IEEE Signal Processing MAGAZINE

[VOLUME 31 NUMBER 6 NOVEMBER 2014]

THE 5G REVOLUTION

ADVANCES AND POTENTIAL CHALLENGES

TELEIMMERSIVE AUDIO-VISUAL COMMUNICATION

ULTRAWIDEBAND SIGNALS IN MEDICINE

REFLECTIONS ON EXCELLENCE IN RESEARCH AND EDUCATION

IEEE Signal Processing Society

IEEE

TINY Wideband Transformers & Baluns!

NOW! **4 kHz - 18 GHz** *From* **99¢** *ea.(qty.20)*

To support an even wider range of applications, Mini-Circuits tiny surface-mount transformers and baluns now cover frequencies up to 18 GHz! Our latest designs achieve consistent performance across very wide frequency bands, and our baluns have demonstrated great utility for chipsets. With over 250 trusted models in stock representing a wide selection of circuit topologies and impedance ratios, chances are, we have a solution for your needs!

Our Low Temperature Co-Fired Ceramic (LTCC) models provide reliable performance in tough operating conditions, tiny size – as small as 0805 – and very low cost. All core-and-wire models are available with our exclusive Top Hat™ feature, improving pick-and-place accuracy and throughput. We even manufacture our own transmission wire under rigorous control and use all-welded connections to ensure reliability and repeatability you can count on.

Visit minicircuits.com and use **Yoni2**™, our patented search engine to search our entire model database by performance criteria and find the models that meet your requirements. Order today and have them in hand as soon as tomorrow! Cost-effective custom designs and simulations with fast turnarounds are just a phone call away!

TC
0.15" x 0.15"

NC
0.08 x 0.05"
Ceramic

NCR2
0.08 x 0.10"
Ceramic

RoHS compliant.

Mini-Circuits®

www.minicircuits.com P.O. Box 350166, Brooklyn, NY 11235-0003 (718) 934-4500 sales@minicircuits.com

528 rev org

CONTENTS

[VOLUME 31 NUMBER 6]

SPECIAL SECTION—THE 5G REVOLUTION

COLUMNS

DEPARTMENT

Digital Object Identifier 10.1109/MSP.2014.2350051

IEEE SIGNAL PROCESSING magazine

[COVER]

5G IMAGE: ©ISTOCKPHOTO.COM/TIMARBAEV.
TECHGRAPHICS: ©ISTOCKPHOTO.COM/SIGAL SUHLER MORAN.

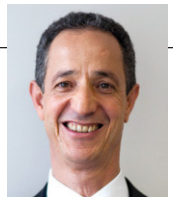
IEEE SIGNAL PROCESSING MAGAZINE (ISSN 1053-5888) (ISPREG) is published bimonthly by the Institute of Electrical and Electronics Engineers, Inc., 3 Park Avenue, 17th Floor, New York, NY 10016-5997 USA (+1 212 419 7900). Responsibility for the contents rests upon the authors and not the IEEE, the Society, or its members. Annual member subscriptions included in Society fee. Nonmember subscriptions available upon request. Individual copies: IEEE Members US$20.00 (first copy only), nonmembers US$201.00 per copy. Periodicals postage paid at New York, NY, and at additional mailing offices. Postmaster: Send address changes to IEEE Signal Processing Magazine, IEEE, 445 Hoes Lane, Piscataway, NJ 08854 USA. Canadian GST #125634188

Printed in the U.S.A.

Digital Object Identifier 10.1109/MSP.2014.2352093

SCOPE: *IEEE Signal Processing Magazine* publishes tutorial-style articles on signal processing research and applications, as well as columns and forums on issues of interest. Its coverage ranges from fundamental principles to practical implementation, reflecting the multidimensional facets of interests and concerns of the community. Its mission is to bring up-to-date, emerging and active technical developments, issues, and events to the research, educational, and professional communities. It is also the main Society communication platform addressing important issues concerning all members.

from the **EDITORS**



Abdelhak Zoubir
Editor-in-Chief
zoubir@spg.tu-darmstadt.de
http://signalprocessingsociety.org/publications/periodicals/spm

A New Era of *IEEE Signal Processing Magazine*

This is my last editorial for *IEEE Signal Processing Magazine* (*SPM*) as my three-year term as editor-in-chief (EiC) expires in December 2014. Usually, in such situations, one tends to revisit his or her achievements with the readers. I will not do that, as I always believe that the readers are in a much better position to judge the quality of the special issues, feature articles, and columns and forum we brought you over the last three years and to draw conclusions based on the work the team has performed. However, I will mention a subject that has been much discussed among authors, readers, and publication boards. It is the impact factor (IF), the average number of citations received per paper published in a journal or magazine during the two preceding years. I am not a great believer of this metric as a sole criterion for judging a journal or magazine, let alone its quality. To measure quality, one has to consider multiple facets that make it up and not reduce these to a single number.

In fact, the IEEE issued a paper [1] (adopted by the IEEE Board of Directors, 9 September 2013) where it recommends "… the use of multiple complementary bibliometric indicators to offer an appropriate, comprehensive, and balanced view of each journal in the space of scholarly publications."

I started my term in January 2012 and, at that time, the magazine enjoyed the highest IF rank of 5.86 among all 247 publications listed in the electrical and electronics engineering category in *Journal Citation Reports*, Science Edition 2010. However, as I just mentioned, I am not a believer in the importance of just a single number, but I knew all along that many colleagues and scientific communities take notice of it. Thus, I knew that it would be a challenge to maintain the high IF rank that *SPM* had accomplished. This past July, the 2013 IF increased to 4.481 after a drop in 2011 and 2012, placing *SPM* ninth among all EE publications. This is a pleasing outcome, but again, please bear in mind that this measure does not reflect quality and cannot be considered as a stand-alone metric. Even more pleasing is the fact that the immediacy index, the eigenfactor score, and the article influence score of *SPM* have also steadily increased over the last three years.

TO MEASURE QUALITY, ONE HAS TO CONSIDER MULTIPLE FACETS THAT MAKE IT UP AND NOT REDUCE THESE TO A SINGLE NUMBER.

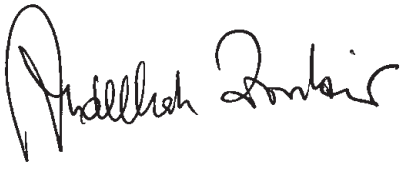
What was the biggest challenge during my tenure? Undoubtedly, it was to make *SPM* as attractive as possible and a most valuable resource for the whole community. It is for that reason that my first priority was to appoint outstanding area editors for all positions, i.e., Fulvio Gini (Special Issues); Marc Moonen (Feature Articles); Andres Kwasinski and Andrea Cavallaro (Columns and Forum); and Z. Jane Wang and Christian Debes (e-Newsletter). They are all dedicated professionals who put their hearts into this venture. It was also of great importance to me to appoint responsible professionals to the Senior Editorial Board. They supported me and especially my colleagues Fulvio and Marc immensely. It was my great pleasure to work with the Senior Editorial Board. I extend my thanks to them for their valuable support. I also wish to thank the IEEE Signal Processing Society staff for their continued support, particularly Rebecca Wollman, Theresa Argiropoulos, Deborah Blazek, Rupal Bhatt, and Denise Hurley, as well as *SPM's* managing editor, Jessica Barragué, and many others working in the background.

