspective by abstracting, isolating and separating logical network behaviors from the underlying physical network resources. Network operators, can exploit network slicing for reducing capital and operations expenditures, allowing also programmability and innovation, necessary to enrich the offered services from simple communications services to a wider range of business services. The separation of different functions by abstractions (e.g. radio resources from packet processing) simplifies the integration challenges especially for applications supporting vertical industries beyond telecommunication.

Network slicing in 5G systems may be performed by abstracting different physical infrastructures into a logical network that contains shared resources, such as radio spectrum or dedicated core network equipment, and virtual network functions obtained by breaking down single physical equipment into multiple instances, which are isolated from each other. Virtualization of network functions allow to decouple network node functions from proprietary hardware appliances in order to create distinct building blocks that can be flexibly chained to create communication services.

The notion of resources in 5G network slicing includes network, compute and storage capacity resources; virtualized network functions; shared physical resources; and radio resources. Service designers can select the optimal control/user plane split, as well as compose and allocate virtualized network functions at particular locations inside the core or radio access network depending on the service requirements. The creation and management of network slicing is a challenging process that poses new problems in service instantiation and orchestration, resource allocation/sharing and assignment procedures as well as network virtualization technologies.

Existing open source, industry and standards developments have given shape to the initial perception of a 5G network slice, while further research activities aim to enhance such new evolving concept by exploring its full potential. The so-called 5G network slice fully supports a particular communication service exploiting the principles of software-defined networks and network function virtualization in order to fulfill the business and regulatory requirements. The achieved networking and service flexibility enables a radical change, beyond network sharing, enabling different mobile operators to offer tailored services and means for network programmability to OTT providers and or vertical market players.

Original contributions are invited on the latest advancements on network slicing for 5G systems considering architecture, network management, orchestration and mechanisms that enable virtualization and multi-tenancy. The topics of interest within the scope of this issue include (but are not limited to) the following:

• Network slicing architectures and deployment practices
• Network slicing and multi-tenancy support in service overlay networks
• Network function (de)composition and allocation considering “atomic” functions
• QoE support management mechanisms in network slices
• Multi-service and multi-connectivity network slicing
• Next generation of orchestration architectures combining SDN and NFV
• Network resource programmability and developments on the Northbound-APIs
• Mobile edge computing and service optimization
• Network slicing and backhaul/fronthaul mechanisms
• Network slicing for converged fixed-wireless 5G networks

SUBMISSIONS

Articles should be tutorial in nature and written in a style comprehensible and accessible to readers outside the specialty of the article. Complete guidelines for prospective authors can be found at http://www.comsoc.org/commag/paper-submission-guidelines. The guest editors reserve the right to reject papers they unanimously deem to be either out of scope of this Feature Topic or otherwise extremely unlikely to be accepted after a peer review process.

It is important to note that IEEE Communications Magazine strongly limits mathematical content, and the number of figures and tables. Mathematical equations should not be used (in justified cases up to three simple equations are allowed). Article length (introduction through conclusions, excluding figures, tables and their captions) should not exceed 4500 words. Figures and tables should be limited to a combined total of six (6). The number of archival references is limited to fifteen (15). Non-archival references (website URLs, web-posted papers and reports, unpublished/to be published/pending papers) should not be included in the “References” section. All articles must be submitted through the IEEE Manuscript Central site (http://mc.manuscriptcentral.com/commag-ieee) to the “May 2017 / 5G Network Slicing” category by the submission deadline according to the following schedule:

IMPORTANT DATES

• Manuscript Submission Deadline: September 15, 2016
• Decision Notification: December 15, 2016
• Final Manuscript Due Date: February 15, 2017
• Publication Date: May 2017

GUEST EDITORS
Konstantinos Samdanis Steven Wright Albert Banchs Antonio Capone
NEC Europe, Germany AT&T Services, US UC3M, Spain Politecnico di Milano, Italy
samdanis@neclab.eu sw3588@att.com banchs@it.uc3m.es capone@elet.polimi.it

Mehmet Ulema Kazuaki Obana
Manhattan College, US NTT Docomo, Japan
mehmet.ulema@manhattan.edu kazuaki.obana.uz@nttdocomo.com
This is the 21st issue of the Network and Service Management series, which is typically published twice a year, in January and July. The Series provides articles on the latest developments in this well established discipline, highlighting recent research achievements and providing insight into both theoretical and practical issues related to the evolution of the discipline from different perspectives. The Series provides a forum for the publication of both academic and industrial research, addressing the state of the art, theory, and practice in network and service management.

A key annual event of the network and service management community was the Network Operations and Management Symposium (NOMS 2016) (http://noms2016.ieee-noms.org/), which took place April 25–29 in Istanbul, Turkey. During NOMS, the prestigious IFIP/IEEE Salah Aidarous was given to Aiko Pras (University of Twente), who was a previous Co-Editor of this series. Olivier Festor (Telecom Nancy) was re-elected as Chair of the IFIP Working Group 6.6, “Management of Networks and Distributed Systems.” The second major annual event of the network and service management community is the Conference on Network and Service Management (CNSM 2016) (http://www.cnsm-conf.org/2016/), which will take place in November in Montreal, Canada. The first key event in 2017 will be the International Symposium on Integrated Network Management (IM 2017), taking place in May 2017 in Lisbon, Portugal.

Using the Internet to provide educational material has become an important factor for many academic institutions. The European FLAMINGO project has started to produce online educational material, such as Wikipedia pages and YouTube videos. Several videos, ranging from online tutorials to quick introductory survey videos, have already been published on the FLAMINGO YouTube channel (https://www.youtube.com/user/lp7flamingo). You can also find interviews with people who have been involved in the creation of various Internet management technologies on the YouTube channel.

We again experienced excellent interest for the 21st issue, having received 16 submissions in total. For each paper we got at least three independent reviews. We finally selected three articles, resulting in an acceptance rate of 18.8 percent. It should be mentioned that the acceptance rate for all the previous issues has ranged between 14 and 25 percent, making this Series a highly competitive place to publish.

The first article, “Geographical Route Design of Physical Networks Using Earthquake Risk Information” by Nga Tran and Saito, investigates the challenges in designing a communication network robust against earthquake-induced disasters, finding appropriate geographical routes under a cost constraint in order to maximize the overall network robustness.

The second article, “Approaches to End-User Applications Portability in the Cloud: a Survey” by Yangui, Glitho, and Wette, presents and evaluates the approaches proposed so far in standardization bodies, research projects, and academia for enabling end-user application portability in cloud environments by using platform as a service (PaaS).

Finally, the third article “Network Slicing as a Service: Enabling Enterprises’ Own Software-Defined Cellular Networks” by Zhou, Li, Chen, and Zhang, presents an approach that introduces hierarchical network slicing as a service (NSaaS), enabling operators to offer a customized end-to-end cellular network as a service.

We hope that readers of this issue again find the articles informative, and we will endeavor to continue with similar issues in the future. We would finally like to thank all the authors who submitted articles to this Series and the reviewers for their valuable feedback and comments on the articles.

GEORGE PAVLOU (g.pavlou@ucl.ac.uk) is a professor of communication networks at the Department of Electronic Engineering, University College London, United Kingdom, where he coordinates networks and services research activities. His research interests focus on networking and network management, including traffic engineering, autonomic networking, information-centric networking, and software-defined networks. He has been instrumental in a number of research projects that have produced significant results with real-world uptake and has contributed to standardization activities in ISO, ITU-T, and the IETF.

JÜRGEN SCHÖNWÄLDER (j.schoenwaelder@jacobs-university.de) is a professor of computer science at Jacobs University Bremen, Germany. His research interests include network management and measurement, network security, embedded systems, and distributed data processing. He is an active member of the Internet Engineering Task Force, where he has edited more than 30 network management related specifications and standards. He has contributed in various roles to the organization of IEEE and IFIP sponsored academic conferences and journals.
Geographical Route Design of Physical Networks Using Earthquake Risk Information

Phuong Nga Tran and Hiroshi Saito

Abstract

In this article, we investigate the challenges in designing a communication network robust against earthquake-induced disasters. Due to the heterogeneity of local geology conditions, an earthquake may cause different devastating effects on network links at different locations. Therefore, the aim of this research was to develop a network design method that, based on actual seismic hazard information, finds appropriate geographical routes for network links under a cost constraint to maximize the robustness of the network. We carried out experiments on local as well as nationwide networks considering many possible earthquake scenarios in Japan. The results suggest that the shortest route is often the optimal route, and, in some cases, long detour routes can significantly reduce the average end-to-end disconnection probability compared to the shortest route solutions. This gives a network operator guidelines to design a network that is resilient against earthquake disasters at the lowest possible cost.

Introduction

Natural disasters in general and earthquakes in particular can damage large portions of communication networks. In March 2011, for example, a severe earthquake near Japan and its associated tsunami destroyed many things, including network facilities. According to a press release from NTT, “Facilities were damaged and commercial power supply was disrupted at exchange offices, among other things, impacting approximately 1.5 million circuits for fixed-line services, approximately 6,700 pieces of mobile base station equipment, approximately 15,000 circuits for corporate data communication services, and others.”

This made the network largely unavailable. Such severe damage and problems in a network cause secondary negative effects over a disaster area. In the above example, emergency calls could not be established due to the congestion caused by a huge number of communication attempts to families and friends in the disaster area.

Similar damage occurs every few years worldwide [1]. Therefore, it is important to improve network robustness against natural disasters [2, 3]. In particular, making society, including communication infrastructures, sustainable to earthquakes is an important role in Japan, due to the increasing possibility of a massive earthquake.

Earthquakes occur in various regions with diverse probabilities and magnitudes. When an earthquake occurs, it causes a shock with different intensities at different locations. Fundamentally, the intensity of the shock at a location $x$ is a function of the magnitude of earthquakes, the distance from the seismic center to $x$, and the geomorphological surface structure at $x$. A network link that goes through an earthquake-affected area will have a certain probability to fail if the earthquake intensity in the area is greater than 5 (on the Japanese scale of 7). Basically, the link failure probability increases with the increasing earthquake intensity and length of the link. However, it is also influenced by other geological conditions. For example, links going through regions with liquefaction effects have a significantly higher failure probability at the same earthquake intensity, or links going along a bridge are more fragile than others during the same earthquake.

Physical networks are networks of fiber cables, and are designed on the basis of estimated traffic demands, costs, provision of future development, and other factors. In physical networks, links are usually deployed along roads and railways. If we take into account the earthquake damage effect, the shortest geographical route may not be the best design in terms of robustness. The reason is that it may go through an area with high earthquake intensity, which causes a higher failure probability than a longer alternative route, which goes through a low earthquake intensity area; hence, it may have a lower failure probability. This is illustrated in Fig. 1. Obviously, if the green routes are chosen instead of the corresponding black ones for deploying network links, the links are likely to have a lower failure probability, which consequently enhances the network robustness.

Formerly, information on earthquakes was not available or not accurate enough. Consequently, network design methods either did not take into account the potential damage caused by earthquakes or were solely based on theoretical models. Nowadays, thanks to advances in seismology, earthquake forecasts are becoming more accurate, and the number of people affected by earthquakes is also decreasing. As a result, the robustness of communication networks has become a priority among network operators. Therefore, the aim of this research was to develop a network design method that, based on actual seismic hazard information, finds appropriate geographical routes for network links under a cost constraint to maximize the robustness of the network.

The authors are with NTT Network Technology Laboratories.
Example of a geographical network design robust against earthquakes. The earthquake intensity map is plotted on the basis of data from the earthquake “ANNKI/MAP/C-V2-ANN21-MAP-CASE1.csv” from J-SHIS (http://www.j-shis.bosai.go.jp/en/), which may affect the western part of Japan.

Dikbiyik et al. proposed a mechanism for re-provisioning connections under the risk of correlated cascading or sequential failures right after the initial impact of disaster, to quickly recover disrupted connections and minimize the risk of further failures [5]. Mukherjee suggested that if the time and location of a disaster can be forecast, some actions should be taken before the occurrence of disasters, such as reallocating network resources so that network elements in the predicted disaster area are used as little as possible [6].

Different from network recovery, which is reactive, robust network design is a proactive method of coping with with disasters. A straightforward approach is resource provisioning to enhance the survival probability during disasters. The work in [7] is a typical example. Another approach, which well complements the provisioning approach, is to plan physical networks to avoid disaster-prone areas. Cao et al. [8] discussed the optimization of the physical route of undersea cables. Given a geometrical route configuration such as a rectangular route, they minimize cable cost while maximizing the probability of connecting two cities that is higher than the threshold. Saito proposed a spatial network design method for determining the geographical shape of a physical network by optimizing a metric such as the probability that every route between two given nodes intersects a disaster area [9]. Unfortunately, these works are not directly applicable if we need to take into account streets and roads to install physical links, and if earthquakes that will occur with high probability are specified, as mentioned above. Hence, in this work, we address the robust physical network design problem on the basis of actual seismic hazard information provided by the Japan Seismic Hazard Information Station (J-SHIS) operated by NIED and geographical street maps to obtain a more practical and realistic solution. The proposed method in this article introduces a new step in disaster management in addition to existing approaches such as provisioning and recovering. It provides a method for geographically placing network cables to make them robust against earthquakes on the basis of quantitative risk evaluation. Here, the network should already contain provisioned alternative routes and spare resources.

\section*{Evaluation of Link Failure Probability}

Let $C$ be the geographical route of a network link. The probability that a link is broken at location $x$ depends on the earthquake intensity, the terrain conditions at $x$ such as a river (along a bridge crossing a river), the construction methods, such as aerial or underground installation, and the network component itself, such as the type of ducts/cables and years of use. NTT has collected data on the failure probability of network components under various earthquake intensities over years; hence, the failure probability of each type of network component under certain conditions including earthquake intensity can be estimated.

To calculate the failure probability $p(x)$ of the physical network design problem on the basis of the seismic hazard information.
The proposed method in this article introduces a new step in disaster management in addition to existing approaches such as provisioning and recovery. It provides a method for geographically placing network cables to make them robust against earthquakes on the basis of quantitative risk evaluation.

Local Pareto Optimal Solution

For each link \( e \in E \) with its set of possible geographical routes \( R(e) = \{ L(e) \} \), we define a local multi-objective optimization problem, where the objective is

\[
\min_{\ell(e) \in R(e), \beta_{\text{ref}}(L(e))} \left\{ l(L(e)), \beta_{\text{ref}}(L(e)) \right\}.
\]

(3)

\( \beta_{\text{ref}}(L(e)) \) is the failure probability of link \( e \) deployed along geographical route \( L(e) \) during earthquake \( a \).

The objective is to simultaneously decrease the length of \( e \) and its failure probability for each earthquake \( a \in A \).

Obviously, the shortest route is included in a Pareto solution set. The other solutions in the set are found on the basis of the set of possible routes \( R(e) \) and the seismic hazard information. In theory, there are many routes connecting two network nodes, which are often located in two different cities. However, in practice, it is not necessary to consider all routes because a small difference in the geographical routes does not change the link failure probability much. Specifically, the number of geographical routes connecting two different cities is often limited. Most of the route differences are in the city area due to there being a large number of streets in the city. However, a city is often relatively small compared to the affected area of an earthquake; therefore, the earthquake intensity in a city is almost homogeneous. In homogeneous earthquake intensity areas, the failure probability of each link segment has a fixed length, the number of segments per segment is smallest. Because link segments have a fixed length, the number of segments on the shortest route is the smallest; hence, a link deployed along the shortest route has the lowest link failure probability. Therefore, we can omit the small route difference inside a city and consider only the large streets or highways connecting two cities.

To find alternative routes to the shortest one, we specify three points that divide a straight line connecting two end nodes of a link into three equal segments. We then move these points perpendicularly along that straight line by distance \( d \). For each new point, we find the shortest route between the two end nodes of the link that goes through the specified point. By changing \( d \), we can obtain different routes. For each route, we compute the link failure probability and determine if it belongs to the Pareto solution set. To avoid unnecessary computations of routes, we evaluate the average earthquake intensity around the new points against the corresponding old points and compute a new route only if the new points fall into the region with a clearly lower average earthquake intensity. The distance \( d \) is increased until the corresponding alternative route reaches a region in which the failure probability is zero, the corresponding alternative route reaches the coast, or the length of the corresponding alternative route exceeds the given cost constraint.
GLOBAL OPTIMAL SOLUTION
Theoretically, if we can assume that link failure probabilities are independent, the global optimal solution will be constructed from the Pareto solution set of each local optimization problem, because the end-to-end disconnection probability is an increasing function of link failure probabilities. In practice, there may be some links that partly share the same street, which leads to dependent failures. However, we have demonstrated through mathematical analysis and practical examples that the global best solution constructed from the local Pareto solutions, which results in the lowest sum of end-to-end disconnection probabilities taking into account the dependent failure of the shared parts, is very close to the true optimal and also very close to the optimal solution under the assumption of independent failures (approximately 2 percent different on average) (details are shown in the supplementary materials at http://www9.plala.or.jp/hslab/supplement.html). Therefore, in practice, the objective function can be approximated to the one when independent failures are assumed. As a result, the optimal solution can be found from the local Pareto solutions.

To find the global optimal solution, we evaluate all combinations of the local Pareto solutions. This includes calculating the sum of the end-to-end disconnection probabilities of every node pair considering the dependent failure of the shared parts. This is the most time-consuming phase. However, we have implemented a program to generate this in advance as a function of link failure probabilities [10]. For each combination of Pareto solutions, we replace the corresponding link failure probabilities and recompute the end-to-end disconnection probabilities by using the given formulation. This significantly reduces the computation time.

COMPUTATIONAL COMPLEXITY
At first glance, the problem seems to be highly complex, and solving it may be very time-consuming because the amount of possible geographical routes can be very large, and a brute force algorithm is usually not efficient. However, the problem can be solved for a real network in a reasonable amount of time due to the following reasons:

• Our proposed method divides the problem into local and global optimizations. Global optimizations can be performed by using the combination of local Pareto solutions.

• As already mentioned above, even though the number of streets in an area is huge, the number of meaningful geographical route candidates connecting two network nodes is limited. Furthermore, network components cannot fail due to an earthquake with an intensity lower than 5 on the Japanese scale of 7. Hence, it is not necessary to consider all routes in areas with earthquake intensity lower than 5, but only the shortest one. The usefulness of geographical routes are therefore restricted within a bounded region.

• For many links, the Pareto solution set consists of just one solution, which is the shortest route solution. The reason is as follows. It is often the case that a network link lies completely inside an area where the earthquake intensity is almost homogeneous (the area affected by a huge earthquake is, e.g., about 10^4 km^2. Even a small earthquake reaches 100 km^2). Hence, a short detour route also lies completely inside such an area. Because the link failure probability in a homogeneous earthquake intensity area increases as the route length of the link increases, the detour route will have a higher failure probability than the shortest one. A longer detour route can have a part going through an area with a significantly lower earthquake intensity. To reach this region, however, a part of this route has to go through an area with an earthquake intensity as high as that in an area through which the shortest route runs. This part alone can already result in a larger link failure probability than the shortest route.

• Because the Pareto solution set contains only one element for many links, while it contains a few elements for other links, and typical networks have about 15–20 links, it is possible to use a brute force algorithm to find the global optimal solution.

• In practice, operators may want to configure only some links while fixing the others due to the existing fibers or other restrictions. In this case, we need to compute the Pareto set for reconfigurable links only, even though the information of the fixed links is still needed to calculate the end-to-end disconnection probabilities. The solution space is therefore relatively small, and hence, the optimal solution can easily be found with a brute force algorithm.

NUMERICAL EXAMPLES
BACKGROUND INFORMATION
In this article, the data in OpenStreetMap (https://www.openstreetmap.org) was used as $R(e)$ to find the geographical routes $L(e)$ for $e \in E$. In practice, if the new route of a fiber is restricted to an existing duct network, we use a duct map instead of a street map.

We assume that all links in an area of earthquake intensity I have the same link failure rate $\beta_I$ [1/km]; 0.05 for $I \in [5, 5.5)$ on the Japanese scale of 7, 0.1 for $I \in [5.5, 6.0)$, 0.12 for $I \geq 6.0$, and 0 otherwise. By using $\beta_I$, $p(x) = \beta_I |dx|$. On the basis of this field data information, we estimated the link failure probability as in Eq. 1.

We performed an experiment on two networks: a regional network of 12 nodes and 14 links with 5 different earthquakes, and a super-regional network of 8 nodes and 9 links with 5 earthquakes. These earthquakes are predicted to occur within the next 30 years with probability of more than 1 percent. One of these earthquakes is very large (strongly affecting the entire western part of Japan). All others are local, affecting areas of tens to a few hundreds of square kilometers, which has an impact on a small part (several links) of the network. Some earthquakes moderately affect the region in which the networks span, with an average intensity of around 5–5.5, while others may have a high average intensity of about 6.5–7 (on the Japanese scale).

In this article, we present the results obtained with the regional network. Results obtained with the artificial network are provided in the supplementary material (http://www9.plala.or.jp/hslab/).
supplement.html), in which we release the network (nodes and links) and earthquake information.

**LOCAL PARETO OPTIMAL SOLUTIONS**

A few links (3 out of the 14 links in the regional network and 3 out of 9 links in the super-regional network) have multiple elements in the Pareto solution set, while the others have only one element in the set, which is the shortest route. The links that have only one local Pareto solution usually lie in the regions that are not affected by local earthquakes, and the intensity of huge earthquakes in those regions is relatively homogeneous.

To better understand why a detour route can reduce the link failure probability, in Fig. 2, we show three elements in the Pareto solution set of a link in the regional network as an example. The local earthquakes affect only a small area, so it is easy to find a detour route that goes through a region less affected by the earthquakes.

Because earthquakes affect various locations differently, the optimal routes in terms of link failure probability strongly depend on the earthquakes. For local earthquake 2, the optimal route is longer than that of local earthquake 1. This is due to the difference of the seismic centers and magnitudes, which leads to the two earthquakes each having a different earthquake intensity distribution. For local earthquake 2, it is possible to find a large detour that makes a part of the network link lie in an area with an earthquake intensity lower than 5, which does not cause network components to fail. However, for local earthquake 1, the surrounding region has a similar earthquake intensity; therefore, the shorter route is the better one.

It is often true that the optimal non-shortest route in terms of failure probability depends on the earthquake. However, there are cases in which one geographical non-shortest route is optimal for all earthquakes. These cases can occur when several conditions are met. First, the shortest routes lie in an area with a weak terrain structure. Thus, the route cannot be optimal. Second, the earthquake intensity of this area is higher than the surrounding region for any earthquake. Consequently, using only a detour route to avoid this area can significantly reduce the link failure probability. Furthermore, a large surrounding area has an almost homogeneous earthquake intensity for all earthquakes. Therefore, it is not possible to improve the link failure probability with a larger detour route. As a result, only one non-shortest route is optimal for all earthquakes.

**PERFORMANCE EVALUATION**

To demonstrate the efficiency of the proposed network design method, we evaluated the end-to-end disconnection probabilities of the optimal network for each individual earthquake and for the combination of all considered earthquakes. In the case of all earthquakes, we assume they have equal weights ($w_a = w \forall a \in A$). The results for the regional network are shown in Fig. 3. The $x$-axis is the coefficient $\eta$ defining the maximum acceptable cost. The $y$-axis is the normalized end-to-end disconnection probability, which is the ratio of $\sum_{(k,m)\in E} P_{k,m}(k,m)$ of the optimal solution and that of the all-shortest-route solution. This can be understood as the gain in the average end-to-end disconnection probability compared to the all-shortest-route solution. In addition,
the optimal geographical network designs in different cases are shown in Fig. 4. The following conclusions can be drawn from these graphs.

The performance gain in terms of average end-to-end disconnection probability strongly depends on the earthquake. From Fig. 3, the gain for the huge earthquake was marginal (approximately 3 percent), while those of local earthquakes could be up to about 40 percent. The huge earthquake will strongly affect the entire network. Therefore, even though we could find a detour route that has a lower failure probability (mainly due to the difference in the terrain structure in the area), this reduction was relatively small, as shown in Table 1. Furthermore, other links have a quite high failure probability, so a little improvement on a few links does not reduce the average end-to-end disconnection probability much. In contrast, the local earthquakes have a devastating effect in just a small area. Therefore, it is possible to find an alternative route that significantly reduces the failure probability (Table 1). Moreover, local earthquakes often affect just a few links while leaving the others risk-free. Therefore, a decrease in the failure probability of just a few links can significantly decrease the average end-to-end disconnection probability of a network. This observation also holds for the super-regional network. Local earthquake 3, which affects just three links, is a typical example. Among these three links, we can find detour routes that notably reduce the failure probabilities of two links. Hence, the average end-to-end disconnection probability during this earthquake can be significantly decreased with new geographical routes.

If the maximum acceptable cost is increased, the gain also increases, but only to a certain extent. In this specific example, if the cost constraint $\eta$ goes beyond 1.15, we do not observe any further improvement. This is because an effective detour route is bounded in a certain region. If the earthquake intensity is lower than 5, the link cannot fail; therefore, a longer detour route does not further reduce the link failure probability. Moreover, it may be not possible to find a long detour route because the region is narrow and surrounded by the sea.

Some areas are more prone to earthquakes than others. As shown in Fig. 4, link 2 lies in an area severely affected by most of the considered earthquakes. Therefore, in practice, just a few links need to use detour routes to reduce the failure probability.

**Conclusion**

We have presented a new method for designing a robust network against earthquake-induced disasters on the basis of actual seismic hazard information and geographical street maps. The aim was to find the optimal geographical routes for network links that minimize the weighted sum of end-to-end disconnection probabilities over all earthquakes under a cost constraint. In many cases, the shortest geographical routes result in the most robust network, while at some links, detour routes can reduce their failure probabilities, hence improving the robustness of the network. Which link can be improved by the detour route and how much the improvement is highly depends on the earthquakes. Furthermore, the higher the available cost, the better the geographical routes that can be found because it allows longer routes. However, this improvement is bounded. At a certain point, increasing

![Figure 3. Performance gain for a regional network.](image)

![Figure 4. Geographical network design for different earthquakes: a) physical network based on the shortest geographical routes; b) optimal physical network for local earthquake 1; c) optimal physical network for the huge earthquake; d) optimal physical network considering all earthquakes with equal weights.](image)
the cost (meaning increasing the route length) does not reduce the link failure probability any further.

Our proposed method adds a new step to the provisioning phase for network survivability against earthquakes and can be executed in an incremental fashion. It can also be combined with other protection and restoration methods to enhance network robustness.

For future work, we want to carry out research towards disaster-free networks, which can cope with not only earthquakes but also other natural disasters such as tsunamis and hurricanes. The disaster-free network is a disaster management concept proposed by Saito [11]. It includes robust physical network design and network avoidance control against spatial disasters [12]. An idea for a new robust physical network design is to integrate our geographical route design method into the design of a network topology (from scratch or adding new links to existing networks) in order to obtain a more survivable network. This research topic is still under-explored and hence deserves more attention.

REFERENCES


Table 1. Optimal route in terms of link failure probability of two links for different earthquakes (length (km); failure probability) (LE: local earthquake; HE: huge earthquake).

<table>
<thead>
<tr>
<th>Link 1</th>
<th>Link 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Shortest route</td>
<td>Optimal route</td>
</tr>
<tr>
<td>LE1</td>
<td>(51; 73%)</td>
</tr>
<tr>
<td>LE2</td>
<td>(51; 99%)</td>
</tr>
<tr>
<td>LE3</td>
<td>(51; 23%)</td>
</tr>
<tr>
<td>LE4</td>
<td>(51; 98%)</td>
</tr>
<tr>
<td>HE</td>
<td>(51; 94%)</td>
</tr>
</tbody>
</table>


BIOGRAPHY

PHUONG NGA TRAN received a B.E. degree in electronics and telecommunications from Hanoi University of Technology in 2002, an M.Sc. in information and communication systems, and a Ph.D. degree in communication networks from Hamburg University of Technology in 2006 and 2010, respectively. From October 2010 to April 2015, she worked at the Institute of Communication Networks at Hamburg University of Technology as a postdoctoral researcher. She joined NTT in April 2015 as a research associate. Her research interests include network planning and future Internet architectures.

Hiroshi Saito [F] graduated from the University of Tokyo with a B.E. degree in mathematical engineering in 1981, an M.E. degree in control engineering in 1983, and a Dr.Eng. in teletraffic engineering in 1990. He joined NTT in 1983. He is currently an executive research engineer at NTT Network Technology Labs. He received the Young Engineer Award from the Institute of Electronics, Information and Communication Engineers (IEICE) in 1990, the Telecommunication Advancement Institute Award in 1995 and 2010, and the Excellent Paper Award from the Operations Research Society of Japan (ORSJ) in 1998. He has served as an Editor and a Guest Editor for technical journals such as Performance Evaluation, IEEE Journal of Selected Areas in Communications, and IEICE Transactions on Communications. He was the Director of Journals and Transactions of IEICE, the organizing Committee/Program Committee Chair of a few international conferences, and a Program Committee member of more than 30 international conferences. He is currently an Editorial Board member of Computer Networks. He is a Fellow of IEICE and ORSJ, and a member of IFIP WG 7.3. His research interests include traffic technologies of communication systems, network architecture, and ubiquitous systems. More information can be found at http://www9.plala.or.jp/hslab.
Abstract

End-user applications are provisioned in cloud settings using PaaS. Portability is key to the success of end-user application providers because it helps avoid vendor lock-in. It makes the components deployment of the same end-user application across multiple PaaS possible. It also enables them to move end-user applications from PaaS to PaaS with minimal adaptation efforts. This survey discusses and evaluates the approaches proposed so far within the standardization bodies, research projects, and academia for enabling end-user applications portability in PaaS. It also identifies the trajectory of research on the topic. An illustrative use case is described in order to identify evaluation requirements, and then the state of the art is reviewed in light of the requirements.

Introduction

Cloud computing is a paradigm for swiftly provisioning a shared pool of configurable resources (e.g., storage, network, application, and services) on demand. It has three key facets: software as a service (SaaS), platform as a service (PaaS), and infrastructure as a service (IaaS) [1]. Applications are offered to end users as SaaS. While IaaS ensures efficient use of the involved physical resources (e.g., memory, processing), PaaS is responsible for provisioning these applications.

This article takes an end-user perspective and focuses solely on end-user cloud applications. The reader should note that there are other applications in the cloud (e.g., IaaS applications) that are not considered in this article. The ultimate goal of cloud computing is to provision (i.e., develop, deploy, and manage) such applications in a rapid and cost-efficient manner.

A key to the success of end-user application providers in cloud settings is portability. Portability aids in preventing vendor lock-in [2]. End-user applications portability means the ability to move end-user applications from PaaS to PaaS [3]. It also means the possibility of provisioning end-user applications across several PaaS, or more precisely, the components deployment of the same end-user application across several PaaS [2].

The potential advantages are illustrated by the example of a travel agency’s end-user booking application designed as a component-based application, here a portal for hotel booking, car rental, and payroll, as shown in Fig. 1. If the end-user application is portable, its provider should be able to deploy each of its components in any target PaaS with minimal integration or adaptation efforts, as shown by Action 1 in the figure. The choice of the target PaaS could be based on criteria such as security and cost.

Some of the components (e.g., the payroll module) could have already been developed and made available in marketplaces. These components could be directly instantiated and deployed in the target PaaS, as shown by Action 2 in the figure, to avoid costly development. The already deployed components could be moved to new PaaS as shown in Fig. 1, Action 3. This could, for instance, be triggered by a change in the price charged by the currently hosting PaaS.

Several approaches that enable cloud end-user applications portability have been proposed in the recent past. This article critically evaluates them and discusses the research directions. It is important to note that another survey has been recently published on a related topic [4]. A key difference is that it deals with applications portability at large, while this article focuses solely on the end-user application portability tackled in more depth. Moreover, unlike [4], this article proposes evaluation requirements and discusses research directions.

The second section details our identified evaluation requirements. The following section critically reviews the specifications proposed by the standards bodies in the light of the evaluation requirements. This is followed by a critical review of the work done outside the standards bodies in the light of the very same requirements. The fifth section is devoted to the research directions. The last section concludes the article.

Requirements for Cloud End-User Applications Portability

In this section, the concept of an end-user application’s life cycle is first introduced. Then, for each phase of the life cycle, the related require-
Cloud End-User Applications’ Life Cycle

Cloud end-user applications provisioning consists of three phases: development, deployment, and management [5]. The development phase consists of developing the end-user application, testing it, and building the application executables. Application executables include all the files needed to execute the application once deployed (e.g., source code, configuration files). The deployment phase consists of:

- Allocating and making ready the PaaS resources needed to host and execute the end-user application
- Uploading its executable over these resources
- Performing needed management operations

Requirement for the Development Phase

The first requirement is the need for standards for end-user application executables. These executables are exported during the development phase. They should be built according to a universal format supported by all involved PaaS clouds and instantiated from remote marketplaces. In the travel agency’s booking application, for instance, the developer should be able to build an executable for all the components (developed or instantiated from marketplaces) and then deploy it as it is on PaaS A and/or PaaS B.

Most existing PaaS support regular programming language formats (e.g., WAR and JAR for Java, .py and .pyc for Python) as executables. However, some of them impose specific structures related to the peculiarities of their runtimes; for example, a WAR built on GAE cannot be executed on another PaaS supporting Java since GAE uses a specific Java runtime environment. Furthermore, some PaaS (e.g., OpenTOSCA) do not even use the regular programming language formats.

Requirement for the Deployment Phase

The allocation of the required PaaS resources during the deployment phase as well as the upload of the executable over these resources are performed with well defined deployment setting descriptors (e.g., deployment manifests for Cloud Foundry) and operations, such as deployment application programming interfaces (APIs) (e.g., PushApp operation for Cloud Foundry). The second requirement is the need to design standards for these descriptors and operations. In fact, each PaaS imposes its proprietary descriptors and deployment operations to developers [7]. This leads to having a specific deployment procedure for each PaaS.

Back to the travel agency’s booking application: to deploy the portal component, for instance, which is implemented as a Java Web application, specific descriptors based on different meta-models and formats (i.e., YAML for Cloud Foundry and json for Jelastic) are needed for each target PaaS. Moreover, the deployment operations differ. Indeed, the deployment of this component in Cloud Foundry consists only of providing the associated executable. The allocation of the hosting Apache tomcat server is implicitly performed by the PaaS based on the type of executables (i.e., WAR archives). However, deploying the same component in Jelastic requires the manual creation of the hosting environment containing a tomcat instance, the upload of the component executables, and the linking of the component to the created environment for the deployment.

Requirement for the Management Phase

In cloud settings, the management operations can be classified into:

- Classical operations, that is, fault, configuration, accounting, performance, and security (FCAPS) operations
- Additional operations such as migration, for example, moving the car rental module from one provider to another

The provided FCAPS operations allow maintenance of the end-user applications and preservation of its performance. The third requirement is the need for standards for the whole set of management operations covering both classical management operations (FCAPS) and the additional operations as well.

Management features are exposed by PaaS through dedicated operations varying from PaaS to PaaS. Furthermore, even when a given management operation is offered by two PaaS, its specification may differ; for example, application scalability settings in GAE are automated and based on a well defined service level agreements (SLA) while still manual in Cloud Foundry.

Work Done within the Standardization Bodies

This section examines the OASIS TOSCA and CAMP specifications. They address cloud application portability at large, including end-user application portability. An overview and a critical evaluation are provided in each case.
Figure 2: TOSCA service templates associated with the travel agency’s booking application.

**Topology and Orchestration Specification for Cloud Applications**

**Overview:** The Topology and Orchestration Specification for Cloud Applications (TOSCA) provides specifications to describe cloud resources and applications topologies as typed graphs to enable portability. TOSCA expresses cloud end-user applications as well as their required hosting environment topologies. The main motivations are simplifying applications management and related components reusability [8].

Cloud end-user applications and associated platform resources that make up the application-hosting environment could be schematized as a set of nodes with well-defined relationships. The nodes and relationships are described in topology templates schematized as a typed graph. Such templates provide the structure of the end-user application as well as its hosting environment in the PaaS.

In addition to the topology template, the service template contains a plan element that describes the process models used to deploy the end-user application and needed information for its management once it is deployed. An appropriate cloud service archive (CSAR) has to be used to package nodes and relationships elements representing the end-user application. The CSAR allows encapsulation of all application TOSCA elements and its plans.

Back to the travel agency’s booking application, the associated TOSCA service template is schematized in Fig. 2. The application’s components are represented as nodes where the bindings between them are represented as relationships. The service templates representing the required PaaS resources, which constitute the application hosting environment (i.e., environment1 and environment2), are also schematized in the same figure. The application’s components, which handle persistent data and thus require a database service (i.e., hotel booking, car rental, and payroll), run on environment1, while the web portal, which requires a message router to route requests to the appropriate back-end component, runs on environment2. In addition, some plans describing the management rules for the service template can be provided; for example, Plan1 describes the application deployment process, and Plan2 describes its scalability strategies.

**Evaluation:** The use of CSAR archives allows the requirement concerning the standardized applications’ executables format to be met. Such archives facilitate applications’ content encapsulation, and they are programming-language-independent. However, it should be noted that most of the existing PaaS do not support current CSAR deployment except for some research prototypes (e.g., Service Offering and Provisioning Platform — SIOPP, OpenTOSCA).

The introduction of the service template and plan concepts allows the requirement concerning standardized deployment descriptors and operations to be met. The .tosca document, part of the CSAR, describes the topology and properties of the service templates related to the application. The plans, which can be implemented as workflows (e.g., BPEL-based, BPMN-based processes), allow the description of deployment information such as hosting resources requirements, configuration steps, and orchestration processes.

The management operations (e.g., scaling an application’s component) have to be described and implemented by the application developer through the plans. This implies that the hosting PaaS supports such management operations, which is not always true. Thus, TOSCA does not meet the requirement concerning the standards for the whole set of management operations. Generally, most of the existing PaaS do not support TOSCA specifications. In order to support TOSCA, the existing PaaS will need to map the service topology concept to the available concrete resources so that they can handle TOSCA-based end-user applications and adapt the management plans accordingly.

**Cloud Application Management for Platforms**

**Overview:** Cloud Application Management for Platforms (CAMP) proposes specifications that enable homogeneous cloud resources deployment and management procedures in heterogeneous PaaS [9]. The CAMP specification is designed for cloud end-user applications and their hosting PaaS. The main entity is the component. A component describes an end-user application component or a PaaS resource. End-user applications’ components have requirements, while PaaS components have capabilities. An end-user application component can be related to a PaaS component if the latter has the required capabilities associated with the requirements of the application component. The component entity is part of an assembly. An assembly describes the whole end-user application and the platform (PaaS) resources it may require once deployed. An assembly is described with a Plan that may refer to the needed PaaS service(s) (e.g., load balancing and storage). An assembly can be included in another assembly to constitute an assemblies entity for reuse purposes. These are encapsulated in specific packages called platform deployment packages (PDPs). The PDP archive consists of end-user application content files (e.g., sources code, scripts, etc.) and plans.

Going back to the travel agency’s booking application, Fig. 3 schematizes its description with CAMP specification. The various applica-

---

*SIOPP is pronounced “shop.”*
Each application component uses one or a set of platform components. Each platform component uses associated PaaS service(s). The deployment and management operations of these services are described with dedicated plans.

**Evaluation:** CAMP meets the requirement concerning standardized applications’ executables formats by introducing the PDP format. However, it should be noted that, as was the case with CSAR for TOSCA, PDP requires adaptation on the PaaS side to support this format. PDP archives are only supported by some PaaS prototypes such as Brooklyn.

CAMP meets the requirement concerning standardized deployment descriptors and operations. Indeed, CAMP introduced vocabulary to describe capacities and requirements of the assemblies and a meta-model defining the structure of the plans (e.g., containers, databases). However, it should be noted that the defined vocabulary is neither rich nor flexible enough to cover the heterogeneity of all existing PaaS resources (e.g., various container distributions, various execution frameworks). It only considers the common concepts and resources of the PaaS providers involved in the project. It should also be noted that there is no information provided on how one can extend this vocabulary.

CAMP does not meet the requirement concerning the standards for the whole set of management operations. The management operations exposed through generic REST-based API do not include FCAPS operations. Generally, the technology vendors behind CAMP (e.g., CloudBees, Cloudsoft Corporation, Huawei, Rackspace, Red Hat) are based on the same common deployment and management operations for end-user applications. However, they only provide operations supported by their relative PaaS and necessarily with the same semantics.

**Work Done within Research Projects and Academia**

The research carried out in projects and academia is important, and their results could be used as input for standards. The relevant work is reviewed as below.

**Work Done within Research Projects**

This section focuses on Cloud4SOA and mOSA-IC, the two projects that have produced the most significant results in this subject area.

**Cloud4SOA EU FP7 Project:** Cloud4SOA provides an architecture that allows to deploy and manage end-user applications in heterogeneous PaaS environments [10]. However, the constraint is that the end-user application should be developed according to service-orient-
Cloud4SOA provides an architecture that allows to deploy and manage end-user applications in heterogeneous PaaS environments. However, the constraint is that the end-user application should be developed according to Service-Oriented Architecture (SOA) specifications and all its components should be modelled using SOA.

However, aggregation of FCAPS operations is advantageous, and the payroll component, which requires continuous integration, will be placed on CloudBees since it supports provisioning of a Jenkins integration server. Then the SLA manager in the governance layer will monitor the running components and send notifications to the Application Migration module to migrate the components from one PaaS to another if needed in order to re-optimize the placement.

Evaluation: Cloud4SOA does not meet the requirement concerning standardized deployment descriptors and operations. The descriptors allow of using the same annotations to describe the applications’ requirements in terms of PaaS resources. The annotations to describe the same container distribution whatever the PaaS is. These annotations are based on semantic ontologies that cover end-user applications’ requirements and their corresponding hosting environments on PaaS.

On the other hand, Cloud4SOA offers common deployment operations through its harmonized API. They allow the allocation of PaaS resources and the deployment of applications in a common way. The exposed operations (e.g., createApp, deployApp, startEnvironment) are aggregated from the integrated PaaS offerings through the adaptors. The adaptors ensure the mapping between the harmonized API and the proprietary PaaS offering APIs.

Cloud4SOA does not meet the requirement concerning the standards for the whole set of management operations. The harmonized API also aggregates management operations following the same principle for deployment. In addition to the aggregated operations from the PaaS offerings, Cloud4SOA implements additional management operations such as moving deployed applications from one PaaS to another. However, aggregation of FCAPS operations is missing.

mOSAIC EU FP7 project: mOSAIC is an open source project that aims to enable data, services, and applications portability and interoperability across multiple clouds [11]. The applications covered by mOSAIC include end-user applications. mOSAIC is based on a brokering mechanism that search for cloud services matching the applications’ requests. The mOSAIC
The mOSAIC framework consists of several layers (SaaS, PaaS, and IaaS) and offers a set of APIs in each one of these layers.

Overview: The mOSAIC framework consists of three layers, as shown in Fig. 5.

The application support layer includes the API implementations and application tools, as well as the Semantic Engine and Service Discoverer modules. The API exposes operations to build and manage cloudlets. In the mOSAIC architecture, a cloudlet is defined as an abstraction of application functionality and is subject to elasticity and monitoring by the software platform layer. Based on this, cloud end-user applications can be represented as Cloudlets. The Semantic Engine is the subsystem supporting the user in selecting the API components and functionalities needed to build cloudlets. The API operations are generic. This ensures a degree of abstraction of the software platform support layer when the end-user application and its resources are described.

The software platform support layer handles the execution and management of the cloudlets related to applications and their hosting environment. This is ensured based on well-defined descriptors. There are two types of descriptors in this layer: application descriptors and deployment descriptors. The former describe the application topology and components, while the latter list the needed cloud resources for each component.

The infrastructure support layer provides concrete cloud resources and services to be provisioned. In this layer, the cloud agency performs required cloud service provisioning for the cloudlets. The selection of the target IaaS is based on brokerage contracts. The cloud adaptors implement the specific IaaS operations to perform the provisioning. Currently, mOSAIC provides adaptors for Amazon EC2, Flexiscale, Eucalyptus, and OpenNebula.

Back to the travel agency’s booking example, the Application Tools module in the application support layer will allow the building of applications with its several components. Each component will be packaged as a cloudlet. Then the API implementations module allows the description of their relative requirements for execution (e.g., required cloud services). Semantic matchmaking between the cloudlets’ requirements and the available cloud services capabilities is performed by the Semantic Engine. For example, the hotel booking cloudlet requires a storage service, and the portal cloudlet requires a routing service. The hosting and execution of the cloudlets are performed in the software platform support layer. This is performed by mapping the cloudlets with the provisioned cloud services. One implementation of this layer is ModaClouds. The harmonized API allows negotiation and provision of the optimal concrete services, for example, the Amazon storage service for the hotel’s booking component and the message router from an OpenNebula instance.