I now welcome my successor Min Wu as the incoming EiC, who knows *SPM* very well and will do all that it requires to maintain the magazine's high quality and reputation, and also e-Newsletter Area Editor Christian Debes, who has energy and new ideas to drive eNews to a higher level.

Most importantly, I wholeheartedly thank all readers of *SPM* and all contributors. Without you, there would be no *SPM*!

REFERENCE

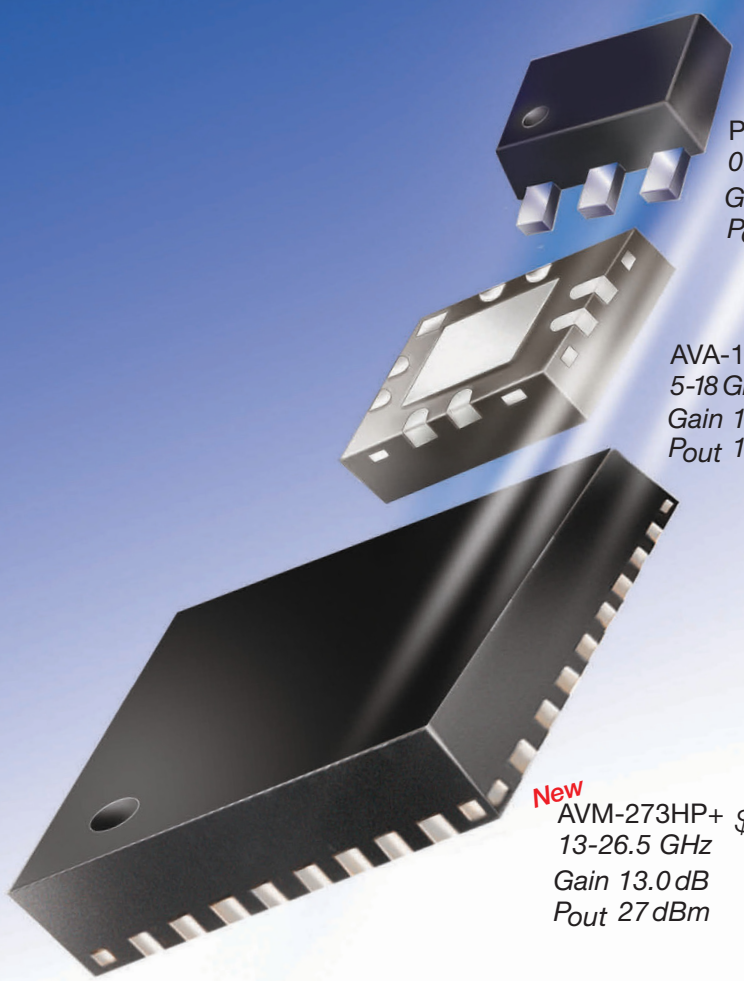
[1] Online. Available: http://www.ieee.org/publications_standards/publications/rights/bibliometrics_statement.html



Digital Object Identifier 10.1109/MSP.2014.2352051
Date of publication: 15 October 2014

50 MHz to 26.5 GHz

THREE AMPLIFIERS COVER IT ALL!



PHA-1+
0.05-6 GHz
$1.99 ea. (qty. 20)
Gain 13.5 dB
P_{out} 22 dBm

AVA-183A+
5-18 GHz
$6.95 ea. (qty. 10)
Gain 14.0 dB
P_{out} 19 dBm

New
AVM-273HP+
13-26.5 GHz
$27.95 ea. (qty. 10)
Gain 13.0 dB
P_{out} 27 dBm

Mini-Circuits' New AVM-273HP+ wideband, 13 dB gain, unconditionally stable microwave amplifier supports applications from 13 to 26.5 GHz with 0.5W power handling! Gain flatness of ±1.0 dB and 58 dB isolation make this tiny unit an outstanding buffer amplifier in P2P radios, military EW and radar, DBS, VSAT, and more! Its integrated application circuit provides reverse voltage protection, voltage sequencing, and current stabilization, all in one tiny package!

The AVA-183A+ delivers excellent gain flatness (±1.0 dB) from 5 to 18 GHz with 38 dB isolation, and 19 dBm power handling. It is unconditionally stable and an ideal LO driver amplifier. Internal DC blocks, bias tee, and microwave coupling capacitor simplify external circuits, minimizing your design time.

The PHA-1+ + uses E-PHEMT technology to offer ultra-high dynamic range, low noise, and excellent IP3 performance, making it ideal for LTE and TD-SCDMA. Good input and output return loss across almost 7 octaves extend its use to CATV, wireless LANs, and base station infrastructure.

We've got you covered! Visit minicircuits.com for full specs, performance curves, and free data! These models are in stock and ready to ship today!

RoHS compliant

FREE X-Parameters-Based
Non-Linear Simulation Models for ADS
Modelithics®
http://www.modelithics.com/mvp/Mini-Circuits.asp

Mini-Circuits®

www.minicircuits.com P.O. Box 350166, Brooklyn, NY 11235-0003 (718) 934-4500 sales@minicircuits.com

478 rev N

Christian Debes
Area Editor, e-Newsletter
CDebes@agtinternational.com
http://signalprocessingsociety.org/publications/periodicals/spm

Inside Signal Processing e-Newsletter

Launched in 2007, *Inside Signal Processing e-Newsletter* has been a source of relevant information for the signal processing community for more than seven years. With the newsletter Web site (http://signalprocessingsociety.org/newsletter) and the monthly e-mail digest, IEEE Signal Processing Society (SPS) members receive the latest news about our Society, updates from our technical committees, signal processing-related trends, and much more.

Regarding the e-Newsletter, we strive to bring relevant and timely information about the Society and beyond to all members. Thus, it should never be seen as a static service but as a continuously evolving effort, always asking the question "What is it that SPS members care about? And how can we best present it to them?"

Z. Jane Wang, the previous area editor for the e-Newsletter for nearly five years, did a great job in redesigning the newsletter and making it more attractive to the readers. It is an honor and pleasure to build upon the work of Jane and her team and to bring the newsletter to the next level. To restructure the e-Newsletter and fill it with the relevant content for SPS members, a new team of associate editors has been appointed. I'm very happy that the following seven individuals agreed to join me in this effort: Csaba Benedek (Hungarian Academy of Sciences, Hungary), Paolo Braca (NATO Science and Technology Organization, Italy), Quan Ding (University of California, San Francisco, United States), Marco Guerriero (General Electric Research, United States), Yang Li (Harbin Institute of Technology, China), Yuhong Liu (Penn State Altoona, United States), and Andreas Merentitis (AGT International, Germany).

AS A MEMBER-CENTRIC SERVICE, THE E-NEWSLETTER IS ALL ABOUT WHAT IS RELEVANT TO YOU.

What can you expect from future e-Newsletters? Our main focus is to bring more relevant content to the readers. This includes a broader coverage of signal processing-related activities from universities, industry, and other organizations around the world. The SPS e-Newsletter will have more reports on industry trends, signal processing applications in emerging fields, new data sets for experimentation, and ongoing competitions. You will find short interviews that put SPS members in the spotlight. This is not all—there are many more ideas in the pipeline that you will see transforming into the monthly e-mail digest and the newsletter Web site in the coming months.