Evaluation: mOSAIC meets our requirement regarding standardized applications executables by introducing the cloudlet concept. The cloudlets are programming-language-independent. In addition, they are composed, instantiated, and managed in a generic fashion from the resources provided by the underlying integrated IaaS.

mOSAIC meets our second requirement concerning standardized deployment descriptors and operations. It allows developers to describe application components requirements and dependencies in terms of communication and data in a unified way through the API implementations module. A generic descriptor will be generated accordingly. The descriptor content is based on a common ontology implemented in the Semantic Engine.

mOSAIC does not meet the requirement concerning the standards for the whole set of management operations. The operations offered by the software platform support layer are limited to application governance. Furthermore, the cloud adaptors used as middleware between mOSAIC and cloud vendors only support deployment operations.

WORK DONE WITHIN ACADEMIA

Two significant academic works focused on cloud end-user applications portability are discussed below.

In [12], the authors use existing concepts from proposed template-based approaches such as TOSCA to propose generic end-user applications deployment and management procedures in the cloud. The proposed work uses resource templates representing reconfigurable entities that can be reused for different end-user applications. The templates as well as the entities are described with descriptors. These descriptors are used to automate and homogenize the deployment procedures using deployment tools (i.e., DevOps technologies like Chef or Puppet). This allows the second requirement to be met. However, it should be noted that the approach impos-
Table 1. Evaluated work synthesis.

<table>
<thead>
<tr>
<th></th>
<th>Standardized application executables</th>
<th>Standardized deployment descriptors and operations</th>
<th>Standardized comprehensive management operations</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOSCA</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>CAMP</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Cloud4SOA</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>mOSAIC</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Loutas et al.</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Satzger et al.</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

Table 1 provides a summary of the critical review provided in this article. The identified research directions are discussed in this section.

**Research Directions**

Cloud end-user applications’ executables’ formats are various and heterogeneous. This may depend on several parameters such as the programming language for building the applications (e.g., Java-based, python) or the specificities of the target cloud runtimes; for example, for Java web applications, the Google App Engine repackages the standard WAR archives in a specific way since it uses a specific Java runtime environment.

Future research could look into the introduction of novel end-user application formats that are programming-language-independent. The Open Virtualization Format* (OVF) may be a good starting point since it is designed for packaging and distributing portable software to be run in virtual machines. OVF is a packaging standard initiated by the Distributed Management Task Force (DMTF) for virtual appliances deployment and management. For example, the Contrail project uses the Open Virtualization Format (OVF) to federate applications between its integrated cloud since such a format is not tied to any particular hypervisor or processor architecture [14]. However, the appropriate standard format for cloud end-user applications should be more refined in order to cover such applications specificities and requirements at the PaaS level (e.g., application communication protocols, required libraries).

**Standardized Deployment Descriptors and Operations**

Most of the reviewed work introduces ontologies for the deployment descriptors. These ontologies allow describing end-user applications’ requirements and PaaS capabilities using the same semantics. However, these ontologies are not powerful enough to cover the huge number of supported resources and their characteristics. Indeed, platform resources can be various and heterogeneous (e.g., service containers, DBMS, monitoring services). Furthermore, platform resources are constantly evolving over time by the providers or even by end users when PaaS are open source (e.g., adding PHP support for Cloud Foundry by the developers community in 2011). This poses a problem with the way in which one can design ontologies flexible enough to easily be enriched by novel concepts in order to cover the full panoply of existing resources. A possible solution could be the use of hierarchical ontology work to describe and support the strong heterogeneity of PaaS resources as done in the previous work with sensors in the wireless sensor network ecosystem [15].

The introduction of hierarchical ontology will also facilitate the aggregation of deployment operations in order to have common deployment procedures. Indeed, this will make possible the mapping of the various proprietary PaaS deployment operations with generic interfaces implementing the proposed common model. Furthermore, hierarchical ontology allows the abstraction of the generic operations that are not supported by the target PaaS (e.g., CreateEnvironment in Cloud Foundry).

**Comprehensive Set of Management Operations**

Aggregating and mapping the APIs’ operations of PaaS management is not always feasible due to the fact that management capabilities vary from one PaaS to another (e.g., supported resources, supported operations on these resources). Furthermore, the operations semantics are not always the same between the two sides. In other terms, the level of abstraction provided by the generic operations might not correspond to the proprietary operations abstraction level. An eventual research direction is the design of multi-level APIs, which are configurable to offer several levels of abstractions, depending on the target PaaS to integrate into the ecosystem.
CONCLUSIONS

Portability is key to the ubiquitous deployment of cloud end-user applications. This article has introduced an illustrative use case and derived requirements for enabling portability in such a context. These requirements have been used to critically review the approaches proposed so far in standards bodies, international projects, and academia. None of these approaches, however, fully meet all the requirements. Research directions for meeting each requirement have also been discussed. We hope that the identified research directions can guide the community in designing solutions that will allow effective application portability.

ACKNOWLEDGMENT

This work is supported in part by Ericsson and the National Science and Engineering Research Council (NSERC) of Canada.

REFERENCES


BIographies

Sami Yangui (http://users.encs.concordia.ca/~s.yangui/) is a postdoctoral research fellow with the Concordia Institute for Information Systems Engineering (CISE) at Concordia University, Montreal, Canada. He is a member of the Telecommunication Service Engineering (TSE) Research Laboratory (http://users.encs.concordia.ca/~tse). He holds a PhD (2014) in computer science from Telecom SudParis and the University of Evry Val d’Essonne, France, and an M.Sc. (2010) in computer science from the University of Tunis-ELManar, Tunisia.

Roch Gutho (http://users.encs.concordia.ca/~rgutho/) holds a Ph.D. (Techn. Dr.) in tele-informatics (Royal Institute of Technology, Stockholm, Sweden) and M.Sc. degrees in business economics (University of Grenoble, France), pure mathematics, and computer science (University of Geneva). He works as an associate professor and Canada Research Chair at CISS, where he leads the Telecommunication Service Engineering (TSE) Research Laboratory (http://users.encs.concordia.ca/~tse). Previously, he worked in industry for almost a quarter of a century and held several senior technical positions at LM Ericsson in Sweden and Canada (e.g., expert, principal engineer, senior specialist). He has also served as a IEEE Communications Society Distinguished Lecturer, Editor-In-Chief of IEEE Communications Magazine, and Editor-In-Chief of IEEE Communications Surveys & Tutorials.

Constant Wette is a senior system developer at the Cloud & IP Business Unit, Ericsson Canada. He holds an M.Sc. in computer engineering from Ecole Polytechnique de Montréal, an M.Eng. in electronics and communication from Ecole Polytechnique de Montréal, and an M.B.A. in management from HEC Montreal. He joined Ericsson in 2000 and has been involved in research, product development, innovation projects, and new business development in different areas including cloud architectures, data management and analytics, IoT/M2M, IMS, and social media.
Network Slicing as a Service: Enabling Enterprises’ Own Software-Defined Cellular Networks

Xuan Zhou, Rongpeng Li, Tao Chen, and Honggang Zhang

ABSTRACT

With the blossoming of network functions virtualization and software-defined networks, networks are becoming more and more agile with features like resilience, programmability, and open interfaces, which help operators to launch a network or service with more flexibility and shorter time to market. Recently, the concept of network slicing has been proposed to facilitate the building of a dedicated and customized logical network with virtualized resources. In this article, we introduce the concept of hierarchical NSaaS, helping operators to offer customized end-to-end cellular networks as a service. Moreover, the service orchestration and service level agreement mapping for quality assurance are introduced to illustrate the architecture of service management across different levels of service models. Finally, we illustrate the process of network slicing as a service within operators by typical examples. With network slicing as a service, we believe that the supporting system will transform itself to a production system by merging the operation and business domains, and enabling operators to build network slices for vertical industries more agilely.

INTRODUCTION

Although the data traffic of mobile terminals is increasing rapidly, the consumer market of mobile broadband services is going to be saturated in North America, Europe, and East Asia [1]. Meanwhile, the growing popularity of machine type communication (MTC) terminals and applications of vertical enterprises poses an increasing demand for diverse services from mobile networks. However, legacy mobile networks are mostly designed to provide services for mobile broadband consumers, and merely consist of a few adjustable parameters like priority and quality of service (QoS) for dedicated services. Therefore, mobile operators find it difficult to get deeply into these emerging vertical services with different service requirements for network design and development. For example, the dedicated network for a railway company just involves the coverage along the railway with high-speed mobility management, but exhibits apparent difference from that of an electricity metering company, which only requires small-volume data transmission but massive connections at static positions. Some vehicle communication services are strictly delay-sensitive, while some video surveillance services require stable and immobile high bandwidth. Recently, network functions virtualization (NFV) technology has been proposed to decouple the software and hardware of network elements so as to simplify service development. A study by the European Telecommunications Standards Institute (ETSI) shows that NFV and software-defined networking (SDN) could shorten time to market and facilitate innovations in the technical field (e.g., saving maintenance cost, auto-scaling, enhancing system resilience) [2]. Nevertheless, currently the products and service types from operators are still limited. In order to enrich operators’ products for vertical enterprises and provide service customization for emerging massive connections, as well as to give more control to enterprises and mobile virtual network operators (MVNOs), the concept of network slicing (NS) is proposed to allow the independent usage of a part of network resources by a group of mobile terminals with special requirements [3, 4].

NS aims to logically separate the set of network functions and resources within one network entity, according to specific technical or commercial demands. Although the concept of NS is still nascent, similar techniques already exist. Among them, IEEE 802.1Q, virtual local area networks (VLANs), which can be regarded as the ancestor of NS, provides a single broadcast domain to bring together a group of hosts possibly having no local and physical connectivity but sharing common interests. Moreover, in the field of fixed networks, Internet Engineering Task Force (IETF) RFC 4026, which is also known as virtual private network (VPN), is another form of NS, which could guarantee the QoS and security requirements for logically independent sessions [5]. However, in cellular networks, the realization of NS faces significant challenges, since more parameters such as mobility and authentication
management in the control plane as well as session and charging management in the user plane need to be customized for a group of connections as a logical network. Fortunately, NFV and SDN can make NS a reality, with NS allowing operators to customize networks according to various requirements of mobile services, thus leading to a more cost-effective way to build dedicated networks. Therefore, NS is attracting significant interest from both industry and academia. For example, the Fifth Generation Infrastructure Public Private Partnership (5G-PPP) project 5GEx introduces a business model of NS among infrastructure owners, wholesale providers, retail providers, content providers, and end users [6]. The Third Generation Partnership Project (3GPP) also initiated a technical study into NS to specify service requirements and operational requirements [7]. Vendors such as Ericsson, Huawei, Nokia, and ZTE have also published white papers about NS to introduce the realization of NS into 5G [8]. In fixed networks, NS has been implemented to logically separate the networks, allowing slice owners to manage their own networks [9]. Another case is for emergency communications, which provides dedicated and priority resources users for emergency communications, even in overwhelming scenarios [10]. However, due to the scattered service models across radio access networks (RANs), core networks (CNs), and transport networks, and complex protocols in tens of 3GPP interfaces, the realization of mobile NS seriously lags behind its counterpart in fixed networks. Specifically, there is still lacking an end-to-end service description of the mobile network for the northbound interface to deploy or manage a multi-vendor network slice across the domains with thousands of parameters. The study and standardization of NS are still at a rudimentary level, and give little insight into the mapping between service providers and consumers, the common demands of users, such as high-bandwidth slice and low-latency slice. The Third Generation Partnership Project (3GPP) also initiated a technical study into NS to specify service requirements and operation; Wanfu Ding et al. [12] presented the design of an open platform for service chain as a service, by using the tangible capabilities of SDN and NFV. In this article, we discuss how operators agilely provide a customized network slice for their customers as a service, which is called NSaaS. According to the relationships between service providers and consumers, the business models of NSaaS can be categorized into three classes as below.

**Business to Business (B2B):** Operators sell the network slice to a company who owns both the network and terminals, such as video surveillance networks for security companies and smart factory networks for manufacturing companies. In the B2B case, operators not only provide customized wireless connections to enterprises, but also release full control of terminals to the enterprise.

**Business to Consumer (B2C):** End consumers are able to purchase customized data pipes from operators for their terminals like smart home devices. In the B2C case, end customers could enjoy the slice once they put subscriber identification module (SIM) cards inside their devices. Generally, customers just use the customized network, but do not possess the network with service separation.

**Business to Business to Consumer (B2B2C):** The operator plays the role of wholesale provider; meanwhile, a broker like an MVNO helps operators to be engaged with end customers. In this case, operators just provide dedicated connections, called MVNO as a service, to the broker, without involving the business part. However, the broker could get more control of the network than traditional MVNOs, who could only get billing files from mobile network operators (MNOs).

From another perspective, there are three service scenarios of NSaaS, which have different life cycles, service objects, and slice scales:

- **Industrial slice:** Customers with the same network service requirements are registered with the same slice, which abstracts the common demands of users, such as high-bandwidth slice and low-latency slice.
- **Monopolized slice:** Anyone (usually an enterprise) who pays for the slice monopolizes and uses it as a private network.
- **Event slice:** A slice is launched for some events with relatively short life cycles, such as sports events, concerts, and even sales promotions inside shopping malls.

NSaaS also demonstrates some advantages to operators. First, confronted with prosperous over-the-top (OTT) services, operators could only provide less competitive OTT-like services and traditional services such as voice, SMS, and data. NSaaS makes a difference by facilitating operators to differentiate their data pipes with various QoS and providing additional promising services. Second, based on NSaaS, the design and configuration experience becomes a simple software reconstitution procedure and shortens the time to market of operators’ products from months to hours. Assuming that infrastructure of network elements has already been virtualized and could be allocated as a simple reconstitution procedure of virtual machines, a
component-based network could be described by configuration files as a service template and orchestrated by combining software packages from a library. Third, NSaaS enriches the products of operators so that operators could agilely offer dedicated network services to small and medium enterprises rather than build expensive dedicated networks case by case only for large enterprises. Moreover, software-based NSaaS facilitates the convergence of operators’ operation support systems (OSSs) and business support systems (BSSs).

### Architectures of Network Slicing
There are several implementations of NS, as illustrated in Fig. 1: CN only, RAN only, and CN and RAN.

**CN Only:** Tenant-level CNs are virtualized as network slices, with component-like functionalities to be programmable and auto-configurable, such as mobility management, session management, and authentication. The network slices only exist in CNs; therefore, neither the RAN nor the user equipments (UEs) need to be specially configured for the sliced CNs. All the interfaces and procedures remain unchanged except the case when the UEs initially attach to the networks, because the UEs should be assigned to the correct slice of the CNs. Here we propose to add a slice selection function (SSF) intervening at the interface between the control plane of the CN and RAN to notify the right slice to activate the bearers with the UEs. Another function of the SSF is to send flow tables to SDN switches to manage connections between the base stations (BSs) and the network slices, due to different coverage requirements of the slices.

**RAN only:** Different from CN only slicing, RAN slices run on the radio hardware and baseband resource pool, called a wireless platform, which exhibit less elasticity than the mature virtualized infrastructure in CNs. With several logical BSs, the slices of a RAN apply various parameters of the air interfaces (e.g., symbol length, sub-carrier spacing, cycle prefix length, and the parameters of hybrid automatic repeat request [HARQ]) to implement slices with different service models. Furthermore, other parameters like cell selection and handover threshold, as well as coordinated transmission policies can be defined for each slice in order to provide a featured wireless experience to the UEs.

**CN and RAN:** In this scenario, each slice of a RAN is connected to a core slice, so operators could offer an end-to-end logical network to clients. The slice selection procedure is the same as that of RAN only, so UEs do not need to select the slice of a CN once they have access to the RAN part. The CN and RAN solution brings the advantages of both the CN only and RAN only solutions, being able to program the functionalities of the CN, as well as customize the air interfaces of the RAN.
FROM NFV TO NSaaS

The NFV technology contains general-purpose processor platform, cloud operating system, hypervisor, distributed computing, and the software of network elements, decouples software and hardware, and shields the hardware details for virtual network functions (VNFs). Based on NFV, NS realizes the service separation for multi-tenancy so as to virtually build an exclusive network for each tenancy. However, NSaaS is a more business-oriented concept than a technological one, with features of mapping service demands automatically from a customer to functionalities, topology, policies, and parameters of a network slice, as well as providing component-based and auto-configured network functionalities for operators to design and launch network services more conveniently. Table 1 lists the abstracted comparison of NFV, network slicing, and NSaaS.

AN EXAMPLE OF NSaaS

Here we take an emergency communication slice as an example to further clarify NSaaS. Usually, an emergency communication slice offers two main functions in an emergency: alert broadcast and distress call. This slice is usually provided by a government to inhabitants for free with the B2B2C business model. Both the CN only and CN and RAN implementations are suitable for the emergency service, and provide dedicated communication resources of top priority when others are congested. The emergency slice could be available once launched without any hardware integration. Moreover, we can load new functions such as push-to-talk on demand just like installing a new software.

SERVICE MODEL AND ORCHESTRATION

In the previous section, we point out that the key technology of NSaaS is service mapping, which translates service requirements into service models of operators and vendors. In this section, in order to better match network slices to various vertical applications, we propose to differentiate necessary service models of mobile network into three levels: application level, network function level, and infrastructure level. Figure 2 illustrates the summarized content of the service models mapped to the NFV architecture of ETSI, as well as the descriptor databases.

APPLICATION LEVEL

At this level, we describe the traffic characteristics of the applications. In view of a single UE, the application requirements could be described by the metrics including arrival rate, average packet length, flow type (burst, periodic, and persistent streaming), download/upload ratio, and so on. Moreover, the application level also contains some additional services, such as location-based service, firewall, and service chain with third-party applications. The RAN part of network slices provides options about the wireless experience of an application, like the mobility of terminals, cell selection preference, and power-saving air interface. The application level service model should be easy to understand even for application developers without any telecommunication background. Therefore, this could be standard-ized as the application descriptors. In view of the entire network slice, another important description of the application requirements is the SLA, which defines some service-specific requirements including capacity, coverage area, QoS requirements, failure duration, network issues, denial of service, and scheduled maintenance, and so on.

The service level agreement (SLA) descriptions are usually included in the business contract, while the traffic characteristics with more technical details represent the operator’s understanding of a vertical industry. Therefore, the traffic characteristics of this level come from a vertical industrial library built by operators, while the customized SLA requirements as well as additional services must be translated from customers’ orders.

NETWORK FUNCTION LEVEL

The network function level shows how VNFs are interconnected and configured with non-vendor-specific descriptions. As we know, the topology of a network describes the connections among the sites of a RAN and network elements of a CN. In terms of the SLA of applications, all these RAN and CN nodes are associated by VLANs of control plane, data plane, and management plane. The other part of the service model at the network function level is the parameters defined by standards, like timers within 3GPP and IETF protocols. For example, the tracking area update timer is set to different values for immobile smart metering devices and high mobility vehicles, and

Table 1. The comparison of NFV, network slicing and NSaaS.

<table>
<thead>
<tr>
<th>Form</th>
<th>NFV</th>
<th>Network slicing</th>
<th>NSaaS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Managed object</td>
<td>Virtual machines</td>
<td>Virtual networks</td>
<td>Customized services</td>
</tr>
<tr>
<td>Value to operators</td>
<td>Better resource utilization</td>
<td>Tenancy separation</td>
<td>Agile product development</td>
</tr>
<tr>
<td>Value to consumers</td>
<td>None</td>
<td>Monopolized network</td>
<td>Customized service</td>
</tr>
</tbody>
</table>

Figure 2. The proposed service models of network slicing.
The control plane requires short latency to access the databases, while the data plane requires high throughput of forwarding modules, and the wireless part needs specific signal processing acceleration. Once a network slice starts to be instantiated, we need to find the suitable infrastructure as well as racks in data centers to bear it.

<table>
<thead>
<tr>
<th>functionalities</th>
<th>Metering</th>
<th>Video surveillance</th>
<th>Automobile</th>
<th>Emergency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mobility management</td>
<td>Static, no handover</td>
<td>Static, no handover</td>
<td>High-speed</td>
<td>Low-speed</td>
</tr>
<tr>
<td>Session management</td>
<td>Light, no user plane</td>
<td>Standard</td>
<td>Multi-session</td>
<td>Broadcast</td>
</tr>
<tr>
<td>Access protocol</td>
<td>3GPP S1-C</td>
<td>3GPP S1 standard</td>
<td>3GPP S1 standard</td>
<td>3GPP MBMS</td>
</tr>
<tr>
<td>QoS policy</td>
<td>Bandwidth limitation</td>
<td>Bandwidth guarantee</td>
<td>Latency guarantee</td>
<td>Top priority</td>
</tr>
<tr>
<td>Security</td>
<td>N/A</td>
<td>Standard</td>
<td>Standard</td>
<td>N/A</td>
</tr>
<tr>
<td>Air interface feature</td>
<td>Power saving</td>
<td>Carrier aggregation</td>
<td>Small TTI</td>
<td>N/A</td>
</tr>
<tr>
<td>Band</td>
<td>800 MHz</td>
<td>6 GHz</td>
<td>900 MHz</td>
<td>450 MHz</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>2 MHz</td>
<td>100 MHz</td>
<td>5 MHz</td>
<td>2 MHz</td>
</tr>
<tr>
<td>HARQ parameters</td>
<td>( N + 14 ), no retransmission</td>
<td>( N + 10 ), no retransmission</td>
<td>( N + 3 ) + ( N + 3 )</td>
<td></td>
</tr>
<tr>
<td>Topology</td>
<td>Centralized</td>
<td>Distributed gateway</td>
<td>Distributed gateway</td>
<td>Centralized</td>
</tr>
<tr>
<td>Auto-scaling</td>
<td>Enabled</td>
<td>Disabled</td>
<td>Enabled</td>
<td>Enabled</td>
</tr>
<tr>
<td>Additional service</td>
<td>Location-based service</td>
<td>Video compression</td>
<td>Push-to-talk</td>
<td>Distress call</td>
</tr>
</tbody>
</table>

Table 2. The functionalities and configurations of NSaaS for typical applications.

the radio resource control inactivity timer also differs between bursty instant messaging service and persistent video streaming.

**Infrastructure Level**

The infrastructure level of CN is maintained by IT engineers of operators who are responsible for ensuring all the virtual machines working properly to satisfy the demands of VNFs. Here, the wireless infrastructure consists of spectrum resource, antennas, BS sites, radio unit, and baseband resource pool, which is plotted as a wireless platform in Fig. 1. This level helps operators to define the resources of infrastructure with parameters like spectrums, CPU cores, memory, and storage. The control plane requires short latency to access the databases, while the data plane requires high throughput of forwarding modules, and the wireless part needs specific signal processing acceleration. Once a network slice starts to be instantiated, we need to find the suitable infrastructure as well as racks in data centers to bear it.

**Service Orchestration**

After describing the service models in three levels in terms of the requirements of the applications, we still need a service orchestrator to bridge the descriptions with an operational system with billing, monitoring, and vendor selection modules. According to the requirements of a customer, the orchestrator instantiates the network slice through assembling functionalities of vendors, such as VNF and BS selection, service chaining, subscriber management, as well as monitoring and billing. One of the most important features of NSaaS is the programmability of the network slice with the component-based network functions. Therefore, operators and slice customers could select different functions from vendors according to their own demands. Table 2 lists the examples of the selected functionalities for four typical applications, so that we can program a network slice conveniently according to the metrics.
to subscribe to element management systems (EMS) of vendors, AT&T proposes to standardize the Vc-Vnfm-vnf interface to control rich real-time data [14].

**Application Level:** According to general application services running on various VNFs of several vendors, there should be some VNF/component selection rules and mapping methods. In other words, operators should translate the requirements of an application into a language that can be understood by the orchestration interfaces of vendors, as studied in projects like Gohan from NTT [15] and OpenMANO from Telefonica. In this regard, the complicated standardization work of network slicing could be shifted to the development of a unified description language and thus interpreting by vendors.

**Mapping of Function, SLA, and Vendor**

In Fig. 2, the service and SLA of the application level is described in the language of slice consumers. However, in order to offer an eligible network slicing service with the right SLA, we have to map the service and SLA into network function level and infrastructure level, which could be executable for operators and vendors. As we know, a network slice could consist of multiple vendors with standard interfaces, although their supported functions and SLAs are different. Figure 3 illustrates how to map the service and SLA on the top level into lower levels, and how to find the matched functions from different vendors. In this figure, the service and SLA are mapped vertically, and the vendor is mapped horizontally. The application level service descriptions are mapped into both network function level and infrastructure level, and some of the application SLAs are mapped as low-level SLAs, while some of them are mapped as functions. All the components developed by vendors have to register their capabilities and functions with a VNF catalog of operators so that they can be selected according to application requirements. For example, an additional service named malicious website protection could be decomposed into VNFs like traffic detection, malicious database, firewall, and web redirection. It is worth mentioning that the inputs of the infrastructure-level SLA description come from the other two levels, as well as the components of vendors, because vendors have to propose the specifications of infrastructure for all the individual components.

**NSaaS Management**

In this section, the operation and process of NSaaS are discussed so that operators could provision the NSaaS with shorter time to market.

**Automated Configuration**

In previous sections, the service requirements have been translated to software packages by vendors’ plug-ins so that VNF managers (VNFMfs) are able to instantiate a network slice according to the order specification. However, there is still some remaining configuration work to do if the goal is to start the NSaaS automatically after obtaining orders from customers. The configurations of a newly instantiated network slice include:

- Infrastructure information: It contains the IP address pool of the control plane, data plane, and management plane of data centers, and the IP addresses and VLANs of BS lists.

- Service information: It describes basic entity identifiers and protocol interfaces, such as public land mobile network (PLMN) code, tracking area code, cell ID, domain name, network element name, access point name (APN), home subscriber server (HSS) address, as well as interface configurations of S1, S11, S6a, and so on.

- Subscription information: It contains the relationship between subscribers and network slices, while one subscriber may belong to several slices and different service chains.

- Slice registration: It connects the newly instantiated network slice to running cellular networks so that UEs can be redirected or assigned to it by SSF.

- Monitoring and billing interfaces: They tell where the KPI data and billing files should be sent.

All the configuration items above are also non-vendor-specific. Vendors are able to generate their scripts by exploiting the plug-ins of the network slicing orchestrator. Different from the
configuration of service models, all the parameters here are related to a real mobile network rather than application requirements. Hence, the most challenging part of automated configuration is to maintain the runtime environment information of the real mobile network, which requires operators to keep updating the environment information to reflect different levels of changes within their network happen (e.g., allocating an IP address or when a BS is offline).

**PRODUCT MANAGEMENT**

As a service or product for consumers and enterprises, the network slicing solution requires full life cycle management, spanning from design, release, order, and operation to disposal. Figure 4 describes the first four steps of product management of NSaaS. First, operators design a network slice for a general vertical enterprise according to the description of the service model, which could be named as the vertical industrial template. However, this general service model is not ready for instantiation, because it still lacks input information such as SLA from buyers. In order to sell the service online, operators need to price the slice service in terms of the subscriber number, QoS, SLA, additional service, and spectrum resources. Second, product managers from the marketing department of operators prepare the introduction and case studies of the slice service to help consumers to understand its value. After some internal review process, the slice service could be released onto the market shelf of operators. Third, customers order the slice service and input requirements to the vertical slice according to the description of the service model. Fourth, the network slice service is deployed and running, while both the operator and customer are able to monitor the status of the connections and service. Finally, if the slice service is not suitable to sell in the marketplace, operators would dispose it and end its life cycle.

**MANAGEMENT APIs FOR CUSTOMERS**

Besides the product management and maintenance of operators, there is another kind of interface for slice customers, management APIs of NSaaS, which could be integrated with third-party systems or platforms. Slice customers could take advantage of this kind of management APIs to add or remove service and connections, as well as monitor the status of slices. For an MVNO retailing connections to its own clients, management APIs also provide an advanced charging policy and service package for each UE. Therefore, management APIs help operators to find new channels as a new type of broker to distribute their service, which is usually difficult for them to touch in traditional architectures. The broker orders a customized network slice just as an enterprise customer does, but with more capabilities like the integration of its own service platform with the slice (e.g., billing file interface and subscriber database) and user behavior data acquisition from OSS as well as components like traffic detection.

**CONCLUSIONS**

In this article, we introduce the concept of NSaaS to help operators to offer customized end-to-end cellular network as a service. Moreover, the service orchestration and SLA mapping for quality assurance are explained to illustrate the architecture of service management across three model levels. Finally, we illustrate the detailed process of NSaaS within operators by typical examples, including the configuration and product management of NSaaS and management APIs for customers.

With the growing maturity of NFV/SDN, we believe that the supporting system will transform itself as a production system by providing a one-stop solution for future wireless connections and services. Particularly in the 5G era, with the merging of the operation domain and business domain, operators will build more and more customized network slices for vertical industries in an agile way.
This article was supported by the National Basic Research Program of China (973Green, No. 2012CB316000) and the Program for the Zhejiang Leading Team of Science and Technology Innovation under Grant 2013TD20.

REFERENCES
CoAP Congestion Control for the Internet of Things

August Betzler, Carla Gomez, Ilker Demirkol, and Josep Paradells

ABSTRACT

CoAP is a lightweight RESTful application layer protocol devised for the IoT. Operating on top of UDP, CoAP must handle congestion control by itself. The core CoAP specification defines a basic congestion control mechanism, but it is not capable of adapting to network conditions. However, IoT scenarios exhibit significant resource constraints, which pose new challenges on the design of congestion control mechanisms. In this article we present CoCoA, an advanced congestion control mechanism for CoAP being standardized by the Internet Engineering Task Force CoRE working group. CoCoA introduces a novel round-trip time estimation technique, together with a variable backoff factor and aging mechanisms in order to provide dynamic and controlled retransmission timeout adaptation suitable for the peculiarities of IoT communications.

INTRODUCTION

The quantity and diversity of devices that are interconnected in the Internet of Things (IoT) are constantly increasing, leading to a large variety of new and appealing application scenarios. More than 25 billion things are expected to be connected over the Internet by the end of 2020.1 A significant pillar for this development is the Constrained Application Protocol (CoAP), a lightweight RESTful protocol recently standardized by the Internet Engineering Task Force (IETF). CoAP is designed as the main application-layer protocol to be used by IoT devices for IP-based HTTP-like interactions [1]. Typically, constrained devices, such as low-power wireless sensor nodes, are used for IoT communications. These devices offer very limited processing and memory capacities. Furthermore, the communication technologies used by these devices exhibit significant limitations such as low data rates and relatively high bit error rates (BER). CoAP is tailored to these extreme resource constraints.

One of the main problems to be handled when designing a new end-to-end communication paradigm is network congestion. This phenomenon occurs when the traffic load offered to a network approaches the network capacity. In many traditional Internet applications, TCP provides end-to-end congestion control. However, CoAP operates over UDP to enable lightweight applications and must handle congestion by itself.

In IoT communications, the traffic patterns are different from the ones in conventional networks. Constrained devices often communicate periodically to notify their sensor measurements. Even when individual devices create small amounts of data, the large number of communicating devices can cause network congestion. Another possible reason for congestion is traffic bursts generated in reaction to events; for example, a large number of notifications sent after a sensor network equipped with accelerometers detects a seismic event. These IoT traffic patterns, together with severe node and link constraints, pose challenges for the design of a congestion control mechanism for CoAP, which should be capable of ensuring safe network operation while using network resources efficiently.

The core CoAP specification offers a simple congestion control mechanism, based on a retransmission timeout (RTO) with binary exponential backoff (BEB), which is, however, insensitive to network conditions. Therefore, default CoAP congestion control may significantly underperform, often being too conservative or too aggressive instead of adapting its behavior on the basis of network status information actually available to CoAP.

In this article we present the CoAP Simple Congestion Control/Advanced (CoCoA) mechanism, defined in a draft specification2 being standardized by the IETF CoRE working group to improve the CoAP congestion control. CoCoA combines the use of round-trip time (RTT) measurements, dynamic RTO backoff calculations, and RTO aging mechanisms to obtain dynamic RTO estimations for the transmission of CoAP messages. CoCoA has been designed to deliver a congestion control that is adaptive to network dynamics and suitable for IoT characteristics. CoCoA has reached matu-
CoAP congestion control is made in the CoCoA draft specification. Results show that, in contrast to the alternative methods considered, CoCoA consistently outperforms default CoAP congestion control in all evaluated scenarios.

### CoCoA Congestion Control

The base CoAP specification provides congestion control by imposing conservative restrictions on the rate of outgoing messages and the number of allowed parallel message exchanges. CoAP defines four types of messages: confirmable (CON), non-confirmable (NON), reset (RST), and acknowledgment (ACK) messages. The restrictions are applied to CON and NON messages, the first one being the limitation of outstanding interactions per destination to 1. An outstanding interaction can be a CON or NON request for which no ACK or reply has been received yet, respectively.

CoAP congestion control is insensitive to network conditions. In fact, it does not adapt the RTO on the basis of RTT information that is actually available to CoAP. Therefore, if the RTO chosen by CoAP congestion control is below the actual RTT, CoAP will incur spurious retransmissions. On the other hand, CoAP is likely to be used in networks with losses due to BER, which can lead to unnecessarily long idle times if the RTO timer overestimates the RTT.

Advanced congestion control mechanisms for CoAP should resolve the aforementioned issues, while ensuring safe behavior in the Internet. The proposal being standardized by the IETF CoRE working group for such advanced congestion control is made in the CoCoA draft specification.

### CoCoA

CoCoA provides a flexible congestion control solution that relaxes the conservative message rate restrictions of the CoAP base specification, while guaranteeing safe protocol operation. A fundamental requirement for the design of CoCoA has been to produce a mechanism that offers performance that is better than, or at least similar to, that of default CoAP. CoCoA comprises three main components: adaptive RTO calculation, a variable backoff factor (VBF), and RTO aging.

#### Adaptive RTO Calculation

In CoCoA, RTT measurement and adapted RTO calculation follow the principles of IETF RFC 6298. This RFC constitutes the basis for RTO computation in most TCP implementations, where the RTO is calculated adaptively by applying an exponentially weighted moving average of RTT and RTT-variation estimates. CoCoA adapts this algorithm for IoT communications.

In TCP, packet loss is assumed to be caused by network congestion. However, in IoT networks, a high packet loss rate is expected due to BER. The RTO estimator detailed in RFC 6298 only uses strong RTTs, that is, RTT measurements from the packets for which an ACK is received before the sender runs into retransmissions. This estimator is referred to as a strong RTO estimator. In CoCoA, a weak RTO estimator is also defined, which uses weak RTTs, that is, RTT measurements taken from packets that have required at most two retransmissions. This increases the chances of obtaining RTT measurements in the presence of packet losses. In CoCoA, when a weak or strong RTO is measured, the corresponding weak or strong RTO (RTO\(_w\)) is updated, respectively, following the same scheme defined in RFC 6298, as

\[
\text{RTO}_{x} = \text{SRTT}_{x} + K_{x} \times \text{RTTVAR}_{x},
\]

where \(x\) is either weak or strong, SRTT and RTTVAR denote the well-known smoothed RTT and RTT-variation estimates. CoCoA adapts this algorithm for IoT communications. The parameter values chosen for RTO\(_w\) are 0.25 for the weak RTO value.

To avoid a steep RTO increase after measuring a weak RTT and to maintain the overall RTO estimation stability, modifications were applied to the weak RTO estimator when compared to the strong RTO estimator:

- Weak RTT measurements are only allowed for up to the second retransmission in order to avoid very large weak RTT measurements (which could overestimate the RTO) and because the probability of obtaining veridical RTT information decreases with every retransmission.
- The value of \(K\) that determines the impact of RTTVAR on the weak RTO estimator is changed from 4 to 1. This reduces the impact of RTTVAR on the weak RTO estimation, since RTTVAR tends to grow large, especially if more than one retransmission is used.
- When calculating the overall RTO, the weak RTO estimator contributes less than the strong RTO estimator by using a reduced weight (0.25) for the weak RTO value. Although considering weak RTT information is necessary, strong RTTs provide more reliable input on the expected RTTs, and deliver a more accurate RTO estimation.

The parameter values chosen for RTO\(_w\) are:

\[
\text{RTO}_{\text{overall}} = \alpha \times \text{RTO}_{x} + (1 - \alpha) \times \text{RTO}_{\text{overall}},
\]

where \(\alpha\) is 0.5 for the strong RTO estimator and 0.25 for the weak RTO estimator.
Since RTO estimation requires the presence of round trip type interaction, CoCoA mandates the use of a controlled fraction of CON messages among the NON messages to be transmitted. For the sake of brevity, in this article we only focus on congestion control for CON messages.

calculations were shown to reduce the fluctuations without compromising the stability [2]. As in default CoAP, CoCoA dithers the initial RTO of a transaction by choosing it from the interval \([\text{RTO}_{\text{overall}} \times 1.5 \times \text{RTO}_{\text{overall}}]\).

**VARIABLE BACKOFF FACTOR**

For small initial RTOs, a BEB may not increase the RTO fast enough to allow the network to recover from congestion, still offering high load to the network and increasing the chance for spurious retransmissions. On the contrary, for large initial RTOs, a BEB may overestimate the RTO, leading to an unnecessary delay increase.

To address these problems, CoCoA applies a VBF that adjusts the backoff factor depending on the initial RTO value of a transmission. If the initial RTO is very small (below 1 s), a larger backoff factor is applied to retransmissions (VBF = 3). If a transaction initiates with a large RTO value (above 3 s), a smaller backoff factor is chosen for retransmissions (VBF = 1.5). For transactions that initiate with an RTO between 1 and 3 s, the VBF is set to 2, corresponding to a BEB.

Several backoff factor values for the VBF have been considered and evaluated [2]. Based on the evaluation results, and in consensus with the CoCoA specification authors and the IETF CoRE working group, the set of backoff factors presented above was chosen.

**RTO AGING**

If estimated RTO values are not updated for an extended period of time, the probability that they are no longer valid becomes high. In IoT networks, network conditions, and thus the RTO, can change fast. To avoid bogus RTO values due to such changes, CoCoA applies an aging mechanism to small and large RTO estimations. If an RTO estimation is small or large (below 1 s or above 3 s, respectively), and no new RTO measurement is made for 16 or 4 times the current RTO, respectively, the RTO value is modified to approach the default initial value.

**CONGESTION CONTROL FOR NON MESSAGES**

CoAP NON messages do not trigger ACKs from the receiver, and therefore are used if end-to-end reliability is not required. Default CoAP does not limit the rate of outgoing NON messages toward a destination endpoint. CoCoA introduces congestion control for NON messages, limiting the rate of outgoing NON messages toward a destination endpoint to one message every RTO s. Since RTO estimation requires the presence of round-trip type interaction, CoCoA mandates the use of a controlled fraction of CON messages among the NON messages to be transmitted. For the sake of brevity, in this article we only focus on congestion control for CON messages. The interested reader is referred to [3] for evaluations of NON-type message traffic.

**ALTERNATIVE CONGESTION CONTROL MECHANISMS**

For a wider understanding of congestion control for the IoT, besides comparing default CoAP with CoCoA, in this article we also analyze other RTO calculation algorithms such as the Linux TCP RTO (Linux-RTO) estimator [4], the peak-hopper TCP RTO estimator (PH-RTO) [5] and also a CoCoA variant that only uses the strong RTO estimator (CoCoA-S).

Linux-RTO adds two mechanisms to the basic TCP RTO algorithm. First, when a new RTT measurement is smaller than the previously gathered RTT information, the RTO is not increased, avoiding peaks in the RTO value when the channel seems to improve. Second, Linux-RTO avoids the RTO estimator to converge into an RTT value after repeatedly measuring constant RTT values [4], which could lead to spurious retransmissions.

PH-RTO reacts to a sudden RTT increase with an RTO increase by using short-term RTT history, which then decays over time toward the value of long-term RTT history. PH-RTO intends to avoid spurious retransmissions by using the long-term history when the channel suffers from sudden delays. RTO dithering is not defined for the Linux and PH-RTO algorithms, which were not designed for IoT scenarios.

Above these two state-of-the-art algorithms used in TCP, we include a minimalist Basic-RTO (B-RTO) estimator in the evaluations as a benchmark, which always sets the initial RTO for a transmission to a random value between 1 and 1.5 times the previously measured RTT. B-RTO can use weak RTT measurements. An overview of the features of all six analyzed congestion control mechanisms is given in Table 1. Neither of the two considered TCP-oriented congestion control mechanisms takes into account the peculiarities of IoT traffic, such as high BER,

<table>
<thead>
<tr>
<th>Per Client</th>
<th>Main Goal</th>
<th>RAM usage</th>
<th>Backoff method</th>
<th>Dithering</th>
<th>Strong RTTs</th>
<th>Weak RTTs</th>
<th>RTO Aging</th>
<th>Use backed-off RTO after no RTT update</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default CoAP</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>BEB</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>2 bytes</td>
</tr>
<tr>
<td>CoCoA</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>VBF</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>29 bytes</td>
</tr>
<tr>
<td>CoCoA-S</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>VBF</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>19 bytes</td>
</tr>
<tr>
<td>Basic RTO</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>BEB</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>2 bytes</td>
</tr>
<tr>
<td>Linux RTO</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>BEB</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>21 bytes</td>
</tr>
<tr>
<td>PH-RTO</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>BEB</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>43 bytes</td>
</tr>
</tbody>
</table>

Table 1. Overview of the features of the different congestion control mechanisms.
and from packet drops due to full buffers in the es in this setup mostly emerge from lossy links, different route lengths and RDC [9]. Packet loss-
erable packet loss rate and RTT variance due to congestion control mechanisms, such as a consid-
Lab imposes additional challenges for the Erbium (Er) CoAP implementation [9].

server motes in FlockLab run the full IPv6-based Cf is used as the CoAP implementation, while the 
save energy in real deployments. On the client side, 
FC is used as the CoAP implementation, while the 
motes run ContikiMAC [8] with radio 
TelosB motes [7], is chosen for this setup. The 
motes are routed toward a PC running a CoAP server. 
When compared to a wired connection, much larger RTTs and much higher RTT jitter are observed over the GPRS link, as well as a higher chance of packet losses. In our testbed we observe uplink/downlink data rates of approximately 15/40 kb/s.

In this setup, both the CoAP clients and server run the Java Californium (Cf) CoAP [6] implementation. The alternative congestion control mechanisms considered in this article are implemented and publicly available in Cf.5 For a fair comparative evaluation, while the PH-RTO and Linux algorithms were designed for TCP, we have implemented these algorithms for CoAP (over UDP). Other TCP features are not present in our evaluation.

In the second setup, the interaction between a cloud service and a multihop low-power wireless network is analyzed. CoAP clients running in the cloud service are connected via Ethernet to an IEEE 802.15.4 testbed where all motes run CoAP servers. FlockLab, a publicly available IEEE 802.15.4 indoor/outdoor testbed composed of 30 TelosB motes [7], is chosen for this setup. The FlockLab motes run ContikiMAC [8] with radio duty cycling (RDC) enabled, which is required to save energy in real deployments. On the client side, Cf is used as the CoAP implementation, while the server motes in FlockLab run the full IPv6-based ContikiOS stack for constrained devices, including the Erbium (Er) CoAP implementation [9].

When compared to the GPRS setup, FlockLab imposes additional challenges for the congestion control mechanisms, such as a considerable packet loss rate and RTT variance due to different route lengths and RDC [9]. Packet losses in this setup mostly emerge from lossy links, and from packet drops due to full buffers in the border router and relay nodes close to it. The border router is the FlockLab node that provides Internet connectivity to the Flocklab motes.

**Traffic Scenarios**

For both testbeds (GPRS and FlockLab), two traffic scenarios are defined to explore the effect of different congestion control mechanisms on the performance of CoAP communications.

**Continuous Traffic:** In this scenario, CoAP clients send CON requests to a CoAP server. When a client receives a reply from the server, the client immediately sends another CON request. Sending messages back-to-back by many clients simultaneously can cause congestion. The number of clients is varied from 10 to 40 (in steps of 10) in order to achieve different degrees of congestion. In the GPRS setup, one server is running on the destination device. In FlockLab, one client is assigned to each CoAP server mote in the testbed. A continuous traffic test lasts 180 s.

**Burst Traffic:** This scenario starts with a low congestion level, where 10 clients (GPRS) or 5 clients (FlockLab) generate continuous traffic of back-to-back CON requests. Then a burst of traffic is generated by a new group of clients that send 50 (GPRS) or 25 (FlockLab) back-to-back CON requests to the servers. Such traffic patterns can correspond to a local event (alerts about presence, temperature, etc.). The burst of messages causes a congestion peak. For the GPRS setup, we vary the number of clients that generate burst traffic. In FlockLab, for each mote that is not a destination of continuous traffic, we create a client that generates burst traffic. Tests are repeated 15 times for each specific configuration.