As a member-centric service, the e-Newsletter is all about what is relevant to *you*. Therefore, I sincerely invite you to e-mail me your thoughts and contributions.

[SP]

Digital Object Identifier 10.1109/MSP.2014.2354091
Date of publication: 15 October 2014

ERRATA

The URL that was included in [6] of the "References" section in the September 2014 "From the Editor" column of *IEEE Signal Processing Magazine* has since expired. We apologize for the inconvenience. Interested readers can view the profiles of unsung engineering heroes that will be highlighted in "Special Report: Dream Jobs 2015" in the March 2015 issue of *IEEE Spectrum*.

Digital Object Identifier 10.1109/MSP.2014.2358731

president's MESSAGE

Alex Acero
2014–2015 SPS President
a.acero@ieee.org

A Chapter's Role in Networking and Continuing Education

The IEEE is a nonprofit organization whose goal is to serve its Members. The IEEE Signal Processing Society (SPS) is one of the IEEE's 38 technical Societies. We are a membership organization, and our primary goal is to serve you, so it is very important for us to understand our members and what they expect from us. We cannot use a one-size-fits-all approach as our membership is very diverse, both in terms on the job function (academia, graduate students, undergraduate students, industry, government) and geography.

While IEEE membership has been steadily growing (from 380,000 in 2000 to 430,000 in 2013), the number of total Society memberships has been decreasing (from 400,000 in 2000 to 350,000 in 2013). Many IEEE Members joined Societies to gain access to their journals or magazines. But when IEEE *Xplore* was introduced in 2001, many IEEE Members could access Societies' publications through their university's or company's institutional membership. This started the decline in Societies' memberships as the value proposition had changed.

After a steady drop in membership (from 20,000 in 2000 to 15,000 in 2009), the SPS started gaining members, ending 2013 with 17,433, as the fourth-largest Society in the IEEE. Most of that growth came from Region 10 (Asia), whereas Society membership in the United States (Regions 1–6) had a slow but steady decline, accounting for 42% of all members in 2013 down from well over 50% a decade ago.

About 30% of those members are faculty in universities and 12% are graduate students. Both faculty and graduate students comprise a large percentage of the attendees to the Society's conferences and also a large percentage of the contributors to papers and articles in the Society's journals and magazine, respectively.

About 50% of the Society's members are in industry and 3% are undergraduate students. Many of these industry members have told us that they cannot easily attend international conferences, but that they would like to attend local events. Attending local Chapter events is something that ranks high in member surveys not only for industry members but also for students, and academics.

The Society has 134 Chapters worldwide. Society Chapters organize many activities for the members in their area, such as hosting talks and networking events. This is a great opportunity for our members to keep up with important new developments in signal processing. Such Chapter events usually offer refreshments either before or after the talk, providing members with an opportunity for networking. Many such talks are held in the evening to make it easier for working professionals to attend.

This past August, I attended a talk by one of the Society's Distinguished Lecturers at an event organized by the Santa Clara Chapter. The talk was inspiring and the networking that took place was energizing. I was thrilled to see that 80 people attended this vibrant meeting. I'd like to thank the leadership of the Santa Clara Chapter for reinvigorating the Chapter, organizing many more events than in the past, all of which were very successful.

The Society's Membership Board understands how important Chapters are in serving many of our members and how they help create a sense of community and pride in our profession. We have started to recognize Chapters that have done a great job through the Chapter of the Year Award: the Malaysia Chapter won the first award in 2011, the Tainan Chapter in 2012, and the Italy Chapter in 2013. We have also started a Chapter certification process to help all our Chapters leverage best practices.

Chapters can also request funding from the SPS Membership Board (sp.bd.membership@ieee.org) for activities such as SPS seasonal schools, meetings featuring industry executives, outreach programs to get students in primary schools interested in signal processing, and more.

Please help us increase the vitality of our Chapters by volunteering to serve on the board of your Chapter. You can contact the SPS Vice President of Membership Kostas Plataniotis, Chapters Committee Chair Sven Longaric, or our Regional Directors: Anthony Kuh (Regions 1–6), Douglas O'Shaughnessy (Regions 7 and 9), Mauro Barni (Region 8), and Mark Liao (Region 10). I'd also love to hear from you—how can we do better?

[SP]

Digital Object Identifier 10.1109/MSP.2014.2352052
Date of publication: 15 October 2014

[special REPORTS]

By John Edwards

Looking at Machine Vision

Before we can enter a world in which cars and trucks drive themselves, autonomous aircraft dot the skies, and robots pitch in to perform a virtually endless array of tasks, these systems will need to have a way of reliably and safely interacting with the surrounding world. Machine vision is the technology that will give future autonomous systems the ability to detect and react to various types of objects, terrains, and situations.

Signal processing lies at the heart of machine vision, opening ways of acquiring, processing, analyzing, and understanding images and other high-dimensional data from the real world. In multiple research areas, today's machine vision developers are pioneering systems that in years ahead promise to make life more faster, safer, healthier, and more convenient in an almost endless number of areas.

CUTTING THROUGH THE CLUTTER

Object recognition is one of the most pressing challenges facing computer vision researchers, since a robot or other type of machine manipulating something in the real world needs to do more than simply recognize an item—it also must be able to perceive the object's precise orientation.

To enhance the ability of robots to determine the orientation of specific objects, researcher Jared Glover (Figure 1) turned to a lesser-known and seminegleted statistical construct known as the *Bingham distribution*. While a graduate student in the Massachusetts Institute of Technology's (MIT's) Department of Electrical Engineering and Computer Science, Glover and coresearcher Sanja Popovic developed a new robot vision algorithm, based on the Bingham distribution, that he says turned out to be 15% more accurate at identifying familiar objects in cluttered scenes than the best previous models. (Glover graduated MIT in May 2014. Popovic, also an MIT graduate, currently works at Google.)

Glover focused his research on a single basic question: How can a robot detect objects within a cluttered environment? "I started working on specific object detection, meaning my system was looking for objects that the robot already has a model of in its database," Glover says. "The robot knows the 3-D (three-dimensional) shape of the object it's looking for, it's just trying to find that shape in the clutter."

SIGNAL PROCESSING LIES AT THE HEART OF MACHINE VISION, OPENING WAYS OF ACQUIRING, PROCESSING, ANALYZING, AND UNDERSTANDING IMAGES AND OTHER HIGH-DIMENSIONAL DATA FROM THE REAL WORLD.

In noisy and jumbled landscapes, accurate orientation detection hinges on precise alignments using multiple cues, such as 3-D point positions, surface normals, curvature directions, edges, and image features. Glover observed that other than brute force optimization, no existing alignment method existed that could merge all of this information together in a meaningful way.

The researcher identified the Bingham distribution as a useful tool because it enables an algorithm to squeeze more information out of each ambiguous, local feature. By connecting the Bingham distribution to the classical least-squares alignment problem, the researchers were easily able to fuse information from both position and orientation information into a principled, Bayesian alignment system that they called the *Bingham procrustean alignment*.