**Congestion Control Evaluation Results**

**Performance Metrics**

In the continuous traffic scenario, the overall throughput as successfully finished transactions per second is chosen as the performance metric, merging delay and packet delivery ratio into a single value.

In the burst traffic scenario, we analyze the settling time of the different congestion control approaches. We define the settling time as the time it takes for the clients to finish at least 50 percent of the burst traffic transactions. This is an important metric, since traffic bursts are expected when CoAP transactions are event-based or transmissions from various senders are synchronized.

Furthermore, the congestion control mechanisms’ behavior is analyzed regarding their fairness in FlockLab, which is challenging given the different path lengths and the tree-like topology of the scenario. Jain’s Fairness Index (FI)\(^\text{1}\) is used as the fairness metric, which ranges between 0 and 1, and a higher FI indicates a higher fairness level.

**Throughput Results**

Nearly all RTT-sensitive mechanisms outperform default CoAP independent of the network setup in terms of throughput in the continuous traffic scenario (Fig. 1).

In the GPRS setup, since the packet loss rate is low, the performance mainly depends on how

---

\(^\text{1}\) https://github.com/eclipse/californium

\(^\text{2}\) http://www.rfc-base.org/rfc-5166.html
the RTO algorithms adapt to the RTT. In this setup, RTT increases with the amount of active clients due to the delay introduced by the queuing of packets in the GPRS modem, which occurs since the overall generated data rate exceeds the uplink capacity. With the average RTT increase, the RTO-sensitive RTO algorithms increase their initial RTO values (Table 2). In FlockLab, the average RTT is small, but the average initial RTO for RTO-sensitive algorithms is larger than in the GPRS setup. This can be ascribed to a greater amount of weak RTTs that increase the RTO value and backed-off RTO values due to packet drops as a consequence of overflowing buffers near the border router, where traffic mostly concentrates.

Default CoAP underperforms independent of the setup (except in the GPRS setup with B-RTO) since it uses a fixed range of initial RTO values and does not adapt to the current RTT. If the real RTT is noticeably below the default RTO range, CoAP reacts slowly to losses. If the RTT lies in the RTO range or even exceeds it, spurious retransmissions are likely to happen, as indicated by the increasing percentage of retransmissions with the number of clients in GPRS (Fig. 2).

CoCoA achieves the highest throughput in the GPRS setup. In the same setup, CoCoA-S does not perform as well as CoCoA since it only allows strong RTT measurements, generally resulting in slightly lower RTO values and thus increasing the probability of spurious retransmissions (Fig. 2). In FlockLab, CoCoA and CoCoA-S perform very similarly. Their features allow them to benefit especially from links with good connections and small RTTs, but they also adapt the RTO and avoid bogus values even in a lossy network: the VBF and the aging mechanisms effectively limit the growth of RTO values when retransmissions are necessary. Without these features, CoCoA would tend to calculate very large RTO values for retransmissions using the BEB, whereas the overall RTO would not be shifted toward the default value of 2 s in the absence of further RTO updates.

The throughput obtained with B-RTO suffers noticeably due to its simplicity. If, after measuring a small RTT, the following transaction RTT is larger, which is likely given the RTT fluctuations in both network setups, the RTO timer will fire ahead of time with a high probability. In fact, B-RTO exhibits the highest retransmit ratio of all tested algorithms in all settings (Fig. 2). On the other hand, when a large RTT is measured, the next RTO used by B-RTO can grow very large due to the random multiplier, potentially leading to low throughput.

While Linux and PH-RTO perform better than default CoAP, they are not able to outperform CoCoA. Linux often calculates smaller RTO values, causing a higher amount of spurious retransmissions (Fig. 2), since contrary to CoCoA it does not increase the RTO when the RTT decreases. The PH-RTO reacts to a sudden RTT increase with a peak in the RTO that then slowly decays in the following transactions. However, given the continuous RTT jittering that is characteristic for both network setups, a sudden RTO increase may not be necessary and can lead to larger idle times if subsequent packets are lost. A disadvantage of both Linux and PH-RTO is their limitation to using only strong RTTs, while weak RTT measurements could provide

![Figure 1. Average throughput with 95 percent confidence intervals achieved by the evaluated congestion control mechanisms in the GPRS and FlockLab setups.](image)

Table 2. Comparison of the average RTT and initial RTO values in milliseconds for different numbers of clients in the GPRS setup and the FlockLab setup.

<table>
<thead>
<tr>
<th></th>
<th>10 GPRS clients</th>
<th>20 GPRS clients</th>
<th>30 GPRS clients</th>
<th>40 GPRS clients</th>
<th>FlockLab</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RTT</td>
<td>RTO</td>
<td>RTT</td>
<td>RTO</td>
<td>RTT</td>
</tr>
<tr>
<td>CoCoA</td>
<td>661</td>
<td>1505</td>
<td>1457</td>
<td>3739</td>
<td>1936</td>
</tr>
<tr>
<td>CoCoA-S</td>
<td>625</td>
<td>1428</td>
<td>1275</td>
<td>2905</td>
<td>1880</td>
</tr>
<tr>
<td>B-RTO</td>
<td>1025</td>
<td>1152</td>
<td>1962</td>
<td>2198</td>
<td>2983</td>
</tr>
<tr>
<td>Linux RTO</td>
<td>682</td>
<td>1325</td>
<td>1550</td>
<td>2801</td>
<td>1863</td>
</tr>
<tr>
<td>PH-RTO</td>
<td>746</td>
<td>1797</td>
<td>1835</td>
<td>3703</td>
<td>1827</td>
</tr>
</tbody>
</table>
additional RTO estimator updates. Instead, the old RTO is maintained and backed off (i.e., large RTO values are reused for new transactions). The reuse of backed-off RTO values by these two algorithms happens frequently in FlockLab due to packet losses, leading to long idle times that reduce throughput.

**Settling Time Results**

Figure 3 shows the average settling times obtained by the congestion control mechanisms in the burst traffic scenario. In the GPRS setup, results reveal that all RTT-sensitive mechanisms, except B-RTO, are able to improve the performance of default CoAP when there is congestion. For 10 burst clients (i.e., low congestion), the different congestion control approaches perform similarly, except for B-RTO since it tends to produce spurious retransmissions. When the amount of burst clients increases to 20 and 30, the RTT-sensitive mechanisms adjust their RTOs to higher RTT values. CoCoA does this most efficiently, followed by Linux-RTO, PH-RTO, and CoCoA-S, which perform slightly worse. While CoCoA-S tends to be too aggressive during the burst, Linux-RTO and PH-RTO need slightly longer time to adjust their RTO timers during congestion, since they do not exploit weak RTT information. The VBF and the aging mechanisms used in CoCoA reduce the chances of spurious retransmissions or long idle times, effectively increasing performance. With few clients and short RTTs, B-RTO deteriorates performance, being too aggressive. However, it gets more conservative with more clients and larger RTTs, eventually increasing performance when compared to default CoAP, which does not adapt the RTO timers at all.

In FlockLab, CoCoA yields an improvement over default CoAP in terms of settling time, and it shows the most stable behavior with the narrowest confidence intervals. Over the course of the tests, CoCoA adapts client RTOs efficiently, so the continuous and burst traffic can be processed in parallel. CoCoA-S behaves similarly, also leading to a minor settling time improvement. In contrast, B-RTO, Linux-RTO, and PH-RTO lead to larger settling times on average. For these algorithms we observe a prevalent behavior of the continuous background traffic that hampers the sudden burst from being processed quickly and abates slowly since few RTT measurements can be made because of packet losses.

**Fairness Evaluations**

In terms of fairness, important differences are observed between the different congestion control algorithms in FlockLab. Figure 4 shows the probability mass function (PMF) for the number of finished transactions per destination node for each congestion control algorithm, measured during the continuous traffic experiments along with the FI.

As seen in the figure, with Linux-RTO, PH-RTO, and B-RTO, a small group of nodes are served with a very high number of transactions (e.g., more than 80), whereas a large number of nodes obtain a very low number of finished transactions (e.g., below 10). This results in much lower FI values compared to those of CoAP, CoCoA, and CoCoA-S. The former methods exploit connections with small hop counts and small RTTs, setting their RTO to small values, behaving aggressively in case of message losses, and increasing throughput. However, connections with multiple hops and larger RTTs do not achieve the same data rates, since larger RTO values and consecutive backoffs to these RTO values are applied, reducing throughput.

While CoCoA and CoCoA-S adapt their RTO values as well, the VBF prevents fast retransmissions for good connections with small RTTs and slow retransmissions for bad connections with large RTTs. Moreover, the use of backed-off values when initiating new transmissions is avoided and the aging mechanism prevents maintaining
very small or very large RTO values in idle periods. Thus, CoCoA(S) does not sacrifice from fairness when compared to default CoAP (Fig. 4), while achieving the performance improvements previously presented. Default CoAP behaves neutrally in terms of fairness, since its RTO computation algorithm is independent of the characteristics of a specific path.

**MEMORY FOOTPRINT CONSIDERATIONS**

There is a trade-off between performance and memory footprint of the congestion control mechanisms. Improving the behavior of default CoAP congestion control requires one order of magnitude more memory consumption per CoAP client (Table 1). However, and despite the limitations of many IoT devices, the additional state required is negligible compared to the one order of magnitude more memory consumption per CoAP implementation such as security support. In fact, Datagram Transport Layer Security (DTLS) is mandatory as per the CoAP specification, and it consumes around 2 kB of RAM [10].

**CONCLUSIONS**

CoAP specifies a conservative and non-adaptive congestion control mechanism. CoCoA, an advanced congestion control mechanism for CoAP, provides a flexible and adaptive solution by combining an adaptive RTO calculation, the use of weak RTTs, a VBF, and an aging mechanism to optimize performance.

In comparison to default CoAP, CoCoA increases throughput and reduces the time it takes for a network to process traffic bursts, while not sacrificing fairness. CoCoA consistently delivers a performance which is better than, or at least similar to, that of default CoAP. In contrast, other approaches may be too simple (B-RTO) or do not adapt well to IoT communications (Linux-RTO, PH-RTO), underperforming default CoAP under certain conditions, and therefore are not recommendable as congestion control mechanisms for CoAP.

**ACKNOWLEDGMENTS**

This work was supported in part by the Spanish Government’s Ministerio de Economía y Competitividad under grant number RYC-2013-13029, through project TEC2012-32531, and FEDER. The authors would like to thank Dr. Carsten Bormann for his major contribution to advanced congestion control for CoAP, and the IETF CoRE WG for providing feedback on CoCoA.

**REFERENCES**


**BIOGRAPHIES**

**August Betzler** is a research engineer at the i2CAT Foundation. He received his Ph.D. degree from Universitat Politècnica de Catalunya in 2015. His work focuses on the design and improvement of IoT communication protocols and the application of software-defined networking in the ambit of mobile networks and fog computing. His contributions have been published in conferences and journals. He is also contributing to the IETF standardization process of congestion control mechanisms for the IoT.

**Carlos Gomez** received his Ph.D. degree from Universitat Politècnica de Catalunya in 2007. He is an associate professor at the same university. He is a co-author of 80+ technical contributions including papers published in journals and conferences, IETF RFCs, and the book Sensors Everywhere − Wireless Network Technologies and Solutions. His research interests focus mainly on the Internet of Things. He serves as Associate Editor of the Journal of Ambient Intelligence and Smart Environments.

**Iker Diaz-Rubio** (SM) is a Ramon y Cajal Research Professor in the Department of Network Engineering at the Universitat Politècnica de Catalunya. His research focus is on communication protocol development and evaluation of wireless networks. He received his B.Sc., M.Sc., and Ph.D. degrees in computer engineering from Bogaziçi University, Istanbul, Turkey. He has held several positions in both academia and industry, such as network engineer, system and database consultant, and lecturer.

**Josep Paradells** is a professor at the Universitat Politècnica de Catalunya and the director of the i2CAT Foundation. He has participated in national and European publicly funded research projects, and collaborated with the main Spanish telecommunication companies. He has published his research results in conferences and journals. His expertise area is wireless Internet combining theoretical studies with real implementations. His main focus is IoT, radio wake-up, and visible light communication.
ABSTRACT

Sensing cost and data quality are two primary concerns in mobile crowdsensing. In this article, we propose a new crowdsensing paradigm, sparse mobile crowdsensing, which leverages the spatial and temporal correlation among the data sensed in different sub-areas to significantly reduce the required number of sensing tasks allocated, thus lowering overall sensing cost (e.g., smartphone energy consumption and incentives) while ensuring data quality. Sparse mobile crowdsensing applications intelligently select only a small portion of the target area for sensing while inferring the data of the remaining unsensed area with high accuracy. We discuss the fundamental research challenges in sparse mobile crowdsensing, and design a general framework with potential solutions to the challenges. To verify the effectiveness of the proposed framework, a sparse mobile crowdsensing prototype for temperature and traffic monitoring is implemented and evaluated. With several future research directions identified in sparse mobile crowdsensing, we expect that more research interests will be stimulated in this novel crowdsensing paradigm.

INTRODUCTION

With the prevalence of rich-sensor equipped smartphones in recent years, mobile crowdsensing (MCS) has become a promising paradigm to facilitate urban sensing applications, such as environment monitoring, traffic congestion detection, hotspot identification, and public information sharing [1–4]. Traditional urban sensing applications rely on the expensive specialized sensing infrastructure (e.g., smartphone energy consumption and incentives) while ensuring data quality. Sparse mobile crowdsensing applications intelligently select only a small portion of the target area for sensing while inferring the data of the remaining unsensed area with high accuracy. We discuss the fundamental research challenges in sparse mobile crowdsensing, and design a general framework with potential solutions to the challenges. To verify the effectiveness of the proposed framework, a sparse mobile crowdsensing prototype for temperature and traffic monitoring is implemented and evaluated. With several future research directions identified in sparse mobile crowdsensing, we expect that more research interests will be stimulated in this novel crowdsensing paradigm.

Sensing cost and data quality are two primary concerns in mobile crowdsensing. In this article, we propose a new crowdsensing paradigm, sparse mobile crowdsensing, which leverages the spatial and temporal correlation among the data sensed in different sub-areas to significantly reduce the required number of sensing tasks allocated, thus lowering overall sensing cost (e.g., smartphone energy consumption and incentives) while ensuring data quality. Sparse mobile crowdsensing applications intelligently select only a small portion of the target area for sensing while inferring the data of the remaining unsensed area with high accuracy. We discuss the fundamental research challenges in sparse mobile crowdsensing, and design a general framework with potential solutions to the challenges. To verify the effectiveness of the proposed framework, a sparse mobile crowdsensing prototype for temperature and traffic monitoring is implemented and evaluated. With several future research directions identified in sparse mobile crowdsensing, we expect that more research interests will be stimulated in this novel crowdsensing paradigm.

The theoretical feasibility of Sparse MCS lies in the fact that high spatio-temporal correlations exist in most urban data (e.g., air quality and noise). Such correlations provide the basis for high-quality missing data inference. Specifically, recent research progress in missing data inference algorithms, such as compressive sensing, could facilitate Sparse MCS to achieve high inferred data quality [5–8].

Actually, Sparse MCS is expected to bring significant benefits for rapidly growing MCS applications in smart cities. From the perspective of organizers, Sparse MCS can obviously reduce the sensing cost for each single application. With the increase of applications, such saving of sensing cost can be dramatic; thus, organizers will be highly motivated to adopt Sparse MCS, as long as it can satisfy their quality requirements. From the perspective of participants, each participant can accept only a limited number of sensing tasks allocated, thus lowering overall sensing cost while ensuring data quality.
due to the constraints in her sensing capacity (e.g., energy consumption, network bandwidth). Thus, if the number of MCS applications is increasing, the number of potential qualified participants for each application is decreasing. Then, even if an organizer wants to ensure full or high coverage, it may be difficult to recruit enough participants. Sparse MCS reduces the required participant number for each application, so more applications can run simultaneously.

As far as we know, this is the first article to explicitly introduce the notion of the Sparse MCS paradigm. Compared to the traditional MCS approaches that ensure high coverage ratio, Sparse MCS explicitly introduces data inference into the MCS process and guarantees the quality of inferred data via only sparsely sensed data. In this article, we aim to identify and address the key challenges in Sparse MCS. To this end, we propose a general framework for Sparse MCS applications, implement a prototype to verify its feasibility, and elaborate future research opportunities beyond the prototype. Starting from this work, we expect that more research efforts could be devoted to the Sparse MCS area.

OVERVIEW OF MOBILE CROWDSENSING PROCESS

An MCS task usually has a life cycle of four stages: task creation, task assignment, individual task execution, and crowd data integration [1]. Figure 1 shows this general process of an MCS task. Assume that the MCS task is for urban-scale air quality monitoring. First, the MCS task is created by an MCS organizer (e.g., a city government or non-governmental organization). Then the MCS task is partitioned into a set of individual tasks that can be allocated to MCS participants. For air monitoring, each individual task corresponds to an activity of sensing air quality at a specific sub-area in a specific time cycle. In this article, we introduce the notion of a spatio-temporal cell to specify the spatial and temporal settings of an individual task. Specifically, a spatio-temporal cell refers to a physical sub-area (spatial) in a specific sensing cycle (temporal). After receiving an individual sensing task associated with a spatio-temporal cell, a participant uses her smartphone to complete the task, that is, sensing the air quality at the corresponding sub-area and cycle, and uploading the sensed data to the MCS server. The server finally integrates all the participants’ sensed data and obtains an overall sensed result, which is, in this case, an urban air quality monitoring map.

Sparse MCS, as a kind of MCS framework, also follows this general MCS process, while its characteristics are presented as follows:

- In order to reduce total sensing cost, Sparse MCS allocates individual tasks to only a small subset of the spatio-temporal cells.
- In order to achieve satisfactory data quality, Sparse MCS introduces missing data inference into the server-side data integration process, which attempts to estimate the missing data of unsensed cells based on the sensed data and spatio-temporal correlations between the cells.

**KEY RESEARCH CHALLENGES**

As Sparse MCS focuses on how to sense only partially selected spatio-temporal cells and still guarantee data quality, several research challenges arise:

- **Missing data inference:** Given a set of sensed data from the sparsely selected spatio-temporal cells, how do we infer the missing data of the remaining unsensed cells with high accuracy?
- **Optimal task allocation:** Given an inference algorithm, how do we select the optimal combination of spatio-temporal cells for task allocation so that the sensing cost is minimized with guaranteed inferred data quality?
- **Data quality assessment:** Given a set of sensed spatio-temporal cells and an inference algorithm, how do we assess the inferred data quality without knowing the ground truth data values of unsensed cells?

**MISSING DATA INFERENCE**

In Sparse MCS, as the participants only sense a small part of the total spatio-temporal cells, the missing data of unsensed cells needs to be inferred based on the sensed data. Thus, the inference algorithm is the core of the Sparse...
MCS applications, which will directly impact the data quality of the overall sensed result. Although a broad spectrum of missing data inference approaches have been proposed, such as compressive sensing [5] and dynamic Bayesian network [9], how to apply them appropriately into the Sparse MCS applications is not trivial. Different inference approaches have their own characteristics and applicable use cases; thus, the best inference algorithm may be different for heterogeneous Sparse MCS applications that have different data types, data sizes, and inherent correlations. For a specific Sparse MCS application, how to design the optimal inference algorithm based on the state-of-the-art inference approaches is both challenging and critical.

**OPTIMAL TASK ALLOCATION**

Given a missing data inference algorithm, another challenge in Sparse MCS is to select the best combination of spatio-temporal cells that need to be sensed within a limited budget in order to achieve the best quality of the overall sensed result (including both sensed and inferred data). This challenge is raised due to the fact that sensed data from different combinations of spatio-temporal cells will incur diverse inferred data quality and sensing cost. In other words, given a missing data inference algorithm, the sensed data from a certain “optimal” combination of cells can boost the inference capability of the algorithm more significantly than that from other combinations of cells. Thus, how to identify this “optimal” set of cells among all the candidate cells is critical. This is a nontrivial problem even if we know the sensed data of all the cells in advance — an intuitive solution will enumerate all the possible cell combinations to determine the best one, which has an exponential time complexity that is not tractable in real-life scenarios. However, the task allocation is a monotonic and unregrettable process — if we have allocated tasks to some cells and collected the data, even if we find that the collected data is not efficient for improving the overall data quality, the allocation cannot be retracted to alter the previous decision. Thus, we need to make a proper real-time decision when allocating each individual task so as to exploit the incentive budget efficiently.

The task allocation problem can be even more challenging as more realistic factors need to be taken into consideration. First, the task allocation mechanism is sensitive to the design of the incentive mechanism. For example, if the sensing cost of different cells is not homogeneous, selecting the optimal cell combination needs to consider the total sensing cost of all the selected cells, rather than just the number of cells. Second, the spatio-temporal coverage of all the candidate participants also affects the task allocation. If the candidates can cover all the spatio-temporal cells, we can freely select any cell for sensing because we can always find participants in the selected cell to upload data; however, if this is not the case, we need to consider user mobility in task allocation to determine whether the selected cell can be covered by any participant or not.

**DATA QUALITY ASSESSMENT**

Satisfactory data quality is an essential requirement for the MCS organizer. However, assessing whether data quality meets a predefined level is exceedingly challenging in Sparse MCS, as we cannot directly compute the quality of inferred data due to the lack of the ground truth data of unsensed cells. Either underestimating or overestimating the quality will lead to inappropriate decision about whether more cells need to be selected for sensing: underestimating the quality would result in sensing redundant cells, thus leading to high sensing cost; while overestimating makes the Sparse MCS application fail to satisfy the desired quality requirement.

Moreover, even for the sensed data uploaded by the participants from the selected cells, the data quality may not be always good due to smartphone sensor errors, forged sensed data, privacy protection mechanisms, and so on. These factors complicate the data quality assessment in real-life Sparse MCS applications.

**GENERAL FRAMEWORK OF SPARSE MOBILE CROWDSENSING AND POTENTIAL SOLUTIONS**

To address the above three research challenges, we design a general framework for Sparse MCS applications, as shown in Fig. 2, with an iterative task allocation process including three steps. Note that for the sake of clarity, the framework focuses on illustrating the running process of Sparse MCS applications in one sensing cycle. Each cycle of the MCS task follows the same process.

After a new sensing cycle starts, the first step is *optimal task allocation*, where we need to decide which spatio-temporal cell(s) should be selected for allocating the next individual task. The output of this step is a “sparse sensing map,” where the selected cells are marked with sensing tasks assigned to the participants, and the other cells are left unsensed. Based on the “sparse sensing map,” the second step, *missing data inference*, attempts to infer the data of unsensed cells to construct an “inferred full

---

1 In this framework, because the selected spatio-temporal cell is for sensing in the current cycle, the selection of a cell is equivalent to the selection of a sub-area for sensing in the current cycle.

2 Although the historical sensed data of the target area in previous cycles is not shown in Fig. 2 for simplicity, the step of missing data inference is based on such historical data, in addition to the sensed data obtained in the current cycle.
sensing data.

Next, we describe some potential solutions to address the issues in each step of the framework.

MISSING DATA INFERENCE

Among a broad spectrum of missing data inference techniques, compressive sensing is proved to be effective in many applications like road traffic reconstruction [6], environmental data inference (e.g., temperature, humidity, noise) [5, 7], and multi-dimensional mobile crowdsourced survey data [8]. For traffic volume estimation based on sparsely observed data, dynamic Bayesian network [9] and Gaussian process regression [10] are also competitive techniques. Furthermore, for a specific type of MCS task, exploiting additional knowledge can significantly ameliorate inference performance. For example, considering spatio-temporal correlations could facilitate environment data inference [7]; congestion states and frequent trajectory patterns may be taken into account to assist traffic estimation [9, 10]; spatial and community influence is able to enhance the information spread prediction among a crowd of people [11].

In addition to the sensed data uploaded by participants, other relevant data sources can also be used to boost the inference performance. For example, for air pollution monitoring, other data sources, such as point of interests, road network, weather, and traffic conditions, can significantly improve the inference performance [12]. This kind of multi-source data fusion algorithm should be considered in real-life Sparse MCS applications to improve the inferred data quality.

OPTIMAL TASK ALLOCATION

To select the optimal combination of cells that can achieve the best inferred data quality, we first estimate the inference uncertainty of each unsensed cell. For certain inference algorithms, the uncertainty could be obtained along with the inferred values. For instance, the inference algorithm in [12] could not only infer the missing air quality data, but also output the entropy (can be seen as uncertainty) of each unsensed cell. In addition, state-of-the-art techniques in active learning, such as query by committee, can also be used to estimate uncertainty. Query by committee applies various algorithms to infer the value of an unsensed cell, and then calculates the variance among these inferred values as the uncertainty for that cell [13].

After obtaining uncertainty, we can select the unsensed cell with the largest uncertainty as the next cell for sensing [13]. Besides, inspired by the air quality station placement algorithm in [12], another way to select the next cell is iteratively assuming the inferred value of the cell with the smallest uncertainty as the “true” sensed value, taking this value into the training dataset, and then re-computing all the data uncertainties for the remaining unsensed cells. This iteration process continues until only one unsensed cell with the largest uncertainty is left, which is then selected as the next cell for sensing.

For real MCS campaigns, the task allocation problem is more complicated than the simplified cell selection problem mentioned above. For example, different cells may have diverse sensing cost due to diverse environmental situations, surroundings, and network statuses. Xu et al. [14] study the task allocation problem under this varying cost scenario (suppose the inference algorithm is compressive sensing), and propose a randomized cell sampling scheme in which the probability of selecting each cell is optimized to ensure that the expected quality of the inferred data can be maximized under a fixed amount of sensing cost. Another practical issue is that due to the limitation of human mobility, an MCS organizer may not be able to find any qualified participant in the selected cell to finish an individual task, which can incur a deadlock in the Sparse MCS framework (Fig. 2) as no sensed data can be obtained from the selected cell. Then modeling the participant mobility may be a must for an effective task allocation mechanism to avoid selecting such “inaccessible” cells, or a cell re-selection procedure can be triggered if no participant is found for the previously selected cell so as to avoid deadlock in the running process of Sparse MCS.

DATA QUALITY ASSESSMENT

To solve the issue of assessing the inferred data quality without knowing the true sensed data of unsensed cells, some statistical techniques can be applied. For example, by using re-sampling methods such as leave-one-out or bootstrap, we could estimate the data quality of the overall sensed result based on the sparsely sensed data.
In this section, we present the Sparse MCS prototype implemented by leveraging some techniques discussed in the previous section. This prototype is evaluated using two representative MCS applications, temperature and traffic monitoring, to show the effectiveness of the Sparse MCS framework.

**Prototype Implementation**

**Missing Data Inference:** According to [6, 7], we turn the data inference in both temperature and traffic monitoring applications into a matrix completion problem (one dimension refers to sub-areas, and the other refers to cycles). To solve the problem, the temperature monitoring application uses spatio-temporal compressive sensing [7], while the traffic sensing application leverages a different algorithm for traffic speed estimation [6].

**Optimal Task Allocation:** Based on query by committee, we use several inference algorithms, including K nearest neighbors (KNN), compressive sensing (CS), spatio-temporal compressive sensing (STCS) [7], and traffic-specialized compressive sensing (TRA-CS) [6], to infer the values of all the unsensed cells, and then allocate the task to the cell that has the largest variance on these inferred values [13].

**Data Quality Assessment:** Via the leave-one-out cross-validation on the sensed cells, first we get several samples of the inference error (each sensed cell refers to one sample); based on these samples, we use Bayesian inference to estimate the probability density function of the inference accuracy [13].

**Experiment Setup**

We conduct the experiments for the two use cases: the temperature\(^3\) and traffic\(^4\) monitoring. Detailed experiment settings are summarized in Table 1. Note that the data quality requirement is defined by the \((\varepsilon, p)\)-quality metric [13] — the mean absolute error (MAE) of the inferred temperature (traffic speed) values should be smaller than \(\varepsilon^\circ\text{C} (\text{m/s})\) in \(p\) · 100 percent of the sensing cycles. In the experiments, \(\varepsilon\) is set to 0.25\(^\circ\text{C}\) (temperature) or 2 m/s (traffic); \(p\) is set to 0.9 or 0.95.

**Performance Analysis**

We first evaluate the performance of the inference algorithms. Figure 3 illustrates that for temperature and traffic scenarios, STCS and TRA-CS achieve less inference error than CS and KNN whatever the sensing coverage (the proportion of sensed cells) is, respectively. From Fig. 3, we also know the data quality given a sensing coverage: the MAE of STCS (TRA-CS) by sensing 10/30/50 percent of the spatial-temporal cells is around 0.25/0.14/0.09\(^\circ\text{C}\) (1.75/1.23/0.81 m/s) per cycle. Note that this is an expected error, not a guarantee. Next, by incorporating optimal task allocation and data quality assessment together, we attempt to verify that our prototype can finish an MCS task with sparsely sensed data and guarantee the data quality.

We compare our prototype with two baselines, RAND and FIX-\(k\). RAND uses random sampling in task allocation, and FIX-\(k\) allocates the fixed number (\(k\)) of tasks in each cycle. As in our prototype, they use STCS/TRA-CS as the inference algorithm. Table 2 shows that for a predefined error bound \(\varepsilon\), in our prototype, the actual fraction of the sensing cycles with errors smaller than \(\varepsilon\) is similar to the required \(p\) in both scenarios. This verifies that our prototype meets the predefined \((\varepsilon, p)\)-quality. So do the two baselines.

Then we compare the number of allocated tasks in Fig. 4. Compared to RAND and FIX-\(k\), our prototype allocates 11.1–26.5 percent

---

\(^3\) http://lcav.epfl.ch/op/edit/sensorscope-en

\(^4\) http://research.microsoft.com/en-us/projects/tdrive/

![Figure 3. MAE (with standard deviation) on two applications: a) temperature; b) traffic.](image-url)
(11.9–28.5 percent) fewer tasks when ensuring the same level of quality in temperature (traffic) monitoring. Specifically, by only allocating tasks to cover 12.9/15.4 percent (18.3/20.0 percent) of spatio-temporal cells, the MAE of temperature (traffic speed) is smaller than 0.25°C (2 m/s) in 90/95 percent of sensing cycles. This proves that by leveraging the idea of Sparse MCS, the sensing cost can be significantly reduced, while the overall data quality is ensured.

**Discussion**

The prototype verifies the feasibility and generality of the Sparse MCS framework, while many opportunities still exist for improvement. Some are listed below.

**Inference Algorithm:** Our prototype includes two algorithms, STCS [7] and TRA-CS [6], to infer missing temperature and traffic speed values, respectively. To tackle more MCS scenarios, different algorithms need to be designed and incorporated in the future. For instance, to infer missing PM2.5 values precisely in air quality monitoring, the features of points of interests, road networks, and weather may be considered in the algorithm design [9].

**User Mobility:** The prototype assumes that a participant can always be found in a selected cell to contribute data, while due to user mobility patterns, this may not be true in reality; future studies on task allocation should consider this issue. Also, if a sub-area lacks sensed data for a long time as no participant goes there, specific strategies (e.g., using higher incentives to steer participants to that sub-area) may be necessary to ameliorate the overall task quality.

**Spatio-Temporal Cell Configuration:** The configuration of sub-area size and cycle length is application-specific; thus, the setting of our experiment needs to be adjusted for other MCS tasks. For instance, in air quality monitoring, the size of sub-area is relatively large, for example, 1 km × 1 km [12]. Besides, how to adaptively configure the spatio-temporal cell in real time to catch the prompt changes of sensing values is another promising research direction (e.g., automatically shortening cycle length when temperature fluctuates frequently).

**Privacy in Sparse Mobile Crowdsensing**

In Sparse MCS, all the participants are requested to report their data with true identities and locations, which can incur serious privacy implications if adversaries hack the communication between the participant and the MCS server, or the MCS server itself is not trustworthy. To tackle participants’ privacy issues, two popular categories of mechanisms are often applied: identity anonymity and location obfuscation.

Identity anonymity mechanisms attempt to hide users’ true identities when uploading data to MCS servers. Anonymizing user identity raises many challenges for MCS applications, such as how to pay the anonymous participants correct incentives and how to identify the malicious participants (who report forged data). Not surprisingly, as a sub-type of MCS applications, Sparse MCS applications also face such problems, and the existing MCS solutions can help to incorporate identity anonymity mechanisms into Sparse MCS applications.

For location obfuscation mechanisms, Sparse MCS brings some new challenges compared to traditional MCS. Location obfuscation mechanisms report participants’ obfuscated locations instead of their actual locations to the server to preserve location privacy. The location inaccuracy incurred by obfuscation will degrade the quality of the sensed data that is uploaded by participants. Specifically, for
Sparse MCS, this issue becomes much more crucial: location obfuscation not only affects the quality of sensed data, but also the quality of inferred data. To reduce data quality loss incurred by obfuscation in Sparse MCS, two possible solutions may be considered. First, while ensuring a desired level of privacy protection, the location obfuscation mechanism can be well tuned to make the obfuscation happen more frequently between two locations with higher sensed data correlation in order to minimize the potential tail quality loss of the uploaded data. For example, when sensing air quality in a city, two parks may have similar air quality values. If a participant is at a park, an obfuscation mechanism that is likely to perturb her location to another park would incur little quality loss. Second, a missing data inference algorithm may be adapted to handle privacy-preserving data. For example, as privacy-preserving data inevitably has some quality loss due to obfuscation, we can model such quality loss and then improve the data inference accuracy by trusting more the data with smaller quality loss.

**CONCLUSION**

In this article, a novel MCS paradigm, Sparse MCS, is proposed to significantly reduce the sensing cost while still guaranteeing the overall data quality for urban sensing. The rationale behind the Sparse MCS framework is the fact that various urban sensing data have spatio-temporal correlations, so the missing data of un sensed spatio-temporal cells can be inferred based on sparsely sensed data. A prototype of the Sparse MCS framework is implemented to show the feasibility of the proposed ideas and techniques. In the future, practical issues in Sparse MCS applications, such as participant mobility, sensed data error, malicious participants, and user privacy, need to be studied.

**ACKNOWLEDGMENTS**

The authors would like to thank the reviewers and the editors for their helpful comments and suggestions. The authors also would like to thank L. Chen, T. Wang, and Z. Tang for insightful discussions and careful proofreading. Correspondence should be sent to D. Zhang (daqing.zhang@telecom-sudparis.eu) or X. Han (xiaohan@mail.shufe.edu.cn). This article was partially supported by the NSFC under Grant No. 61572048, the Fundamental Research Funds for the Central Universities (No. 106112015CDJXY180001), the Open Research Fund Program of Shenzhen Key Laboratory of Spatial Smart Sensing and Services (Shenzhen University), and the Chongqing Basic and Frontier Research Program (No. cstc2015cjyjA00016).

**REFERENCES**

Index Modulation Techniques for 5G Wireless Networks

Ertugrul Basar

ABSTRACT

The ambitious goals set for 5G wireless networks, which are expected to be introduced around 2020, require dramatic changes in the design of different layers for next generation communications systems. Massive MIMO systems, filter bank multi-carrier modulation, relaying technologies, and millimeter-wave communications have been considered as some of the strong candidates for the physical layer design of 5G networks. In this article, we shed light on the potential and implementation of IM techniques for MIMO and multi-carrier communications systems, which are expected to be two of the key technologies for 5G systems. Specifically, we focus on two promising applications of IM: spatial modulation and orthogonal frequency-division multiplexing with IM, and discuss the recent advances and future research directions in IM technologies toward spectrum- and energy-efficient 5G wireless networks.

INTRODUCTION

Unprecedented levels of spectrum- and energy efficiency are expected from fifth generation (5G) wireless networks to achieve ubiquitous communications between anybody, anything, and anytime [1]. In order to reach the challenging objectives of 5G wireless networks, researchers have envisioned novel physical layer (PHY) concepts such as massive multiple-input multiple-output (MIMO) systems and non-orthogonal multi-carrier communications schemes. However, the wireless community is still working day and night to come up with new and more effective PHY solutions toward 5G networks. There has been a growing interest in index modulation (IM) techniques over the past few years. IM, in which the indices of the building blocks of the considered communications systems are used to convey additional information bits, is a novel digital modulation scheme with high spectral and energy efficiency. Spatial modulation (SM) and orthogonal frequency-division multiplexing with IM (OFDM-IM) schemes, where the corresponding index modulated building blocks are the transmit antennas of a MIMO system and the subcarriers of an OFDM system, respectively, appear as two interesting as well as promising applications of the IM concept.

After the pioneering work of Mesleh et al. [2], SM techniques have attracted significant attention over the past few years. Although having strong and well established competitors such as vertical Bell Labs layered space-time (V-BLAST) and space-time coding (STC) systems, SM schemes have been regarded as possible candidates for spectrum- and energy-efficient next generation MIMO systems. On the other hand, researchers have started to explore the potential of the IM concept for subcarriers of OFDM systems in recent times, and it has been shown that the OFDM-IM scheme [3] can provide attractive advantages over classical OFDM, which is an integral part of many current wireless standards.

The aim of this article is to present the basic principles of these two promising schemes, SM and OFDM-IM, which are still waiting to be explored by many experts, and review some of the recent interesting results in IM techniques. Furthermore, we discuss the implementation scenarios of IM techniques for next generation wireless networks and outline possible future research directions. Particularly, we shift our focus to generalized, enhanced, and quadrature IM schemes and the application of IM techniques for massive multi-user MIMO (MU-MIMO) and cooperative communications systems.

INDEX MODULATION FOR TRANSMIT ANTENNAS: SPATIAL MODULATION

SM is a novel way of transmitting information by means of the indices of the transmit antennas of an \( n_T \times n_R \) MIMO system in addition to the conventional \( M \)-ary signal constellations, where \( n_T \) and \( n_R \) denote the number of transmit and receive antennas, respectively. In contrast to conventional MIMO schemes that rely on either spatial multiplexing to boost the data rate or spatial diversity to improve the error performance, the multiple transmit antennas of a MIMO system are used for a different purpose in an SM scheme. More specifically, there are two information carrying units in SM: indices of transmit antennas and \( M \)-ary constellation symbols. For each signaling interval, a total of

\[
\log_2(n_T) + \log_2(M) \tag{1}
\]

bits enter the transmitter of an SM system, where \( M \) is the size of the considered signal.
Generalized, Enhanced, and Quadrature SM Schemes

The generalized SM (GSM) scheme improves the spectral efficiency of SM by increasing the number of active transmit antennas [4]. In the GSM scheme, multiple antennas are selected as active to transmit the same data symbol. Denoting the number of active transmit antennas by \( n_A \) where \( n_A < n_T \), information bits can be transmitted for each signaling interval in addition to the \( \log_2(M) \) bits transmitted by the \( M \)-ary data symbols, where \( \lfloor \cdot \rfloor \) is the floor operation. Since \( \log_2(n_T) \leq \log_2(2^n) \) for \( n_T = 2^h (n = 1, 2, \ldots) \), the spatial domain can be used more effectively by the GSM scheme. In [6], the concept of GSM has been extended to multiple-active spatial modulation (MA-SM), where different data symbols are transmitted from the selected transmit antennas to further boost spectral efficiency. Consequently, the spectral efficiency of the MA-SM scheme becomes \( \log_2(2^n) + n_A \log_2(M) \) bits per channel use (bpcu), which is considerably higher than that of SM.

Enhanced SM (ESM) is a recently proposed and promising variant of SM [7]. In the ESM scheme, the number of active transmit antennas can vary for each signaling interval, and the information is conveyed not only by the indices of active transmit antennas but also by the selected signal constellations used in transmission. In other words, the ESM scheme considers multiple signal constellations, and the information is transmitted by the combination of active transmit antennas and signal constellations.

Quadrature SM (QSM) [8] is yet another clever modification of classical SM to improve the spectral efficiency while maintaining its advantages such as operation with single RF chain and inter-channel interference (ICI) free transmission. In the QSM scheme, the real and imaginary parts of the complex \( M \)-ary data symbols are separately transmitted using the SM principle. For a MIMO system with \( n_T \) transmit antennas, the spectral efficiency of QSM becomes \( 2\log_2(n_T) + \log_2(M) \) bpcu by simultaneously applying the SM principle for in-phase and quadrature components of the complex data symbols. Even if the number of active transmit antennas can be one or two for the QSM scheme, a single RF chain is sufficient at the transmitter since only two carriers (cosine and sine) are used during transmission.

In Fig. 1, we compare the minimum squared Euclidean distance between the transmission vectors \( \left( d_{\text{min}} \right) \), which is an important design parameter for quasi-static Rayleigh fading channels to optimize the error performance of single-input multiple-output (SIMO), SM, ESM, and QSM schemes. In all considered configurations, the average total transmitted energy is normalized to unity to make fair comparisons. It is interesting to note that ESM and QSM schemes achieve the same \( d_{\text{min}} \) value for 4 and 6 bpcu transmissions. However, as seen from Fig. 1, QSM suffers a worse minimum Euclidean distance, and as a result worse error performance, compared to ESM for higher spectral efficiency values, while the ESM scheme requires a more complicated transmitter with two RF chains. Finally,
Although the spectral efficiency of SM systems cannot compete with that of traditional methods such as V-BLAST for massive MIMO systems, the use of the IM concept for the transmit antennas of a massive MIMO system can provide an easy as well as cheap implementation solution thanks to the inherently available advantages of SM systems.


The results of Fig. 1 also prove that the relative $d_{\text{min}}$ advantage of IM schemes over the classical SIMO scheme increases with increasing spectral efficiency, that is, IM techniques become more preferable for higher spectral efficiency values.

**MASSIVE MULTI-USER MIMO SYSTEMS WITH SM**

The massive MIMO concept, in which BSs have tens to hundreds of antennas, is considered as one of the potential key technologies for 5G wireless networks due to its appealing advantages such as very high spectral and energy efficiency. While the initial studies on MIMO systems generally focus on point-to-point links where two users communicate with each other, practical MU-MIMO systems are gaining more attention to exploit the multiple antennas of a MIMO system to support multiple users simultaneously.

The extension of MIMO systems into massive scale provides unique opportunities for SM systems since it becomes possible to transmit a higher number of information bits in the spatial domain with massive MIMO systems, even if the number of available RF chains is very limited. Although the spectral efficiency of SM systems cannot compete with that of traditional methods such as V-BLAST for massive MIMO systems, the use of the IM concept for the transmit antennas of a massive MIMO system can provide an easy as well as cheap implementation solution thanks to the inherently available advantages of SM systems [9].

In Fig. 2a, a massive MU-MIMO system is considered where $K$ users employ SM techniques for uplink transmission. Compared to user terminals with single antennas, additional information bits can be transmitted using SM without increasing the system complexity. GSM, ESM, and QSM techniques can be implemented at the users to further improve the spectral efficiency. At the BS, the optimal (ML) detector can be used at the expense of exponentially increasing decoding complexity (with respect to $K$) due to the inter-user interference. Low-complexity near-optimal detec-
implementation methods can be implemented as well by sacrificing the optimum error performance. Alternatively, SM techniques can also be used at the BS for downlink transmission as shown in Fig. 2b. To support a high number of users, the massive antennas of a BS can be split into subgroups of fewer antennas where SM techniques can be employed for each user [10]. For the specific case of two users, the data of User 1 can be mapped into antenna indices, while the data of User 2 can be conveyed with M-ary signal constellations. GSM techniques can also be implemented at the BS to transmit the data of different users with either antenna indices and/or constellation symbols.

**COOPERATIVE SM SYSTEMS**

Considering the effective solutions provided by SM techniques and cooperative communications systems, the combination of these two technologies naturally arises as a potential candidate for future wireless networks. Due to recent technological advances, more than one antenna can be employed at mobile and relay terminals, and cooperative SM systems can provide new implementation scenarios, additional diversity gains, and higher data rates without increasing the cost and complexity of the mobile and relay terminals.

Figure 2. Massive MU-MIMO systems with SM: a) an uplink transmission scenario where User $k$ has $n_T^k$ transmit antennas available for SM and the BS has $n_R^k$ ~ 10–100 receive antennas; b) a downlink transmission scenario where User $k$ has $n_R^k$ receive antennas and the BS has $n_T^k$ ~ 10–100 transmit antennas available for SM.

relaying techniques. The scenario of Fig. 3a is generally observed in practical networks where S and D cannot communicate directly due to distance or obstacles. In this relaying scenario, SM techniques can improve the energy and spectral efficiency of S compared to the single antenna case. In Fig. 3b, a direct link from S to D exists, and R can cooperate by employing different relaying methods.