[FIG1] **As an MIT graduate student, Jared Glover developed a new robot vision algorithm based on the Bingham distribution. (Photo courtesy of Jared Glover.)**

In his research, Glover used a Microsoft Kinect camera to identify locations in an image where color or depth values change abruptly—likely object edge locations. The work was then narrowed down to taking two sets of points—the model and the object—and determining whether one could be superimposed on the other.

Most algorithms, including Glover's, will make an initial immediate attempt at aligning the points. If both sets of points really do describe the same object, they can be quickly aligned by rotating one of them around the right axis. For any given pair of points—from the model and the

Digital Object Identifier 10.1109/MSP.2014.2343983
Date of publication: 15 October 2014

object—one can effectively determine the probability that rotating one point by a particular angle around a particular axis will align it with the other. The challenge is that the same rotation might also move other pairs of points farther away from each other. Glover, in his research, showed that the rotation probabilities for any given pair of points can be described as a Bingham distribution, which can then be merged into a single, comprehensive Bingham distribution.

Getting noise under control proved to be one of the researcher's major challenges "If you have noise, say, noisy estimates on the object's depth, and, if that noise is different from the first time you saw it to the second time you saw the object—because you see it from a different view or under different lighting—then the system might not have an accurate model for the noise, and so it will get confused," he says.

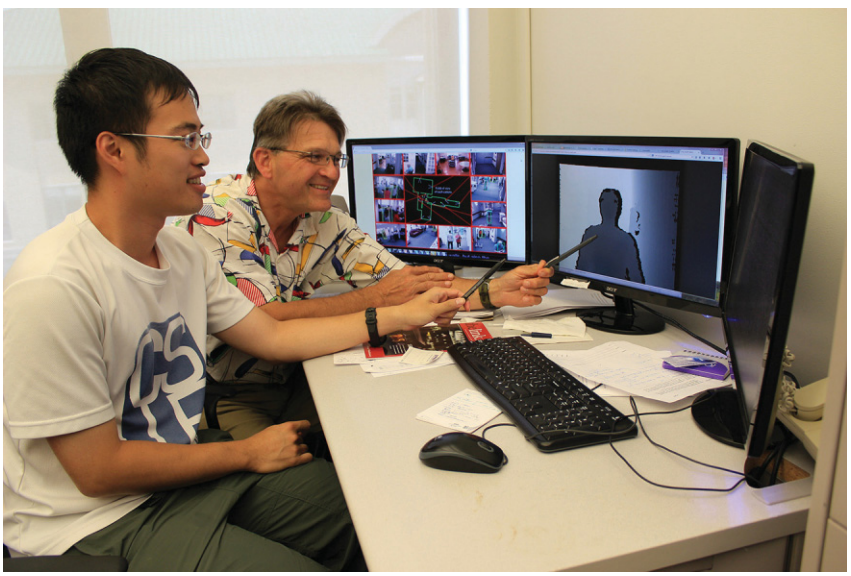
Nonetheless, in experiments using visual data about particularly cluttered scenes, the algorithm identified 73% of the objects in a given scene, compared to 64% from the best existing algorithm. With further research and sources of information, Glover believes the algorithm's performance can be improved even more.

"Besides increasing accuracy and robustness, the biggest challenge is relationship understanding," he explains. "If a robot can understand, for example, that the bowls are on top of each other, or things are touching each other in a certain way, or wrapped around each other, that's the kind of information that is going to be necessary for it to manipulate objects in the real world."

HAUPTMANN'S RESEARCH TEAM DEVELOPED MATHEMATICAL MODELS THAT LET THEM COMBINE CRITICAL INFORMATION, SUCH AS APPEARANCE, FACIAL RECOGNITION, AND MOTION TRAJECTORIES.

IDENTIFYING FACES IN THE CROWD

Multicamera, multiobject tracking has been an area of intense research for over a decade. Yet few automated techniques have been tested on objects located outside of well-controlled lab environments. To make tracking technology more useful in a potentially wide range of commercial and civic applications, researchers at Carnegie Mellon University recently developed an algorithm designed to track the locations of multiple individuals in complex, indoor settings via a network of video cameras.



[FIG2] **Alexander Hauptmann, principal systems scientist in the Carnegie Mellon Computer Science Department (right), and a student view motion-tracking system images. (Photo courtesy of Carnegie Mellon University.)**

Alexander Hauptmann (Figure 2), principal systems scientist in the Carnegie Mellon Computer Science Department, notes that developing an effective motion tracking system required overcoming a number of challenges. Something as apparently simple as tracking a person based on the color of the clothing worn proved to be frustratingly difficult because the apparel color can appear different to cameras in assorted locations due to lighting variations. Likewise, a camera's view of an individual can be blocked by people passing in hallways, by furniture or other stationary objects, or when someone enters a room or other area not covered by cameras. All of these situations, and others, make it necessary for individuals to be regularly reidentified by the system.

Hauptmann's research team developed mathematical models that let them combine critical information, such as appearance, facial recognition, and motion trajectories. Using all of this information is key to successful tracking, Hauptmann says, but facial recognition provided the greatest help. "The core tracking was a particle filter tracker, based on appearance," Hauptman remarks. When the researchers removed facial recognition data from the tracking, accuracy collapsed from 88% to 58%, not significantly better than existing tracking algorithms.

"The idea of particle filter tracking is that you don't commit to any one thing, so what you're tracking could be anywhere in the space, yet it's more likely to be here and less likely to be there," Hauptmann says. "So you always have these distributions of possible places for each particle that you're tracking and in the end you find that, overall, this is the most likely place for a particular person," he adds.

The algorithm's input consists of a set of person detection results at each time instant. "The person detection results from different camera views mapped to a common 3-D coordinate system using camera calibration and ground plane parameters provided," Hauptman says. Each person detection result is described by a color histogram. "Our algorithm's main task is to predict a label for each result," Hauptmann

explains. "To perform the prediction task, our algorithm incorporates two main innovative components, which are manifold learning in appearance space, with spatiotemporal constraints, and trajectory inference by nonnegative discretization.

The algorithm was able to automatically follow 13 individuals within a nursing home, with the residents' consent, despite the fact that people occasionally moved out of camera range. The researchers used 6 min of footage recorded by 15 cameras in a nursing home in 2005 to develop the algorithms and test the system. The team took advantage of multiple cues within the video, including trajectory, clothing color, person detection, and, most critically, facial recognition. "We thought it would be easy," Hauptmann said of multicamera tracking, "but it turned out to be incredibly challenging."

After working on the project for nearly a decade, Hauptmann notes that a series of relatively small technology and technique advancements can have as big an impact in an area like object tracking as a major breakthrough. "There's a big disconnect in computer vision between things that are published and things that work," he says. "What tends to get published are really novel ways of thinking about it—novel theories and novel algorithms."

But real life isn't quite that simple. Camera angle, for instance, can make a big difference in results. "Most research is done in a lab with a good camera position, so if a person eating is directly facing the camera, you've got high enough resolution to see their mouth moving, track points around the corners of their mouth and so on," Hauptmann says. "This approach is really impressive in that sort of laboratory situation, but when you take it into the real world it's a different story, and that's why this project took us so long."