In Fig. 3c, we take into account the two-way communications of S and D, which is accomplished via R. Without network coding, the overall transmission between S and D requires four transmission phases, which considerably reduce the overall spectral efficiency. However, two-way communications between S and D can be accomplished at two phases with network coding where in the first transmission phase, S and D simultaneously transmit their signals to R using SM techniques. In the second transmission phase, R combines the signals received from S and D, then forwards this combined signal to S and D. The use of SM provides some opportunities for R such as transmitting one user’s data with antenna indices and the other one’s with constellation symbols. Finally, in Fig. 3d, we consider a distributed cooperation scenario with $N$ relay nodes ($R_1$, …, $R_N$). In the first transmission phase, S can use SM techniques to transfer its data to relays. In the second transmission phase, one or more relays cooperate, and the indices of the activated relays can be considered as an additional way to convey information. This flexibility allows the relays to cooperate even if they have single antennas ($n_R = 1$). Furthermore, opportunistic relay selection is also an option for the network topology of Fig. 3d, where the adaptively selected best relay takes part in transmission. For all cooperation scenarios described above, S and/or R can use GSM/ESM/QSM techniques to further improve the spectral efficiency as well as to obtain more flexibility in the system design.

Due to recent technological advances, more than one antenna can be employed at mobile and relay terminals, and cooperative SM systems can provide new implementation scenarios, additional diversity gains and higher data rates without increasing the cost and complexity of the mobile and relay terminals.
The IM concept can be considered for other communications systems apart from MIMO systems. For instance, IM techniques can be efficiently implemented for the subcarriers of an OFDM system. OFDM-IM is a novel multi-carrier transmission scheme that has been proposed with inspiration from the IM concept of SM.

The IM concept can be considered for other communications systems apart from MIMO systems. For instance, IM techniques can be efficiently implemented for the subcarriers of an OFDM system. OFDM-IM is a novel multi-carrier transmission scheme that has been proposed with inspiration from the IM concept of SM [3]. Similar to SM, in the OFDM-IM scheme, the incoming bitstream is split into index selection and M-ary constellation bits. According to the index selection bits, only a subset of available subcarriers are selected as active, while the remaining inactive subcarriers are not used and set to zero. On the other hand, the active subcarriers are modulated according to the M-ary constellation bits. In other words, the information is conveyed not only by the data symbols as in classical OFDM, but also by the indices of the active subcarriers, which are used for the transmission of the corresponding data symbols for the OFDM-IM scheme.

Considering an OFDM system with \( N_F \) subcarriers, one can directly select the indices of active subcarriers similar to the IM technique used for the transmit antennas of an MA-SM system. However, the massive OFDM frames can provide more flexibility for the employment of IM techniques for OFDM-IM schemes compared to SM schemes. On the other hand, keeping in mind that \( N_F \) can take very large values, such as 512, 1024 or 2048 as in Long Term Evolution-Advanced (LTE-A), there could be trillions of (actually more than a googol \( 10^{100} \) in mathematical terms) possible combinations for active subcarriers if index selection is applied directly. As an example, assume that we want to select the indices of 256 active subcarriers out of \( N_F = 512 \) available subcarriers; then there could be 472.55 \( \times 10^{150} \) possible combinations of active subcarriers, which turn the selection of active subcarriers into an almost impossible task. Therefore, for the implementation of OFDM-IM, the single and massive OFDM-IM block should be divided into \( G \) smaller and more manageable OFDM-IM subblocks, each containing \( N \) subcarriers to perform IM, where \( N_F = G \times N \). For each subblock, \( K \) out of \( N \) available subcarriers can be selected as active according to the \( p_1 = \lfloor \log_2(\frac{K}{N}) \rfloor \) index selection bits where typical \( N \) values could be 2, 4, 8, 16, and 32 with \( 1 \leq K < N \). Please note that classical OFDM becomes a special case of OFDM-IM with \( K = N \), that is, when all subcarriers are activated.

The block diagrams of OFDM-IM scheme’s transmitter and receiver structures are illustrated in Figs. 4a and 4b, respectively. As seen from Fig. 4a, for each OFDM-IM frame, a total of

\[
m = pG = \left\lfloor \log_2 \left( \frac{N}{K} \right) \right\rfloor + K \log_2 M \cdot G
\]

bits can be transmitted where \( p = p_1 + p_2 \) and \( p_2 = K \log_2 M \). In Fig. 4a, \( j_p \) and \( j_g \) denote the vector of selected indices and M-ary data symbols with dimensions \( K \times 1 \), respectively. First, the OFDM-IM subblock creator forms the \( N \times 1 \) OFDM-IM subblocks \( x_p, g = 1, \ldots, G \); then the OFDM-IM block creator obtains the \( N_F \times 1 \) main OFDM-IM frame \( x \) by concatenating these \( G \) OFDM-IM subblocks. After this point, \( G \times N \) block interleaving can be performed to ensure that the subcarriers of a subblock are affected by uncorrelated wireless fading channels. Finally, classical OFDM procedures such as inverse fast Fourier transform (IFFT), cyclic prefix (CP) insertion, and digital-to-analog conversion (DAC) are applied.

Two different index selection procedures are available for OFDM-IM: reference look-
Due to its flexible system design with an adjustable number of active subcarriers and its attractive advantages over OFDM, OFDM-IM can be a possible candidate not only for high-speed wireless communications systems but also for M2M communications systems of 5G wireless networks which require low power consumption.

**Recent Advances in OFDM-IM**

The subcarrier IM concept for OFDM has recently attracted significant attention from researchers, and has been investigated in some up-to-date studies that deal with the error performance and capacity analysis, generalization, enhancement, and optimization of OFDM-IM, and its adaptation to different wireless environ-

**Figure 4.** OFDM-IM system at a glance: a) transmitter structure; b) receiver structure; c) two different index selection procedures.

up tables for smaller \( N \) values and combinatorial number theory for higher \( N \) values, where examples of these two methods are provided in Fig. 4c. Similar to SM, the receiver of OFDM-IM has to determine the active subcarriers and the corresponding data symbols in accordance with the index selection procedure used at the transmitter. After applying inverse operations, the received signals are separated since the detection of different subblocks can be carried out independently. The optimum but high-complexity ML detector makes a joint search over possible subcarrier activation combinations and data symbols, while the low-complexity log-likelihood ratio (LLR) calculation-based near-optimal detector determines the indices of the active subcarriers first, and then detects the corresponding data symbols. The LLR detector calculates a probabilistic measure on the active status of a given subcarrier by considering the fact that the corresponding subcarrier can be either active (carrying an \( M \)-ary constellation symbol) or inactive. This detector is classified as near-optimal since it does not know the set of all possible subcarrier activation combinations.

It has been shown that OFDM-IM provides an interesting trade-off between error performance and spectral efficiency, and it offers some attractive advantages over classical OFDM. Unlike classical OFDM, the number of active subcarriers of an OFDM-IM scheme can be adjusted accordingly to reach the desired spectral efficiency and/or error performance. Furthermore, due to the information bits carried by IM, which have lower error probability compared to ordinary \( M \)-ary constellation bits, OFDM-IM can provide better bit error rate (BER) performance than classical OFDM for low-to-medium spectral efficiency values, while it exhibits comparable decoding complexity using the near-optimal LLR detector. Furthermore, it has been recently proved that OFDM-IM also outperforms classical OFDM in terms of ergodic achievable rate [12]. Consequently, we conclude that due to its flexible system design with adjustable number of active subcarriers and its attractive advantages over OFDM, OFDM-IM can be a possible candidate not only for high-speed wireless communications systems but also for machine-to-machine (M2M) communications systems of 5G wireless networks, which require low power consumption.
ments. Interested readers are referred to [13, references therein] for an overview of these studies. In this section, we focus on two recently proposed and promising forms of OFDM-IM: generalized OFDM-IM and MIMO-OFDM-IM systems.

**Generalized OFDM-IM Schemes**

Two generalized OFDM-IM structures (OFDM-GIM-I and OFDM-GIM-II) have been recently proposed by modifying the original OFDM-IM scheme [14]. In the OFDM-GIM-I scheme, the number of active subcarriers is no longer fixed, and it is also determined according to the information bits. Considering the case of $N = 4, K = 2$ with binary PSK (BPSK) modulation ($M = 2$), according to Eq. 2, (OFDM), $\log_2(4) + \log_2(4) = 4$ bits can be transmitted per OFDM-IM subblock, which is a total of $4 \times 2^2 = 16$ subblock realizations that can be obtained. On the other hand, considering all activation patterns ($K \in \{0, 1, 2, 3, 4\}$), which means that the number of active subcarriers can take values from zero (all subcarriers are inactive, $K = 0$) to four (all subcarriers are active, $K = 4$), as well as considering all possible values of $M$-ary data symbols, a total of $\sum_{K=0}^{4} \binom{K}{2} M^K = 81$ possible subblock realizations can be obtained for which $\log_2(81) = 6$ bits can be transmitted per OFDM-GIM-I subblock. As a result, the OFDM-GIM-I scheme can provide more flexibility for the selection of active subcarriers and can transmit more bits per subblock compared to OFDM-IM.

The OFDM-GIM-II scheme aims to further improve the spectral efficiency by applying IM independently for in-phase and quadrature components of the complex data symbols similar to the QSM scheme. In other words, a subcarrier can be active for one component, while it can simultaneously be inactive for the other component. Considering the case of $N = 16, K = 10$ with quadrature PSK (QPSK) modulation ($M = 4$), according to Eq. 2, $\log_2(16) + 10\log_2(4) = 32$ bits can be transmitted per OFDM-IM subblock. On the other hand, the OFDM-GIM-II scheme allows the transmission of

$$\log_2 \left( \frac{16}{10} \binom{16}{K} M^K \right)$$

= 44 bits per subblock, which is 37.5 percent higher than that of OFDM-IM. Please note that the in-phase and quadrature components of a complex $M$-QAM symbol are the elements of a $\sqrt{M}$-ary pulse amplitude modulation (PAM) constellation, where a total of

$$\binom{N}{K} M^K$$

realizations are possible per component.

**From SISO-OFDM to MIMO-OFDM**

The first studies on OFDM-IM generally focused on point-to-point single-input single-output (SISO) systems, which can be unsuitable for some applications due to their limited spectral efficiency. More recently, MIMO transmission and OFDM-IM principles are combined to further boost the spectral and energy efficiency of the OFDM-IM scheme [13]. Specifically, the transmitter of the MIMO-OFDM-IM scheme consists of parallel concatenated SISO-OFDM-IM transmitters (Fig. 4a) to operate over $n_T \times n_R$ MIMO subcarrier frequency selective fading channels. At the receiver of the MIMO-OFDM-IM scheme, the simultaneously transmitted OFDM-IM frames are separated and demodulated using low-complexity minimum mean square error (MMSE) detection and an LLR calculation-based detector that considers the statistics of the MMSE filtered received signals. It has been shown via extensive computer simulations that due to its improved error performance and flexible system design, MIMO-OFDM-IM can be a strong alternative to classical MIMO-OFDM, which has been included in many current wireless standards. In Fig. 5, the un-coded BER performance curves of the MIMO-OFDM-IM and classical V-BLAST type MIMO-OFDM schemes are given for three MIMO configurations where the same spectral efficiency values are obtained for both schemes. As observed from Fig. 5, significant signal-to-noise ratio (SNR) improvements can be obtained by the MIMO-OFDM-IM scheme compared to classical MIMO-OFDM to reach a target BER value. On the other hand, the generalization of MIMO-OFDM-IM for massive MU-MIMO systems remains an interesting research problem toward 5G wireless networks.

Another recently proposed IM scheme, which is called generalized space-frequency index modulation (GSM) [15], combines the OFDM-IM concept with the GSM principle by exploiting both spatial and frequency (subcarrier) domains for IM. It has been shown that GSM can also provide improvements over MIMO-OFDM in terms of achievable data rate and BER perfor-
formance with ML detection for BPSK and QPSK constellations. However, the design of low-complexity detector types is an open research problem for the GSFIM scheme.

**Conclusions and Future Work**

IM is an up and coming concept for spectrum-and energy-efficient next generation wireless communications systems to be employed in 5G wireless networks. IM techniques can offer low-complexity as well as spectrum- and energy-efficient solutions toward the single/multi-carrier, massive MU-MIMO, and cooperative communications systems to be employed in 5G wireless networks. In this article, we have reviewed the basic principles, advantages, recent advances, and application areas of SM and OFDM-IM systems, which are two popular applications of the IM concept. In Table 1, the pros and cons of the reviewed IM schemes in terms of spectral efficiency, ML detection complexity, and error performance are provided. We conclude from Table 1 that IM schemes can be considered as possible candidates for 5G wireless networks due to the interesting trade-offs they offer among error performance, complexity, and spectral efficiency, while there are still interesting as well as challenging research problems that need to be solved in order to further improve the efficiency of IM schemes. These research challenges can be summarized as follows:

- The design of novel generalized/enhanced IM schemes with higher spectral and/or energy efficiency, lower receiver complexity, and better error performance
- The integration of IM techniques (e.g., SM, GSM, ESM, QSM, and OFDM-IM) into massive MU-MIMO systems to be employed in 5G wireless networks and the design of novel uplink/downlink transmission protocols
- The adaptation of IM techniques to cooperative communications systems (e.g., dual/multihop, network-coded, multi-relay, and distributed networks)
- The investigation of the potential of IM techniques via practical implementation scenarios

**Acknowledgment**

This work is supported in part by the Scientific and Technological Research Council of Turkey (TUBITAK) under grant number 114E607.

**References**


**Table 1. Pros and cons of several index modulation schemes.**

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Spectral efficiency</th>
<th>ML detection complexity</th>
<th>Error performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIMO</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>SM</td>
<td>Moderate</td>
<td>Low*</td>
<td>Moderate</td>
</tr>
<tr>
<td>GSM</td>
<td>Moderate</td>
<td>Moderate*</td>
<td>Moderate</td>
</tr>
<tr>
<td>MA-SM</td>
<td>High</td>
<td>Moderate*</td>
<td>Moderate</td>
</tr>
<tr>
<td>ESM</td>
<td>High</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>QSM</td>
<td>High</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>V-BLAST</td>
<td>High</td>
<td>High*</td>
<td>Moderate</td>
</tr>
<tr>
<td>OFDM-IM</td>
<td>Low</td>
<td>Moderate*</td>
<td>Moderate</td>
</tr>
<tr>
<td>OFDM-IM-I</td>
<td>Moderate</td>
<td>High*</td>
<td>Moderate</td>
</tr>
<tr>
<td>OFDM-IM-II</td>
<td>Moderate</td>
<td>High*</td>
<td>Moderate</td>
</tr>
<tr>
<td>SIMO-OFDM-IM</td>
<td>High</td>
<td>High*</td>
<td>High</td>
</tr>
<tr>
<td>GSM</td>
<td>High</td>
<td>Moderate*</td>
<td>Moderate</td>
</tr>
<tr>
<td>ESM</td>
<td>High</td>
<td>Moderate*</td>
<td>Moderate</td>
</tr>
<tr>
<td>QSM</td>
<td>High</td>
<td>Low*</td>
<td>Moderate</td>
</tr>
<tr>
<td>V-BLAST-OFDM</td>
<td>Low</td>
<td>Moderate*</td>
<td>Moderate</td>
</tr>
</tbody>
</table>

*Lower-complexity near/sub-optimal detection is also possible.

**Biography**

Enesil Basar [S’09, M’13, SM’16] (basarer@itu.edu.tr) received his B.S. degree with high honors from Istanbul University, Turkey, in 2007, and his M.S. and Ph.D. degrees from Istanbul Technical University in 2009 and 2013, respectively. He spent the academic year 2011-2012 at the Department of Electrical Engineering, Princeton University, New Jersey. Currently, he is an assistant professor at Istanbul Technical University, Electronics and Communication Engineering Department, and a member of the Wireless Communication Research Group. He was the recipient of the Istanbul Technical University Best Ph.D. Thesis Award in 2014 and has won three Best Paper Awards. He is an Associate Editor for IEEE Access, a regular reviewer for various IEEE journals, and has served as a TPC member for several conferences. His primary research interests include MIMO systems, index modulation, cooperative communications, OFDM, and visible light communications.
ADVERTISERS’ INDEX

ADVERTISING SALES OFFICES
Closing date for space reservation: 15th of the month prior to date of issue

NATIONAL SALES OFFICE
James A. Vick
Sr. Director Advertising Business, IEEE Media
EMAIL: jv.ieeemedia@ieee.org

Marion Delaney
Sales Director, IEEE Media
EMAIL: md.ieeemedia@ieee.org

Mark David
Sr. Manager Advertising & Business Development
EMAIL: m.david@ieee.org

Mindy Belfer
Advertising Sales Coordinator
EMAIL: m.belfer@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: (773) 283-4528
FAX: (774) 283-4527
EMAIL: je.ieeemedia@ieee.org

SOUTHEAST
Scott Rickles
TEL: (770) 664-4567
FAX: (770) 740-1399
EMAIL: rickles@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

COMPANY .................................................................................... PAGE

IEEE Communications Magazine • July 2016

176
Unleashing 5G mm-Waves - A Test & Measurement Perspective

In the past 12 months, research around 5G wireless communications has gained significant momentum, yet it is still unclear what the new physical layer may look like. This poses a challenge for transceiver components such as filters, mixers, and amplifiers. This webinar discusses the associated test and measurement aspects and challenges that have been identified and demonstrates initial measurement solutions for 5G.

Sponsored content provided by:

ROHDE & SCHWARZ

5G Coexistence in a 3G / 4G / Satellite World?

5G applications may utilize the crowded sub 6 GHz spectrum, as well as centimeter-wave and millimeter-wave spectrum.

Operating in crowded spectrum poses questions such as:
- How will sub 6 GHz candidate 5G waveforms coexist and interact with legacy 3G, 4G, and PAN waveforms?
- Can the 28 GHz frequency band be shared between satellite waveforms and candidate 5G waveforms?

This webcast will discuss some of the candidate 5G waveforms being proposed and examine 5G coexistence case studies for sub 6 GHz applications and 28 GHz applications. Millimeter-wave applications will also be discussed, as well as an LTE / Radar coexistence case study.

Sponsored content provided by:

KEYSIGHT TECHNOLOGIES

Limited Time Only at >> www.comsoc.org

For this and other sponsor opportunities contact Mark David // 732-465-6473 // m.david@ieee.org
Robust, low latency infrared LED link for mobile backhaul

Specifications
- Infrared LED based
- Easy alignment:
  - 500 Mbps over 100 m
  - 250 Mbps over 200 m
- Bidirectional data exchange
- Dynamic rate adaptation
- Latency: < 2 ms
- 1 GbE chipset and interface
- Footprint and weight:
  - 240 mm x 230 mm x 130 mm, 3 kg

Benefits
- Low cost optical wireless link based on infrared LEDs

Applications
- Wireless point-to-point communication in industrial environments
- Backhauling for WiFi and LTE
- Building to building connectivity
- Redundancy for fixed line connection

E-Mail: products-pn@hhi.fraunhofer.de
Web: www.hhi.fraunhofer.de/LED-Backhaul
Patents and Standardization, Part 2: Duties to Disclose Patents to Standards Development Organizations
Characterization of Web Single Sign-On Protocols
From Network Sharing to Multi-Tenancy: The 5G Network Slice Broker
Internet of Things Standards: Who Stands Out from the Crowd?
Standard-Based IoT Platforms Interworking: Implementation, Experiences, and Lessons Learned
If your 5G idea works here...

50,000 soccer fans streaming live-action mobile video. All at once.


HARDWARE + SOFTWARE + PEOPLE = 5G INSIGHTS

5G
Get the latest app notes, white papers and tutorials
www.keysight.com/find/5G-Insight

© Keysight Technologies, Inc. 2016

Agilent’s Electronic Measurement Group is now Keysight Technologies.
Machine to machine communication is not new. The basics of control systems have been with us for decades in areas like factory automation and power control. Machine to machine applications and standards have developed to meet specific needs in specific industries. When we start to connect things together more flexibly, we open a whole new range of applications and the need for new standards.

The Internet of Things (IoT), then, envisages the design and implementation of Internet-based systems that connect devices that interact with the physical environment. These devices can share information by connecting to the Internet, giving us much more freedom to use and re-use the information they provide for a wide range of use cases we can see today, and many more not yet thought of in the future. Several industries stand to be transformed by moving to the Internet of Things, such as the automotive, healthcare, and utility industries. Looking back at the changes that the Internet and mobility have brought in the last decade hints that the Internet of Things is going to make life very different in the years ahead.

By 2020 it is expected that 5G networks will be deployed with IoT applications and machine type communication representing key use cases. 5G networks will be multi-purpose and will support different virtual networks with different characteristics. Using a common infrastructure will stimulate many new business models by removing the need for separate infrastructure investment for different verticals. Standardization will be fundamental in defining this common infrastructure to influence mass market adoption of IoT.

The importance of standards to the work and careers of communications practitioners is the basis of this publication. It is a platform for presenting and discussing standards related topics in the areas of communications, networking, research, and related disciplines.

This issue of the Communications Standards Supplement contains a number of open call papers, several of which are focused on IoT and 5G. Readers will notice the ongoing Commentary section with a recurring view from the IEEE-SA President. The IEEE Standards Education section returns to create awareness and understanding of the importance of standardization and the critical role standards play in industry and society. The Standards News section offers the current status of standards work in various SDOs relevant to service management and information modeling, as well as pointers to SDO material. I trust that the reader will find these informative and illustrative of the fundamental role standards play in the communications networking ecosystem.

The first article in this issue is the second in a series on patents and standardization by Krista Jacobsen. In Part 2, Jacobsen provides us with an overview of the IEEE’s patent disclosure policy and identifies several issues the policy raises for standards development organization participants and parties who own patents and patent applications that are subject to the policy. The article presents an example of the consequences of failing to disclose the existence of relevant patent assets during the development of a standard. The article also identifies several issues the IEEE’s patent policy raises for patent right owners and SDO participants.

Single Sign On (SSO) protocols are today integrated into millions of Web services so end users can authenticate to a third-party identity provider (IdP) for accessing multiple services. The article by Beltran presents a conceptual characterization of Web SSO protocols through their assertions and their features that help preserve the privacy of the user resources involved in SSO. It is shown that cumbersome protocol specifications result in developers without a clear view of the underlying SSO protocol.

Next, the article by Samdanis et al. offers an introduction to network sharing paradigms being studied toward the deployment of future 5G mobile networks. This article provides an overview of the 3GPP standard evolution from network sharing principles, mechanisms, and architectures to future on-demand multi-tenant systems. In particular, it introduces the concept of a network slice broker in 5G systems, which enables mobile virtual network operators, over-the-top providers, and industry vertical market players to dynamically request and lease resources from infrastructure providers via signalling means. Finally, it reviews the latest standardization efforts, including the remaining open issues, for enabling advanced network slicing solutions.

While time-to-market constraints are accelerating the deployment of a variety of fragmented and proprietary Internet of Things (IoT) products, there is still a lack of understanding of what an IoT service is meant to be, what
its consequences are, and how to promote standard IoT services. The article by Meddeb gives a concise and comprehensive survey of IoT service definition, regulation, and standardization activities. Global standards are critical to the success of the IoT. The article summarizes mainstream standards as well as emerging, independent, and state-funded projects for IoT ecosystem.

Following on with IoT standards activities, this issue concludes with an article by Song et al., which introduces standardized interworking interfaces and procedures based on oneM2M standards. The article then evaluates these through use cases with multiple IoT service platforms. The interworking is about smart city applications/services run on multiple IoT service layer platforms interworking with each other. The article concludes that such an interworking function proves that the global IoT standards specifications can foster implementations of a service layer, which enables services and interoperability between devices/device networks and cloud-based applications.

Looking to the remainder of 2016, the Communications Standards Supplement will continue quarterly publication. The next issue will contain articles based on the recent ITU Kaleidoscope, “Trust in the Information Society,” offering insight into the means of building information infrastructures deserving our trust. The final issue of the year will feature another M2M related topic on the “Next Generation of IoT Applications and Architectures.” Proposals for future standards feature topics are welcome as we look forward to evolving into the new IEEE Communications Standards Magazine in 2017.

Biography

Glenn Parsons (SM) (glenn.parsons@ericsson.com) is an internationally known expert in mobile backhaul and Ethernet technology. He is a standards advisor with Ericsson Canada, where he coordinates standards strategy and policy for Ericsson, including network architecture for LTE mobile backhaul. Previously, he has held positions in development, product management and standards architecture in the ICT industry. Over the past number of years, he has held several management and editor positions in various standards activities including IETF, IEEE, and ITU-T. He has been an active participant in the IEEE-SA Board of Governors, Standards Board and its Committees since 2004. He is currently involved with mobile backhaul standardization in MEF, IEEE and ITU-T and is chair of IEEE 802.1. He is a Technical Editor for IEEE Communications Magazine and has been co-editor of several IEEE Communications Society Magazine feature topics. He graduated in 1992 with a B.Eng. degree in electrical engineering from Memorial University of Newfoundland.
IEEE's Industry Connections Program:  
A Collaborative Environment for Developing Shared Results

By Bruce Kraemer, President, IEEE Standards Association

http://standards.ieee.org/develop/index.html

The IEEE Standards Association is primarily known for its extensive set of published standards, for example 802.3 Ethernet, and the process used to bring together experts from around the world to contribute to standards development.

But there are occasions where industry wants to discuss technology trends and options without an immediate plan to produce a standard.

The IEEE Standards Association (IEEE-SA) Industry Connections (IC) program provides a forum for collaboration among organizations and individuals that want to discuss new technology challenges and opportunities. The collaboration within these Industry Connection groups has results in white papers, technical guides, software, databases, conferences, and workshops.

Rather than attempting to generically describe the virtues of the program, I refer you to the fascinatingly diverse range of activities already underway. I present below a brief description of each of the existing programs. I encourage you to obtain further details about an existing program and how to participate at:

http://standards.ieee.org/develop/index.html

Similarly, I will not attempt to use this article to describe details about the process for forming a new Industry Connection Program. That information can be found on the website:

http://standards.ieee.org/develop/groups.html

Note that the project list is provided in chronological order with the most recently started projects listed at the end.

IC09-001 Computer Security Group (ICSG): Computer security experts from around the globe, working together to combat malware and other computer security threats.

IC12-003 IEEE Intercloud Testbed: Bringing industry and research institutions together to gain real-world experience in cloud-to-cloud interoperability with the IEEE Intercloud Testbed.

IC12-006 IEEE Actionable Data Book for Education: Creating and demonstrating prototypes for new methods of activity-based mobile learning via an open-standard e-book format with new forms of interactive learning technology and functionality.

IC13-001 IEEE-SA Symposium on EDA Interoperability: Organizing an annual symposium to help members of the electronics/semiconductor design and verification community better understand the landscape of Electronic Design Automation and semiconductor intellectual property standards, as well as the role of these standards to address industry interoperability challenges.

IC13-002 Electric Vehicle Wireless Power Transfer: Pre-standardization efforts for Electric Vehicle Wireless Power Transfer, with a particular focus on dynamic wireless charging to address issues such as range limitation of electric vehicles and energy storage costs.

IC13-004 IEEE-SA Ethernet & IP @ Automotive Technology Day: Organizing an annual conference and exhibition on the application of Ethernet and Internet standards in the automotive environment.

IC13-005 DC in the Home: Exploring the issues and work to be done to ensure DC electricity can be safely and conveniently accessed in the home, improving energy efficiency by eliminating the wasteful conversions between AC and DC.

IC14-001 Green Power Generation: Bringing together enterprises and research institutions to overcome major issues and discuss potential technical standards relating to cleaner fossil fuel and especially coal-fired power generation in China and beyond.

IC15-001 Fiber Optics Sensors: Companies that manufacture fiber optic sensing systems or components for fiber optic sensing systems working together to identify gaps in standards related to fiber optic sensors and develop a plan to address these gaps.

IC15-002 Smart Glasses Roadmap: Adoption and acceptance of Augmented and Virtual Reality will be dependent on the readiness of technologies to provide a positive and cost effective user experience. This project aims to analyze the markets, use cases, and technology considerations that must be addressed to accelerate enhanced reality technology readiness and adoption.

IC15-003 Smart City Compliance Indicators: Define the factors that determine the “smartness” of a city. Identify the key indicators of “smartness” that are required for a city to be called a smart city and develop a smart city rating index.

IC15-004 3D Body Processing: Bringing together an ecosystem of players to propose new standards around enabling 3D body processing, which includes the capture, processing, storage, sharing, and (augmented) representation for “Of-the-body” and “On-the-body” technologies.

IC15-005 Next Generation Enterprise/Campus/Data Center Ethernet: The growing diversity of applications within enterprise, campus, and data center networks requires new Ethernet standards to be developed. This is evident by recent standardization activities related to 2.5 Gb/s, 5 Gb/s and 25 Gb/s Ethernet, as well as subsequent conversations related to introducing new Ethernet solutions at these rates. The goal of this activity is to assess emerging requirements for enterprise, campus, and data center networks, identify gaps not currently addressed by IEEE 802.3 standards, and facilitate building industry consensus toward proposals to initiate new standards development efforts.

IC16-001 Open Data: Stakeholders assessing existing Open Data initiatives and proposing new standards for making data “open.” Creating a globally accepted definition, format, and structure for Open Data to enable interoperability, whereby Open Data sets from different sources can be easily combined into larger and more valuable data sets.

IC16-002 The Global Initiative for Ethical Considerations in the Design of Autonomous Systems: Bringing together experts in fields relating to autonomous systems (e.g. Robotics, Artificial Intelligence, Computational Intelligence, Machine Learning, Deep Learning, Cognitive Computing, Affective Computing) to identify and address the ethical considerations related to the design of autonomous systems and the issues they involve.

IC16-003 Internet of Things Interest Group: Bringing together participants active in the Internet of Things (IoT) space for collaboration toward mutually beneficial outcomes and deliverables, with special focus on connecting with industry players.

IC16-004 Augmented Reality in the Oil/Gas/Electric
Industry: Stakeholders collaborating to identify requirements, standards needs, and other issues, to help enable augmented reality solutions, as well as potentially mixed and virtual reality solutions, that can benefit the oil, gas, and electric industries.

IC16-006 International Roadmap for Devices: This Industry Connections activity will focus on an International Roadmap for Devices and Systems (“IRDS”) via establishment of an interest group closely aligned with the new electronics industry ecosystem. Activity members will collaborate in the development of this roadmap, as well as engaging with other segments of the IEEE in complementary activities (Rebooting Computing Initiative, Computer Society, and other supporting IEEE societies) that help assure alignment and consensus across a range of stakeholders.

**ATIS: Enabling ICT Industry Transformation**

*By Susan Miller, President and CEO, ATIS*

Since its inception as a standards organization to advance the telecommunications industry’s transformation from regulated monopoly to competitive communications services, the Alliance for Telecommunications Industry Solutions (ATIS) has evolved at the velocity of the information and communications technology (ICT) industry. Today, in addition to the thousands of standards, specifications, and requirements delivered, ATIS generates business use cases, software toolkits, application program interfaces, interoperability guidelines, best practices, and more. Our solutions embrace the opportunity of open source, but also recognize that open source and standards will co-exist, both playing a role in our industry’s transformation. Educating regulators on evolutionary topics such as the 5G future is an important area where ATIS also engages.

ATIS brings together the industry’s top technology leadership. Our board of directors consists of a diverse group of C-level executives from the leading global ICT companies. Our success critically depends on attracting innovative companies that are disrupting how the more traditional industry players do business. ATIS’s historical roots are in defining standards for traditional telecommunications service providers and equipment vendors. As competition changed the industry landscape, Internet and cable operators joined ATIS, expanding the organization’s diversity and its suite of solutions. This expansion has continued in 2016, with ATIS welcoming Google and Facebook to its membership ranks. Combined with our Board’s insight and guidance, aligning the broad and diverse ecosystem of stakeholders and engaging them in dialogue are what help us to deliver the full opportunity of the future. Our approach positions ATIS to address technologies that will disrupt ICT, specifically the game-changing technologies that will impact compute, storage, and the network, to better understand how they will drive future innovation.

Accelerating change, innovating, and advancing industry transformation are central to ATIS’s mission. We embrace and define forward-looking opportunities through our Innovation Agenda. The Agenda’s objective is to set the stage for industry alignment on priorities that reflect the industry’s direction over the next 18 months. One of its first major initiatives is a roadmap for the 5G future. The Agenda has also launched ATIS’s cybersecurity work, which focuses on emerging threats. It is defining the evolutionary path from today’s IP-based routing network to a future network that leverages the increasingly important role of content. Slated for future engagement is the creation of a technology roadmap to guide Smart Cities planners and accelerate investments in infrastructure.

**A History of Advancing Transformation**

ATIS has a 30-year history in bringing the industry together to align on and speed transformation. For example, ATIS developed the SONET-SDH synchronous optical standards, including the European version that later was adopted by the ITU-T as the first-ever global optical standard. Although the original SS7 specification pre-dates ATIS, much of the ongoing evolution of the standard was driven by ATIS and then adopted by ITU-T. This work laid the foundation for today’s optical networks and intelligent services that have helped our industry handle the exponential increase in data rates.

ATIS was also a catalyst in advancing high-speed Internet. ATIS’s work contributed some of the major broadband Internet standards, including those guiding successful mass-market deployment of services such as DSL and VDSL.

Not only has our work advanced network speeds, it also has transformed quality of service. Innovation in devices is now pushing the need for an expanded definition of quality. In response, ATIS has contributed a wide range of performance, reliability, and quality of service standards to improve the processing of voice, audio, data, image, and video signals, as well as their multimedia integration.

**The Future Unfolding Now**

With the emergence of mobility, ATIS was a member of the visionary team that founded the Third Generation Partnership Project (3GPP) in 1998, a pivotal force in enabling global wireless connectivity and roaming. As the founding North American Organizational Partner of 3GPP, ATIS created standards and solutions for wireless systems and services, from the creation of the first domestic Personal Communications Service standards, to development of global specifications for 3G, 4G, and now pointing to 5G. ATIS has become a major contributor and leader in the development of global wireless specifications.

ATIS is also an industry leader in advancing the all-IP network. Through a joint task force with the SIP Forum, ATIS delivered the first standardized, industry-based specification designed to achieve IP-based interconnection of all service provider networks. ATIS is also leading a testsbeds initiative to validate key aspects of the all-IP network transformation. This work addresses emerging solutions in areas including numbering evolution, IP-NNI routing, and authenticated caller-ID. Already, an assessment of key testing requirements has been conducted and has identified nine business use cases of interest for an ATIS testsbed. These cover number provisioning, IP-NNI routing, and secure caller-ID.

The ATIS-SIP Forum IP-NNI Task Force is currently developing a specification for a “verified token” in cooperation with the Internet Engineering Task Force (IETF). The verified token leverages service provider infrastructure to provide an important enabling for more effective approaches to combat caller-ID spoofing. This work is specifically geared toward adding value to service provider offerings while giving consumers important tools to combat the problem.

As ATIS continues to transform the communications eco-
Commentary

system for users, it is also advancing machine-to-machine communications. ATIS is a founding partner of oneM2M, the global organization for standards for M2M and the Internet of Things. oneM2M is developing a common service layer that can benefit vertical markets such as the automobile industry, the healthcare industry, and the smart grid, all of which will be central to the surge in M2M communications. This work is critical as these communications are expected to exceed that generated by the sum of all human voice conversation on the world’s wireless networks. oneM2M has a strong security focus, which is vital to protect M2M applications and avoid the type of high-profile security lapses we have seen in other systems. Also, oneM2M has named ATIS as the initial Registry Management Authority for the oneM2M app-ID registry, developed by iconectiv, which is essential for accelerating the adoption of open M2M systems.

In the transformation from hardware to software-centric networks and the growing virtual network operator movement, ATIS plays a key role. Our Network Functions Virtualization (NFV) Forum leverages new technology to enable innovative business models based on the use cases that matter most to ATIS members. These include the use of NFV services between providers to deliver cheaper and more efficient mobile roaming, CDNs, and other applications. Underlying these is the use of virtualization to dynamically combine resources from several providers. Beyond a traditional network-to-network interface, open standards for exposure of NFV capabilities enable operators, enterprises, and web companies to construct services using virtual resources chosen, on demand, from a catalog of available functions.

NFV will create both new opportunities and security challenges. It is critical that the transition to NFV preserves the good operational security achieved in today’s carrier networks. ATIS has taken the initiative to work with operators to define security approaches that are preparing the network for the NFV world.

This brings us to the present, which, at ATIS, is always focused toward the future. In 2016, more than ever, ATIS is helping the industry advance the opportunities brought our way by new technology, member business imperatives, and consumer preference. In an industry that is constantly changing, ATIS has been adept in transforming itself. As our industry evolves, ATIS will continue to pivot with it, and serve as a critical force in advancing industry transformation.

“The best way to predict the future is to invent it.”

-Alan Kay
CALL FOR PAPERS AND PROPOSALS

IEEE ICC 2017 will be held at Palais des Congrès - Porte Maillot, Paris, France, 21-25 May 2017. Located in the heart of the City of Lights, IEEE ICC 2017 will exhibit an exciting technical program, complete with 13 Symposia highlighting recent progress in all major areas of communications. IEEE ICC 2017 will also feature high-quality Tutorials and Workshops, Industry Panels and Exhibitions, as well as Keynotes from prominent researchers and industry leaders.

Prospective authors are invited to submit high-quality original technical contributions for presentation at the conference and publication in the IEEE ICC 2017 Proceedings and IEEE Xplore. Proposals for Tutorials, Workshops, and Forums are also invited. Visit http://icc2017.ieee-icc.org for more details.

TECHNICAL SYMPOSIA

SELECTED AREAS IN COMMUNICATIONS

Access Systems and Networks
Steven Horowitz, McMaster University, Hamilton, Canada

Big Data Networking
Shu Yu, Deakin University, Australia

Cloud Communications and Networking
Ioannis Papaparmpilou, Purdue University, USA

Communications for the Smart Grid
Deepa Kundur, University of Toronto, Canada

Data Storage
Onur Ozturk Kaykobad, University of Arizona, USA

E-Health
Jaime I. Lorent Mañé, Polytechnic University of Valencia, Spain

Internet of Things
Antonio Jara, University of Applied Sciences Western Switzerland (HES-SO), Switzerland

Molecular, Biological, and Multi-Scale Communications
Umberto Ona, University of Southern Californa, USA

Satellite and Space Communications
Igor Blažek, University of Sera, Italy

Social Networking
Damir Turogled, University of Central Florida, USA

Software Defined Networking & Network Function Virtualization
Laszlo Lelied, Luxembourg University, Luxembourg

Ad-Hoc and Sensor Networking
Shibbo He, Zhejiang University, China
Aden Kandala, University of Rennes, France
Cheng Li, Memorial University of Newfoundland, Canada

Cognitive Radio and Networks
Oliver Holland, King’s College London, UK
Wald Saad, Virginia Tech, USA

Communication and Information Systems Security
Peter Mueller, IBM, Switzerland
Cong Wang, University of Hong Kong, China

Communications Software, Services and Multimedia Applications
Maria-G. Martin, Kingston University London, UK
Said Henein, Universite Paris Saclay, France
Said Rami, Western University, Canada

Communication Theory
Ender Ayanoglu, University of California, Irvine, USA
Fulvio Babin, University of Trieste, Italy
Steven Weber, Drexel University, USA

Green Communications Systems and Networks
Czech Cavdar, Royal Institute of Technology Stockholm, Sweden
Michelae Hoa, Politecnico di Torino, Italy

Communications QoS, Reliability and Modeling
Dzmitry Kliazevich, Luxembourg University, Luxembourg
Samir Souit, UPEC, France
Kohel Shimoda, NTT, Japan

Next Generation Networking and Internet
Shihen Mao, Auburn University, USA
Mahesh K. Marin, University of Edinburgh, UK
Sidi-Mohammed Senouci, University of Bourgogne, France

Signal Processing for Communications
Michael Buehrer, Virginia Tech, USA
Tomohisa Taniguchi, Fujitsu Laboratories Limited, Japan

Optical Networks and Systems
Grzegorz Danelwicz, Poznan University of Technology, Poland
George Roukas, North Carolina State University, Raleigh, USA

Wireless Communications
Mohammad Assaad, CentraleSupelec, France

Azeddine Benslimane, University of Ottawa, Canada
Davide Dardari, University of Bologna, Italy
Hamid Reza, Aalto University, Finland
Yahong Rasa Zheng, Missouri University of Science and Technology, USA

Mobile and Wireless Networking
Mobayed Alii, University of Pittsburgh, USA
Menfi Bennis, University of Ou, Finland
Jaeil Ben-Offman, University of Paris 13, France
Shuaiwei Wang, Nanyang University, China

INDUSTRY FORUMS AND EXHIBITION PROGRAM

IEEE ICC 2017 will feature several prominent keynote speakers, major business and technology forums, and a large number of vendor exhibits. Submit your proposals to the IFIE Co-Chairs Luis M. Carreira (luc.carreira@vtt.fi) and Jamshid Khan-Jush (khanjush@ntelos.com).

TUTORIALS

Proposals are invited for half- or full-day tutorials in all communication and networking topics. For inquiries, please contact the Tutorials Program Co-Chairs Hanna Bogucka (hanna.bogucka@jpl.nasa.gov) and Luc Vandendorpe (luc.vandendorpe@louvain.be).

WORKSHOPS

Proposals are invited for half- or full-day workshops in all communication and networking topics. For inquiries, please contact the Workshops Program Co-Chairs Abbas Jamali-Pour (a.jamali-pour@ieee.org) and Constantinos Papadis (cpap@ait.gr).

Full details of submission procedures are available at icc2017.ieee-icc.org

#IEEEICC17
Wireless Power Transfer: “Look Ma, No Hands, No Wires!”

By Yatin Trivedi, Member, IEEE Standards Education Committee

Today, most of us take access to Wi-Fi (IEEE Std. 802.11n™) for granted. That is what makes our daily lives go around at home, in the office, or on the road with access to internet access across the Internet. Compared to the ‘tethered’ Internet access of a few years ago, we became so much more mobile and productive having “cut the cord.” But one aspect of our technology usage that is still tied to cords is the need for power, even when you have battery operated devices. Think about all of the devices in our lives that require batteries to be recharged: wireless phones, laptops, electric vehicles, medical devices, all the fun wearables, and so much more. What if you didn’t need to plug in a device to recharge its battery? Or what if your device didn’t even need a battery for its power supply? There are a number of standards currently available and in development that support that capability through wireless power transfer (WPT) that can allow us to “cut the cord” to traditional power sources.

WPT can help make our lives much easier and can help support cleaner, greener energy options. There is a lot to consider in this area such as compatibility, spectrum availability, and health/safety concerns. The articles in this Communications Standards Supplement explore the work that has been done recently and is currently taking place around the world for WPT standards in a variety of fields.

In consumer electronics, there are currently a number of competing standards for WPT, even within the same Standards Development Organization (SDO). The question is then raised for manufacturers as to which standard they should choose for their products? As some of you may have experienced in the last year or two, several popular coffee shops offer you the ability to recharge your smartphone. However, it requires that your smartphone has certain wireless charging capability (hardware features) that are compatible with the charger built into the table or counter top at the coffee shop. Such compatibility requirements are defined by the existence and product implementation of the WPT standards. So which WPT standard to implement is a critical decision that can affect how well a product may perform in the market and be adopted by consumers. Similar concerns arise for car manufacturers, as the technology needs to be compatible and work properly with devices installed at charging locations. Applications of WPT are numerous and go beyond what has already been seen in the market today. Understanding the role of standards in this area can help to facilitate advancements in current and new technologies.

As you can see, industry agreement on standards plays a critical role in the development of technologies that will work as promised and be successfully accepted by consumer markets. Our hope is that as you read this issue of the Communications Standards Supplement, you will learn about the importance of collaboration in standards development in wireless power transfer, the impact of standards on these technologies, and why education about standards is critical for moving technologies in this area forward. If you are wondering how WPT affects you, just think about the recharging convenience offered at your favorite coffee shop, or the time you forgot the charger for your phone. If you are among those who drive/own an electric vehicle, you know the need for recharging your vehicle’s batteries. Wouldn’t it be great if you could recharge while you are driving (dynamic wireless charging)? In order to make these applications possible, the industry needs standards and products built on (in compliance with) these standards.

Happy reading!

The IEEE Standards University e-Magazine

This online publication, available at http://www.standardsuniversity.org/e-magazine/, is sponsored by the IEEE Standards Education Committee, a joint committee of the IEEE Educational Activities Board and the IEEE Standards Association. Serving the community of students, educators, practitioners, developers, and standards users, we are building a community of standards education for the benefit of humanity.