After years of hard work, Hauptmann regards the project as a success. "Our algorithm exhibited the robust localization and tracking of persons-of-interest not only in outdoor scenes, but also in a complex indoor real-world nursing home environment," he says.

Hauptmann acknowledges, however, that the algorithm still has some limitations. "Our objective function does not have a spatial locality constraint on a trajectory," he says. "Therefore, our algorithm is not effective in very crowded sequences where each person wears the same color clothes." Another challenge is that optimization converging to a severe local optima, making initialization crucial. "Bad initialization may cause the performance to degrade," Hauptmann says.

While real-world deployment of the technology is still years away, Hauptman sees identification applications in venues beyond nursing homes, ranging from casinos to prisons. "We're still improving the accuracy," he says. "We're trying to get it so that we can easily apply it to other places."

SPEEDY CELL SORTING

Machine vision can also be used to recognize and differentiate objects as small as a biological cell. Researchers at the University of California, San Diego (UCSD), say that with the assistance of computer vision and hardware optimization they are now able to analyze and sort cells up to 38 times faster than with previous methods.

The approach, based on research originated at the University of California, Los Angeles (UCLA), improves imaging flow cytometry, a technique that uses a microscope-mounted camera to capture the morphological features of up to thousands of cells per second. The technology sorts cells into different categories, such as benign or malignant cells, based on their shape and structure. "The idea is, can we, at 50,000 frames per second, accurately identify each cell?" says Ryan Kastner, a UCSD professor of computer science (Figure 3).

Algorithms currently used take anywhere from 10 s to 0.4 s to analyze a single frame, making imaging flow cytometry far too slow for routine clinical use. The researchers' new approach promises to speed processing rates up to between 11.94 ms and 151.7 ms, depending on the hardware used. For enhanced performance, the team created a custom field-gate programmable array (FPGA). Low-range performance results, still significantly faster than currently achievable rates, were obtained by using an off-the-shelf graphics processing unit (GPU).

[FIG3] Ryan Kastner, a UCSD professor of computer science, leads a project aimed at speeding cell analysis and sorting. (Photo courtesy of UCSD.)

Four stages are necessary to perform the morphological analysis necessary for high-speed cell sorting: Blob Search, Image Interpolation and Adjustment, Find Center, and Coordinate Conversion/Radius Extraction. "Each module had to be carefully designed to achieve our performance targets," Kastner says. Yet reaching for maximum speed also required making some tradeoffs. "For example, at the end of the process, histogram equalization works better than image adjustment for contrast enhancement," Kastner explains. "Histogram equalization requires more complex processing leading to a lower throughput. Therefore, we sacrificed quality for performance."

The Blob Search module analyzes the images to detect the cell's area. The module then transforms the monochrome cell image into a binary digital image (only the pixels representing the cell are highlighted). The module then creates a histogram and crops a 20×20 pixel image around the cell.

To improve the fidelity of the analysis, the selected cell area from the Blob Search module is resized by a factor of ten. The Interpolation step also generates a higher contrast image by linearly adjusting the brightness level. The resized 200×200 image is input to the Find

(continued on page 117)

from the **GUEST EDITORS**

Robert W. Heath, Jr., Geert Leus, Tony Q.S. Quek, Shilpa Talwar, and Peiying Zhou

Signal Processing for the 5G Revolution

Cellular communication systems are continuing to incorporate advanced signal processing techniques. Third-generation cellular systems are already widely deployed and are being followed by fourth-generation (4G) systems. Since 4G cellular technology development is considered to have concluded in 2011, the attention of the research community is shifting toward what will be the next set of innovations in wireless communication technologies that are now broadly known as fifth-generation (5G) technologies.

Given a historical ten-year cycle for every generation of cellular advancement, it is expected that networks with 5G technologies will be deployed around 2020. While 4G standards were designed to meet requirements issued by the International Telecommunication Union-Radio, no definition for 5G is currently available. Experts vary in opinion whether the next generation of cellular networks will continue to enhance (peak) service rates further, focus on spectral efficiency enhancements, or move to newer metrics such as energy efficiency, cost- and utilization-efficiency, or even define new metrics around service quality experience. There is also the possibility that 5G will enable digital sensing, communication, and processing capabilities to be ubiquitously embedded into everyday objects, turning them into the Internet of Things (IoT) or machine-to-machine (M2M). In this new paradigm, smart devices will collect data, relay the information or context to each other, and process the information collaboratively over the 5G cellular networks. No matter what the eventual metric or system, it is certain that signal processing will play an important role in the features that define 5G.

Digital Object Identifier 10.1109/MSP.2014.2345430
Date of publication: 15 October 2014

This issue of *IEEE Signal Processing Magazine* (*SPM*) provides an overview of recent advances in signal processing for communication with an emphasis on signal processing techniques that will be relevant for 5G cellular systems. It covers a wide range of topics including modulation, beamforming, cross-layer optimization based on different performance metrics, location-aware communication, cloud computing, and cloud radio access networks. The articles provide a diverse perspective on the potential challenges in 5G cellular systems.

NO MATTER WHAT THE EVENTUAL METRIC OR SYSTEM, IT IS CERTAIN THAT SIGNAL PROCESSING WILL PLAY AN IMPORTANT ROLE IN THE FEATURES THAT DEFINE 5G.

The first set of articles addresses challenges related to the optimization of 5G systems.

The article by Björnson et al. considers the problem of operating a 5G system with conflicting performance metrics including higher peak rates, improved coverage with uniform user experience, higher reliability, lower latency, and better energy efficiency. The authors review a mathematical framework known as multiobjective optimization that can be used to solve problems with multiple competing objectives. The article concludes with an example application to massive multiple-input, multiple-output (MIMO) systems.

Cavalcante et al. consider optimizations related to energy efficiency, arguing that reducing the transmit energy per bit may increase the total energy consumption in the network. Traffic patterns and interference calculus are used to suggest algorithms for energy efficiency. The article concludes with a detailed example where an interference calculus approach is used to adaptively select a subset of active base stations based on prior traffic history, using the majorization–minimization algorithm.

The second set of articles addresses challenges related to cloud computing and cloud radio access networks.

The article by Wübben et al. reviews the benefits that cloud computing can offer in 5G cellular networks. Signal processing issues related to the cloud implementation of three representative parts of the signal processing chain are described in detail: hybrid automatic repeat request, forward error correction, and multiuser detection.

In the article by Barbarossa et al., the authors approach cloud computing from the perspective of offloading computations. They provide a mathematical formulation of a computation offloading problem aimed at jointly optimizing the communication and computation resources subject to latency and energy constraints. They consider computation offloading strategies and different ways to jointly optimize communication and computation resources.

Baligh et al. look at cloud computing from the perspective of interference management and network provisioning. They propose a cross-layer optimization framework for joint user admission, user base station association, power control, user grouping, transceiver design as well as routing and flow control, suggesting that they should be treated in a unified way for 5G networks.

Park et al. consider the topic of fronthaul compression for cloud radio access

networks. The fronthaul is the connection between remote radio heads and the centralized control unit with cloud computing technology. This article surveys work on fronthaul compression, especially multiterminal compression and structured coding, leveraging insights derived from network information theory.