IEEE’s Standards Education Program is committed to:

• Promoting the importance of standards in meeting technical, economic, environmental, and societal challenges.
• Disseminating learning materials on the application of standards in the design and development aspects of educational programs.
• Actively promoting the integration of standards into academic programs.
• Providing short courses about standards needed in the design and development phases of professional practice.

The June issue of the IEEE Standards Education eMagazine explores the work that has been done and is currently taking place around the world for wireless power transfer (WPT) standards in a variety of fields. Several of the articles are included here, and the abstracts are shown for the rest.

Abstract

Smartphone technologies are the key indicators of consumer electronic trends. New technologies and features continue to emerge such as USB type C, 4K video, and immersive VR gaming. Wireless charging is one of these technologies that has been in use since late 2000, but is still not used by the masses in smartphones. The answer to the compatibility question is the eco-system; the standard. There are multiple standards in the world, but their technologies are incompatible at present. This article addresses current wireless power standards for mainstream consumer electronics, the challenges related to the different technologies, and new developments in this area.

Reprint of article, “Wireless Charging of Consumer Electronics: Rubbish Heap or Mass Adoption?” from Wireless Week

By John Perzow

Abstract

In this reprint of the 19 February 2016 article from Wireless Week, the author explores the possible commercial demand for wireless charging technologies in consumer electronics. Factors that drive mass adoption of the technology are discussed.

“IEEE Brings Standards Education to Capstone Design Conference”

By Jennifer Mcclain

Abstract

“IEEE Brings Standards Education to Capstone Design Conference”
Abstract
The Capstone Design Conference, held every two years, offers a forum for faculty, administrators, industry representatives, and students to share ideas about improving design-based capstone courses. Capstone courses, also referred to as senior design courses, are for undergraduate engineering students in their last year of study. This article describes an innovative new workshop for educators that utilizes an interactive consensus-building exercise that simulates the activities of a standards working group tasked with developing a new technical standard.

Student Application Paper: “Implementation of a Qi Compliant Wireless Power System for an Underwater Probe”
By Heitor M. Santos, Luiz F. O. Chamon, and Cassio G. Lopes, Signal Processing Lab, Electronic Systems Engineering Department, University of Sao Paulo, Brazil
Abstract
Underwater sensing probes are used in a myriad of applications, for instance, oceanographic profiling. These probes present a design challenge as they are submitted to harsh marine environments and large pressures, which makes their waterproofing intricate. A solution to this issue would be to completely and permanently seal the probe's shell. This, however, complicates the access to the internal circuit, which poses the problem of how the equipment’s batteries would be charged. This work addresses this issue by implementing a Qi-compliant wireless power transfer system. Introducing a layer model of the standard, each part of the power transmitter and receiver modules are designed, mounted, and tested. Even though the purposes of this system is battery charging, the wireless power modules described in this white paper could also be used in other applications.

Featured Video: Meet the Author video featuring Ron Schneiderman
Abstract
This IEEE.tv video features Ron Schneiderman, the author of the book, Modern Standardization: Case Studies at the Crossroads of Economics, Technology and Politics. The full video can be viewed at https://ieeetv.ieee.org/meet-the-authors

Featured Courses: Practical Ideas from Professors
Abstract
To help professors around the world include more information about technical standards and standardization in their academic curricula, in 2013 the IEEE Standards Education Committee created the “Practical Ideas from Professors” series. Integrated in 2016 as part of the new IEEE Standards University library, the series features two-page installments authored by university educators who have experience successfully incorporating standards education into their classes. Best practices and methodologies for introducing and exploring standards topics in academic environments are discussed, and examples of ways to include standards in lectures, projects, and assignments are provided. The “Practical Ideas” series highlights input from professors representing a diverse geographic area as well as a wide range of technical subjects, including biomedical engineering; environmental assessments; communications; history and technology studies; and electromagnetics. In early 2016 the number of installments in the library nearly doubled with the addition of the latest contributions. View the full library of Practical Ideas from Professors at http://www.standardsuniversity.org/courses/practical-ideas-from-professors/
Standards Education

Research LED teaching

Standards and regulations aware research

Standards and regulations aware teaching

Figure 1. A possible scheme for enhancing the diffusion of standards and regulations into teaching.

The term research-led is given to the aspiring engineer. In this way, standards and regulations can be more easily integrated into coursework instead of teaching a general course dedicated to standards and regulations (Fig. 1).

Utilization of ISM and Other Frequency Bands

Frequencies are not given for free. In fact, spectrum as a scarce resource is quite expensive, and in most cases frequencies are licensed with a substantial fee. However, when the interference potential of a device is low (e.g., a short range device) then this device can operate as a secondary user without paying frequency fees. There are some specific frequency bands that are intended for industrial, scientific, and medical (ISM) uses. The use of these frequencies requires no licensing, is in general straightforward, and thus popular. On the other hand, as frequency of operation goes higher, the interference potential gets stronger, and subsequently an allocation of a frequency band for dedicated use as a separate service becomes necessary. Many recent realizations of WPT circuits are in ISM frequency bands. A brief survey of publications in IEEE and other journals reveals a preference for the lower frequencies up to 13.56 MHz (Table 1).

Developments for WPT Devices in the ITU

The ITU is the global ICT regulator. Whenever a new wireless technology appears, frequency allocation is the most important issue to consider. This allocation is greatly facilitated by the input of interested stakeholders and countries that develop these technologies. For example, countries with a large automotive industry (Japan and Korea in particular) have already developed national standards and regulations for near field systems [3]. A global frequency allocation goes through the ITU and in specific ITU-R (radio communications), one of the three ITU divisions. A global effort is necessary for spectrum harmonization across the globe, which greatly facilitates economies of scale.

ITU-R deals with new technologies by employing specialized teams of experts called study groups (SGs). WPT matters fall under the jurisdiction of SG1 (study group 1). In order to expedite the developments, subgroups are formed whose task is to answer specific mandates of a formal document called Question. In the case of WPT, this Question bears the code 230/1 and was issued in 2012 to be handled by technical subgroup working party 1A. The proceedings of this ongoing effort can be found online [12].

A Question is considered resolved when recommendations

<table>
<thead>
<tr>
<th>No</th>
<th>Frequency (MHz)</th>
<th>ID</th>
<th>Application</th>
<th>2.45 GHz band</th>
<th>5.8 GHz band</th>
<th>900 MHz band</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>0.020</td>
<td>1.</td>
<td>Wireless powered sensor network</td>
<td>•</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>0.125</td>
<td>2.</td>
<td>Wireless charger of mobile devices</td>
<td>•</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>0.200</td>
<td>3.</td>
<td>WPT in a metal pipe</td>
<td>•</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>6.78</td>
<td>4.</td>
<td>Microwave buildings</td>
<td>•</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>13.56</td>
<td>5.</td>
<td>2D (surface) WPT</td>
<td>•</td>
<td>•</td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td>27.12</td>
<td>6.</td>
<td>Wireless charging for electric vehicles</td>
<td>•</td>
<td>•</td>
<td></td>
</tr>
<tr>
<td>7.</td>
<td>40.88</td>
<td>7.</td>
<td>Point-to-point WPT</td>
<td>•</td>
<td>•</td>
<td></td>
</tr>
<tr>
<td>8.</td>
<td>433.92</td>
<td>8.</td>
<td>WPT to moving/flying target</td>
<td>•</td>
<td>•</td>
<td></td>
</tr>
<tr>
<td>9.</td>
<td>915</td>
<td>9.</td>
<td>Solar power satellite</td>
<td>•</td>
<td>•</td>
<td></td>
</tr>
<tr>
<td>10.</td>
<td>2,450</td>
<td>10.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11.</td>
<td>5,800</td>
<td>11.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Most popular frequency bands in use by WPT systems.

Table 2. Applications of wireless power transfer in far field and potential frequency bands [14].
and reports are developed. The text with the regulatory force is the recommendation, with much more technical detail given in an associated report. In the area of near field (or non-beam as they are termed by the ITU), the ITU report has now been published [13]. Recent contributions are aiming toward a report and a recommendation for FF, e.g. see [12, 14, 15], and [16] from Japan. The input from industry and research institutes is significantly eased by alliances such as the Wireless Power Transfer Consortium for Practical Applications (WiPoT) of Japan [17].

Besides charging, there are many interesting applications in the far field which cannot be covered by near field, e.g., the possibility of a space solar power satellite [11], as well as other possibilities which are shown in Table 2. The preferred frequency bands of interest for WPT-FF are 2.45 GHz and 5.8 GHz.

Conclusion

There are a multitude of parallel activities in the regulatory and standards organizations in the field of WPT. Diffusion of these activities and awareness in research programs in the field of WPT is necessary. EU COST program WiPE was described as a case study of how standards and regulations can be explicitly integrated in research. Furthermore, it was stressed that WPT-FF requires more standardization work and regulations that in turn need to be informed by research studies. Contributions by Japan in the area of far field WPT have been described in the context of recent ITU regulatory activities.

Acknowledgment

The work of C. Kalialakis, N. Carvalho, and A. Georgiadis was supported by EU COST Action IC1301 WiPE. The work of C. Kalialakis and A. Georgiadis was further supported by EU H2020 Marie Sklodowska-Curie grant agreements No. 661621 and No. 654743.

References


Biographies

Christos Kalialakis (christos.kalialakis@cttc.es) was awarded the Ph.D. Degree in electronic and electrical engineering from the University of Birmingham, UK, in 1999. Since 2002, he has been employed at the National Regulatory Authority of Greece for ICT as an expert on wireless communications, and has been a member of the national delegation in International regulatory and standardization organizations (ITU, ECC). He is currently a senior researcher at CTTC, Spain. He serves on the editorial board of Wireless Power Transfer (Cambridge University Press) and the Journal of Communications. He has been an IEEE Senior Member since 2009.

Nuno Borges Carvalho (SP’07, M’00, SM’05, F’15) (nborreges@cttc.es) received the Doctoral degree in electronics and telecommunications engineering from the University of Aveiro, Portugal, in 2000. He is currently a full professor and a senior research scientist with the Institute of Telecommunications, University of Aveiro. He is a co-author of Intermodulation in Microwave and Wireless Circuits (Artech House, 2003), Microwave and Wireless Measurement Techniques (Cambridge University Press, 2010), and White Space Communication Technologies (Cambridge University Press, 2014). He is associate editor of the IEEE Transactions on Microwave Theory and Techniques, IEEE Microwave Magazine, and the Wireless Power Transfer Journal. He is co-chair of the IEEE MTT-23 Technical Committee. He is the co-inventor of four patents and the recipient of the 2000 IEEE Measurement Prize.

Noriko Shinohara (shino@rish.kyoto-u.ac.jp) received the Ph.D. (Eng.) degree in electrical engineering from Kyushu University, Japan, in 1998. She has been a professor in the Research Institute for Sustainable Humanosphere, Kyoto University, since 2010. He is an IEEE Distinguished Microwave Lecturer, IEEE MTT-S Technical Committee 23 (Wireless Power Transfer and Conversion) vice-chair, the International Journal of Wireless Power Transfer (Cambridge Press) executive editor, past technical committee chair for IEEE Wireless Power Transfer, and is chair of the Wireless Power Transfer Consortium for Practical Applications (WiPoT).

Apostolos Georgiadis (apostolos.georgiadis@ieee.org) received the Ph.D. degree in electrical engineering from the University of Massachusetts, Amherst, in 2002. He is senior researcher and microwave systems and nanotechnology group leader at CTC-T, Spain. He is vice-chair of EU COST IC1301 on wireless power transfer for sustainable electronics. He is associate editor of IEEE Microwave Wireless Components Letters, editor-in-chief of Cambridge Wireless Power Transfer, past chair of IEEE MTT-24 on RFID Technologies, and a member of IEEE MTT-16 on wireless energy transfer and conversion. He is vice-chair of USRT Commission D. and a distinguished lecturer and VP of conferences of the council on RFID.

SAE publishes TIR J2954, Wireless Power Transfer EV/PHEV

In 2010, SAE International launched its Wireless Power Transfer (WPT) task force for light duty EVs and PHEVs for the purpose of standardizing both the vehicle and ground assembly infrastructure to enable commercialization. The task force, chaired by Jesse Schneider (BMW), published its first document SAE TIR J2954 on May 31st, 2016 called “Wireless Power Transfer for Light-Duty Plug-In: Electric Vehicles and Alignment Methodology.” A link to download SAE TIR J2954 is found below:

http://standards.sae.org/j2954_201605/

The SAE WPT task force is comprised of multiple automakers, EV infrastructure companies, local and national governments, laboratories, universities, wireless technology suppliers, and OEM (original equipment supplier) first tier suppliers (Tier 1) from around the globe (EU, Asia, North America, etc.) as shown in Fig 1. Wireless charging is starting to become mainstream for consumer electronic devices in low power applications. This technology has also the potential to transfer high power in the kilowatt range to enable charging of electric and plug-in electric vehicles. In order to achieve this commercially, standardization is needed.

Wireless power transfer using inductive charging enables electrified vehicles to be seamlessly and automatically charged without operator interaction. Essentially, the two coils from the vehicle and ground need to be aligned to a minimum degree (by parking properly in a spot) to transfer energy through inductance. The customer can charge the vehicle by simply parking an electric vehicle in a SAE J2954 WPT parking space. No additional action is required of the customer to charge as everything is automated. This could be very attractive to electric vehicle (EV/PHEV–plug-in hybrid electric vehicle) customers, by offering charging along with high efficiencies similar to conductive charging (above 85%).

As an analogy to conductive charging (e.g. SAE J1772), wireless charging can be interoperable and charge using the SAE J2954 specific common specific frequency band, coil guideline and topology with a high coupling factor. The com-
component needed to make this happen wirelessly from the vehicle to the ground assemblies are shown in Fig. 2.

The TIR J2954 guideline establishes guidance for the inductive coils, electromagnetic limits, and interoperability for the testing and demonstration phase of the technology. This entails specifying a common frequency of operation at the 85 kHz band (81.39 kHz–90 kHz) for interoperability. In addition, there are power levels in four classes up to 3.7 kW (WPT 1), 7.7 kW (WPT 2), 11 kW (WPT 3), and 22 kW (WPT 4).

In order to harmonize electromagnetic field (EMF) limits, the latest international levels for electric and magnetic fields limits from ICNIRP (International Commission on Non-Ionizing Radiation Protection) were utilized. The American Association of Medical Instrumentation (AAMI) has also been consulted regarding EMF compatibility related to medical devices such as pacemakers. The EMF limits published in TIR J2954 have been harmonized by combining the AAMI limits with ICNIRP values.

One of the key components of standardization is the inductive coil geometry and topology specification for both the vehicle assembly (VA), mounted under the vehicle pan, and the ground assembly. The inductive vehicle/ground coil set enables a performance-based interoperability and reference point for charging that is automated for the customer.

The SAE J2954 task force found consensus to develop a common coil geometry specification and circuit specification for both the vehicle and ground assembly coil set to ensure interoperability of wireless charging. The team has already established a close to commercial master vehicle/ground assembly coil set for the WPT 1 (3.7 kW) power class as a baseline for both the OEMs and infrastructure providers. For the WPT 2 (7.7 kW) power class, the SAE J2954 team has created two reference coil options for the testing phase of the SAE J2954 guideline.

The SAE J2954 team developed a testing protocol for safety and performance of wireless charging using a standardized test stand for bench and full vehicle testing (Fig. 2). The SAE J2954 testing procedure confirms the system electromagnetic field limits, power transfer performance, and safety levels (such as ICNIRP magnetic field levels and temperature limits).

The SAE Wireless Power Transfer task force is working very closely with the U.S. Department of Energy (DOE) in relation to testing of both the bench and vehicle systems (Fig. 3). The US DOE has assigned a number of its laboratories (Idaho National Lab, Argonne National Lab, etc.) to assist in the data validation projects to test industry equipment.

In addition, the SAE J2954 team is in close contact with the U.S. Federal Communications Commission (FCC) in regard to frequency and EMC levels, as well as the international task force CISPR (International Special Committee on Radio Interference) group. The Draft ANSI C63.30 team, which is being led by the FCC, is...
also aligned with the SAE team, which references the charging coils from J2954. Draft ANSI C63.30 is the American national standard of EMC procedures for compliance testing of wireless power transfer products.

For the next standardization phase, planned in 2017, there will be a SAE J2954 WPT recommended practice, which will include a refined definition of the master and reference coils.

In order to support the impending wireless charging vehicle and infrastructures roll out worldwide (slated for 2020), SAE J2954 will be finalized as a standard in 2018 using the results of both the bench testing and actual vehicle field testing.

Book Review
Modern Standardization: Case Studies at the Crossroads of Technology, Economics and Politics, by Ron Schneiderman

In today’s global economy, the importance of the formal study of standards has been highlighted by the new demands of international trade. Politics of standards development and adoption is becoming a complex affair in an era of intellectual property rights. In today’s global market place, the major challenges are: How do standards development organizations (SDOs) keep pace with the creation and development of products driven by new and emerging technologies? How to teach engineering and technology workforce and students the importance and applications of standards?

Globally, more than a half a million engineering and technology standards are considered as the basic building blocks for the development of new products that help drive the processes of compatibility and interoperability. Standards make it easier to compare competing products. As new standards are developed and adopted, they promote international trade and allow technical cooperation between organizations and countries.

Recently published book Modern Standardization — Case Studies at the Crossroads of Technology, Economics, and Politics by Ron Schneiderman (published by IEEE Press/Wiley) address these key challenges. It was written as a result of efforts by the IEEE Standards Education Committee (SEC) to gain a better understanding of what products and services would be useful for standards education at the university level. The book covers up-to-date issues related to ethics, policies, and business strategies in standards developments.

The book is a useful resource for faculty, students, engineers and entrepreneurs. The major strength of the book is its collection of standards-specific case studies which offer an opportunity to combine professors’ teaching preferences with real-world insights into the technical, political, and economic domains of engineering. Students can appreciate how standards experts and SDO working groups institute policies, procedures, and guidelines to develop and establish standards. Students can learn how to select and apply standards in new product design and service. The book is primarily designed to be used as a supplemental resource for a course in standards. There is a curriculum guide available for educators to help them design and implement courses effectively.

The book is also a good reference resource for engineers and entrepreneurs, as it covers a survey of national and international standards development needs for a wide array of technologies such as smart energy grid, cyber security, wireless technologies, vehicles’ black box, electronic design automation (EDA), the Internet and Internet of Things (IoT), and explores the push for open standards. The book presents a discussion on the collaboration efforts between U.S. and European standards organizations for promoting trade through Transatlantic Trade and Investment Partnership (TTIP).

In short, the book is a welcome addition to standards literature and serves as an important resource for standards stakeholders in academia and industry. It informs them how standards experts and SDO working groups establish the policies, procedures, and guidelines to ensure the integrity of the standards development process, thus enabling organizations to develop new products and promote global trade.

Biography
Ron Schneiderman’s (jesse.schneider@bmw.de) work has led to a number of firsts to further electric and fuel cell vehicles such as establishing the emergency response guide, fueling validation, and numerous standards. For over 20 years, he worked in both the U.S. and Germany in automotive management ranging from conventional series development to electric and fuel cell electric vehicles. At BMW North America, he is the manager of fuel cell electric vehicle development. He relocated from the Munich office of BMW, where he was the first program manager of the hydrogen storage system, as well as requirements management. Mr. Schneider established the SAE wireless power transfer standardization for the PHEVs team in 2010 and continues to chair the task force.

Book Review
Modern Standardization: Case Studies at the Crossroads of Technology, Economics and Politics, by Ron Schneiderman

In today’s global economy, the importance of the formal study of standards has been highlighted by the new demands of international trade. Politics of standards development and adoption is becoming a complex affair in an era of intellectual property rights. In today’s global market place, the major challenges are: How do standards development organizations (SDOs) keep pace with the creation and development of products driven by new and emerging technologies? How to teach engineering and technology workforce and students the importance and applications of standards?

Globally, more than a half a million engineering and technology standards are considered as the basic building blocks for the development of new products that help drive the processes of compatibility and interoperability. Standards make it easier to compare competing products. As new standards are developed and adopted, they promote international trade and allow technical cooperation between organizations and countries.

Recently published book Modern Standardization — Case Studies at the Crossroads of Technology, Economics, and Politics by Ron Schneiderman (published by IEEE Press/Wiley) address these key challenges. It was written as a result of efforts by the IEEE Standards Education Committee (SEC) to gain a better understanding of what products and services would be useful for standards education at the university level. The book covers up-to-date issues related to ethics, policies, and business strategies in standards developments.

The book is a useful resource for faculty, students, engineers and entrepreneurs. The major strength of the book is its collection of standards-specific case studies which offer an opportunity to combine professors’ teaching preferences with real-world insights into the technical, political, and economic domains of engineering. Students can appreciate how standards experts and SDO working groups institute policies, procedures, and guidelines to develop and establish standards. Students can learn how to select and apply standards in new product design and service. The book is primarily designed to be used as a supplemental resource for a course in standards. There is a curriculum guide available for educators to help them design and implement courses effectively.

The book is also a good reference resource for engineers and entrepreneurs, as it covers a survey of national and international standards development needs for a wide array of technologies such as smart energy grid, cyber security, wireless technologies, vehicles’ black box, electronic design automation (EDA), the Internet and Internet of Things (IoT), and explores the push for open standards. The book presents a discussion on the collaboration efforts between U.S. and European standards organizations for promoting trade through Transatlantic Trade and Investment Partnership (TTIP).

In short, the book is a welcome addition to standards literature and serves as an important resource for standards stakeholders in academia and industry. It informs them how standards experts and SDO working groups establish the policies, procedures, and guidelines to ensure the integrity of the standards development process, thus enabling organizations to develop new products and promote global trade.

Biography
Ron Schneiderman’s (jesse.schneider@bmw.de) work has led to a number of firsts to further electric and fuel cell vehicles such as establishing the emergency response guide, fueling validation, and numerous standards. For over 20 years, he worked in both the U.S. and Germany in automotive management ranging from conventional series development to electric and fuel cell electric vehicles. At BMW North America, he is the manager of fuel cell electric vehicle development. He relocated from the Munich office of BMW, where he was the first program manager of the hydrogen storage system, as well as requirements management. Mr. Schneider established the SAE wireless power transfer standardization for the PHEVs team in 2010 and continues to chair the task force.

Biography
Ron Schneiderman’s (jesse.schneider@bmw.de) work has led to a number of firsts to further electric and fuel cell vehicles such as establishing the emergency response guide, fueling validation, and numerous standards. For over 20 years, he worked in both the U.S. and Germany in automotive management ranging from conventional series development to electric and fuel cell electric vehicles. At BMW North America, he is the manager of fuel cell electric vehicle development. He relocated from the Munich office of BMW, where he was the first program manager of the hydrogen storage system, as well as requirements management. Mr. Schneider established the SAE wireless power transfer standardization for the PHEVs team in 2010 and continues to chair the task force.

Biography
Ron Schneiderman’s (jesse.schneider@bmw.de) work has led to a number of firsts to further electric and fuel cell vehicles such as establishing the emergency response guide, fueling validation, and numerous standards. For over 20 years, he worked in both the U.S. and Germany in automotive management ranging from conventional series development to electric and fuel cell electric vehicles. At BMW North America, he is the manager of fuel cell electric vehicle development. He relocated from the Munich office of BMW, where he was the first program manager of the hydrogen storage system, as well as requirements management. Mr. Schneider established the SAE wireless power transfer standardization for the PHEVs team in 2010 and continues to chair the task force.

Biography
Ron Schneiderman’s (jesse.schneider@bmw.de) work has led to a number of firsts to further electric and fuel cell vehicles such as establishing the emergency response guide, fueling validation, and numerous standards. For over 20 years, he worked in both the U.S. and Germany in automotive management ranging from conventional series development to electric and fuel cell electric vehicles. At BMW North America, he is the manager of fuel cell electric vehicle development. He relocated from the Munich office of BMW, where he was the first program manager of the hydrogen storage system, as well as requirements management. Mr. Schneider established the SAE wireless power transfer standardization for the PHEVs team in 2010 and continues to chair the task force.
IoT Standards toward Its Next Stage
By Prof. Song JaeSeung, Sejong University, Seoul, South Korea

Internet of Things (IoT) is moving toward the most significant revolution that humanity has ever experienced, changing everything around people, from the way people live to how they work via connecting everything to the Internet. Therefore, IoT has been attracting researchers in both industry and academia. As the pace of IoT deployments accelerates, IoT standards are poised to move to their next stage of development with several pillars of new technology.

For instance, supporting global interoperability is becoming a major hurdle to achieve mass IoT adoption. oneM2M, which is a global IoT/M2M standard for service layer platforms, has been developing a semantics enabling service interoperability via annotating semantic information to IoT resources. ITU-T has created a new activity to coordinate various IoT standards through empirical gap analysis.

In addition, several SDOs are now looking at the possibility to integrate AI technologies to IoT in order to support intelligent IoT services. As IoT will produce a large amount of data from various resources around people, there is a strong need to provide a standardized mechanism analyzing the deluge of such data and finding valuable information that can, for example, help cities predict accidents. Many existing IoT standards have been developed based on the RESTful architecture style, so that IoT applications are usually designed to understand RESTful APIs. In order for IoT applications to utilize AI-driven features such as machine learning algorithms and big data analytics, standardized ways to use such features need to be developed.

Finally, as IoT has already pointed to the idea of a fourth industrial revolution that will trigger a paradigm shift in manufacturing, IoT is getting the majority of attention from various industrial domains. IoT standards are now also moving toward standardizing new features that will help accelerate the adoption of IoT across various industries. For example, oneM2M is now preparing to develop a new standard for DDS binding to leverage the advantages of the DDS protocol such as scalability, performance, and QoS. The IEEE P2413 standards group has established a collaboration with the Industrial Internet Consortium (IIC) in order to develop a reference architecture that can be referenced by various industrial IoT architectures.

3GPP System Architecture News
By Andreas Kunz, 3GPP SA2 Group

The current Release 14 is in the final phase for most of the service and system aspects (SA) working groups in 3GPP. The services working group SA1 completed its “5G” study on services and markets technology enablers (SMARTER) and now is moving toward a normative work item that specifies the implementation of the use cases and the related requirements. The SMARTER study was split into four documents covering the following areas: massive Internet of Things (mIoT); critical communications (CriC); enhanced mobile broadband (eMBB); and network operation (NEMO). Moreover, SA1 finalizes within Release 14 the work items on LTE support for V2X services, i.e., vehicle to vehicle (V2V), vehicle to passenger (V2P) and vehicle to infrastructure (V2I), and public safety communications related work items for mission critical group based video communications (MCVideo) and mission critical data (MCData) over LTE.

The system architecture working group (SA2) finished the work item on cellular Internet of Things (CIoT) at an exceptionally late stage for Release 13. This work was also carried out in various other working groups in 3GPP and introduces a new device type, supporting a narrow band (NB) radio, enabling the transmission/reception of low bitrate data via the control plane (CP) or user plane (UP). Furthermore, those CP/UP optimizations were also made available for the normal LTE devices, called wide band (WB) CIoT. A new study for the “5G core” in SA2 within the Architecture for Next Generation System (NextGen) follows up the SA1 study SMARTER on the requirements and use cases. At the moment the relevant key issues were identified for a basic system, and the first high-level interim agreements were reached. Due to the complexity of some of the key issues, they were further split into smaller work tasks to then be addressed by the proposed solutions. The intended plan for completion of the first phase of the NextGen study is December 2016.

The telecom management group SA5 is currently specifying Release 14 management-related procedures (performance, configuration, life cycle, fault) for mobile networks that include virtualized network functions and cooperates with ETSI ISG NFV on those topics. Further, SA5 will start studies on the management and orchestration architecture of the next generation network and services for 5G-related aspects, including, for example, the management of network slicing.

SA6, the working group for mission-critical applications, is currently working on various topics in Release 14. The main topic is to introduce two new mission critical services: mission critical video (MCVideo) and mission critical data (MCData). Other Release 14 work includes studies of mission critical system migration and interconnect between mission critical push to talk (MCPTT) systems and the mission critical communication interworking between LTE and non-LTE legacy systems, such as, for example, Project 25 (governed by the TIA-102 standards) and TETRA (governed by ETSI standards), and enhancements for mission critical push to talk.

Building Cohesion in the Development of IoT and Smart Sustainable Cities
By Nasser Al Marzouqi, Telecommunications Regulatory Authority of United Arab Emirates, and Chairman of ITU-T Study Group 20 and Chairman of the Advisory Board on Smart Sustainable Cities

“Cities are engines of dynamism and creativity. In many respects, cities are the proving ground for our efforts to combat climate change, build resilience and achieve faster, more equitable development progress.” United Nations Secretary-General, Ban Ki-moon

Smart sustainable cities have the capacity to contribute to the achievement of the sustainable development goals (SDGs) by leveraging information and communication technologies (ICTs) to set cities on a development course characterized by environmental sustainability, resilience, and equitable social and economic growth.

To rally the international community around a common understanding of the concept, the International Telecommunication Union (ITU) and the United Nations Economic Commission for Europe (UNECE) have together formulated a definition of smart sustainable cities:

“A smart sustainable city is an innovative city that uses information and communication technologies (ICTs) and other means to improve quality of life, efficiency of urban operation and services, and competitiveness, while ensuring that it meets the needs of present and future generations with respect to economic, social, environmental, as well as cultural aspects.”
ICTs play a crucial role in smart sustainable cities by increasing environmental efficiency across industry sectors, and enabling innovations such as intelligent transport systems and smart water, energy, and waste management. The implementation of ICTs in the interests of sustainable urbanization will assist in improving the quality of urban life, helping us to achieve Goal 11 of the UN Sustainable Development Goals (SDGs): to “make cities and human settlements inclusive, safe, resilient, and sustainable”.

Smart sustainable cities have become a key application area for Internet of Things (IoT) technologies. Integrating IoT technologies into city systems will map these systems in the virtual world, helping us to improve our understanding of how complex urban ecosystems behave and where behavioral change can improve the efficiency and sustainability of city operations.

The Importance of Standardization to the Success of IoT and Smart Cities

In recent years, the convergence of technologies and industry sectors has seen the ICT sector gain a diverse range of new stakeholders. ICT standards bodies are now faced with the challenge of addressing the standardization requirements of the many vertical industries applying ICTs as enabling technologies. This is particularly evident in the field of IoT, where IoT platforms are being developed independently, according to the specific needs of various stakeholders.

The vision of a smart sustainable city is one of a highly efficient “system of systems”, built on the horizontal and vertical integration of city processes, making full use of the data generated by IoT-enabled systems. In the context of smart cities, we cannot afford to maintain ICT infrastructure fragmented in several silos. Smart sustainable cities must have the capability to combine and exploit information drawn from a diverse range of sources in order to build more integrated and holistic management systems, as well as more cost-efficient smart-city solutions. Effective standardization will be essential in achieving this aim, helping us to ensure that we build an inclusive Internet of Things rather than a patchwork of “Intranets of Things”.

ITU-T Study Group 20: Internet of Things and Its Applications, Including Smart Cities and Communities

ITU-T Study Group 20 (SG20) provides government, industry, and academia with a unique global platform to engage and collaborate in the development of international IoT standards. SG20 is building on more than 10 years of ITU-T experience in IoT standardization. The group is tasked with the development of international standards to enable the coordinated development of IoT technologies including machine-to-machine (M2M) communications and ubiquitous sensor networks. An integral part of this study is the standardization of end-to-end architectures for IoT, and mechanisms for the interoperability of IoT applications and datasets employed by various vertical industries. An important part of SG20’s work is the development of standards that leverage IoT technologies to address urban-development challenges, in keeping with international conventions, the UN SDGs and the UN new urban agenda.

SG20 is taking an innovative approach to standardization by placing ITU’s technical expertise in IoT standardization at the service of national and local governments, city planners, and a wide range of vertical industries. This multi-stakeholder approach intends to pave the way to trusted IoT technologies that fully address and anticipate end-user and market expectations.

The Value of SG20 as a Platform for Collaboration

The transition to smart sustainable cities will require trusted information infrastructure capable of supporting an enormous volume and diversity of ICT applications and citizen-driven services. The foundational ICT infrastructure of a smart sustainable city should ensure openness and interoperability, achieved with coordinated adherence to international standards. SG20’s top priority is to build greater cohesion in IoT standardization; success in this endeavor will improve the coordination of smart-city development.

The development of smart sustainable cities requires efficient multi-stakeholder collaboration. ITU-T is in a unique position to achieve this, with its work driven by a diverse membership representing governments, industry, and academic and research institutes.

SG20 has already demonstrated the strengths of ITU-T standardization, building cooperation among various stakeholders to develop standards that provide an equitable basis for the development of IoT and smart sustainable cities.

Update on the Progress of SG20

SG20 has held two meetings in addition to the regular e-meetings of our working groups. The first meeting was held in Geneva in October 2015, and the second in Singapore in January 2016. We match our physical meetings with forums open to both members and non-members of ITU to ensure that our standardization work is informed by the requirements of as broad a range of stakeholders as possible.

SG20 is responsible for the maintenance of the existing ITU standards on IoT, and two new IoT standards developed by SG20 have gained ITU members’ approval.

The new Recommendation ITU-T Y.4702 on “Common requirements and capabilities of device management in the Internet of Things” identifies common parameters for remote activation, diagnostics, software upgrades, and security management to improve the efficiency with which IoT devices and applications are managed. The standard is expected to provide the basis for the development of further standards to enable the large-scale deployment of IoT and M2M communications.

The new Recommendation ITU-T Y.4553 on “Requirements of the smartphone as sink node for IoT applications and services” provides for smartphones to collect IoT data such as monitored health parameters, device status, and video and audio feeds. Smartphones provide Internet connectivity for wearable technologies and home-monitoring devices, giving this standard the potential to support a range of smart healthcare initiatives.

SG20 has also produced a series of documents targeted toward smart cities. A range of supplements to ITU standards includes guidance relevant to establishing the framework for the ICT architecture of a smart city; the integrated management of smart-city systems; intelligent sustainable buildings; and the means to encourage the engagement of a broad range of stakeholders in smart-city projects. Many more standards are nearing completion as we continue to develop 47 draft texts outlined in the SG20 work program.

Advocating for a Worldwide Transition to Smart Sustainable Cities

ITU has teamed up with cities including Dubai, Singapore, Montevideo, Buenos Aires, Manizales, Valencia and Rimini to undertake a two-year trial of the key performance indicators (KPIs) for smart sustainable cities detailed by the ITU standards. Recommendations ITU-T L.1600, L1601, and L.1602. This pilot project will be of great assistance in proving the value of these indicators and encouraging their application by city stakeholders worldwide. The project will also ensure that...
any future refinement of these indica-
tors is undertaken on the basis of cities' experiences with their application.

The new Recommendation ITU-T L.1603 was developed in collaboration by ITU, UNECE, and other UN agencies. It is in its first stage of approval and is soon to be endorsed by the KPI portfolio by highlighting the applicability of these KPIs to assessments of cities’ progress in their pursuit of the urban-development targets of the UN SDGs.

In addition, ITU and UNECE have launched the United for Smart Sustainable Cities (U4SSC) global initiative to offer a platform for city stakeholders to advocate for public policy to encourage the use of ICTs in facilitating the transition to smart sustainable cities.

The work of U4SSC will be supplemented by an Advisory Board on Smart Sustainable Cities consisting of representatives of cities and UN agencies. This Advisory Board will advise standardization experts on the best course of action in refining the ITU-UN Economic Commission for Europe (UNECE) KPIs for smart sustainable cities and aligning these KPIs with relevant international conventions. It is our hope that U4SSC will soon be embraced as a UN-wide initiative on smart sustainable cities. The first meetings of U4SSC and the Advisory Board on Smart Sustainable Cities will be held in Geneva, on 21 and 22 July 2016, respectively.

IOT Standards Vision
By Christian Légaré, President and Chairman, IPSO Alliance, EVP and CTO, Micrium

The majority of the discussion around the IoT is about the cloud and networking. There are many standards development organizations (SDOs) and special interest groups (SIGs) dedicated to these domains. However, for the “Things” in the IoT, there is not a formal organization looking at the standards. The IPSO Alliance is also not a standards organization. However, we do agree that standards are necessary for the Internet of Things (IoT) to truly come into its own.

It is interesting that there are already many standards related to the IoT. In fact, there are probably just a few areas that need more. From the IPSO Alliance’s perspective, the challenge is that there is a system of systems, and currently, each system has its own standards. What is needed, therefore, is the “how to” piece of the puzzle: how to bring all of the available standards together in a functional manner from the point of view of the device, the “thing.”

This is why the IPSO Alliance exists and how we relate to the standards discussion. Our focus is to publish guidelines and best practices, and ultimately, a roadmap to help designers understand where to use which standard and how. The industry must decide if it prefers closed systems, such as iOS, or open systems to allow community participation, more along the lines of Google’s activities. If open systems, this is where standards need to be clearly defined and proven so that they can be used by the rest of the world. Standards work is not complete, and existing ones need to be functionally tested to determine their usability and usefulness as well as any gaps that exist when the various standards are put together.

Interestingly, the standards discussion really starts with the Internet itself. It was built for human communication and simply was not designed for machine communication needed for the IoT. Machine to machine communication revolves around the very frequent transmission of small data packets. Fortunately, this work is being addressed by organizations such as OpenFog and others. Then, the understanding extends to the backend system, web servers, etc., and the data. We know how to store data, but what about real "big data"?

Fundamentally, the IoT is not monolithic; it is the sum of all of these things, which need to be interconnected. If we look at the big picture, there are not only hardware-based standards, but also IT-based ones, and all must be harmonized. As an agnostic and objective organization, we are able to make sure all standards work together, or, if not, identify the pain points that need to be addressed. Identifying the missing links is what will allow the IoT to reach its potential, and provide designers with the right, standards-driven toolkit to achieve the benefits of the IoT.

Update on Wireless Access Standards and Spectrum in ITU-R
By Jose Costa, Ericsson, Ottawa, Canada

Since the article published in the March 2015 issue of IEEE Standards News, significant progress has been made in ITU-R. The ITU Radiocommunication Assembly 2015 (RA-15), 26-20 October 2015, and the World Radiocommunication Conference 2015 (WRC-15), 2-27 November 2015, set the pace for the new study period, 2015-2020. In particular, Resolution ITU-R 56 “Naming for International Mobile Telecommunications” was updated to include IMT-2020, the next generation of mobile systems, referred to as 5G, and Resolution ITU-R 65 “Principles for the process of future development of IMT for 2020 and beyond”, defines the process on the basis of submissions from members and invited external organizations, to be evaluated against the requirements for the future development of IMT.

The ITU Radiocommunication Study Group 5 and its working parties and task group have already held one meeting in 2016. As a result of the work conducted in Working Party 5D (WP 5D), ITU-R issued Circular Letter 5/LCE/59 with an invitation for submission of proposals for candidate radio interface technologies for the terrestrial components of the radio interface(s) for IMT-2020, and an invitation to participate in their subsequent evaluation. The evaluation guidelines, including the criteria and methodology, are to be evaluated by WP 5D in June 2017, in accordance with the detailed timeline and process for IMT-2020 in ITU-R.

WRC-15 considered additional spectrum allocations to the mobile service on a primary basis, and some bands below 6 GHz were newly identified for IMT in various footnotes in Article 5 of the ITU Radio Regulations for those countries wishing to implement IMT (including 470-698 MHz, 1427-1518 MHz, 3300-3400 MHz, 3600-3700 MHz, and 4800-4990 MHz, in a number of countries). Consequentially, WP 5D initiated the development of a draft revision of Recommendation ITU-R M.1036-5 to include the frequency arrangements for the new bands.

WRC-15 also agreed on an agenda item for WRC-19 to study the following bands above 6 GHz for IMT identification: 24.25-27.5 GHz, 37-40.5 GHz, 42.5-43.5 GHz, 45.5-47 GHz, 47.2-50.2 GHz, 50.4-52.6 GHz, 66-76 GHz, and 81-86 GHz, which have allocations to the mobile service on a primary basis; and 31.8-33.4 GHz, 40.5-42.5 GHz, and 47-47.2 GHz, which may require additional allocations to the mobile service on a primary basis. Study Group 5 established Task Group 5/1 to rapidly progress this work. Furthermore, some administrations see opportunities in other bands, for example in the 28 GHz band, and may pursue it outside the WRC process on a national and regional basis.

Other important agenda items for WRC-15 that are relevant for wireless standards include radio local area networks (RLANs) between 5150 and 5925 MHz, intelligent transport systems (ITS), and railway radiocommunication systems. This work has been assigned to Working Party 5A (WP 5A).
Special Member Rates
50% off - Membership for new members. Offer valid through 15 August 2016.

Member Benefits
IEEE Communications Magazine
(electronic & digital delivery)
IEEE Communications Surveys and Tutorials
(electronic)

Member Discounts
Valuable discounts on conferences, publications, IEEE WCET Certification program, IEEE Training courses and other exclusive member-only products.

Join Now!
http://bit.ly/1sdsNno

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.

www.comsoc.org
The author provides an overview of the patent disclosure provisions of two standards development organization (SDO) patent policies and identifies issues those disclosure provisions raise for SDOs, participants and parties who own patent applications and issued patents subject to the patent policies. The author also presents an example to illustrate the possible consequences of failing to disclose the existence of relevant patent assets during the development of a standard.

Krista S. Jacobsen

ABSTRACT

This article provides an overview of the patent disclosure provisions of two standards development organization (SDO) patent policies and identifies issues those disclosure provisions raise for SDOs, participants and parties who own patent applications and issued patents subject to the patent policies. The article also presents an example to illustrate the possible consequences of failing to disclose the existence of relevant patent assets during the development of a standard.

INTRODUCTION

Many formal standards development organizations (SDOs) require, or at least strongly encourage, disclosure of intellectual property assets. But exactly what must SDO participants disclose, when must they disclose it, and what are the consequences of failing to disclose?

Different SDOs have different disclosure policies, typically with some commonalities. This article provides an overview of the patent disclosure provisions of two such policies and identifies some of the issues these disclosure provisions raise for meeting participants and for parties who own patent assets subject to the patent policies. The article also discusses the potential consequences that may flow when an SDO participant knows of relevant issued patents or pending patent applications during the development of a standard, but fails to disclose their existence to the SDO. As one court case from the United States illustrates, the consequences, which may be imposed years after the nondisclosure as a result of a patent litigation lawsuit, can be severe, even if the inventions claimed in the known undisclosed patents or applications turn out not to be essential to implement the standard.

SDO Disclosure Requirements

SDOs provide a forum for companies, often competitors, to agree on common technological solutions so that products complying with a standard will interoperate. While an SDO develops a standard, “there is often vigorous competition among different technologies for incorporation into that standard.” But after a standard has been finalized, the potential for “patent hold-up” arises. As one court stated:

“An SDO may complete its lengthy process of evaluating technologies and adopting a new standard, only to discover that certain technologies essential to implementing the standard are patented. When this occurs, the patent holder is in a position to ‘hold up’ industry participants from implementing the standard. Industry participants who have invested significant resources developing products and technologies that conform to the standard will find it prohibitively expensive to abandon their investment and switch to another standard. They will have become ‘locked in’ to the standard. In this unique position of bargaining power, the patent holder may be able to extract supra-competitive royalties from the industry participants.”

Many SDOs, including the IEEE and the ITU/ISO/IEC, have adopted intellectual property right (IPR) policies that require, or at least strongly encourage, participants to disclose the existence of patent applications and issued patents having claims that are or might be necessary to implement a standard. Moreover, most SDO IPR policies require the owner of any relevant patent applications or issued patents to state its intentions regarding licensing such patents and applications to parties wishing to implement the standard.

Two IPR disclosure policies, discussed below, are the IEEE’s patent disclosure policy and the Common Patent Policy of the ITU-T/ITU-R.
ISO/IEC [2]. All readers, and especially those involved in these or other SDOs, are encouraged to seek out applicable IPR disclosure policies and gain an understanding of SDO participants’ obligations under those policies.

**Overview of the IEEE’s Patent Disclosure Policy**

The IEEE’s patent policy, which is set forth in the IEEE Standards Association (IEEE-SA) Standards Board Bylaws, states explicitly that SDO participants have a duty to disclose certain information about issued patents and pending patent applications:

“In order for IEEE’s patent policy to function efficiently, individuals participating in the standards development process: (a) shall inform the IEEE (or cause the IEEE to be informed) of the holder of any potential Essential Patent Claims of which they are personally aware and that are not already the subject of an Accepted Letter of Assurance, that are owned or controlled by the participant or the entity the participant is from, employed by, or otherwise represents; and (b) shall inform the IEEE (or cause the IEEE to be informed) of any other holders of potential Essential Patent Claims that are not already the subject of an Accepted Letter of Assurance.”