The third set of articles focuses on more traditional physical-layer signal processing techniques relevant to 5G cellular systems.

The article by Banelli et al. considers the topic of modulation, speculating that 5G may employ another modulation strategy besides orthogonal frequency-division multiplexing. It reviews other competing modulation strategies including filter bank multicarrier, faster-than-Nyquist/time-frequency packing, and single-carrier modulations. The article concludes with a review of the potential interactions between the choice of modulation and other 5G requirements: high data rates, small cells, IoT, low latency, and energy efficiency.

Razavizadeh et al. examine a multiple antenna technique known as three-dimensional beamforming, where antenna elements in both horizontal and vertical directions are used. It reviews the concepts of two- and three-dimensional beamforming and discusses various challenges including channel modeling and array design.

The final article by Di Taranto et al. considers how 5G networks might benefit from precise location information at different layers of the protocol stack. Different applications of location are discussed including radio channel prediction and the ways that it can be used at the physical, medium access control, and higher layers.

In summary, we received many contributions in response to the call for papers of this special issue. Based on relevance and fit, many high-quality papers were not invited for full-paper submission. We would like to express our appreciation to all the authors who submitted white papers and full articles to this special issue. We would also like to thank all the reviewers who provided critical reviews of the diverse set of papers that we received. Finally, we would like to acknowledge Abdelhak Zoubir, *SPM's* editor-in-chief, who was very supportive of our issue, and Fulvio Gini, the special issues area editor who provided assistance and encouragement along with countless reminders, and, of course, Rebecca Wollman for her assistance with the entire process. We hope that you will enjoy the articles in this special issue of *SPM*.

[SP]

Emil Björnson, Eduard Jorswieck, Mérouane Debbah, and Björn Ottersten

Multiobjective Signal Processing Optimization

The way to balance conflicting metrics in 5G systems

The evolution of cellular networks is driven by the dream of ubiquitous wireless connectivity: any data service is instantly accessible everywhere. With each generation of cellular networks, we have moved closer to this wireless dream; first by delivering wireless access to voice communications, then by providing wireless data services, and recently by delivering a Wi-Fi-like experience with wide-area coverage and user mobility management. The support for high data rates has been the main objective in recent years [1], as seen from the academic focus on sum-rate optimization and the efforts from standardization bodies to meet the peak rate requirements specified in IMT-Advanced. In contrast, a variety of metrics/objectives are put forward in the technological preparations for fifth-generation (5G) networks: higher peak rates, improved coverage with uniform user experience, higher reliability and lower latency, better energy efficiency (EE), lower-cost user devices and services, better scalability with number of devices, etc. These multiple objectives are coupled, often in a conflicting manner such that improvements in one objective lead to degradation in the other objectives. Hence, the design of future networks calls for new optimization tools that properly handle the existence of multiple objectives and tradeoffs between them.

THE 5G REVOLUTION

©ISTOCKPHOTO.COM/ZONADEARTE

In this article, we provide a review of multiobjective optimization (MOO), which is a mathematical framework to solve design problems with multiple conflicting objectives [2]–[6]. In contrast to conventional heuristic approaches where some objectives are converted into constraints, MOO enables a rigorous network design. MOO has been applied in many engineering and economic related fields but has received little attention from the signal processing and wireless communication communities. We provide a survey of the basic definitions, properties, and algorithmic tools in MOO. This reveals how signal processing algorithms are used to visualize the inherent conflicts between 5G performance objectives, thereby allowing the network designer to understand the possible operating points and how to balance the objectives in an efficient and satisfactory way. For clarity, we provide a case study on massive multiple-input, multiple-output (MIMO) systems, which is one of the key enablers of 5G cellular networks.

INTRODUCTION

We are currently at a point in time when many researchers in industry and academia are trying to formalize their expectations and requirements on next-generation wireless communication networks. These views are expressed in various magazine articles, white papers, and plenary talks. To get a sense of the range of expectations, one can take a look at the project Mobile and Wireless Communications Enablers for the 2020 Information Society (METIS), where telecommunications manufacturers, network

Digital Object Identifier 10.1109/MSP.2014.2330661
Date of publication: 15 October 2014

1053-5888/14©2014IEEE

operators, and academic partners are gathering their 5G requirements (for more information, see http://www.metis2020.com). The following summarizes their main objectives [7]:

- *Higher user data rates*: 10–100 times higher average user rates are expected, at least in urban scenarios.
- *Higher area data rates*: 1,000 times higher average rates per unit area are anticipated.
- *More connected devices*: With the respective expected increases in user and area rates, 10–100 times more devices can be accommodated per unit area.
- *Higher EE*: The throughput should be improved without increasing the operational cost or the energy consumption, thus greatly improving the EE. If EE is measured as area data rate per power expenditure, this requires a 1,000 times EE improvement.

Furthermore, *heterogeneity* appears as a keyword that can be tied to a variety of network aspects:

- *Heterogeneous networks*: The combination of access points with different ranges, traffic loads, radio access technologies, licensed/unlicensed spectrum, and hardware capabilities makes the network highly heterogeneous. The same deployment strategy cannot be used everywhere and the same resource management scheme cannot be used throughout the day.
- *Heterogeneous user conditions*: As the performance requirements become tighter, the mobility and path loss of a specific user determines its quality of service, unless the network is designed to counteract these effects.
- *Heterogeneous devices*: The differences in functionality and hardware capability of user devices are expected to grow. Large handheld devices can, for example, achieve high data rates by spatial multiplexing and advanced signal processing, while small sensors seek low data rates under extremely tight energy constraints.
- *Heterogeneous service requirements*: Some cyberphysical systems and public-safety applications require very fast and reliable response times, while best-effort delivery is fine for other types of data services. Similarly, certain multimedia applications have tight and continuous quality-of-service requirements, while other services are bursty in nature.

There are apparently many different requirements, or objectives, to keep in mind when designing future wireless networks. Unfortunately, these objectives cannot be treated separately because they are coupled; sometimes in a consistent fashion, but often in conflicting ways such that improvements in one objective lead to deterioration of other objectives. This is because the same network resources (for example, time, frequency, space, power, and hardware) play key roles in all these requirements/objectives, but in incompatible ways. As a simple example, higher peak user rates can be achieved by using more power (which affects the EE), allocating more transmission resources to users with good channels (which means less uniform user experience and higher latencies), or making use of intricate signal processing algorithms (which increases the complexity and cost of user devices).

To achieve the ambitious 5G goals, efficient network operation with respect to all the conflicting 5G objectives is required. This calls for a design framework that handles multiple objectives and supports the search for the best attainable operating point. But can we really formulate and solve multiobjective problems rigorously or is heuristic trial and error the only option? Is there even any optimal solution? These are questions that we address in this article.

CONVENTIONAL SINGLE-OBJECTIVE OPTIMIZATION

The conventional approach to physical-layer system optimization is that of selecting a scalar network utility function that is maximized under a set of constraints [8], [9]. A common problem formulation is that of maximizing the weighted sum of the users' data rates under transmit power constraints [6], [10], [11]. Alternatively, one can minimize the transmitted power under the constraint of guaranteeing certain data rates to each user [12], [13]. In recent years, the EE (in bit/Joule) has also arisen as a utility function [14]–[16].

In essence, the conventional approach is to select one of the objectives listed above as the sole objective, while the other objectives are transformed into constraints. The inherent heuristic assumptions are 1) one of the objectives is of dominating importance and 2) it is known beforehand what are good values for the constraints related to the other objectives. Moreover, the short-term values of the different objectives are usually considered in these network utility problems and not the long-term values, which are of main importance in the network design. Given the increased complexity due to heterogeneity, the need for long-term network optimization, and the diverse expectations on 5G networks, the conventional approach is no longer viable. However, we later show how to construct more appropriate single-objective problems.

NEW PARADIGM: MULTIOBJECTIVE OPTIMIZATION

Instead of assuming that one of the objectives is the sole objective, the fundamental approach is to recognize the existence of multiple objectives [2]: $g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_M(\mathbf{x})$, where M is the number of objectives. These objective functions can, for example, be area throughput, guaranteed rates for different classes of users, number of simultaneously active users, EE, etc. Explicit examples are given later in this article, while the theory is applicable for any arbitrary functions. The notation $\mathbf{g}(\mathbf{x}) = [g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_M(\mathbf{x})]^T$ is used to emphasize that the objective is vector valued.

The available resources (for example, time, frequency, space, power, and hardware) are modeled by a compact set $\mathcal{X} \subset \mathbb{R}^D$, which is called the *resource bundle* and has any finite dimension D. Each vector $\mathbf{x} \in \mathcal{X}$ represents a feasible way of utilizing the network resources. The satisfaction of this resource utilization equals $g_m(\mathbf{x}) \in \mathbb{R}$ with respect to the mth objective function. A larger value corresponds to higher satisfaction. For tractability, we assume that $g_m(\mathbf{x})$ is a bounded continuous function of $\mathbf{x}$ and nonnegative. We also assume that there exists a point $\mathbf{x}_0 \in \mathcal{X}$ such that $g_m(\mathbf{x}_0) = 0$ for all m. This operating point is the dissatisfaction of turning off the network and makes the satisfaction (for each objective) become a number from zero and upward. Not all practical objectives satisfy these conditions by nature; for example,

latency and error probability are typically to be minimized. However, there are standard transformations that reformulate such metrics into objective functions in our framework [3]–[6].

A key assumption is that the M objectives are not ordered and therefore studied without any preconceptions—all doors are kept open. In contrast to game theory, where each objective belongs to one of the competing agents, we assume that there is a network designer that would like to design the network to maximize all the M objectives simultaneously:

$$\begin{aligned} &\underset{\mathbf{x}}{\text{maximize}} \quad \mathbf{g}(\mathbf{x}) = [g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_M(\mathbf{x})]^T \\ &\text{subject to} \quad \mathbf{x} \in \mathcal{X}. \end{aligned} \tag{1}$$

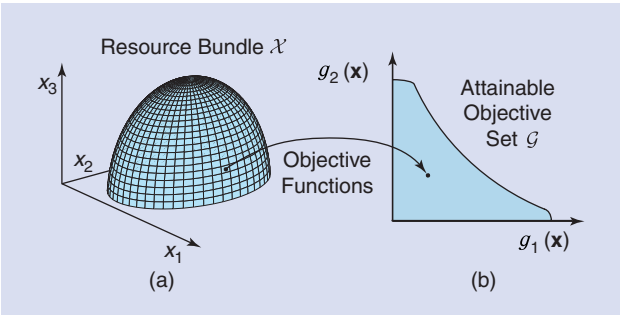
Note that (1) is the maximization of the vector $\mathbf{g}(\mathbf{x})$ containing the M objectives, which is defined as maximizing all elements simultaneously. This is known as a *MOO problem* (*MOOP*) or, alternatively, as a multicriteria or vector optimization problem [2]–[6]. These types of problems arise in many engineering fields because of the difficulty to find a scalar metric that exactly describes what we would like to achieve. We review the main concepts and properties related to MOOPs in this article. We provide the basic tools to understand the structure of MOOPs and how to solve these problems in practice. The properties are stated without proofs, while we recommend [3]–[5] for further details and [6] for a recent survey aimed at communication applications.

PROPERTY 1
The M objectives in (1) are conflicting and since there is no total order of vectors, there is (generally) no global optimum to the MOOP in (1).

This is the first important insight from the multiobjective framework; we cannot solve (1) in any globally optimal sense because there are only subjectively optimal solutions. Therefore, we turn the attention to the attainable objective set

$$\mathcal{G} = \{\mathbf{g}(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}, \tag{2}$$

which contains all the combinations of objective values $g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_M(\mathbf{x})$ that are simultaneously attainable under the available resources. The relationship between the resource bundle $\mathcal{X}$ and the attainable objective set $\mathcal{G}$ is visualized in Figure 1. Note that the origin is always in the objective set, $\mathbf{0} = [0 \ldots 0]^T \in \mathcal{G}$, due to the assumptions made earlier.



[FIG1] An illustration of a MOOP with a three-dimensional resource bundle $\mathcal{X}$ and a two-dimensional attainable objective set. For each resource utilization $\mathbf{x} = [x_1\ x_2\ x_3]^T \in \mathcal{X}$, the objective functions $g_1(\mathbf{x})$ and $g_2(\mathbf{x})$ assign a vector $g(\mathbf{x}) \in \mathcal{G}$.

When formulating the MOOP, the resource bundle $\mathcal{X}$ is selected to minimize the preconditions made on the utilization of network resources. This keeps all the options on the table, because it is generally difficult to articulate the network requirements a priori—at least in a strict mathematical sense. Nevertheless, the resource bundle can include certain fundamental network performance constraints (for example, that the M metrics should be better than in previous network generations).

PARETO OPTIMAL OPERATING POINTS

The shape of the attainable objective set $\mathcal{G}$ depends on the objective functions and the resource bundle $\mathcal{X}$, but it is usually a compact set with the property that $\mathbf{g} \in \mathcal{G}$ implies $c\mathbf{g} \in \mathcal{G}$ for all $c \in [0, 1]$ (that is, the performance can be uniformly degraded). The set $\mathcal{G}$ can be convex or nonconvex. Although Property 1 expresses that there is no global optimum, most points in $\mathcal{G}$ are strictly suboptimal. In fact, any point in the interior of $\mathcal{G}$ can be discarded because there exist other points in $\mathcal{G}$ that are more preferable with respect to all M objectives. The remaining points belong to the Pareto boundary.

DEFINITION 1 (PARETO BOUNDARY)
The strong Pareto boundary, $\partial\mathcal{G}$, consists of all points $\mathbf{g} \in \mathcal{G}$ for which there does not exist any $\mathbf{g}' \in \mathcal{G} \backslash \{\mathbf{g}\}$ with $g'_m \geq g_m$ for $m = 1, \ldots, M$.

The strong Pareto boundary consists of the attainable operating points that cannot be objectively dismissed, because none of the objectives can be improved without degrading other objectives. Evidently, any point that is not on the strong Pareto boundary is suboptimal because there exist other operating points that are better or at least as good for every objective. The strong Pareto boundary is as close to global optimality as one can get in multiobjective optimization; the operating points in $\partial\mathcal{G}$ are mutually unordered and can only be compared by subjective means. Each point $\mathbf{g} \in \partial\mathcal{G}$ describes a particular tradeoff between the M objectives. Hence, the Pareto boundary describes the set of (Pareto) efficient potential operating points from which we, as network designers, should select the one that is subjectively preferable to us.