The IEEE defines an Essential Patent Claim as any claim, in an issued patent or pending patent application, “the practice of which was necessary to implement either a mandatory or optional portion of a normative clause of the IEEE Standard when, at the time of the IEEE Standard’s approval, there was no commercially and technically feasible non-infringing alternative implementation method for such mandatory or optional portion of the normative clause” [1]. A Letter of Assurance is a document stating the position of the owner of the issued patent or pending patent application containing Essential Patent Claim(s) regarding the ownership, enforcement, or licensing of the Essential Patent Claim(s) [1, 3]. An Accepted Letter of Assurance is a Letter of Assurance that the IEEE-SA has determined is complete in all material respects and has been posted to the IEEE-SA website [1].

Thus, the IEEE’s patent policy imposes an affirmative duty on SDO participants to disclose when they are personally aware of Essential Patent Claims owned by their employers or the companies they represent that are not already subject to an Accepted Letter of Assurance [1]. The policy also encourages SDO participants to disclose when they are personally aware of other potential Essential Patent Claims, such as those owned by third parties. The policy does not, however, impose any duty on SDO participants or their employers to search for issued patents or pending patent applications having claims that may read on an IEEE standard [1].

Essential Patent Claims include only claims for technologies explicitly required by or expressly set forth in the IEEE standard [1]. Claims covering technologies that might be desirable or necessary to make or use standardized products (e.g., semiconductor manufacturing technology or operating system technology) are not Essential Patent Claims [1].

The definition of Essential Patent Claims raises the issue of whether the IEEE’s patent policy extends to pending provisional applications. As explained in [4], the United States allows applicants to file provisional patent applications to establish an invention’s priority date, and provisional applications can, but need not, contain claims [5]. On its face, the IEEE’s patent policy appears to extend to pending provisional applications that include claims.

The IEEE provides a Letter of Assurance template for use by patent right owners [3]. For example, using this template, a patent right owner may state that, after a “reasonable and good-faith inquiry,” the patent right owner is not aware of any patent claims it owns, controls, or has the ability to license that might be or might become Essential Patent Claims [1]. A patent right owner has engaged in a “reasonable and good-faith inquiry” if it has made reasonable efforts to identify and contact the patent right owner’s representatives in a sponsor ballot concerning the IEEE standard identified in the Letter of Assurance or at meetings at which that IEEE standard was developed [1]. If the patent right owner did not or does not have such SDO participants, the patent right owner should use reasonable efforts to contact individuals who are from, employed by, or represent the patent right owner and whom the patent right owner believes are most likely to have knowledge about the technology covered by the IEEE standard [1]. A patent right owner that has submitted a Letter of Assurance stating that it has made a “reasonable and good-faith inquiry” is not Essential Patent Claims will likely find itself in an awkward position if it later sues another party for allegedly infringing Essential Patent Claims that the patent right owner owned or was in the process of prosecuting on the date it submitted such a Letter of Assurance.

Alternatively, a Letter of Assurance may state that the patent right owner may own, control, or have the ability to license claims that might be or might become Essential Patent Claims [3]. In this case, the IEEE asks the patent right owner to state its position regarding the licensing and enforcement of its Essential Patent Claims [3]. The patent right owner may state either that:

a. It will license its Essential Patent Claims without compensation;

b. It will license its Essential Patent Claims under fair, reasonable, and nondiscriminatory terms;

c. It will not enforce its Essential Patent Claims against anyone making a standard-compliant product;

d. It will not grant licenses as provided by option (a) or (b), and it will not agree to enforce its Essential Patent Claims as provided by option (c) [3].

---

8 The word “shall” indicates a requirement, whereas the word “should” indicates encouragement.
When the patent right owner has determined that it may own, control, or have the ability to license claims that might be or might become Essential Patent Claims, the patent right owner’s Letter of Assurance must either identify (e.g., by patent number or application number) its issued patents and pending patent applications that do or may contain Essential Patent Claims, or the patent right owner may provide a “blanket” assurance that “all Essential Patent Claims that the [patent right owner] may currently or in the future have the ability to license shall be available” under the indicated terms [3]. Patent right owners who own pending-but-unpublished applications that include Essential Patent Claims may choose to submit a “blanket” letter of assurance to avoid having to provide identifying information about their pending unpublished applications.

Regardless of the format a Letter of Assurance takes, the IEEE patent policy specifies that a patent right owner should provide a Letter of Assurance as early as reasonably feasible in the SDO process and in no event after the Standards Board approves the applicable standard. An Accepted Letter of Assurance is irrevocable and remains in effect at least from the date of the IEEE standard’s approval to the date of the standard’s transfer to inactive status [1]. Moreover, an Accepted Letter of Assurance is binding not only on the patent right owner, but also on any party to whom the patent right owner transfers the patents or applications containing the Essential Patent Claims [1].

After receiving notice that an IEEE standard may require the use of one or more Essential Patent Claims that are not already subject to an Accepted Letter of Assurance (e.g., directly from a patent right owner through its SDO participant’s announcement at a working group meeting, or from a person who is aware of a patent right owner’s Essential Patent Claims but is not affiliated with the patent right owner), the IEEE contacts the patent right owner to request a Letter of Assurance [1]. In response to the IEEE’s request, the patent right owner may submit a Letter of Assurance promising either to license or not enforce its Essential Patent Claims against parties offering standard-compliant products, or stating that it has performed a “reasonable and good-faith inquiry” and has determined that it does not own, control, or have the ability to license any Essential Patent Claims [3]. Alternatively, of course, the patent right owner could ignore the IEEE’s request or decline to provide any licensing assurance, in which case the IEEE patent policy states that the Essential Patent Claim(s) “shall be referred to the Patent Committee” [1].

9 Patent applications in most countries are kept confidential for a period of 18 months after their priority dates. Moreover, in the United States, applicants who do not intend to seek patent protection in any other country may elect to choose to submit a “blanket” letter of assurance that “all Essential Patent Claims that the [patent right owner] may currently or in the future have the ability to license shall be available” under the indicated terms [3]. Patent right owners who own pending-but-unpublished applications that include Essential Patent Claims may choose to submit a “blanket” letter of assurance to avoid having to provide identifying information about their pending unpublished applications.


To encourage disclosure of the fullest available information, the Common Patent Policy provides that “any party participating in the work of ITU, ISO, or IEC should, from the outset, draw the attention of the [SDO’s management] to any known patent or to any known pending patent application, either their own or of other organizations...” [2]. If a standard is developed, and information about any known patent or pending patent application has been disclosed, the patent right owners must provide a written statement to the SDO’s management stating that either a. The patent right owner is “willing to negotiate licenses free of charge with other parties on a non-discriminatory basis on reasonable terms and conditions,” b. The patent right owner is “willing to negotiate licenses with other parties on a non-discriminatory basis on reasonable terms and conditions,” c. The patent right owner is not willing to provide any licensing assurance, in which case the Common Patent Policy states that the standard “shall not include provisions depending on the patent” [2].


The patent disclosure provisions of patent policies such as the IEEE’s patent policy and the Common Patent Policy raise issues and considerations for patent right owners and SDO participants. One issue for SDO participants is that the information patent right owners are required to disclose often does not give SDO participants complete technical information. For example, if SDO participants wanted to minimize the inclusion of patented technology in a developing standard, they likely would be unable to do so based on patent right owners’ disclosures because the policies do not require patent right owners to submit or disclose the claims that they believe are or may be essential, nor must patent right owners disclose the contents of the applications...
or patents containing such claims [1]. Furthermore, as explained above, a patent right owner may simply submit a “blanket” letter of assurance committing to license its patents under the terms specified in the letter of assurance [3], which gives SDO participants no information about the contents of patent assets that are or may be essential. Even if the patent right owner identifies (e.g., by patent number or application number) its issued patents or pending patent applications containing claims that are or may be essential, SDO participants do not know which claims the patent right owner believes are or may become essential. And if the patent right owner identifies pending, but unpublished, applications, SDO participants cannot even determine to what aspect of the standard the allegedly-essential claims are directed. As a result, even if SDO participants wanted to avoid a patent right owner’s patented or patent-pending technology by selecting an alternative approach, they would likely lack the information needed to do so from the disclosure required by the SDOs.

Moreover, whether a claim in a patent application is essential to implement a standard may change over time. As explained in [4], the claims in a pending application may be amended during prosecution. Therefore, even if the claims originally filed in an application would have been essential to a standard if those claims were in an issued patent, the claims that eventually may not be essential to that standard. Furthermore, the standards themselves also evolve during their development, which means that claims of an application or patent that may have been essential to implement a draft of the emerging standard on the date of submission of a letter of assurance may not, in fact, turn out be essential to implement the issued standard. As a result of both of these factors, some patents and applications disclosed expressly (e.g., by patent or application number) and listed in an SDO’s database as containing “essential” patent claims may not be essential at all.10 “Blanket” letters of assurance do not result in non-essential patents and applications being erroneously identified as essential, but, as explained above, “blanket” letters do not provide helpful technical information to SDO participants.

Although patent right owner disclosures under SDO patent policies do not provide perfect, or possibly even helpful, technical information about a patent right owner’s patented technologies and essential patent claims, a key objective of patent policies is to seek assurances from patent right owners regarding their intentions to enforce or license any essential patents they hold.11 Thus, even if the precise contours of the allegedly-essential patent claims are unknown, and various aspects of a standard are still fluid or only partially defined, the SDOs and SDO participants at least know whether the patent right owner is willing to license whatever essential claims it may own, and under what general terms (e.g., without compensation or for reasonable rates).

Another issue with SDO patent policies is how SDO participants even determine that they are aware of patents and applications that should be disclosed. For example, the IEEE defines an Essential Patent Claim as any claim in an issued patent or pending application that an implementation must use to implement a mandatory or optional portion of a normative clause of an IEEE standard [1]. It is well understood in patent law that the interpretation of patent claims is the exclusive role of judges.12 In contrast, most SDO participants are engineers without legal training who cannot be expected to perform a correct assessment of whether a claim in a patent application or issued patent is essential to implement a draft standard or an issued standard. For this reason, and because both patent application claims and emerging standards can change, patent holders may choose to submit “blanket” letters of assurance.

Possible Consequences of Nondisclosure

There have been many arguments in legal proceedings in the United States, that an SDO participant failed to comply with the disclosure requirements of an SDO’s patent policy.13 This section discusses one such case to illustrate the consequences that may flow from an SDO participant’s failure to disclose information about known relevant patents during the development of a standard. The case demonstrates how one court in the United States assessed whether an SDO participant breached an SDO’s duty to disclose, and, finding that it did, what the appropriate remedy was for that failure to meet the SDO disclosure obligation.

In Qualcomm Incorporated v. Broadcom Corporation, the Federal Circuit14 considered whether Qualcomm had waived its right to assert two of its patents by failing to disclose those patents to an SDO.15 Qualcomm had participated in the Joint Video Team (JVT) during the development of the H.264 standard. Although Qualcomm owned the two asserted patents at the time it participated in the JVT, Qualcomm did

10 Thus, many SDO patent databases merely provide access to received letters of assurance or information extracted from received letters of assurance. The author is unaware of any SDO patent database that makes a representation that the listed patents and patent applications are, in fact, essential to implement a standard.
11 Microsoft Corp. v. Motorola, Inc., 795 F.3d at 1031.
14 The United States Court of Appeals for the Federal Circuit, often referred to as “the Federal Circuit,” is the court that hears appeals from cases arising under the patent laws in the United States.
15 545 F.3d at 1008.
16 Id. at 1009.
not disclose those patents to the JVT while the H.264 standard was in development.\textsuperscript{16}

In 2005, Qualcomm sued Broadcom for patent infringement, alleging that Broadcom’s products, compliant with the H.264 standard, infringed both patents. A jury found that Broadcom’s standard-compliant products did not infringe either of the two patents. The district court declared the patents, and related applications, including continuation, continuation-in-part, and divisional applications, to be unenforceable against the world if based on Qualcomm’s failure to disclose the patents to the JVT.\textsuperscript{18} Qualcomm appealed.

The Federal Circuit began its analysis by examining the JVT’s disclosure policy, which said:

“According to the ITU-T and ISO/IEC IPR policy, members/experts are encouraged to disclose as soon as possible IPR information (of their own or anyone else’s) associated with any standardization proposal (of their own or anyone else’s). Such information should be provided on a best effort basis.”\textsuperscript{19}

Qualcomm argued that although the JVT policy encouraged disclosures from participants not making proposals, it only required disclosures from those making technical proposals.\textsuperscript{20} The court found that the word “encouraged” applied to the timing of the disclosure, not the disclosure duty itself. Furthermore, the court highlighted the “best effort” requirement and noted that “[b]y Qualcomm’s own admission, it did not present evidence of any efforts, much less best efforts, to disclose patents associated with the standardization proposal (of their own or anyone else’s) to the JVT prior to the release of the H.264 standard.” Moreover, the court found that Qualcomm had a disclosure obligation under the rules of the JVT’s parent organizations, the ITU-T and ISO/IEC. Specifically, the JVT rules required a “final IPR declaration to the ITU TSB and ISO/IEC, which must be done in the approval process for the ITU-T Recommendation and ISO/IEC International Standard.”\textsuperscript{21}

Thus, the court concluded that all JVT participants, including Qualcomm, had a duty to disclose, before final approval of the H.264 standard, all patents that “reasonably might be necessary” to practice the H.264 standard. The court added that a patent reasonably might be necessary to practice the H.264 standard when a reasonable competitor would not expect to practice the H.264 standard without a license under the undisclosed claims. It was irrelevant to the court that the claims of the asserted patents did not turn out to be essential to implement H.264; what mattered was that, when Qualcomm participated in the development of H.264, the patent claims might reasonably have been necessary, and therefore Qualcomm had a duty to disclose the patents containing those claims.\textsuperscript{22}

Having found that Qualcomm had a disclosure duty, the court turned to whether Qualcomm had breached that duty. The court found “that JVT participants understood the JVT IPR policies as imposing a disclosure duty, that Qualcomm participated in the JVT prior to release of the H.264 standard, and that Qualcomm was silent in the face of its disclosure duty.”\textsuperscript{23} Furthermore, the court found that Qualcomm had “intentionally organized a plan to shield [its] patents from consideration by the JVT, planning to demand license fees from those seeking to produce H.264-compliant products.”\textsuperscript{24} The Federal Circuit affirmed the district court’s finding that Qualcomm’s SDO representatives had a duty to disclose Qualcomm’s patents to the JVT, and that they breached this duty.\textsuperscript{25} The court found that Qualcomm had impliedly waived its right to enforce the two asserted patents, and that the appropriate remedy was to declare the patents unenforceable, but only against all H.264-compliant products, including Broadcom’s accused H.264-compliant products.\textsuperscript{26}

Thus, as a result of Qualcomm’s participation in the development of the H.264 standard but its failure to disclose the existence of patents that “reasonably might be necessary” to practice that standard, Qualcomm is precluded from asserting those patents against any H.264-compliant product. Qualcomm also lost the ability to collect royalties based on those patents from companies making H.264-compliant products.

**Conclusion**

The Federal Circuit has said that “[w]hen direct competitors participate in an open standards committee, their work necessitates a written patent policy with clear guidance on the committee’s intellectual property position.” Otherwise, “members form vaguely defined expectations as to what they believe the policy requires — whether the policy in fact so requires or not.”\textsuperscript{27}

\textsuperscript{17} An unenforceable patent is presumed to be a valid patent, but it cannot be asserted in at least some circumstances. A patent that is unenforceable against the world cannot be asserted against any party and is, effectively, valueless.

\textsuperscript{18} Qualcomm Inc. v. Broadcom Corp., 548 F.3d at 1009.

\textsuperscript{19} Id. at 1013.

\textsuperscript{20} Id. Although the case does not say so, the implication of Qualcomm’s argument is that Qualcomm did not make any technical proposals that would have required it to disclose the two patents.

\textsuperscript{21} Id. at 1014.

\textsuperscript{22} Id. at 1018.

\textsuperscript{23} Id. at 1021.

\textsuperscript{24} Id. at 1022.

\textsuperscript{25} Id. at 1019.

\textsuperscript{26} Id. at 1022, 1026. The court’s holding left Qualcomm free to assert the patents against non-H.264 products.

\textsuperscript{27} Rambus Inc. v. Infineon Techs. AG, 318 F.3d at 1102.
This article provided an overview of the patent disclosure requirements of two such policies: the IEEE’s patent policy and the Common Patent Policy of the ITU/ISO/IEC. The IEEE’s policy requires, and the Common Patent Policy encourages, individuals participating in the standards development process to inform the SDO when they are aware of patent applications or issued patents that might be essential to practice a standard.

The consequences of SDO participants failing to disclose information about relevant issued patents and patent applications may not be known for several years, but, as suggested by the case of Qualcomm Incorporated v. Broadcom Corporation, those consequences can be severe. To avoid undesirable consequences, SDO participants should ensure that they understand and are able to fulfill their patent disclosure obligations to the SDOs in which they participate.

References

Biography
KRISTA S. JACOBSEN (M) (krista@jacobseniplaw.com) is a patent attorney based in Campbell, California, whose solo practice includes expert witness services, patent litigation support, patent preparation and prosecution services, and intellectual property counseling services. She is also a Lecturer in Law at Santa Clara University School of Law. She received a B.S. degree in electrical engineering from the University of Denver in 1991, M.S. and Ph.D. degrees in electrical engineering from Stanford University in 1993 and 1996, respectively, and a J.D. degree from Santa Clara University School of Law in 2009. From 1996 to 2006, she worked as an engineer in the digital subscriber line (DSL) industry. At various times between 1993 and 2006, she was an active participant in meetings of IEEE 802.14, IEEE 802.3ah, ITU-T Study Group 15/Kuestion 4, ETSI TMC, and ANSI T1E1.A. She is the co-editor of two books on DSL technology, a named inventor on eleven United States patents, and the author or co-author of multiple standards contributions, conference papers, magazine articles, technical journal papers, and law journal papers.
CHARACTERIZATION OF WEB SINGLE SIGN-ON PROTOCOLS

Single Sign On (SSO) protocols are today integrated in millions of Web services so end users can authenticate to a third-party identity provider (IdP) to access multiple services. IdPs normally provide integration tools that hide almost all implementation details and allow developers to implement SSO in some minutes. Such integration tools along with cumbersome protocol specifications result in developers without a clear view of the underlying SSO protocol.

Victoria Beltran

ABSTRACT

Single Sign On (SSO) protocols are today integrated in millions of web services so end users can authenticate to a third-party identity provider (IdP) to access multiple services. IdPs normally provide integration tools that hide almost all implementation details and allow developers to implement SSO in some minutes. Such integration tools along with cumbersome protocol specifications result in developers without a clear view of the underlying SSO protocol. This article presents a conceptual characterization of web SSO protocols through their assertions and their features that help preserve the privacy of the user resources involved in SSO.

INTRODUCTION

Web services need to verify the identity of their users to guarantee user security and privacy, or offer a personalized user experience. Web services have traditionally had to replicate the same personal and contextual data (e.g. name, address, bank account, telephone number, etc.) to create their user accounts. Such site-centric identity management forces users to replicate their identities all over the web and to authenticate to services, each with its own credentials, one after the other.

The well known Single Sign-On (SSO) approach permits users to log in once and to access multiple services by delegating user authentication from service providers to authentication services [1]. Service providers and authentication services form identity federations that enable importing identity information from one domain to another [2].

On the web, identity federation protocols like OpenID allow web services to redirect users to third-party services, which are known as identity providers (IdPs) for authentication [3]. The IdP tells the web service who the present user is by delivering an identity assertion that contains at least the user’s unique identifier and may include other identity-related attributes such as name, address, and email.

In SSO, web services only rely on identity information to grant users access. Although the HTTP flows of identity federation protocols have already been extensively studied, these protocols’ data model for assertions has not yet been sufficiently analyzed. Only in the framework of Web Real-Time Communications (WebRTC), has the impact of identity assertions on user privacy been addressed [4].

This article provides a characterization of SSO assertions and interprets four protocols through their assertions: OpenID, OAuth2.0, OpenID Connect, and BrowserID. This article also conceives a high-level view on how these protocols’ design choices impact user privacy.

THE WEB FLAVOR OF IDENTITY FEDERATION

Identity federation and SSO originally came into being in the security domains of corporate and governmental networks. Corporate users can transparently access any service in the security domain of the authentication server to which they have authenticated. Kerberos is the de facto authentication protocol for centralized authentication services. For inter-domain scenarios, the SAML/SOAP paradigm is deeply integrated in corporations, government networks, and verticals such as healthcare, banking and education. Similar to SAML, WS-* (e.g. WS-Federation and WS-Security) is a loose set of OASIS specifications for identity federation promoted by Microsoft.

Although SAML and WS-* specifications enable importing identity assertions from one domain to another, their complexity and their strong requirements on authorization rules and trust relationships between services and authentication servers have not fit in with the ever-changing World Wide Web (WWW). Identity federation for Internet services has a different nature than its corporate counterpart. There are no centralized security servers but millions of APIs, mashups, and applications that liberally communicate through basic exchanges of HTTP requests and responses, based on JSON or XML. Thus, flexible REST-based protocols such as OpenID1 and OAuth2.0 [5], that simplify loosely-coupled identity federation and speed up application development, have found their way onto the web. Web identity federation protocols mainly define how to deliver identity information from IdPs to web services that are called relying parties (RPs) in web SSO terminology. These protocols are independent from the authentication process carried out by the IdP. Indeed, they do not provide SSO by themselves. The SSO experience is given by the user browser’s HTTP cookie storage. As long as the IdP’s session cookie is active, the user will not have to re-authenticate to log on to a web service. Nevertheless, for the sake of brevity, in the rest of this article I refer to web identity federation protocols as Web SSO protocols.

Although OpenID was the first Web SSO protocol to be widely adopted by Internet service providers by 2005, OAuth gained ground rapidly due to its ability to delegate access to protected APIs along with the user’s identity. In fact, the biggest driver of OAuth adoption was the release of a set of OAuth-based Facebook APIs, known as Facebook Connect, that allow web services to log users in and get access to social network

1 OpenID, http://openid.net/specs/openid-assertion-2_0.html
features based on the users’ Facebook identities. OAuth is a generic delegated authorization protocol that provides an access token to a client so that it can access a protected resource, based on the permission of the resource owner. Although OAuth is not a SSO protocol and leaves the resources being authorized out of scope, it has been massively used to build authentication and identity protocols.

The fact that OAuth does not specify how to convey authentication and identity information has led to interoperability and even security problems. Each IdP defines their own customized identity layer on top of OAuth to provide delegated access to identity information. Authorization and user authentication are often confused, resulting in vulnerable implementations. To address these limitations, OpenID Connect (OIDC) emerged to provide an interoperable authentication and identity framework on top of OAuth. Finally, BrowserID is a browser-side API specification for SSO based on user certificates.

**Native Apps**

OAuth has become the de facto protocol for implementing federated authentication and authorization functionality in mobile applications. Since OAuth evolved during the same period that the adoption of mobile applications exploded (by 2007–2008), this protocol was the natural choice for merging authentication and authorization in mobile platforms.

Mobile RPs can integrate third party login natively (by means of the SDKs provided by some IdPs) or by the same web-based protocol flows (through the mobile browser or WebView) implemented by web RPs. On the web, SSO relies on the browser’s cookie storage to verify whether the user has an active session with the IdP. On mobile platforms, mobile browsers also keep a cookie storage that can be shared by mobile applications for SSO. However, redirecting users to mobile browsers for login provides a poor user experience. Furthermore, if something goes wrong with the login flow, the user might not be redirected back to the application. Native or WebView-based implementation of third party login offers a much better user experience. However, mobile applications are isolated and do not share any space of user authentication proofs, and neither do WebViews.

A shared space of authentication proofs for mobile SSO can be provided by native SSO applications or platform-integrated SSO. Some IdPs are already providing native SSO applications that handle protocol flows and authentication proofs on behalf of mobile applications. When a mobile app wishes to authenticate a user, it will invoke the SSO native application of the user’s IdP, instead of implementing a third party login by itself (by an SDK, WebView, or the mobile browser). Nevertheless, to install one SSO app for each IdP to which the user wishes to authenticate can be tedious. Thus, there already exist third party providers that offer a single native app for SSO with multiple IdPs. For the sake of standardization, the recent Native Applications (NAPPS) Working Group is defining a profile of OpenID Connect (OIDC) for a native token-issuing agent. This profile will provide native SSO through standardized interfaces to mobile applications. Regarding platform-integrated SSO, some solutions already exist. Samsung Knox and Apple iOS (from iOS7 releases) implement Kerberos to address the mobile security needs of enterprise users. For customer services, both Google and Microsoft integrate OAuth into Android and Microsoft’s Windows Phone 8.1 to enable authenticating to their respective IdPs (i.e. Google Sign-In and Microsoft Live Connect SSO).

**SSO Assertions**

The ultimate goal of any SSO system is the delivery of identity assertions that allows users to sign into web services without re-authenticating. The format of identity assertions depends on the underlying SSO protocol. OpenID encodes user identity information as standard plaintext key-value attributes. OIDC defines two different JSON objects for user identity information: ID tokens and UserInfo objects. Both options are JSON Web Token (JWT) objects that can include the same identity attributes, which are called claims in OIDC terminology. Both OpenID and OIDC define a set of standard “attributes” and “claims,” respectively, but also allow defining non-standard identity information. In BrowserID, user identity information is encoded in a JSON Web Signature (JWS) object that only contains an email address that identi-

---

2 J. Richer, User authentication with OAuth 2.0 http://oauth.net/articles/authentication/
3 OpenID Connect, http://openid.net/connect/
5 NAPPS working group, http://openid.net/wg/napps/charter/
7 OpenID Attribute Exchange 1.0, http://openid.net/specs/OpenID-attribute-exchange-1_0.html
Assertion characteristics in SSO protocols (in immediate (I) and pre-authorized (A) delivery modes).

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Name</th>
<th>Type</th>
<th>Format</th>
<th>Delivery mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>OpenID</td>
<td>Positive assertion(^1)</td>
<td>Identity</td>
<td>(key, value) pairs</td>
<td>I</td>
</tr>
<tr>
<td>OAuth2.0</td>
<td>Authorization code</td>
<td>Authorization</td>
<td>Opaque string</td>
<td>I</td>
</tr>
<tr>
<td></td>
<td>Access token</td>
<td>Authorization</td>
<td>Opaque string</td>
<td>I</td>
</tr>
<tr>
<td></td>
<td>Authorization code</td>
<td>Authorization</td>
<td>Opaque string</td>
<td>I</td>
</tr>
<tr>
<td></td>
<td>Access token</td>
<td>Authorization</td>
<td>Opaque string</td>
<td>I</td>
</tr>
<tr>
<td>OpenID</td>
<td>ID token</td>
<td>Authentication and identity</td>
<td>JWT/JWS/JWE</td>
<td>I</td>
</tr>
<tr>
<td>Connect</td>
<td>UserInfo</td>
<td>Identity</td>
<td>JWT/JWS/JWE</td>
<td>A</td>
</tr>
<tr>
<td>BrowserID</td>
<td>Backed identity assertion</td>
<td>Identity</td>
<td>JWS</td>
<td>I</td>
</tr>
</tbody>
</table>

\(^1\) The “positive assertion” term is used to refer to the set of attributes that results from a successful user authentication event.

Table 1. Assertion characteristics in SSO protocols (in immediate (I) and pre-authorized (A) delivery modes).

SSO protocols differ in their assertions’ types, formats, and delivery modes. Figure 2 shows how the user browser’s software components, and the IdP’s and RP’s servers, interact to deliver SSO assertions to the RP once the user has authenticated, and authorized the RP. An authentication assertion serves as a pass given by the IdP to an RP to retrieve another assertion. SSO assertions can be exchanged authentication and authorization assertions. An authentication assertion tells a web service about the user’s authentication event. For example, OIDC ID tokens can specify the exact time when the user authenticated and the authentication method. An authentication assertion serves as a pass given by the IdP to an RP to retrieve another assertion. SSO assertions can be classified, no matter their nature (i.e. authentication, identity, and authorization), based on their delivery mode, i.e. immediate or pre-authorized modes, as depicted in Figure 1. Table 1 summarizes the characteristics of the assertions provided by OpenID, OAuth, OIDC, and BrowserID.

**A View of Web SSO Protocols Through Their Assertions**

SSO protocols differ in their assertions’ types, formats, and delivery modes. Figure 2 shows how the user browser’s software components, and the IdP’s and RP’s servers, interact to deliver SSO assertions to the RP once the user has authenticated, and authorized the RP. An authentication assertion serves as a pass given by the IdP to an RP to retrieve another assertion. SSO assertions can be classified, no matter their nature (i.e. authentication, identity, and authorization), based on their delivery mode, i.e. immediate or pre-authorized modes, as depicted in Figure 1. Table 1 summarizes the characteristics of the assertions provided by OpenID, OAuth, OIDC, and BrowserID.

As shown in Fig. 2a, OpenID delivers immediate identity assertions by means of an HTTP redirection response that contains the identity information in the response body. OAuth2.0 only defines how to obtain authorization assertions that are called access tokens, which can be used to retrieve server resources based on the resource owner’s authorization. Thus, OAuth-based IdPs only provide pre-authorized identity assertions by requiring the RP to possess an access token. Access tokens can be immediate, or pre-authorized by another (immediate) assertion called authorization code. The latter case is referred as the authorization code flow and is depicted in Fig. 2b. An authorization code is attached as a parameter of the redirection URL in the HTTP redirect response sent by the IdP. The RP exchanges the authorization code for an access token; the pre-authorized access token is included in a JSON object in the HTTP response’s body. When an access token is immediate, it is directly sent to the Web service (without a previous authorization code). Figure 2c shows this flow, which is referred to as the OAuth2.0 implicit flow. The IdP includes the token as a fragment of the HTTP response’s redirection URL. The redirection request will then download a script that retrieves the token from the redirection URI and sends it to the server-side RP.

OIDC defines an identity layer on top of OAuth2.0 that enables web services to obtain immediate and pre-authorized identity assertions. UserInfo objects are identity assertions that are always pre-authorized by an access token (see right part of Fig. 2d) and are included as JSON objects in the IdP’s response to the RP. As in OAuth2.0, an access token can be immediate or pre-authorized by an authorization code. ID tokens can also contain identity assertions and can be retrieved in a pre-authorized or immediate mode. Figure 2d shows the OAuth authorization code flow that provides ID tokens and access tokens pre-authorized by authorization codes. Other OIDC flows combine immediate and pre-authorized access tokens and ID tokens. An immediate access token is sent immediately after an authorization request as a URL fragment of the HTTP redirection response (as in the OAuth2.0 implicit flow for access tokens).

BrowserID flows for delivering identity assertions are different from those of the other three protocols, as shown in Fig. 2e. Although BrowserID identity assertions are immediate, they are not sent directly to the RP. The IdP sends a normal response to the user browser rather than a redirect response toward the RP. The IdP and RP are therefore completely disconnected. Indeed, the IdP does not reply with the final user identity assertion for the RP, but with a user certificate. The user certificate is a JWS object that contains the user’s email address and public key signed by the IdP. When the user browser receives the user certificate, it generates another JWS object that includes the target service’s URL. The browser signs this object, which is called the user assertion, with the user’s private key and passes this assertion along with the user certificate to the client-side RP code on the browser. A BrowserID identity assertion (so called user backed identity assertion) is therefore composed of two JWS objects: one generated by the IdP (the user certificate) and another generated by the browser (the user assertion).
An IdP asserts the veracity of a user’s identity assertion, thereby ensuring that such assertion belongs to the user that has been correctly authenticated. While the user authenticates to the IdP, generally by proving possession of something that they know (e.g., a password) and/or have (e.g., a SIM card), an identity assertion is the only authentication proof that the web service needs to grant the user access, as depicted in Fig. 3.

Compared to traditional client-server authentication systems, third party login entails more security risks, i.e., there are simply more resources to protect. Figure 3 shows that in a single SSO authentication event the user gets access to two private resources, i.e., their accounts at the IdP and at the web service. Moreover, identity assertions themselves can entail sensitive information, which can motivate attackers to steal or eavesdrop on them.

User privacy does not only concern the confidentiality of the user information stored in the RP’s and IdP’s user accounts and in identity assertions, but also the protection of user activity on the web. User privacy in a SSO system depends on the underlying SSO protocol’s functionality, architecture, and user identification scheme.

**User Accounts, Assertions, and the User: The Three Points of Attack**

SSO systems take part in a complex web ecosystem that exposes them to a variety of attacks that can break user privacy. Web application vulnerabilities (e.g., no transport layer protection, Cross-Site Request Forgery (CSRF), and Cross-site scripting (XSS)), the misuse of opaque SDKs provided by IdP’s, and the misunderstanding of the underlying SSO protocol, result in insecure SSO implementations [6–8].

Web attackers can take advantage of insecure SSO systems to break user privacy by illegitimately obtaining user information from some of the user information sources in a SSO system: the user accounts at the RP and IdP, the identity assertion, and the user. Web attacks can be classified into user impersonations toward the RP and phishing attacks through fake RPs, through fake IdPs, or even through fake RP user accounts. Table 2 shows the sources of user information that are compromised in each possible attack.

User impersonation attacks aim to get control
of the RP’s user account by reusing a stolen or eavesdropped user identity assertion against a benign RP. If an attacker succeeds in impersonating the user, it will then be able to read any personal information stored on the user account. User impersonation has proven its success in SSO systems mainly due to application vulnerabilities [6, 8]. In OpenID, man-in-the-middle vulnerabilities can lead to user impersonation attacks [9]. BrowserID has also been proven vulnerable to identity forgery and user impersonation [8] [11]. The OAuth2.0 implicit flow is inherently vulnerable to user impersonation attacks since access tokens are not bound to the RP to which the token was issued [5]. Thus, an RP cannot be aware of whether an attacker replaced the real token in the IdP’s HTTP response with another one (i.e. one previously stolen or eavesdropped, or even legitimately emitted to a malicious RP). For mobile applications, bad practices can make OAuth-based third party login vulnerable to user impersonation. In [12], almost 60 percent of considered implementations are found to be faulty and vulnerable to attacks, mainly due to the local storage of RP secrets, the usage of the OAuth2.0 implicit grant, or unsecure mobile redirections. If an attacker succeeds in stealing the secret from the client-side RP, it may be able to retrieve the user’s access token from the IdP’s user account and use it for user impersonation attacks. iOS custom schemes and Android Intent mechanisms simulate the HTTP redirections that are fundamental for any web SSO protocol. These mechanisms allow mobile applications to register custom URI schemes for themselves. Thus, an IdP can send assertions to the URI with the custom scheme of the RP that sent the request. Although these mechanisms resemble web browser redirection, most of the time they are unsecure because mobile applications cannot be globally and uniquely identified. Thus, an IdP cannot verify that an access token is sent to the RP that requested it. Mobile OAuth implementations are only secure in the Android OS under the condition that they are native and the IdP verifies the RP mobile application’s developer key hash [12]. Web-based implementations in Android are unsecure because the developer key hash can only be verified using a native mobile application. For iOS, third party login is always vulnerable to user impersonation attacks, since it is not possible to guarantee that access tokens are sent to the RP that requested them. The redirection behavior for a custom scheme is undefined if two applications registered to it. A fake IdP can imitate the appearance of a benign IdP’s web site and trick the user into giving their account credentials, which is well known as phishing. Having a look at Fig. 3, it is easy to realize that the loss of control on the IdP’s user account is the worst security risk in a SSO scenario. The attacker will be able to sign into the IdP as the victim user and hence to impersonate the user in any RP. Phishing attacks have largely exploited web vulnerabilities such as XSS. The redirection mechanism of OpenID [9] and the characteristics of BrowserID windows [10] make these protocols especially vulnerable to IdP impersonation attacks.

A fake RP attack aims to redirect the user, after they have authenticated to their IdP, to a malicious RP’s web site instead of the RP to which the user is authenticating. If the user does not realize that the website is fake, the malicious RP can trick the user by getting them to enter confidential information. The OpenID realm spoofing attack abuses the redirection mechanism in OpenID to perform this attack [9]. The malicious RP, rather than the RP to which the user thinks they have signed in, will receive the identity assertion. Thus, the malicious RP may be able to impersonate the user toward other RPs in the presence of vulnerabilities between the benign RPs and IdPs (e.g. the lack of contextual binding such as nonce between authentication request and responses).

Finally, a fake RP user account is the result of a user session swapping attack by which the attacker supplants the victim user’s SSO credentials by its own credentials. Thus, the victim user logs in the RP without noticing that he is registered on an account under the attacker’s control rather than their own account. Any personal information entered by the user will therefore be accessible to the attacker. In OAuth2.0 and OpenID, the lack of contextual binding in assertions makes session swapping attacks possible [6, 8]. In BrowserID, vulnerabilities of login persona can be exploited to inject an attacker’s certificate [11].

**Table 3** shows user privacy properties on user information and activity, and the fulfillment of these properties by Web SSO protocols.

**Privacy on User Information:** The first property, i.e. free choice of IdP, will determine the identity federation protocol to use. Thus, it affects the overall privacy of both user information and activity. All the protocols except OAuth provide freedom of choice of IdP since they support IdP discovery. Regarding private user information, the first four properties are concerned with the contents of identity assertions. OpenID, OIDC, and BrowserID guarantee the integrity of the exchanged assertion. In OpenID, IdPs rely on a shared symmetric key to sign the identity assertion. This key is negotiated between the IdP and RP by the Diffie-Hellman algorithm. OIDC and BrowserID sign identity assertions using JWS. OIDC is the only protocol that enables assertion confidentiality by encrypting assertions using JSON Web Encryption (JWE) with the cryptographic material registered by the RP. Although the other protocols do not
provide built-in solutions for data confidentiality, transport layer encryption (i.e., HTTPS certificates) may be used. None of the protocols support anonymous identity assertions that allow users not to disclose their identities while asserting that they have authenticated. Recyclable user identities can also compromise the privacy of user information. When a user abandons an identifier, it may become available to others. The new owner of the recycled identifier then will be able to authenticate as the previous owner to all the RPs that only rely on user identifiers to identify user accounts, thereby getting access to all user information stored in the previous owners’ accounts at these RPs. Both BrowserID and OpenID rely on public identifiers (emails and OpenID URLs, respectively) that can potentially be recycled by IdPs [9, 10].

The last two privacy properties of user information are concerned with protocol functionality rather than the content of identity assertions. Selective disclosure refers to the protocol’s capacity to allow the user to select the information to send out to the web service. The OpenID attribute exchange specification determines that the user can choose from a series of optional attributes to be disclosed. OIDC and OAuth2.0 provide scopes that indicate the information requested to the IdP. Nonetheless, the RP rather than the user decides these scopes. BrowserID does not provide more user information apart from their identity, and hence this property is not applicable. Single Log Out is the protocol capability to synchronize logout events between RPs and IdPs. If users are not aware that their sessions at RPs are still active when they sign out of their IdP (or vice versa), user privacy may be compromised in shared devices. When the user leaves, any other user that can access the same browser and will be able to access the user accounts that have not been terminated [13]. Only OIDC tackles the process of logging the user off when they finish the session at the RP or IdP. The behavior of logout in BrowserID has been especially criticized[10, 11]. When the user logs out of the IdP, the user certificate installed on the user browser is not deleted. Since this certificate is used by BrowserID to automatically generate user assertions, anyone that has access to the same browser can sign into web services as the user certificate’s owner. Furthermore, since IdPs are not aware of the RPs to which assertions are sent, they cannot notify the RPs when the user logs out.

On mobile applications, the back-end nature of OAuth and OpenID authorization brings out privacy concerns. OAuth and OpenID access tokens are independent from the mobile app and device that requested them. Once a user authorizes an app (i.e. the client-side RP), the app’s back-end servers (i.e. the server-side RP) can retrieve the user’s information as long as the token does not expire, even after the user uninstalls the application. Front-end authorization should therefore be a future privacy property of SSO protocols for mobile platforms in order to limit the disclosure of user information to mobile apps [14].

Privacy of User Activity: Privacy of user activity depends on how users are identified in identity assertions and how these assertions are delivered. Unobservability means that the IdP is incapable of tracking the users as they visit web services. BrowserID is the only one to provide unobservability, since it does not involve the web service communicating with the IdP to obtain identity assertions. Identity unlinkability is the inability of web services to track user activity based on the user’s identity assertions across communications [15]. It is possible to differentiate between identity unlinkability by the same RP and by different RPs. In the former, a web service cannot correlate a user’s identity assertions across different login events. Conversely, identity unlinkability by different RPs means that a web service is able to correlate a user across login events but different web services cannot collaboratively link the same user’s identity assertions. Identity assertions must contain anonymous and pseudonymous identifiers in order to provide identity unlinkability by the same RP and by different RPs, respectively. None of the considered protocols can fulfill identity unlinkability by the same RP because none of them provide anonymous user identifiers. OIDC is the only protocol that defines the use of pairwise pseudonymous identifiers that are different for every web service. Thus, the unlinkability by different RPs’ property is marked for OIDC in Table 3 (analysis techniques to correlate assertions based on their contents are not considered). BrowserID identity assertions can always be correlated based on the user’s email address. Likewise, OpenID user identifiers are public URLs or XRI identifiers [15]. It is possible to differentiate between identity unlinkability by different RPs. Nevertheless, since this type of

<table>
<thead>
<tr>
<th>Property</th>
<th>OpenID</th>
<th>OAuth2.0</th>
<th>OIDC</th>
<th>BrowserID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free choice of IdP</td>
<td>✓</td>
<td>N/A</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>User Information</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assertion integrity</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Anonymous identity assertion</td>
<td>x</td>
<td>N/A</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Non-recyclable user identifiers</td>
<td>x</td>
<td>N/A</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Selective disclosure</td>
<td>✓</td>
<td>RP request</td>
<td>RP request</td>
<td>N/A</td>
</tr>
<tr>
<td>Single log out</td>
<td>x</td>
<td>N/A</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td>User Activity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unobservability</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>✓</td>
</tr>
<tr>
<td>Unlinkability by different RPs</td>
<td>x</td>
<td>N/A</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Unlinkability by the same RP</td>
<td>x</td>
<td>N/A</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

Table 3. User privacy properties in SSO protocols. Since OAuth is an authorization rather than an identity federation protocol, many of the privacy features are not applicable (N/A) to this protocol (i.e., they depend on each IdP’s OAuth-based identity layer).

11 https://groups.google.com/forum/#!topic/mozilla.dev.identity/Wm0Cd-rjczc

12 Pseudonymous identifiers conceal a user’s identity but allow them to be recognised on a return visit

13 XRI 2.0 is an OASIS specification aimed to be a standard syntax and discovery format for abstract identifiers that was discontinued in 2008, https://www.oasis-open.org/committees/te_home.php?wg_abbrev=xri

As of now, OIDC is the protocol that can satisfy the most privacy and hence developers should always choose OIDC in preference to OAuth2.0 and OpenID. Although BrowserID lacks many desirable privacy properties, it is the only one that provides unobservability and hence it should be chosen if this property is paramount.

CONCLUSIONS

This article has provided an assertion-driven analysis of web SSO protocols to help developers assessing the nature of the flows and data items of these protocols. The article highlights the importance for service developers to be aware of the user resources involved in SSO login events and how these resources may be compromised. Before developing a SSO system, the requirements for user privacy should be determined to choose the underlying SSO protocol that best meets them.

None of the studied protocols address all the user privacy concerns. Indeed, some privacy properties are contradictory and hence can hardly be implemented by the same protocol. The most illustrative example is unobservability and Single Log Out. The former prevents the IdP from tracking the user’s activity through their login events to web services. This property involves the IdP and RP being incapable of knowing each other. Thus, unobservability makes Single Log Out impossible since the IdP cannot communicate user logouts to the RP (and vice versa). None of the protocols provide anonymous identifiers and hence web services are always able to track users through login events (i.e. unlinkability property is not satisfied). SSO in mobile apps is especially vulnerable. IOS mobile apps are not secure because apps cannot be uniquely identified. Moreover, none of the protocols offer a suitable authorization model for mobile environments that grants authorization to mobile apps rather than back-end servers.

The article aims to bring out the need for further improvements in user privacy in SSO on the web. As of now, OIDC is the protocol that can provide the most privacy and hence developers should always choose OIDC before OAuth2.0 and OpenID. Although BrowserID lacks many desirable privacy properties, it is the only one that provides unobservability and hence it should be chosen if this property is paramount.

ACKNOWLEDGMENT

I thank Emmanuel Bertin for his valuable suggestions on this work.

REFERENCES


Biography

VICTORIA BELTRAN (vicbelma@gmail.com) has been a senior researcher on Internet services and identity protocols at Orange Labs in France. She holds a Ph.D. in telematics engineering from the Universitat Politècnica de Catalunya in Barcelona. She has approximately 10 years experience in telecommunications research and practice in different research institutions such as Bell-Labs Alcatel-Lucent (Murray-Hill), Institut Mines-Telecom (Paris), and Columbia University (New York). She is currently an associate teacher and senior researcher at the University of Murcia.
FROM NETWORK SHARING TO MULTI-TENANCY: THE 5G NETWORK SLICE BROKER

The authors provide an overview of the 3GPP standard evolution from network sharing principles, mechanisms, and architectures to future on-demand multi-tenant systems. In particular, they introduce the concept of the 5G Network Slice Broker in 5G systems, which enables mobile virtual network operators, over-the-top providers, and industry vertical market players to request and lease resources from infrastructure providers dynamically via signaling means.

Konstantinos Samdanis, Xavier Costa-Perez, and Vincenzo Sciancalepore

ABSTRACT

The ever-increasing traffic demand is pushing network operators to find new cost-efficient solutions toward the deployment of future 5G mobile networks. The network sharing paradigm was explored in the past and partially deployed. Nowadays, advanced mobile network multi-tenancy approaches are increasingly gaining momentum, paving the way toward further decreasing capital expenditure and operational expenditure (CAPEX/OPEX) costs, while enabling new business opportunities. This article provides an overview of the 3GPP standard evolution from network sharing principles, mechanisms, and architectures to future on-demand multi-tenant systems. In particular, it introduces the concept of the 5G Network Slice Broker in 5G systems, which enables mobile virtual network operators, over-the-top providers, and industry vertical market players to request and lease resources from infrastructure providers dynamically via signaling means. Finally, it reviews the latest standardization efforts, considering the remaining open issues for enabling advanced network slicing solutions, taking into account the allocation of virtualized network functions based on ETSI NFV, the introduction of shared network functions, and flexible service chaining.

INTRODUCTION

Network sharing has evolved from a novel concept a few years back with the arrival of 3G networks to a fundamental feature of the emerging 5G systems. Mobile operators are facing tremendous traffic increases with the introduction of smartphones and tablets, especially due to content rich multimedia and cloud applications, and the upcoming vertical market services in automotive, e-health, etc. [1]. The challenge for mobile operators is to accommodate such traffic volumes without significantly increasing operational and infrastructure costs. The trend toward network densification for increasing network capacity and the practice of overprovisioning to accommodate peak demands including future traffic volumes adds additional burdens to operational complexity and cost, diminishing the return on investment (ROI).

Indeed, 50 percent of the radio access sites carry traffic that yields less than 10 percent of the revenue [3]. Consequently, there is a need for mobile operators to exploit new revenue sources and break the traditional business model of a single network infrastructure ownership. Network sharing can recover up to 20 percent of operational costs for a typical European mobile network operator, and can reduce by at least half the infrastructure cost of passive radio access network (RAN) components, which account for up to 50 percent of the total network cost [2]. An overview of network sharing CAPEX/OPEX savings on different parts of the network is depicted in Fig. 1, considering the RAN, backhaul, and core networks.1

Mobile operators can share network infrastructures, accelerating network roll-outs and offering services to customers with reduced costs. In urban areas network sharing can help avoid complex and lengthy processes for site acquisition due to regulation issues, especially in high-populated regions where dense deployments restrict the available space, while for rural areas sharing can reduce the network investment pay-back period. Two different roles can be defined for network sharing solutions:

• Infrastructure provider (InP) responsible for the physical network deployment and maintenance. Mobile network operators (MNOs) or third parties that interact with other “players” but not with end users directly can take the InP role.
• Mobile virtual network operator (MVNO) lacks network infrastructure or has limited capacity and/or coverage, and leases resources from an existing InP.

Future multi-tenant systems are envisioned to expand the aforementioned roles to also include:

• Over-the-top (OTT) service providers operating on top of a network infrastructure belonging to an MNO and based on a pre-defined service level agreement (SLA) set of requirements.
• Vertical industries exploiting an MNO network infrastructure for services complementary to the telecommunication industry.

In both cases the allocated network slices can be provided on a permanent basis or on-demand, i.e. opportunistically or periodically.

This article provides an overview of the 3GPP standardization activities on network sharing, focusing on the business requirements, architectures, and network management framework. In addition, it introduces the main enablers for realizing future flexible, on-demand multi-tenant networks. The main contribution of this article is the analysis and design of a signaling-based, i.e. with no human intervention, on-demand multi-tenant network, building on the top of the 3GPP network sharing management architecture. In the core of our proposed on-demand multi-tenant network architecture lies a logically centralized monitoring and control entity defined as a 5G Network Slice Broker providing admission con-
control for incoming requests (placed by MVNOs, OTTs, and verticals), and resource assignment by means of an enhancement of the 3GPP network sharing management architecture interfaces and service exposure capability function (SECF).

The rest of this article is organized as follows. We present the fundamental network sharing scenarios and their corresponding business requirements. We describe the 3GPP network sharing standardization efforts and architectures. We introduce the network sharing management architecture for supporting mobile virtual operators along with the ongoing enhancements for supporting vertical industries. We analyze the proposed 5G Network Slice Broker architecture. We present the 3GPP efforts on evolved network slicing toward the realization of full multi-tenancy considering open standardization issues. Finally, we provide the conclusions.

**Network Sharing Scenarios and Business Requirements**

The adoption of network sharing and multi-tenancy from the business perspective aims to address different strategic and commercial targets for each participant player. For InPs, network sharing results in additional revenue sources and thus better return on CAPEX/OPEX investments. MVNOs exploit network sharing as a means to enhance service provisioning in regions with low or no network footprint, where the payback period is estimated to be greater than the expected business targets.

In general, the adoption of network sharing in mature markets concentrates on increasing RoI and capacity enhancement. In developing markets network sharing usually focuses on coverage expansion. A significant aspect that influences MNOs’ decision of whether enabling network sharing is beneficial for their business relies on the purpose of sharing and on the risk of reducing their competitive advantage. For instance, allowing coverage enhancements of their competitors is sensitive for emerging mobile markets where coverage is a significant service attribute, but it becomes more relaxed in cases where QoS provisioning and service innovation is the key business differentiator. The 3GPP Services Working Group SA1 specified five main business scenarios for network sharing in [5], summarized here:

**Multiple core networks sharing a common RAN**: An early scenario considered in 3GPP Release 99, where operators share RAN elements, but not the spectrum. In this spectrum. In this case operators connect directly to their own dedicated carrier layer in the shared radio network controller (RNC) in the shared RAN.

**Operator collaboration to enhance coverage**: In this scenario two or more operators with individual frequency licenses and respective RANs that cover different parts of a country, together provide coverage for the entire country.

**Sharing coverage in specific regions**: In this scenario one operator provides shared coverage in a specific geographical area, with other operators allowed to use it for their subscribers. Outside such an area, coverage is provided by each operator independently.

**Common spectrum sharing**: This scenario corresponds to common spectrum RAN sharing considering the following two variations:

- One operator has a frequency license and shares the spectrum with other operators.
- A number of operators decide to pool their allocated spectrum and share the total.

**Multiple RANs share a common core network**: In this scenario multiple RANs share a common core network, with each RAN belonging to different network operator.

Challenging the traditional mobile communication paradigm by considering the evolution toward multi-tenancy on an open regulation environment provides the opportunity for commercial and operational separation of the mobile infrastructure from service layers. In this way they can evolve independently according to different business needs and performance characteristics.

**Early Network Sharing Standardization and Architectures**

In the early GSM and UMTS stage, network sharing support was not included; the mobile network design was concentrating on a single MNO. 3GPP Rel.99 introduced the UMTS
network the first generation of network sharing concentrating on simple solutions in terms of commercial exploitation, with passive sharing and network roaming being the two main pillars. Passive sharing is defined as the sharing of site locations or physical supporting infrastructure of radio equipment, such as masts. Site sharing allows mobile operators to share space and optionally share certain support equipment such as shelters or power supply, but with separate installations of masts, antennas, cabinets, and backhaul equipment. Such approaches did not gain significant interest from the industry until the early 2000s.

A step further was accomplished with mast sharing, where mobile operators can co-locate their sites and even share the antenna frame, but still install their own radio equipment, maintaining separate coverage. An overview of passive network sharing is illustrated in Fig. 2 highlighting the main components. As for network roaming, certain mobile subscribers can use the network of other MNOs based on contractual agreements without imposing any particular network sharing requirements, so in that sense it is not exactly a form of infrastructure sharing. With 3GPP Rel-6 (UMTS), Rel-8 (LTE), and Rel-10 (LTE-A), new requirements were needed to shed light on the potential of network sharing.

Active RAN sharing followed the first generation of network sharing, which focused on sharing access network equipment including base stations, antennas, and even mobile backhaul equipment. In active RAN sharing, MNOs can pool spectrum resources, which are shared alongside other RAN equipment based on fixed, contractual agreements. The interest in sharing the resources dynamically introduced new requirements, beyond the original RAN sharing concepts, where MNOs share core transmission equipment, billing platforms, and core network equipment.

3GPP specified two distinct active RAN sharing architectures as illustrated in Fig. 3 in the Architecture Working Group SA2 in Rel.11 – Rel.12, as documented in [4]:

- **Multi-operator core network (MOCN),** where each operator has its own enhanced packet core (EPC) providing a strict separation among the core network and RAN. Shared base stations, i.e. evolved Node Bs (eNBs), are connected to core network elements of each different operator, i.e. mobility management entity (MME) and serving/packet-gateway (S/P-GW), using a separate S1 interface. This enables customization, for example allowing load balancing policies to be provided within each operator’s core network. MOCN offers benefits regarding service differentiation and interworking with legacy networks.
- **Gateway core network (GWCN),** where operators additionally share the MME, an approach that further enables cost savings compared to MOCN, but at the price of reduced flexibility, i.e. restricting mobility for inter-radio access technology (RAT) scenarios and circuit switching fallback for voice traffic.

In general, MOCN requires a higher investment but is considered to be more flexible, addressing easier conventional MNOs’ needs. The user equipment (UE) behavior in both MOCN and GWCN is identical with resource sharing being transparent. In both cases, UEs can distinguish up to six different MNOs that share the RAN infrastructure based on broadcast information, i.e. public land mobile network (PLMN)-id, and can signal to obtain connectivity or perform a handover irrespective of the underlying RAN sharing arrangement. Specifically, the S1 interface supports the exchange of PLMN-ids between eNBs and MMEs in order to assist the selection of the corresponding core network, as documented in TS 36.413, while the X2 interface supports a similar PLMN-id exchange among neighboring eNBs for handover purposes, as per TS 36.423. Regarding broadcasting, the Uu interface supports the PLMN-ids, enabling the UEs to perform the network selection as specified in TS 36.331.

**INCORPORATING VIRTUAL OPERATORS AND VERTICALS IN 3GPP NETWORKS**

The 3GPP Telecom Management Working Group SA5 has extended the legacy network management architecture to accommodate network sharing based on long term contractual agreements [6]. Such a network sharing paradigm considers that an InP, referred to as a master operator in the 3GPP terminology, facilitates resource sharing to participant MVNOs or otherwise sharing operators through the InP network manager system, using the Type 5 interface.

The Type 5 interface is established upon an agreement between MNOs to provide connectivity among the network manager systems across different organizations, e.g. for roaming purposes. The master operator can then forward performance monitoring information through the network manager system, referred in 3GPP as the master operator-network manager (MO-
NM), to the participant sharing operator-network manager (SO-NM). For monitoring and configuration operations on network elements, the SO-NM can use:

- **The Type 2 interface or Itf-N** between the MO-NM and network element manager. In LTE the element manager is co-located at the eNB, while in UTRAN it is located on the master operator-sharing RAN domain manager at the RNC. This interface is used for performance monitoring, reporting, and control of network elements to the network manager system.

- **The Type 1 interface or Itf-B** between the master operator-shared RAN domain manager and a NodeB. Typically, the master operator-shared RAN domain manager serves a number of shared RAN NodeBs. This interface is also used for network management purposes.

Vertical industries and OTT providers, which do not own a network infrastructure, need to interact with InPs to request network resources and to negotiate SLAs, a process that is achieved by allowing exposure of the 3GPP service capabilities to third parties. In this way operators are no longer merely suppliers of communication services, but business enablers. The 3GPP service capability exposure function (SCEF) [7], located at the operator trust domain, provides a mean to securely expose selected service capabilities via network application programming interfaces (APIs). The SCEF abstracts service capabilities related to the communication type, network elements, policy control, and network resource allocation from the underlying 3GPP network. Effectively, such service capability abstraction can also assist third parties to issue a network resource request toward an InP.

### 5G Network Slice Broker: Architecture

To enhance the existent RAN sharing flexibility, the authors of this article introduced in the 3GPP Services Working Group SA1 the concept of the on-demand capacity broker [5]. Different from the SCEF, which exposes service capabilities, the on-demand capacity broker facilitates on-the-fly resource allocation by allowing a host RAN provider, i.e. InP, to allocate via signaling means an indicated portion of network capacity for a particular time period to an MVNO, OTT provider, or vertical market player.

In this article we build on top of the 3GPP SA3 network sharing management architecture, introducing a novel concept of capacity broker with a more generic objective, in order to address dynamic resource sharing scenarios by establishing network slices. A network slice refers to an isolated amount of network capacity customized to best suit specific service require-
Figure 4. Service Capability Exposure Function (SCEF) architecture.

The proposed capacity broker, namely 5G Network Slice Broker, can facilitate on-demand resource allocation and perform admission control based on traffic monitoring and forecasting, including mobility, based on a global network view. In addition, it configures RAN scheduler to either follow a two-layer paradigm, with the higher layer operating an inter-slice resource allocation and the lower layer enabling tenants to customize scheduling on the allocated spectrum in isolation, or configure policies to enable the selection of resource blocks from a shared pooled spectrum, taking into account the service SLA and the size of the network slice across the core network.

To accomplish this task, we propose to co-locate the 5G Network Slice Broker at the MO-NM, which monitors and controls the shared RAN, while interacting with the sharing operator network manager (SO-NM). In this way, the 5G Network Slice Broker can gain access to network monitoring measurements such as load and various key performance indicators (KPIs), e.g., mobility optimization, failures, SLA violations, etc., as well as obtain network infrastructure capability information. In addition, it can receive on-demand network resource requests from MVNOs, via the Type 5 interface, for allocating network slices based on SLAs. Upon performing the corresponding admission control decisions, the 5G Network Slice Broker can then take advantage of the existing MO-NM interfaces, i.e., Itf-N and Itf-B, to configure the desired network slice on specific RAN network elements.

Besides MVNO requests, the 5G Network Slice Broker can also handle requests with a specified SLA from a range of vertical industries and OTT providers, through a close interaction with the SCEF or exploiting the co-location of the SCEF at the MO-NM. The interface of verticals or OTT providers toward the SCEF is under discussion, with 3GPP adopting APIs developed in other standardization bodies, e.g., the Open Mobile Alliance (OMA) API focusing on sensor/machine type applications, and the GSM Association (GSMA) Open API designed for application providers. In this way the SCEF and the corresponding API is not only exposing information about devices, but can also provide control to vertical industries and OTT providers through the 5G Network Slice Broker and the MO-NM interfaces, i.e., Itf-N and Itf-B, to allocate the desired SLAs.

An overview of the proposed 3GPP-compliant network slice broker management architecture is illustrated in Fig. 5, showing the 5G Network Slice Broker and the SCEF co-located at the MO-NM. The SCEF provides access to OTT providers, e.g., video, voice, and social applications, etc., as well as to vertical industries, e.g., electricity utility, automotive, e-health, etc. A direct connection through the network managers, i.e., MO-NM and SO-NM, enables various tenants to easily access RAN resources. Hence, the 5G Network Slice Broker acts as a mediator, mapping the SLA requests of multiple tenants with the physical network resources through the interfaces provided by the MO-NM.

Interestingly, the proposed 5G Network Slice Broker management architecture supports on-demand resource allocation operations. This can be achieved by enhancing the existing interfaces. In particular, enhancements should differentiate tenants in order to handle the corresponding data traffic and provide performance monitoring information toward each participant operator through the Type 5, Itf-N, and Itf-B interfaces. To enable this, a tenant identifier, e.g., PLMN-id, can be included in each data packet corresponding to different slices, as well as in performance measurement reports to enable the MO-NM to provide feedback toward the corresponding SO-NM. Such performance feedback should involve only the allocated slice resources for privacy and competition reasons. For supporting verticals and OTT providers, the Itf-N and Itf-B should also be enhanced to distinguish these types of tenants by introducing a corresponding service identifier to each data packet and performance monitoring report.

The Type 5 interface as well as the vertical industries/OTT provider APIs should be extended to accommodate on-demand network slice requests with a particular SLA and timing requirements. The Itf-N and Itf-B interfaces should also be extended to carry out the configuration of network slices by introducing a new type of signaling considering MVNOs and vertical industries/OTT providers. Such interface enhancement and signaling should contain a set of additional information, including:

- The amount of resources allocated to a network slice, e.g., physical resources or data rate.
- Timing, e.g., starting time, duration, or periodicity of a request and time window.
- The type of resources and QoS, e.g., the radio/core bearer type, prioritization, delay, jitter, and loss.
- The size of the file to be downloaded or data to be communicated to particular device/user or application server.
- Service related information, e.g., mobility (stationary, low, medium, high), data off-loading policies, and service disruption tolerance.

Besides the service characteristics of a network slice, the set of cells that should accommodate the network slice request is an additional parameter that can be considered. Effectively, such a parameter can be either explicitly provid-
ed by the MVNO via the Type 5 interface or it can be determined by the InP, considering the location of the corresponding devices/users in combination with tailored service information provided by the 5G Network Slice Broker. The set of cells that need to accommodate a network slice should be contacted via the Itf-N interface toward the master operator-shared RAN domain manager, which in turn would configure the appropriate cells using the Itf-B interface.

**The Network Slicing Road Toward Full Multi-Tenancy**

The evolution of network sharing toward full multi-tenancy relies on virtualization mechanisms and software-based capabilities which are progressively introduced into 3GPP networks, influencing its standardization roadmap. These capabilities enhance the notion of network slicing for supporting particular communication services. Such emerging network slicing will be realized by allocating not only network capacity, but also virtual network functions (VNFs), computing resources, per slice tailored control/user-plane splits, shared network functions across different slices, and RAT settings, as described by NGMN in [8]. Network slicing can further be enriched accommodating particular applications, which can be located at the network edge to improve end-users’ performance. 3GPP SA1 emphasizes support for vertical industries via network slicing in Release 14 considering terminal operations and configuration management via suitable APIs. The main attributes to realize the aforementioned network slicing extensions in 3GPP toward full multi-tenancy are described as follows.

**Network Function Virtualization:** 3GPP has adopted ETSI NFV MANO [9], shedding light on the potential impact of virtualized networks on the existing 3GPP SA5 network management architecture [10], considering partially and entirely VNFs with respect to macro-base stations and core network elements. The objective is to identify requirements, interfaces, and procedures that can be re-used or extended for managing virtualized networks. In Release 14, 3GPP has introduced a specification on architecture requirements for virtualized network management [11], considering complementary specifications on configuration, fault performance, and life-cycle management. An equivalent study focusing on small cells and on the adoption of flexible Centralized-RAN has been performed at the Small Cell Forum [12].

**Dedicated Core Networks (DCNs):** In an effort to support devices/customers with different service characteristics, including vertical market players, 3GPP SA2 introduced in Release 13 the support of separate DCNs [13], with different operation features, traffic characteristics, policies, etc. Each DCN is assigned to serve different types of users based on subscription informa-
tion, assuring resource isolation and independent scaling, offering specific services and network functions. Effectively, the 5G Network Slice Broker may allocate a collection of shared network resources and VNFs among particular slices that fulfill the requirements of certain communication services.

User/Control-Plane Separation and Service Chaining: 3GPP has initiated a study on user/control-plane separation in TR 23.714, analyzing potential architecture enhancements for core network elements, e.g., packet data network gateway (P-GW), traffic detection function (TDF), etc., to further enable flexibility for network deployment and operation, allowing a unified network management across different RATs. An equivalent process focusing on services, e.g., firewall, deep packet inspection (DPI), etc., should also be considered when establishing network slices. 3GPP has performed in TR 23.718 (Rel.13) a study on flexible mobile service steering focusing on policy provisioning and on instantiating dynamic services in SGi-LAN, a service-oriented network connected to the P-GW. A comprehensive study on flexible service chaining on mobile networks, considering a range of different service chaining mechanisms, is provided in [14].

Mobile Edge Computing (MEC): Many evolving 5G services are envisioned to be offered closer to the user at the network edge in order to reduce latency, and in general end-user perceived performance, e.g., adopting the ETSI MEC paradigm [15]. Hence, flexible service chaining should also be enhanced to establish dynamic services considering edge network locations and potentially be combined with VNFs, in order to enable a joint optimization of services and networking. Edge server locations can also be exploited for storage, computation, and dynamic service creation by verticals/OTT providers, introducing in this way another multi-tenancy dimension.

Figure 6 illustrates an example of different network slices operating on the same infrastructure:

- A network slice that accommodates mobile broadband services.
- An automotive network slice wherein latency and reliability are critical parameters.
- A massive Internet of Things (IoT) network slice where scalability is essential for efficiently handling huge amounts of small data.

To accommodate strict latency goals and scalability, network functions can be instantiated at the edge cloud as necessary, optimizing radio and core networks with respect to particular services. Different RATs should be associated with distinct types of network slices, since they can best serve the requirements of particular services.

**Conclusion**

This article reviewed the path from network sharing toward multi-tenancy, describing business requirements and standardization efforts with a focus on 3GPP. In particular, it analyzed the evolutionary path from early passive sharing to on-demand multi-tenant networks considering:

- 3GPP network sharing architectures.
- Network management extensions for supporting mobile virtual operators.
- The service exposure capability function that allows vertical market players to gather information about mobile network resources.

The notion of the 5G Network Slice Broker has been introduced, which resides inside the infrastructure provider, detailing the required interfaces and functional enhancements for supporting on-demand multi-tenant mobile networks based on the latest 3GPP network sharing...
management architectures. Finally, our work provided an overview of the 3GPP Rel.14 standardization efforts related to multi-service support and network virtualization, as well as other relevant standardization efforts outside 3GPP addressing how to enrich a 5G network slice by flexibly provisioning virtualized network functions and services.

Acknowledgement

This work has been performed in the framework of the H2020-ICT-2014-2 project 5G NORMA. The authors would like to acknowledge the contributions of their colleagues. This information reflects the consortium’s view, but the consortium is not liable for any use that may be made of any of the information contained therein.

References

[12] Small Cell Forum, Network Aspects of Virtualized Small Cells, Rel. 5.1, Document No.05.1.01, Jun. 2015.

Biographies

Konstantinos Samdanis (samdanis@neclab.eu) is a senior researcher and backhaul/broadband standardization specialist with NEC Europe. He is involved in research for 5G architectures participating in 5G-NORMA, and is active in BBF on 5G and network virtualization. He has provided a number of tutorials in IEEE conferences for green communications, and he is the editor of the book Green Communications: Principles, Concepts and Practice, published by Wiley. He received his Ph.D. and M.Sc. degrees from Kings College London.

Xavier Costa-Pérez (xavier.costapere@neclab.eu) is head of 5G networks R&D at NEC Laboratories Europe, where he manages several projects focused on 5G mobile core, backhaul/fronthaul, and access networks. His team contributes to NEC projects for products roadmap evolution as well as to European Commission R&D collaborative projects, and he has received several R&D awards for successful technology transfers. In addition, the 5G team contributes to related standardization bodies, including 3GPP, BBF, ETSI NFV, ETSI MEC, and IETF.

Vincenzo Sciancalepore [S’11, M’15] (vincenzo.sciancalepore@neclab.eu) received his M.Sc. degree in telecommunications engineering and telematics engineering in 2011 and 2012, respectively, and in 2015 he received a double Ph.D. degree. From 2011 to 2015 he was research assistant at IMDEA Networks, focusing on inter-cell coordinated scheduling for LTE-Advanced networks and device-to-device communication. Currently he is a research scientist at NEC Laboratories Europe in Heidelberg, focusing his activity on network virtualization and network slicing challenges.
While time-to-market constraints are accelerating the deployment of a variety of fragmented and proprietary IoT products, there is still a lack of understanding of what an IoT service is meant to be, what its consequences are, and how to promote standard IoT services. The author gives a concise but comprehensive survey of IoT service definition, regulation, and standardization activities. The author discusses mainstream standards as well as emerging, independent, and state-funded projects.

Aref Meddeb

ABSTRACT

While time-to-market constraints are accelerating the deployment of a variety of fragmented and proprietary IoT products, there is still a lack of understanding of what an IoT service is meant to be, what its consequences are, and how to promote standard IoT services. This article gives a concise but comprehensive survey of IoT service definition, regulation, and standardization activities. We discuss mainstream standards as well as emerging, independent, and state-funded projects.

INTRODUCTION

So far, the implications and impact of IoT are not fully understood and should be explored by standards organizations and industry consortia so that strategic roadmaps can be developed [1]. In particular, there are significant gaps between legislative mechanisms concerning privacy between many countries such as Europe, Canada, and the US. Fragmented IoT implementations exacerbate this issue and could potentially drive users away from IoT.

Since its inception, there were several movements and actions opposing IoT. In particular, RFID deployment in schools is raising a lot of opposition (rfidinschools.com). Also, some seem to be frightened by IoT. There are fears that a sort of “global government” will become able to control the world population to the extent where individuals may be remotely governed, tracked, and even “alienated.”

Perhaps the most challenging obstacles on the way toward IoT are trust, security and privacy (TSP). Questions include: Will customers feel comfortable having their bank account, insurance, location, and health information available on the net and accessible by or via objects? How can we protect ourselves, and our homes, computers, cars, etc. from illegal monitoring and/ or remote control? How to authenticate sensed data and tag reading? With IoT devices deployed everywhere, successful attacks can be very embarrassing.

On the other hand, IoT applications in industrial environments bring a plethora of concerns [2]. Industries may be skeptical when it comes to introducing communicating devices in their premises such as offices, plants, and supply chains. For example, businesses may be worried about fraudulent access to their inventories, work orders, and strategic business plans. Remote monitoring and/or control can have drastic consequences.

Another challenging issue is regulation. A recent consultation that drew over 600 responses from industry, government, academia, civil society, and consumer groups shows that opinions differ on what policymaking is needed for IoT. Industry stakeholders argue that unnecessary regulation could cripple innovation and compromise IoT business models. Consumers, on the other hand, are worried about the potential impact of IoT with regard to data protection and privacy. Innovation cannot be achieved at the expense of fundamental human rights. However, it may not be fair to place the emphasis solely on the potential problems of IoT. In fact, IoT is expected to grow mainly through users’ trust.

Further, while there are claims about the need for common IoT standards, there are actually an overwhelming number of standards for IoT, emanating from mainstream standards development organizations (SDOs), mainly IETF, ITU-T, IEEE, ETSI, ISO/IEC, and the International Society of Automation (ISA), as well as other state-funded and international projects [3]. Therefore, without unified effort, instead of converging toward common standards, this overwhelming number of proposals might contribute to further exacerbating the confusion about services and regulation. In fact, IoT standardization is a very crowded area, to the extent that it is often compared to a war [4]. As such, we believe it is necessary to synthesize all ongoing activities to help researchers and stakeholders get a clear view of what is going on in terms of IoT standardization actions worldwide.

We first give a comprehensive overview of some of the main challenges IoT needs to overcome in terms of service definitions and regulation. These are discussed later. We illustrate that there is no general consensus on nor common understanding about, IoT services. We provide a survey of IoT standards. State-funded and independent IoT projects, as well as initiatives, clusters, forums, etc., are described. We finally conclude the article by summarizing the most important findings and by “connecting the dots” between the standardization activities.

IoT SERVICE DEFINITIONS

The term Internet of Things was first introduced back in 1999 by the Auto-ID Labs (autoidlabs.org) (formerly the Auto-ID Center of MIT), primarily for networked radio frequency identification (RFID) devices. Since then the concept has evolved, and nowadays IoT encompasses many other technologies including wireless sensor networks (WSN), near field communications (NFC), biotechnology and body area networks (BAN), machine-to-machine (M2M) communications, and other “legacy” personal area networks (PAN) such as WiFi, Bluetooth, etc. [5].
Inexpensive, highly personalized, ubiquitous, and instantaneous services constitute key success factors for IoT. Unfortunately, so far there is no general consensus on what IoT is exactly meant to be. There is seemingly a “race” between IoT stakeholders to the extent that driven by time-to-market constraints and business requirements, most proposals made so far are fragmented and fail to consider all facets required to deliver a globally accepted IoT service. These proposals often lead to ambiguous and even contradictory definitions. For example, some proposals assimilate IoT to Web 3.0; others claim it is primarily based on RFID and similar systems; others associate IoT with machine-to-machine (M2M) communications; while others focus on WSN. Perhaps IoT is all of this. However, between web semantics, radio frequency identification, and sensor networks, there is a huge vacuum that needs to be filled by adequate standards and definitions to pave the way toward an IoT ecosystem.

Figure 1 depicts a high-level IoT stand-alone or “silo” architecture with most commonly used verticals including IPv6 over low-power wireless personal area networks (6LoWPAN), Zigbee (www.zigbee.org), Z-Wave (www.z-wave.com), wireless highway addressable remote transducer (WirelessHART http://en.hartcomm.org), ISA-100.11a, Bluetooth Low Energy (BLE), RFID, Developers Alliance for Standards Harmonization of ISO/IEC 18000-7 (DASH7), Building Automation and Control Network (BACNet, http://bacnet.org), Local Operating Network (LonWorks, www.onmark.org), KNX (http://knx.org), and INSTEON (www.insteon.com). These are interconnected via an IPv6 backbone to familiar or legacy technologies such as 3G/4G/5G, WiFi, LiFi, Home Plug, vehicular networks (VANETs), as well as PSTN and IPv4 enterprise networks.

Although most of these technologies are streamlined and widely deployed, in an IoT ecosystem where secure end-to-end interoperability will be needed, global standards and regulatory rules will be required [6]. For instance, there is a myriad of home automation “industry standards.” Popular wired technologies include X-10 (X-10.com), universal powerline bus (UPB), KNX, LonWorks, and BACnet. On the other hand, Z-Wave and ZigBee are some of the most popular wireless solutions, while INSTEON is a combined wireless/wired solution. Each of these solutions has its own advantages and drawbacks, but the most important concern is that often they do not interoperate.

Similarly, WirelessHART and ISA-100.11a are two of the most popular technologies used for smart plant, field process, and automation systems, yet they do not interoperate at all, although they both rely on the IEEE 802.15.4 MAC layer.

Beyond technical and syntactic interoperability, semantic interoperability constitutes another challenge to IoT systems. In a fully heterogeneous environment, semantic interoperability uses a common ontology for describing resources across fragmented IoT systems [7]. For instance, the ontology of the World Wide Web Consortium (W3C) semantic sensor networks (SSN) incubator group aims to overcome the limitations of XML formats and the fragmentation of sensor ontologies. However, this ontology describes sensors, observations, and related concepts, but not domains. Thus, it was further extended to specific IoT domains and applications such as smart cities and smart homes [8].

Further, the open source cloud solution for the Internet of Things (www.openiot.eu) project provides a common semantic layer enabling the annotation of sensors and devices as W3C SSN compliant sensors. In fact, OpenIoT is an open source cloud-based platform supporting components such as sensor middleware and cloud data storage.

Furthermore, as a result of the proliferation of big data, cloud computing, and sensor networks, “sensing as a service” is emerging as a promising IoT paradigm aiming at leveraging the sensed “big” data [9]. In fact, if all gathered big data is not processed, it may be useless to sense it. For example, in the smart city domain, the huge amount of video streams, environmental monitoring, surveillance, and traffic control may be very difficult to track. If this sensed information is made available on the net, then there is almost surely someone who might exploit at least part of this information.

**IoT Service Regulation**

While regulation of the traditional Internet is primarily driven by service, regulation of IoT is primarily driven by trust, security, and privacy. However, there are some fears that regulation might cripple the development of IoT.

So far, a large part of the Internet community is opposing internet regulation. As a matter of fact, the Internet Society (intersociety.org) works to ensure that three key aspects of the Internet are retained: permissionless innovation, open access, and collaboration. One of the most important arguments is that regulators may not be able to anticipate how IoT will evolve.
According to some views, specific IoT regulations may not be needed. Only issues specific to IoT and that cannot be adequately addressed by existing rules may require specific policies. For instance, the European General Data Protection Regulation provides a regulatory framework that may address IoT-specific privacy concerns.

and that regulatory rules could have unintended consequences. As a matter of fact, according to Vinton Cerf, “Regulation is tricky...we’re going to have to experience the problems before we understand the nature of the problems.”

Besides, IoT regulation might be more beneficial for consumers than for governments. For instance, in the USA, state and federal regulators are working to restrict government and private-sector control of IoT.

On the other hand, several technology companies are promoting the virtues of self-regulation when it comes to managing consumer data [10]. They further stress the benefits of leveraging large amounts of big data to simplify daily tasks and give consumers the option to make conscious decisions. According to the Consumer Technology Association (CTA, cesweb.org), “Big data innovation would be stifled by governmental regulation.” AT&T also prefers “proactive, industry-led initiatives and best-practices guidelines” to ensure appropriate handling of sensitive consumer data.

However, without regulation, security features provided by vendors can be very poor. According to a recent technical risk assessment performed on 43 healthcare mobile applications (privacyrights.org), only 15 percent of the apps encrypted all of the transmitted data. Further, none of the apps encrypted data stored on the users’ mobile device. Furthermore, most apps connect to third-party sites without the user’s knowledge. Another important finding was that 72 percent of the apps were considered to be presenting medium (32 percent) to high (40 percent) risk with respect to privacy. Further, the apps that presented the lowest privacy risk were paid apps.

In fact, according to some views, specific IoT regulations may not be needed. Only issues specific to IoT and which cannot be adequately addressed by existing rules may require specific policies. For instance, the European General Data Protection Regulation provides a regulatory framework that may address IoT-specific privacy concerns.

**IoT Mainstream Standardization Activities**

SDOs such as IETF, ITU-T, ETSI, ISO/IEC, and IEEE are actively working on service definitions, architectures, and security aspects. While each of these SDOs may have a particular perspective regarding IoT, a significant effort is currently being carried out to bring these perspectives closer.

**IETF Standardization Activities**

Back in 2006, the IETF started working on a set of IoT standards. The work originally focused on running IP over various resource-constrained networks. In particular, a series of drafts are being considered to investigate the transmission of IPv6 packets over NFC, Bluetooth Low Energy (BLE), multi slave twisted pair (MS/TP), i.e. BACnet networks, and ITU-T G.9959 networks.

**ITU-T Standardization Activities**

While the ITU-T has been primarily working on switched, connection oriented networks, its involvement in the Internet, primarily through the Next Generation Network (NGN) framework, assures a converged approach in developing IoT architectures. Note that collaboration between ITU-T and IETF is not new; back in 1998, RFC 2436 set the scope of this collaboration.

In 2005, the ITU-T published several reports on IoT as one of the ITU-T Internet report series. It covers topics including “enabling technologies, business opportunities, public policy challenges, and implications for the developing world.” The ITU-T’s IoT Global Standards Initiative (IoT-GSI) is intended to promote a unified approach and develop recommendations “enabling IoT on a global scale” and to “act as an umbrella for IoT standards activities worldwide” (www.itu.int/en/ITU-T/gsi). This will in turn give service providers the means to offer a large variety IoT services. ITU-T also created the Joint Coordination Activity on Internet of Things (JCA-IoT) to coordinate ITU-T’s work on IoT, including network aspects of Identification of Things and ubiquitous sensor networks (USN).

In 2012, ITU-T Study Group 13 (SG 13) approved a set of recommendations that define IoT, characterize its emerging environment, and outline the “functional requirements of machine-oriented communication applications in an NGN context.” Among these recommendations, Y.2060 defines IoT as “a global infra-
structure for the Information Society, enabling advanced services by interconnecting (physical and virtual) things based on existing and evolving, interoperable information and communication technologies. “An overview and a detailed description of an IoT Reference Model are also provided in rec. Y.2060.

On the other hand, ITU-T SG 15 is actively working on communication aspects related to smart grids. Recommendations of particular importance include G.9903 and G.9905, which normatively reference 6lowPAN and its header compression mechanisms. In fact, Rec. G.9903 adopted the IEEE 802.15.4 and the IETF 6LoWPAN MAC and adaptation layers, respectively.

Finally, the ITU-T SG17 is working on cyber security and identity management for IoT and related environments, including cloud computing, smart grids, web services, etc. Among the recommendations approved by SG17, Rec. X.1205 provides an overview of cyber security. A list of selected IoT-related ITU-T recommendations is given in Table 1.

ISO/IEC Standardization Activities

The ISO/IEC (www.iec.ch) Special Working Group 5 (SWG 5) of the ISO/IEC Joint Technical Committee 1 (JTC 1) was established in 2012 as a result of a growing interest in the field of IoT by other SDOs. JTC 1 has tight relationships with ITU-T SG-17 on various security aspects. The relationship is at various levels including joint work (level 1), technical collaboration by liaison mechanism (level 2), and information liaison (level 3).

ISO/IEC JTC 1/SWG 5 does not actually develop standards, but rather consolidates standardization activities and identifies current and future IoT trends and needs. A number of documents were issued by the SWG, including a collection of definitions and a mind map describing technologies related to IoT, as well as application domains, requirements, and stakeholders. A list of definitions collected from various standards organizations is divided into four categories: IoT, M2M, machine type communications (MTC), and cyber-physical systems (CPS).

The ISO/IEC NP 19654 standard introduces a reference architecture (RA) as a “generalized system-level architecture of IoT systems that share common domains.” The IoT RA also provides rules, guidance, and policies for building a specific IoT system architecture. IoT RA also aims to help develop interoperable IoT systems that interact seamlessly. The IoT RA includes three key enabling technology areas:
1) IoT system of interest.
2) Communications technology.
3) Information technology.

An IoT system of interest includes smart health care, agriculture, environment, grid, building, transportation, city, etc. The IoT RA standard also describes a conceptual model where seven IoT domains are defined:

1) IoT System: System that is to be developed, implemented, and operated. This domain includes descriptions of target applications and services of the IoT system. These include health care, grid, home, etc.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y.2060</td>
<td>Overview of the Internet of Things</td>
</tr>
<tr>
<td>Y.2061</td>
<td>Requirements for support of machine-oriented communication applications in the NGN environment</td>
</tr>
<tr>
<td>Y.2062</td>
<td>Framework of object-to-object communication for ubiquitous networking in next generation networks</td>
</tr>
<tr>
<td>Y.2063</td>
<td>Framework of the web of things</td>
</tr>
<tr>
<td>Y.2064</td>
<td>Energy saving using smart objects in home networks</td>
</tr>
<tr>
<td>Y.2065</td>
<td>Service and capability requirements for e-health monitoring services</td>
</tr>
<tr>
<td>Y.2066</td>
<td>Common requirements of the Internet of Things</td>
</tr>
<tr>
<td>Y.2067</td>
<td>Common requirements and capabilities of a gateway for Internet of Things applications</td>
</tr>
<tr>
<td>Y.2068</td>
<td>Functional framework and capabilities of the Internet of Things</td>
</tr>
<tr>
<td>Y.2069</td>
<td>Terms and definitions of the Internet of Things</td>
</tr>
<tr>
<td>Y.2070</td>
<td>Requirements and architecture of home energy management systems and home network services</td>
</tr>
<tr>
<td>Y.2074</td>
<td>Requirements for Internet of Things devices and operation of Internet of Things applications during disaster</td>
</tr>
<tr>
<td>Y.2213</td>
<td>NGN service requirements and capabilities for network aspects of applications and services using tag-based identification</td>
</tr>
<tr>
<td>F.771</td>
<td>Service description and requirements for multimedia information access triggered by tag-based identification</td>
</tr>
<tr>
<td>G.9903</td>
<td>Narrow-band OFDM power line communication transceivers for G3-PLC networks</td>
</tr>
<tr>
<td>G.9905</td>
<td>Centralized metric based source routing</td>
</tr>
<tr>
<td>H.621</td>
<td>Architecture of a system for multimedia information access triggered by tag-based identification</td>
</tr>
<tr>
<td>X.1691</td>
<td>Security framework for cloud computing</td>
</tr>
<tr>
<td>X.1205</td>
<td>Overview of cyber security</td>
</tr>
<tr>
<td>X.1312</td>
<td>Ubiquitous sensor network middleware security guidelines</td>
</tr>
<tr>
<td>X.1313</td>
<td>Security requirements for wireless sensor network routing</td>
</tr>
<tr>
<td>X.1314</td>
<td>Security requirements and framework of ubiquitous networking</td>
</tr>
<tr>
<td>X.1171</td>
<td>Threats and requirements for protection of personally identifiable information in applications using tag-based identification</td>
</tr>
<tr>
<td>X.672</td>
<td>Object identifier resolution system (ORS)</td>
</tr>
<tr>
<td>X.660</td>
<td>General procedures and top arcs of the international object identifier tree</td>
</tr>
</tbody>
</table>

Table 1. ITU-T recommendations related to IoT.
2) Sensing Devices: Domain representing all physical entities such as sensors, tag readers, etc.

3) Things/Objects: Physical (things) and virtual (objects) domain that includes entities that are part of the IoT system domain that do not have sensors. These include displays, alarms, smartphones, etc.

4) Control/Operations: Domain representing organizations that manage the system activities of an IoT system.

5) Service Providers: Domain representing organizations providing IoT services.

6) Customers: Domain representing the end user of goods (both tangible and intangible) and services provided by the organizations in the service providers domain or by the IoT system in IoT systems of interest.

7) Markets: Domain representing operators and participants in the IoT system and service provider markets.

Various ISO/IEC standards related to IoT have been published or are under development.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Title</th>
<th>Status (Year)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISO/IEC NP 19654</td>
<td>Internet of Things reference architecture (IoT RA)</td>
<td>Under development (2014)</td>
</tr>
<tr>
<td>ISO/IEC 30101</td>
<td>Sensor network and its interfaces for smart grid system</td>
<td>Published (2014)</td>
</tr>
<tr>
<td>ISO/IEC 14543</td>
<td>Information technology – home electronic systems (HES) – Part 3-10: wireless short-packet (WSP) protocol optimized for energy harvesting – architecture and lower layer protocols</td>
<td>Published (2012)</td>
</tr>
<tr>
<td>ISO/IEC 24767</td>
<td>Information technology – home network security</td>
<td>Published (2008)</td>
</tr>
<tr>
<td>ISO/IEC 29180</td>
<td>Information technology – telecommunication and information exchange between systems – security framework for ubiquitous sensor networks</td>
<td>Published (2012)</td>
</tr>
<tr>
<td>ISO/IEC 15961-1</td>
<td>Information technology – radio frequency identification (RFID) for item management: data protocol – Part 1: application interface</td>
<td>Published (2013)</td>
</tr>
<tr>
<td>ISO/IEC 15962</td>
<td>Information technology – radio frequency identification (RFID) for item management – data protocol: data encoding rules and logical memory functions</td>
<td>Published (2013)</td>
</tr>
<tr>
<td>ISO/IEC 15963</td>
<td>Information technology – radio frequency identification for item management – unique identification for RF tag</td>
<td>Published (2009)</td>
</tr>
<tr>
<td>ISO/IEC 18000-x</td>
<td>Information technology – radio frequency identification for item management – Part x</td>
<td>Published</td>
</tr>
<tr>
<td>ISO/IEC 24791-x</td>
<td>Information technology – radio frequency identification (RFID) for item management – software system infrastructure – Part x</td>
<td>Published</td>
</tr>
</tbody>
</table>

Table 2. ISO/IEC specifications related to IoT.
standard has become a de facto MAC and PHY layer protocol for various IoT technologies.

The IEEE 802.15 WG is also actively working on other IoT standards: Part 15.5 dealing with Mesh Topology Capability in WPAN; Part 15.6 dealing with MAC and Physical Layer Specifications for WPANs used in or around a body; and Part 15.7 dealing with Short-Range Wireless Optical Communication Using Visible Light.

Moreover, IEEE 802.15.4e Time Slotted Channel Hopping (TSCH) recently amended the MAC layer of the IEEE 802.15.4 standard for industrial automation and process control networks. TSCH evolved directly from WirelessHART and ISA100.11a. To further consolidate standardization efforts, the IETF’s 6TISCH WG is currently working on IPv6 over TSCH in order to facilitate the adoption of IPv6 in industrial environments.

**ETSI Standardization Activities**

ETSI (www.etsi.org) has just introduced the low throughput network (LTN) as a WAN, bidirectional wireless networking technology. Key different features compared to common WSNs include long-range transmission (up to 40 km in line of sight) and communication with buried underground equipment.

There are three new ETSI group specifications (GS) related to LTN: GS LTN 001, GS LTN 002, and GS LTN 003, dealing with, respectively, use cases, LTN functional architecture, and definitions of protocols and interfaces. The functional architecture makes use of an IP-based transport network such as 3rd Generation Partnership Project (3GPP), Telecommunications and Internet Converged Services and Protocols for Advanced Networking (TISPAN), and 3GPP2.

ETSI also published a set of technical specifications (TS) covering service requirements, the M2M functional architecture, and CoAP interoperability. Table 3 gives a list of selected ETSI group and technical specifications, as well as technical reports (TR) related to IoT. Note that current ETSI M2M standardization activities are primarily done under the oneM2M umbrella (see below).

**oneM2M Standardization Activities**

In 2012, the oneM2M standardization body emerged as a unified effort of SDOs: ETSI from Europe, ATIS (Alliance for Telecommunications Industry Solutions) and TIA (Telecommunications Industry Association) from the US, CCSA (China Communications Standards Association), TTA (Telecommunications Technology Association of Korea), and ARIB (Association of Radio Industries and Businesses) and TTC (Telecommunication Technology Committee) from Japan. TSDSI (Telecommunications Standards Development Society, India) also joined oneM2M in May 2015.

As of December 2015, 224 companies and six universities had joined oneM2M. The goal is to develop common specifications for “M2M services that have been vertically integrated” so far and to propose a “common M2M service layer that can be embedded in various hardware and software.” Table 4 gives a list of oneM2M technical reports and technical specifications (TS) related to IoT. Note that oneM2M TS are common with ETSI, ATIS, TTA, and TTC.

**3GPP Standardization Activities**

Cellular technologies are being adapted to meet IoT requirements. In particular, LTE Release-12 (Rel-12) of the 3GPP (www.3gpp.org) introduces a power save mode and simplified signalling procedures to provide additional battery savings. Rel-12 also allows LTE modems to be significantly less complex and cheaper than current modems. Further, 3GPP has identified ways to increase the coverage of LTE, making it possible to communicate with objects in difficult to reach locations.

LTE Rel. 12 is paving the way toward 5G, and may constitute a major step toward ubiquitous access to IoT and M2M services. The advent of LTE enabled IoT devices may constitute a key milestone in the development of IoT mobility and long reach features.

### Table 3. ETSI specifications related to IoT

<table>
<thead>
<tr>
<th>Reference</th>
<th>Title</th>
<th>Status (Year)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GS LTN 001</td>
<td>Low throughput networks (LTN); use cases for low throughput networks</td>
<td>Published (2014)</td>
</tr>
<tr>
<td>GS LTN 002</td>
<td>Low throughput networks (LTN); functional architecture</td>
<td>Published (2014)</td>
</tr>
<tr>
<td>GS LTN 003</td>
<td>Low throughput networks (LTN); protocols and interfaces</td>
<td>Published (2014)</td>
</tr>
<tr>
<td>TS 102 689</td>
<td>Machine-to-machine communications (M2M); M2M service requirements</td>
<td>Published (2013)</td>
</tr>
<tr>
<td>TS 103 104</td>
<td>Machine-to-machine communications (M2M); interoperability test specification for CoAP binding of ETSI M2M primitives</td>
<td>Published (2013)</td>
</tr>
<tr>
<td>TS 102 690</td>
<td>Machine-to-machine communications (M2M); functional architecture</td>
<td>Final draft for proposal (2014)</td>
</tr>
<tr>
<td>TR 102 996</td>
<td>Machine-to-machine communications (M2M); interfacing between the M2M architecture and M2M area network technologies</td>
<td>Published (2014)</td>
</tr>
<tr>
<td>TR 102 995</td>
<td>Machine-to-machine communications (M2M); applicability of M2M architecture to smart grid networks; impact of smart grids on M2M platform</td>
<td>Published (2012)</td>
</tr>
<tr>
<td>TR 102 688</td>
<td>Machine to machine communications (M2M); use cases of automotive applications in M2M capable networks</td>
<td>Published (2013)</td>
</tr>
<tr>
<td>TR 102 687</td>
<td>Machine-to-machine communications (M2M); use cases of M2M applications for connected consumer</td>
<td>Published (2013)</td>
</tr>
<tr>
<td>TR 102 732</td>
<td>Machine-to-machine communications (M2M); use cases of M2M applications for eHealth</td>
<td>Published (2013)</td>
</tr>
<tr>
<td>TR 102 725</td>
<td>Definitions</td>
<td>Published (2013)</td>
</tr>
<tr>
<td>TR 101 584</td>
<td>Machine-to-machine communications (M2M); study on semantic support for M2M data</td>
<td>Published (2013)</td>
</tr>
<tr>
<td>TR 102 449</td>
<td>Telecommunications and internet converged services and protocols for advanced networking (TISPAN); overview of radio frequency identification (RFID) tags in the telecommunications industry</td>
<td>Published (2005)</td>
</tr>
<tr>
<td>TR 102 644</td>
<td>Electromagnetic compatibility and radio spectrum matters (ERM); RFID pluggable to investigate the interoperability of tags manufactured by different vendors;</td>
<td>Published (2009)</td>
</tr>
</tbody>
</table>
The goal of a standard is typically to unify interfaces, protocols, and services so that various systems can be interconnected. After this overview of IoT standards, a legitimate question is: Do we have a common and clear understanding of a standard IoT service? The answer is quite mitigated. In fact, to some extent, the overwhelming number of standards might have contributed in further exacerbating the ambiguity about service and may deepen the interoperability issues. In fact, most of these standards limit their scope to specific domains (M2M, WSN, RFID, etc.) and stakeholders yielding isolated and/or redundant solutions.

We believe that SDOs should further unify their efforts around IETF’s streamlined protocols, namely IP and its variants, i.e. primarily 6LoWPAN in the network layer and CoAP in the application layer; and on the IEEE 802.x.y standard (primarily 802.15.4) in the MAC and PHY layers, be it for M2M, WSN/USN, or tag communications. All smart home/grid/city/transportation services may use these streamlined standard protocols. As further emphasized in an IETF liaison statement “lessons learned about cooperation between SDOs” (http://datatracker.ietf.org/liaison/549/), “SDOs should minimize potential duplicate work, and minimize this via collaboration, not competition.”

As a matter of fact, ETSI and the Internet Protocol for Smart Objects (IPSO) alliance organized a successful CoAP “Plugtest” interoperability in March 2012. Also, in July 2013, a 6LoWPAN Plugtests event gave vendors the opportunity to assess the level of interoperability of their products and verify the correct interpretation of IETF based specifications. The tests were carried out using the 2006-2.4 GHz release of the IEEE 802.15.4 PHY/MAC standard. While almost all implementations have exhibited a great level of basic compatibility, i.e. data was sent and interpreted correctly, compliance with RFC 6775 (6LoWPAN-ND) was very low.

The IETF is also actively working on consolidating standardization efforts of various bodies. The 6Lo WG is currently working on how to transport IPv6 packets over NFC, BLE, MS/TP, and ITU-T G.9959 networks. On the other hand, the 6TiSCWG WG is working on IPv6 over IEEE 802.15.4e TSCH in order to accelerate the adoption of IPv6 in industrial environments.

As another collaboration initiative, the Global Standards Collaboration (GSC) encompasses 11 SDOs including ARIB, ATIS, CCSA, ETSI, IEEE-SA, ITU-T/R, TIA, TSDSI TTA, and TTC. So far, the IETF, a key player in IoT development, is not a GSC member, while ISO and IEC are expected to join during the 20th meeting in India, in 2016. One of the main goals of GSC is to harmonize standards to guarantee compatibility of communication equipment. In July 2014, the 15th meeting of the GSC was hosted in Kolkata where IoT, critical communications (CC), big data, M2M, and software defined networking (SDN) and their applications in various domains, primarily in emergency communication systems, were discussed. Note that ISO/IEC JTC 1 was among the guests during that meeting. In July

---

**Table 4.** oneM2M technical reports and specifications related to IoT.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Title</th>
<th>Status (Year)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TR0001</td>
<td>oneM2M use cases collection</td>
<td>Published (2013)</td>
</tr>
<tr>
<td>TR0002</td>
<td>Architecture analysis – part 1: analysis of architectures proposed for consideration by oneM2M</td>
<td>Published (2013)</td>
</tr>
<tr>
<td>TR0003</td>
<td>Architecture analysis – part 2: study for the merging of architectures proposed for consideration by oneM2M</td>
<td>Published (2013)</td>
</tr>
<tr>
<td>TR0006</td>
<td>Study of management capability enablement technologies for consideration by oneM2M</td>
<td>Published (2013)</td>
</tr>
<tr>
<td>TR0008 (ETSI TR118508)</td>
<td>Analysis of security solutions for the oneM2M system</td>
<td>Published (2014)</td>
</tr>
<tr>
<td>TR0009</td>
<td>Protocol analysis</td>
<td>Published (2014)</td>
</tr>
<tr>
<td>TS 0001 (ETSI TS 118 110)</td>
<td>Functional architecture</td>
<td>Published (2015)</td>
</tr>
<tr>
<td>TS 0002 (ETSI TS 118 102)</td>
<td>Requirements</td>
<td>Published (2015)</td>
</tr>
<tr>
<td>TS 0003 (ETSI TS 118 103)</td>
<td>Security solutions</td>
<td>Published (2015)</td>
</tr>
<tr>
<td>TS 0004 (ETSI TS 118 104)</td>
<td>Service layer core protocol specification</td>
<td>Published (2015)</td>
</tr>
<tr>
<td>TS 0005 (ETSI TS 118 105)</td>
<td>Management enablement (OMA)</td>
<td>Published (2015)</td>
</tr>
<tr>
<td>TS 0006 (ETSI TS 118106)</td>
<td>Management enablement (BBF)</td>
<td>Published (2015)</td>
</tr>
<tr>
<td>TS 0008 (ETSI TS 118 108)</td>
<td>CoAP protocol binding</td>
<td>Published (2015)</td>
</tr>
<tr>
<td>TS 0009 (ETSI TS 118 109)</td>
<td>HTTP protocol binding</td>
<td>Published (2015)</td>
</tr>
<tr>
<td>TS 0010 (ETSI TS 118110)</td>
<td>MQTT protocol binding</td>
<td>Published (2015)</td>
</tr>
<tr>
<td>TS 0011 (ETSI TS 118 111)</td>
<td>Common terminology</td>
<td>Published (2015)</td>
</tr>
</tbody>
</table>

---

**INDEPENDENT AND STATE-FUNDED PROJECTS**

While the mainstream SDOs are actively working on IoT service definitions and architectures, many independent and state-funded projects are being carried out to support and promote IoT worldwide, ranging from alliances (ipso-alliance.org, allseenalliance.org), architectures (iot-a.eu), consortia (iotthings.org, industrial-internetconsortium.org, openinterconnect.org), forums (iot-forum.eu, wireless-iot.org), groups (threadgroup.org), initiatives (iot-i.eu, home-gatewayinitiative.org), projects (iot6.eu, probe-it.eu, openiot.eu, iot-iotest.eu, iot-at-work.eu, iot-icore.eu), research clusters (internet-of-things-research.eu), and the list goes on! Despite the crowd, “newcomers” keep stepping in, such as the Wireless IoT Forum recently.

Among the state-funded projects, the European Internet of Things Architecture (IoT-A) seems to be standing out from the crowd and gaining reasonable acceptance. On the other hand, with more than 100 stakeholders, the Industrial Internet Consortium (IIC) and the AllSeen Alliance may stand out from the crowd as industry-led activities. IIC works on an architectural framework for the Industrial Internet and plays a key role in the development of IoT. The other IoT “activities” are summarized in Table 5.

---

**CONCLUSION: CONNECTING THE DOTS**

The goal of a standard is typically to unify interfaces, protocols, and services so that various systems can be interconnected. After this overview of IoT standards, a legitimate question is: Do we have a common and clear understanding of a standard IoT service? The answer is quite mitigated. In fact, to some extent, the overwhelming number of standards might have contributed in further exacerbating the ambiguity about service and may deepen the interoperability issues. In fact, most of these standards limit their scope to specific domains (M2M, WSN, RFID, etc.) and stakeholders yielding isolated and/or redundant solutions.

We believe that SDOs should further unify their efforts around IETF’s streamlined protocols, namely IP and its variants, i.e. primarily 6LoWPAN in the network layer and CoAP in the application layer; and on the IEEE 802.x.y standard (primarily 802.15.4) in the MAC and PHY layers, be it for M2M, WSN/USN, or tag communications. All smart home/grid/city/transportation services may use these streamlined standard protocols. As further emphasized in an IETF liaison statement “lessons learned about cooperation between SDOs” (http://datatracker.ietf.org/liaison/549/), “SDOs should minimize potential duplicate work, and minimize this via collaboration, not competition.”

As a matter of fact, ETSI and the Internet Protocol for Smart Objects (IPSO) alliance organized a successful CoAP “Plugtest” interoperability in March 2012. Also, in July 2013, a 6LoWPAN Plugtests event gave vendors the opportunity to assess the level of interoperability of their products and verify the correct interpretation of IETF based specifications. The tests were carried out using the 2006-2.4 GHz release of the IEEE 802.15.4 PHY/MAC standard. While almost all implementations have exhibited a great level of basic compatibility, i.e. data was sent and interpreted correctly, compliance with RFC 6775 (6LoWPAN-ND) was very low.

The IETF is also actively working on consolidating standardization efforts of various bodies. The 6Lo WG is currently working on how to transport IPv6 packets over NFC, BLE, MS/TP, and ITU-T G.9959 networks. On the other hand, the 6TiSCWG WG is working on IPv6 over IEEE 802.15.4e TSCH in order to accelerate the adoption of IPv6 in industrial environments.

As another collaboration initiative, the Global Standards Collaboration (GSC) encompasses 11 SDOs including ARIB, ATIS, CCSA, ETSI, IEEE-SA, ITU-T/R, TIA, TSDSI TTA, and TTC. So far, the IETF, a key player in IoT development, is not a GSC member, while ISO and IEC are expected to join during the 20th meeting in India, in 2016. One of the main goals of GSC is to harmonize standards to guarantee compatibility of communication equipment. In July 2014, the 15th meeting of the GSC was hosted in Kolkata where IoT, critical communications (CC), big data, M2M, and software defined networking (SDN) and their applications in various domains, primarily in emergency communication systems, were discussed. Note that ISO/IEC JTC 1 was among the guests during that meeting. In July
2015, the 19th meeting of the GSC was hosted by ITU and also focused on IoT, CC, and 5G. The various SDOs agreed to further increase collaboration and leverage existing efforts such as oneM2M.

Finally, policymakers may not impose unnecessary regulatory rules on IoT and should identify carefully what information actually requires protection. In order to instill consumer confidence in IoT without slowing its growth, awareness, good practice, and voluntary codes of conduct that stop short of regulation may constitute more viable options.

REFERENCES


BIOGRAPHY

Aref Meddeb (Aref.Meddeb@infcom.rnu.tn) obtained his engineer’s degree from ENIT, Tunisia, in 1992, and both his M.S. and Ph.D. degrees from Ecole Polytech-

nique, Montreal, Canada, in 1995 and 1998, respectively. From 1993 to 2002 he worked with Alcatel, INRS-Telecom, Teleglobe, and Nortel. He is currently a full professor at the National School of Engineering, University of Sousse. His research interests include Internet of Things, wireless sensor networks, and RFID systems, with a focus on security, QoS, routing, and design.

<table>
<thead>
<tr>
<th>Name</th>
<th>Number of partners/members (Jan. 2015)</th>
<th>Year</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>IOT-A</td>
<td>13 companies</td>
<td>2010</td>
<td><a href="http://www.iot-a.eu">http://www.iot-a.eu</a></td>
</tr>
<tr>
<td>IPSO Alliance</td>
<td>39 companies</td>
<td>2008</td>
<td><a href="http://www.ipso-alliance.org">http://www.ipso-alliance.org</a></td>
</tr>
<tr>
<td>AllSeen Alliance</td>
<td>106 companies</td>
<td>2014</td>
<td><a href="https://allseenalliance.org">https://allseenalliance.org</a></td>
</tr>
<tr>
<td>Internet of Things Consortium</td>
<td>48 companies</td>
<td>2014</td>
<td><a href="http://i0ththings.org">http://i0ththings.org</a></td>
</tr>
<tr>
<td>Industrial Internet Consortium</td>
<td>127 companies</td>
<td>2014</td>
<td><a href="http://www.iiconsortium.org">http://www.iiconsortium.org</a></td>
</tr>
<tr>
<td>Open Interconnect Consortium</td>
<td>49 companies</td>
<td>2014</td>
<td><a href="http://openinterconnect.org">http://openinterconnect.org</a></td>
</tr>
<tr>
<td>Internet of Things International Forum (IoTForum)</td>
<td>9 companies</td>
<td>2013</td>
<td><a href="http://iotforum.org/">http://iotforum.org/</a></td>
</tr>
<tr>
<td>Wireless IoT Forum</td>
<td>6 companies</td>
<td>2015</td>
<td><a href="http://www.wireless-iot.org/">http://www.wireless-iot.org/</a></td>
</tr>
<tr>
<td>Thread Group</td>
<td>46 companies</td>
<td>2014</td>
<td><a href="http://threadgroup.org/">http://threadgroup.org/</a></td>
</tr>
<tr>
<td>The Internet of Things Initiative</td>
<td>11 companies</td>
<td>2009-2012</td>
<td><a href="http://www.iot-i.eu/">http://www.iot-i.eu/</a></td>
</tr>
<tr>
<td>Home Gateway Initiative (HGI)</td>
<td>47 companies</td>
<td>2004</td>
<td><a href="http://www.homegatewayinitiative.org/">http://www.homegatewayinitiative.org/</a></td>
</tr>
<tr>
<td>Pursuing Roadmaps and Benchmarks for the Internet of Things (PROBE-IT)</td>
<td>1 company</td>
<td>2012-2014</td>
<td><a href="http://www.probe-it.eu/">http://www.probe-it.eu/</a></td>
</tr>
<tr>
<td>Open Source cloud solution for the Internet of Things (OpenIoT)</td>
<td>4 universities</td>
<td>2012-2014</td>
<td><a href="http://www.openiot.eu/">http://www.openiot.eu/</a></td>
</tr>
<tr>
<td>Internet of Things at Work (IoT@Work)</td>
<td>5 companies</td>
<td>2010-2013</td>
<td><a href="https://www.iot-at-work.eu/">https://www.iot-at-work.eu/</a></td>
</tr>
<tr>
<td>iCore</td>
<td>12 companies</td>
<td>2011-2014</td>
<td><a href="http://www.iot-icore.eu/">http://www.iot-icore.eu/</a></td>
</tr>
</tbody>
</table>

Table 5. IoT independent activities.
Standard-Based IoT Platforms Interworking: Implementation, Experiences, and Lessons Learned

The authors introduce standardized interworking interfaces and procedures based on oneM2M global standards, and tests them through use cases involving multiple IoT service platforms. The main purpose of the interworking experiment is to show how M2M/IoT service providers are using oneM2M compliant service layer platforms to deliver services more efficiently across multiple technology domains such as smart city.

Jaeho Kim, Jaeseok Yun, Sung-Chan Choi, Dale N Seed, Guang Lu, Martin Bauer, Adel Al-Hezmi, Konrad Campowsky, and Jaeseung Song

Abstract

The Internet-of-Things (IoT) provides a great opportunity to many vertical industries because IoT interconnects various devices such as sensors and actuators and collects/processes data from them in order to improve services and reduce costs. As there exists many IoT technologies in the market, global standards and interworking mechanisms are critical to the success of the IoT. This article introduces standardized interworking interfaces and procedures based on oneM2M global standards, and tests them through use cases involving multiple IoT service platforms. The interworking involves smart city applications/services running on multiple IoT service layer platforms interoperating with each other. The main purpose of the interworking experiment is to show how machine-to-machine (M2M)/IoT service providers are using oneM2M compliant service layer platforms to deliver services more efficiently across multiple technology domains such as smart city. Because the deployment configurations of this interworking experiment span multiple domains, and the IoT devices and platforms are from different companies, we believe that this interworking experiment clearly proves that global IoT standards specifications can foster implementations of a service layer that enables services and interoperability between devices/device networks and cloud-based applications.

Introduction

The Internet-of-Things (IoT) and machine-to-machine (M2M) communications provide a great opportunity to many industries. However, so far many industries have relied on vertically developed IoT/M2M solutions. These solutions typically use specifically customized hardware and software for a specific industry, resulting in an increase in capital expenditure (CAPEX) and operating cost (OPEX). As there exists already many IoT technologies and solutions in the market, global standards and interworking mechanisms are critical to the success of the IoT [1–4]. The oneM2M Global Initiative is a global standardization body for M2M and IoT in order to co-operate in the development of globally applicable, access-independent and technology agnostic IoT/M2M service layer specifications [5, 6]. This standardized service layer is intended to ensure that IoT devices can communicate and interwork with each other on a global scale. As oneM2M has recently issued its first release of IoT standards specifications, we take this opportunity to test interworking between multiple IoT service layer platforms implemented by different manufacturers. In this article, we realize a standardized interworking interface between oneM2M based IoT platforms. We also develop various interworking procedures with real-world smart services and various IoT devices. The interworking procedures incorporate three oneM2M compliant platforms developed by different vendors. These procedures demonstrate successful interworking between the different platforms.

Although our experiments show the feasibility of different IoT platforms interworking, we realize that there still are several standards issues that need to be tackled. Lessons that we have learned through the experiments, e.g., the implementations of standard-based IoT platforms, conducting interworking testing, and integrating smart city application to IoT platforms, are as follows.

Latency and connectivity loss. Architectural support for real time IoT applications such as a drone control service has not been the main subject of many currently available IoT standards (including oneM2M). As more real world applications will be operated on top of standardized IoT infrastructure, this needs to be considered in future IoT standards.

Pre-released specifications. In order to avoid unnecessary and costly interworking/interoperability testing with pre-release versions of a specification, a pre-analysis phase, which performs up-front analysis and test planning, should be exercised.

The need for an open API. To facilitate application implementation and market adoption of global IoT standards, API bindings for relevant programming languages should be defined as part of the standard.

Data interoperability. To support sharing and interworking of data between IoT devices, a standardized negotiation process regarding data format or data size needs to be considered.

This article discusses limitations of vertical IoT systems followed by an introduction to oneM2M standards activities, including their interworking architecture for different IoT service layer platforms. We then introduce an interworking configuration design and use cases. Next, the interworking of three oneM2M compliant service platforms and a smart city application is illustrated. Finally, we discuss lessons learned from the interworking testing and conclude this article.
Vertical IoT Systems and Limitations

Various existing vertical IoT systems have been deployed. Each vertical system has unique characteristics and requirements and targets different ecosystems. For example, intelligent transportation deals with vehicle networks and traffic information, while smart home focuses on intelligence enabled by connected appliances. Smart eHealth handles sensitive body condition data and personal health records. But those vertical systems are not completely independent of each other (e.g., body condition data could be leveraged in both smart eHealth and smart home) and separate deployments are not cost-effective. As such, a horizontal platform that can integrate or interwork those vertical systems becomes crucial [6, 7].

In the vertical IoT systems, the entities (e.g., the IoT device, gateway, and the infrastructure platform) composing an IoT eco-system are all provided and managed by the same company. From the end-user perspective, this vertical model can provide compatibility among the various IoT entities and a simple way to handle all management problems. However, this vertical model has limitations in that the IoT entities of one vertical cannot be easily shared or interworked with another vertical. As such, a horizontal platform that can integrate or interwork those vertical systems becomes crucial [6, 7].

OneM2M and IoT Interworking Architecture

OneM2M [3] is a global standardization body for M2M and IoT initiated by seven leading ICT standards organizations: ARIB and TTC (Japan); ATIS and TIA (USA); CCSA (China); ETSI (Europe); and TTA (Korea). The objective of oneM2M is to define a service layer that is network independent and provides interworking to different existing M2M vertical systems.

OneM2M High-Level Architecture Overview

As shown in Fig. 1, the oneM2M functional architecture also defines different types of nodes in various specified configurations that can interconnect and communicate with one another in specified topologies. Each node can be comprised of three types of logical entities.

• An application entity (AE) that consists of M2M application logic.
• A common services entity (CSE) that consists of common M2M services.
• A Network services entity (NSE) that interfaces services from underlying networks to the CSEs. An example of an AE is a smart metering application. Examples of CSE services include data management, device management, subscription/notification services, location services, and charging services. An example of an underlying network...
service provided to a CSE via an NSE is device triggering.

As shown in Fig. 1, an M2M service provider domain is divided into the field domain and the infrastructure domain. The field domain consists of application dedicated nodes (ADN), application service nodes (ASN), and middle nodes (MN), which can be physically implemented as sensors, M2M devices, and M2M gateways respectively. The infrastructure domain contains an infrastructure node (IN), which is physically mapped to the M2M service server. The nodes consist of at least one CSE or one AE. According to the type of node that hosts the CSE, the CSE can be categorized as an IN-CSE, MN-CSE, or ASN-CSE.

Communication between oneM2M entities takes place over the Mca, Mcc, Mcc', and Mcn, which are oneM2M defined reference points. For example, AEs and CSEs communicate over Mca; CSEs in the same M2M service provider domain communicate over Mcc; CSEs in different M2M service provider domains communicate over the Mcc'; and CSEs communicate with NSEs over the Mcn.

**oneM2M Standards Specifications**

oneM2M has published a suite of research reports and technical specifications for its Release 1 to enable a horizontal M2M service layer. From the protocol perspective, the M2M service layer sits between applications and the application protocol layer (e.g., HTTP). The M2M service layer interworks different application protocols, different vertical applications, different M2M service platforms, and different existing service functions.

As a horizontal service layer, oneM2M provides various interworking features, for example, interworking of vertical applications (via Mca reference points), interworking of various M2M service platforms, interworking of various existing service functions, interworking of various application protocols (such as CoAP, HTTP, and MQTT), and interworking with other oneM2M service platforms (via the Mcc' reference point). In the next section we illustrate further details on the interworking features of the oneM2M system.

**INTERWORKING ARCHITECTURE FOR IOT PLATFORMS**

By building IoT platforms based on the oneM2M standard, interworking of IoT platforms from multiple service providers, operators, and vendors is now more achievable than ever. Until now, IoT platforms from different vendors needed to implement their own proprietary service layer, which stood in the way of interworking between platforms from different vendors/providers. However, the standards based oneM2M service layer solves this interworking issue. Now IoT platforms can reap the same interoperability and interworking benefits from standardization at the M2M service layer they have come to rely on at the underlying network and transport layers.

In oneM2M, Mcc' plays an important role to provide an extension of service reachability from one service provider domain to multiple service provider domains. As shown in Fig. 2, for interworking between applications and a device/gateway/server, the messages delivered over the Mca reference point need to be oneM2M interoperable between AE and CSE. In addition, for interworking between device and gateway or between gateway and server,
the messages delivered over the Mcc reference point should follow oneM2M interoperability. Inter-M2M service provider domain communication over Mcc′ is to be processed between IN-CSEs in both M2M service provider domains. Integrating multi-vendor solutions in the domain of IoT is common, essential, and complicated. The standards based interworking architecture can serve to reduce interworking complexity.

**INTERWORKING DESIGN AND IMPLEMENTATIONS**

As discussed earlier, interworking is an important feature of oneM2M to foster the growth of the IoT market and innovation. In this section, in order to test as well as demonstrate the effectiveness of the oneM2M standard and the interworking feature, we introduce a design of an interworking testing system with multiple oneM2M compliant service platforms and smart services running through various oneM2M service platforms.

This system interworks these individually and independently developed oneM2M compliant platforms together with one another, as shown in Fig. 3. Each platform consists of a combination of oneM2M-based sensors, gateways, servers, and/or applications. The interworking experiments targeted in this article include:

- Registration procedures of sensors and actuators to a oneM2M compliant service layer platform.
- Mutual registrations of multiple adjacent service layer platforms.
- Smart city applications that discover the registered devices and introduce semantic data, thus enabling fast auto-configuration and big-data analytics.

Since the deployment configurations used in this interworking testing span multiple domains (i.e. field domain for end devices and infrastructure domain for infrastructure nodes) and the node entities (i.e. ADN, ASN, MN, and IN) are from different companies, the interworking clearly demonstrates that the oneM2M specifications can provide a service layer that enables services and interoperability between devices/device networks and cloud-based applications.

**Interworking Use Case:** The use case selected for this interworking experiment is that of a smart city, where sensors are deployed throughout a city at various locations. Sensors report sensor readings to gateways over Mca reference points. These gateways host oneM2M MN-CSEs, which in turn are connected to servers that host oneM2M IN-CSEs. Sensor readings are propagated from the gateways to the servers via oneM2M subscription/notification services over Mcc reference points spanning multiple oneM2M IN-CSEs from different vendors. Similarly sensor readings are further propagated between servers over Mcc′ reference points. Finally, these sensor readings can be propagated to a smart city application registered to one of the servers, again using oneM2M subscription/notification services over the Mca reference point.

**INTERWORKING PROCEDURES**

As described in the previous section, the interworking use case requires various procedures to be performed such as mutual registration, resource announcement, and subscription/notification. These procedures are key enablers for allowing data interworking between multiple IoT platforms. Mutual registration is the process in which a oneM2M node (e.g. a CSE in ASN) and a remote node (e.g. a CSE in MN) register to each other, while announcement is a process to create a resource at a remote CSE that is linked to the original resource that has been announced. Subscription and notification procedures allow monitoring of events involving changes to specific resources.

Figure 4 shows a high-level information flow of mutual registration, resource announcement, and subscription/notification that are used in our interworking experiments. The issuer CSE (i.e. IN-CSE #1) sends a registration message to the

---

**Figure 4. Mutual registration and resource announcements.**

![Diagram](image-url)
The smart city application is a commercial product that allows aggregating and processing of data coming from various sensors deployed in a smart city with the purpose of visualizing and comparing the data in a graphical user interface called the Dashboard.

Implementations and Experiments

In this section, we first illustrate the three oneM2M compliant IoT service layer platform implementations and a smart city application used as part of this interworking experiment.

Santander Smart City: The smart city application was run on top of the federated oneM2M platform consisting of the three oneM2M platforms. The smart city application is a commercial product that allows aggregating and processing of data coming from various sensors deployed in a smart city with the purpose of visualizing and comparing the data in a graphical user interface called the Dashboard. The application is shown with real smart city data coming from Santander. The SmartSantander European research project has deployed an infrastructure with about 15,000 individual sensors installed in an area of approximately 13.4 square miles [11] in the city of Santander. There are a large number of stationary sensors providing light and temperature information, but also noise, parking space, traffic sensors, a few weather, irrigation, and pollution sensors. In addition, there are a number of mobile sensors mounted on taxis and buses.

oneM2M Platform Implementations: For multiple oneM2M service layer platforms, three different platform implementations were used: Mobius [8], oneMPOWER [9], and openMTC [10].

oneMPOWER (PF2) is a commercial oneM2M service platform developed by Korea Electronics Technologies Institute (KETI). oneMPOWER (PF2) is a commercial oneM2M compliant service layer platform developed by InterDigital. Finally, openMTC (PF3) is an open IoT platform developed by Fraunhofer FOKUS. The smart city application running on top of these three IoT platforms is provided by NEC.

Interverking Configuration: For interworking experiments, a particular configuration composed of three oneM2M-compliant platforms (PF1, PF2, and PF3), and a smart city application, as shown in Fig. 3, were used. In addition, various prototype IoT devices developed by the participating companies were also attached to this configuration.

The Mcc′ interfaces interconnecting the three IN-CSEs are implemented with oneM2M messages bound to the HTTP protocol. The Mcc interfaces interconnecting the CSEs within the same infrastructure domain are implemented with both HTTP and CoAP protocol bindings. In addition, the MQTT broker residing in the Mobius platform allows the CSEs in the KETI′s infrastructure domain to support Mcc and Mca interfaces bound to the MQTT protocol.

Such interworking scenarios between IoT devices across multiple platforms clearly show the benefit of developing IoT applications based on global IoT standards.

With this configuration, four interworking scenarios were demonstrated:

• Collecting data of numerous sensors from the smart city and running smart city applications through interworking between two oneM2M-compliant service layer platforms, PF1 and PF2.

• Running the smart city application synced with sensor data streamed from the PF1 and PF3 platforms via PF2 in real-time.

• Actuating an LED-motor device registered with the Mobius platform triggered by an illumination sensor registered with the PF3.

• Actuating a drone registered with the PF1 triggered by a smart device registered to PF2.

In the interworking testing, the information from all these sensors was pushed to PF1 and stored in different containers according to the type of information, e.g. one container for temperature and light sensors and another for parking sensors. The application read the information from there through PF2 using redirection across the Mcc′ interface.

To be able to show dynamic changes, we added sensors at the testing demo site, which were virtually placed in Santander, i.e. they appeared in the Dashboard application on the Santander map (Fig. 5b), but the actual sensors were located at the testing demo site. Two sensors were connected to PF1 and another sensor was connected to PF3, all storing their information in individual containers on the respective platforms. All sensors were accessed through PF2 using the Mcc′ interface. Dynamic changes were shown by covering the attached light sensors.

Apart from visualizing the individual sensors with a color-coded value on the map, as shown in Fig. 5, the smart city application also showed averages on a section/district level as well as a comparison in a table and column chart, enabling city personnel to get a fast overview of the situation in the city. In addition, there are more elaborate evaluations for the traffic and parking situation put into correlation with CO2 emissions.

As a result of these interworking experiments, we can consider that the viability of the oneM2M standard as a mechanism to enable service layer interworking and interoperability to be validated.
LESSONS LEARNED

In this section we illustrate lessons learned through implementing standard compliant IoT platforms, testing interworking between different platforms as well as the smart city application.

LESSON LEARNED #1: LATENCY AND CONNECTION LOSS

Applications using the service provided by the subscription-notification CSF of a oneM2M node may be affected by the potential problem of latency or connection loss in case of receiving notification messages. In particular, for a real-time application, if a connection is lost unexpectedly between the application and node that the application subscribed, the notification message might be accumulated in the subscription-notification CSF of the node. In this case, when the application re-connects to the node and then receives the notification message, the message is out of date. This may result in unsafe actions on real-time applications (subscriber). As more real-world applications will be operated on top of standardized IoT infrastructure, this needs to be considered in future IoT standards.

LESSON LEARNED #2: PRE-RELEASED SPECIFICATIONS

The majority of challenges encountered during this interworking testing involved working with a pre-release version of the specification that was still under development. As is to be expected with any pre-release version of a specification, it can miss and/or specify incorrect functionality. In an attempt to avoid this, we did some up-front work to try to align each platform from different companies on which features and functionality within the specification would be used as part of the interworking test. For example, we drew detailed call flows describing which oneM2M features, messages, and resources would be targeted. We also agreed upon a version of the oneM2M XSDs to use as well as how oneM2M requests and responses would be bind to HTTP. All of this up-front work did help in saving valuable debug time when trying to inter-work each other’s platforms, devices, and applications with one another.

LESSON LEARNED #3: OPEN API

oneM2M interfaces are not easy for developers to use if they do not have a basic knowledge of the standards. Thus we need a common API library to accelerate development of new services with oneM2M standards. The oneM2M standard currently provides bindings to HTTP, CoAP, and MQTT communication protocols. However developing an application and implementing these bindings is still a challenge. Therefore, both oneMPOWER and Mobius provided a Java API to ease the implementation. The API provided by the onePOWER platform was more generic and provided more flexibility, whereas the Mobius API was targeted to specific situations, making it easier to use, but requiring changes and additions during the demo integration. To facilitate application implementation and thus support the uptake of oneM2M, API bindings for relevant programming languages should be defined as part of the standard.

LESSON LEARNED #4: DATA INTEROPERABILITY

Another important lesson learned was that when interworking nodes residing in two different service provider domains exchange messages between these nodes, the message data format and payload data size must be interoperable and conform to the format used within each service provider domain. Therefore, a negotiation process regarding data format or data size should be considered essential.

The oneM2M platform currently is agnostic to the information stored in the content instances within a container. This makes it easy to store any information using base64 encoding, e.g. the JSON representation of the OMA NGSI-10 [12] data format used by the smart city application. However, as a result the consuming application needs to know a-priori the specific information representation as it is not explicitly provided in the oneM2M platform. Also, the discovery functionality supported by oneM2M is limited to the matching of specific labels. Thus, a-priori agreement between producers and consumers of information is required, limiting the support of dynamic adaptation based on changing producers of relevant information over time. The second release of oneM2M is working to address this by adding semantic capabilities.

DISCUSSION AND CONCLUSIONS

In this article we presented an overview of a global IoT/M2M standardization effort currently being performed and standardized in oneM2M. In particular, an interworking feature between common service layer platforms and its advantages has been investigated. To justify a strong need for the interworking feature in oneM2M, we realize standardized interworking interfaces based on oneM2M global standards, and test it through use cases with multiple IoT service platforms. The interworking test experiments are designed with smart city applications/services run on multiple IoT service layer platforms interworking with each other.

With these activities such as the implementations of oneM2M compliant platforms and preparations of interoperability testing, we learned several important facts that need to be considered in the future for conducting interworking test-
ing. For example, in order to avoid unnecessary costs associated with interworking testing with pre-release versions of a specification, a pre-analysis phase is required. For future work, we plan to extend our interworking test framework to cover other interworking features supported by oneM2M such as interworking with an underlying network, in particular for 3GPP and security.

**ACKNOWLEDGEMENTS**

J. Kim, J. Yun, and S.-C. Choi were supported by an Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No.B0184-15-1003). Professor Song was supported by the Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the MSIP (No. 2012M3C4A7033348). This work has also received funding from the European Union’s Horizon FP7 research program within the project FI-CORE under grant agreement No 632893. We would like to thank Chonggang Wang (from InterDigital) for his valuable comments and suggestions.

**REFERENCES**


**BIographies**

Jairo Koi (M) (jhkim@keti.re.kr) is a managerial researcher and a team leader at the IoT Platform Research Center at the Korea Electronics Technology Institute. He is currently the Technical Leader of the NICTA and the Device Working Group chair of the IoT and Device Working Group of the Korea IoT Association. He received the B.S. and M.S degrees from the Hankuk University of Foreign Studies. Currently, he is a Ph.D. candidate in electronic and electrical engineering at Yonsei University.

JaeSeon Yun (M) (jeeseon@keti.re.kr) is a senior researcher at the IoT Platform Research Center at the Korea Electronics Technology Institute (KETI). Prior to his current position, he worked as a postdoctoral research scientist in the Ubiquitous Computing Research Group in the School of Interactive Computing at Georgia Institute of Technology, GA, USA. He earned his M.S. and Ph.D. in mechatronics from Gwangju Institute of Science and Technology (GIST).

Sinz-Culan Onic (M) (socsteanu@retek.com) is senior researcher at the IoT Platform Research Centre at the Korea Electronics Technology Institute (KETI). He received the B.S. and M.S. degrees in electrical and electronic engineering from Yonsei University, South Korea, in 2006 and 2008, respectively. Before he joined KETI, he worked on the research staff at the Samsung Advanced Institute of Technology in Gyeonggi, South Korea for five years.

Dale N. Sero (M) (dale.sero@interdigital.com) is a principal engineer focused on M2M/IoT research, development, and standardization at InterDigital. He actively participates in oneM2M standards definition. Prior to his current position, he worked for Intel and Lucent Technologies in various hardware and software research and development roles. He received his B.S. and M.S. degrees in electrical engineering and computer science from Pennsylvania State University.

Guang Lu (M) (guang@proverbal.com) is currently a business and technology advisor on unified communications for ProVerbal Telecommunications. Previously, she worked in management and senior engineering positions at InterDigital, focusing on Internet of Things (IoT) and other emerging technologies. Early in her career, she worked at Nortel and Unisys. She holds B.S. and M.S. degrees in computer science.

Martin Baier (M) (martin.baier@telefonica.de) is a senior researcher focused on IoT and context management related research and standardization at the NEC Laboratories Europe in Hedelberg, Germany. He participates in oneM2M standardization, focusing on aspects related to semantics. Prior to his current position he worked as a researcher at the University of Stuttgart, from where he also received his Ph.D. He holds M.S. degrees from both the University of Oregon and the University of Stuttgart.

Aitzai Al-Hetzi (M) (aizelizmi@gmail.com) is a senior researcher at the Next Generation Network Infrastructure Competence Center at Fraunhofer FOKUS. He started his working experience in 2003 at Fraunhofer FOKUS. He then joined Saudi Telecommunication Company from 2008 to 2010. Thereafter, he worked as a senior solutions architect at the Qatar Mobility Innovations Center in Doha until 2012. He received a diploma (master) in electrical engineering in 2003, and a Ph.D. (Dr.-Ing.) in engineering from the Technische Universität Berlin.

Konrad Campowski (M) (konrad.camposowski@fokus.fraunhofer.de) is a researcher at the M2M group at Fraunhofer FOKUS where he coordinates the technical development of the OpenMTC software. Prior to his current position, he worked at the NGNI Competence Center at Fraunhofer FOKUS. He received his Dipl. Ing. degree from the Technical University Berlin.

JaeSeong Sino (M) (jsijoong@sejong.ac.kr) is an assistant professor in the Computer and Information Security Department at Sejong University. He holds the position of oneM2M Test Working Group chair. Prior to his current position, he worked for NEC Europe Ltd. and LG Electronics in various positions. He received a Ph.D. from Imperial College London in the Department of Computing, United Kingdom. He holds B.S. and M.S. degrees in computer science from Sogang University.
Now...

2 Ways to Access the IEEE Member Digital Library

With two great options designed to meet the needs—and budget—of every member, the IEEE Member Digital Library provides full-text access to any IEEE journal article or conference paper in the IEEE Xplore® digital library.

Simply choose the subscription that’s right for you:

IEEE Member Digital Library
Designed for the power researcher who needs a more robust plan. Access all the IEEE content you need to explore ideas and develop better technology.

- 25 article downloads every month

IEEE Member Digital Library Basic
Created for members who want to stay up-to-date with current research. Access IEEE content and rollover unused downloads for 12 months.

- 3 new article downloads every month

Get the latest technology research.

Try the IEEE Member Digital Library—FREE!
www.ieee.org/go/trymdl

IEEE Member Digital Library is an exclusive subscription available only to active IEEE members.
The 2016 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.

We invite you to submit original technical papers in the following areas:

- Ad Hoc and Sensor Networking
- Communication & Information System Security
- Cognitive Radio and Networks
- Communication QoS, Reliability and Modeling
- Communication Theory
- Communications Software, Services and Multimedia Applications
- Green Communications
- Mobile and Wireless Networking
- Next-Generation Networking
- Optical Networks and Systems
- Signal Processing for Communications
- Wireless Communications

Selected Areas in Communications:
- Access Networks & Systems
- Big Data
- Cloud Networks
- Data Storage
- e-Health
- Internet of Things
- Power Line Communications
- Satellite & Space Communications
- Social Networks

Please address questions regarding the Technical Symposia to TPC Chair Sherman Shen (sshen@uwaterloo.ca) and/or TPC Vice Chairs Maite Brandt-Pearce (mb9q@eservices.virginia.edu) and Sennur Ulukus (ulukus@umd.edu).

Accepted and presented technical and workshop papers will be published in the IEEE GLOBECOM 2016 Conference Proceedings and submitted to IEEE Xplore®. See the website for author requirements of accepted authors. Full details of submission procedures are available at www.ieee-globecom.org.

**IMPORTANT DATES:**
- Symposia Papers - **15 April 2016**
- Workshops Proposals - **6 February 2016**
- Tutorial Proposals - **18 March 2016**

**COMMITTEE**

**General Chair**
Bijan Jabbari, George Mason University

**TPC Chair**
Sherman Shen, University of Waterloo

**TPC Vice Chairs**
Maite Brandt-Pearce, University of Virginia
Sennur Ulukus, University of Maryland

**Tutorial Co-Chairs**
Brian Mark, GMU
Yahong Rosa Zheng, University of Missouri-Columbia

**Workshop Co-Chairs**
Shuguang Cui, Taxes A&M
Stefano Secci, University of Paris 6

**Executive Chair**
Paul Cotae, University of District of Columbia

**General Vice Chair**
Jerry Gibbon, US Department of Commerce

**GIMS Advisor**
Chi-Ming Chen, AT&T

**GITC Advisor**
Peter Rost, TU Darmstadt

- www.ieee-globecom.org -