The strong Pareto boundary is a subset of the upper boundary of $\mathcal{G}$. The complete upper boundary is referred to as the *weak* Pareto boundary and also contains points were some of the objectives (but not all) can be improved without degrading other objectives. This is illustrated in Figure 2, where the strong Pareto boundary either equals the complete upper boundary as in (a) or is a strict subset thereof as in (b). Figure 2 also shows the utopia point, which is defined as

$$\mathbf{u}_{\text{utopia}} = [u_1 \ldots u_M]^T = \begin{bmatrix} \max_{\mathbf{x} \in \mathcal{X}} f_1(\mathbf{x}) \\ \vdots \\ \max_{\mathbf{x} \in \mathcal{X}} f_M(\mathbf{x}) \end{bmatrix}. \tag{3}$$

This is the ideal operating point that simultaneously maximizes all M objectives. If $\mathbf{u}_{\text{utopia}} \in \mathcal{G}$, the MOOP is trivial because the strong Pareto boundary consists of only the utopia point, $\partial\mathcal{G} = \{\mathbf{u}_{\text{utopia}}\}$, and it is the unique global optimum.

PROPERTY 2
Any MOOP with multiple conflicting objective functions is nontrivial in the sense that $\mathbf{u}_{\text{utopia}} \notin \mathcal{G}$ and, consequently, there is no global optimum.

Single-objective optimization problems are MOOPs with $M = 1$ and are thus trivial from the MOO perspective. Since the Pareto boundary consists of all tentative effective operating points, we need to find the network parameters (that is, the resource utilizations) that attain these points.

DEFINITION 2 (PARETO OPTIMAL POINT)
A point $\mathbf{x}^* \in \mathcal{X}$ in the resource bundle is a Pareto optimal point if $\mathbf{g}(\mathbf{x}^*) \in \partial\mathcal{G}$.

The mapping from a Pareto optimal point $\mathbf{x}^*$ to the Pareto boundary is given by the vector-valued multiobjective function $\mathbf{g}(\mathbf{x}^*)$ and is, hopefully, given in closed form. The inverse mapping is, on the other hand, hard to derive in most cases. The multiobjective function might not be bijective, which means that multiple points in $\mathcal{X}$ can give exactly the same objective point. This happens frequently when transmitting from multiantenna arrays, where the beamforming coefficients are only unique up to a common phase rotation [6].

SOLVING A MOOP BY VISUALIZATION

In practice, we would like to go beyond the Pareto boundary and actually solve the MOOP, in the sense of selecting a single Pareto optimal point $\mathbf{x}^*$ and its corresponding operating point $\mathbf{g}(\mathbf{x}^*) \in \partial\mathcal{G}$. To this end, we need to bring in the subjective preference of the network designer to compare different operating points at the Pareto boundary. This is not as simple as it might seem, because neither the Pareto boundary $\partial\mathcal{G}$ nor the objective set $\mathcal{G}$ are known beforehand. Simple closed-form expressions are seldom available. In fact, one needs to spend considerable computational resources on learning the objective set. For example, one can characterize $\mathcal{G}$ by computing a discrete set of sample points, which enables the network designer to visualize the different possibilities and make an informed decision. This is known as the *a posteriori method*, because the network designer formulates its subjective preference after the numerical computations have taken place [2].

We describe two approaches to compute sample points:

1) Traverse the resource bundle $\mathcal{X}$ by computing $\mathbf{g}(\mathbf{x})$ over a finite grid of $\mathbf{x} \in \mathcal{X}$. For example, if $0 \le x_m \le 1$, then we can limit ourselves to the six discrete values $x_m \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. If the same number of discrete values are taken for all D resource variables in $\mathcal{X}$, we have 6^D grid points to consider.

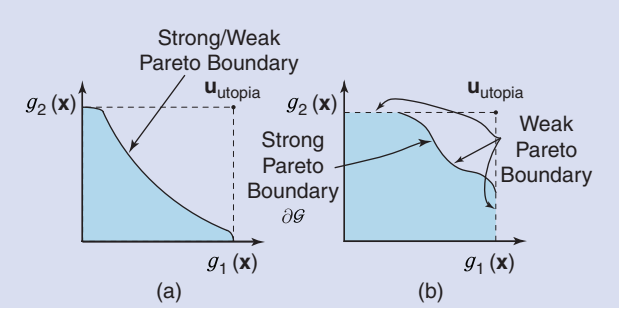
2) Traverse the strong Pareto boundary $\partial\mathcal{G}$ by searching for the outermost point in $\mathcal{G}$ in different directions. The search directions can be represented by vectors $\mathbf{v} = [v_1 \ldots v_M]^T$ that point out (nonnegative) geometric directions from the origin (recall that $\mathbf{0} \in \mathcal{G}$ by definition). Each search corresponds to solving the single-objective optimization problem

$$\begin{aligned} &\underset{\mathbf{x},\lambda}{\text{maximize}} \quad \lambda \\ &\text{subject to} \quad g_m(\mathbf{x}) \ge \lambda v_m, \quad m = 1, \ldots, M, \quad \mathbf{x} \in \mathcal{X}, \end{aligned} \tag{4}$$

which is a referred to as a weighted Chebyshev problem in the MOO literature [3]–[6] (in fact, it is the epigraph form of it [17]). If λ^* is the optimal value for a given $\mathbf{v}$, we can be sure that $\lambda^*\mathbf{v} \in \mathcal{G}$ and that this point lies on the weak Pareto boundary (upper boundary). If needed, one can guarantee to attain the strong Pareto boundary by slightly modifying (4); see [3] for details. By solving (4) for a finite set of search directions (for example, equally spaced in the angular sense), one can obtain a set of sample points that characterizes the weak/strong Pareto boundary.

These two approaches have their respective pros and cons. The first approach is computationally efficient, assuming that the function values $\mathbf{g}(\mathbf{x})$ are easy to evaluate. The main limiting factor might be the memory storage, since the number of samples scales exponentially with D. Extensive postprocessing might also be required because most sample points will be in the interior of $\mathcal{G}$ and can be discarded since there are other samples that are better with respect to all M objectives. The resource bundle can sometimes be parameterized more efficiently by exploiting the objective functions. This can be used to improve the resolution of the objective set $\mathcal{G}$ using fewer samples. For example, transmit beamforming can be represented by one parameter per user [6, Sec. 3.2], which removes redundancy in multiantenna wireless communications where the number of beamforming coefficients equals the number of users times the number of transmit antennas.

The second approach guarantees a high resolution because every sample point lies on the weak Pareto boundary. The downside is the computational complexity, which is proportional to the complexity of solving the search problem in (4). Indeed, this approach can only be utilized if there is a tractable way of solving (4). This is the case whenever there exists an efficient way to make a membership test; that is, to determine if a given point $\tilde{\mathbf{g}} \in \mathbb{R}^M$ belongs to the objective set or not. We elaborate on this in "Finding the Pareto Boundary by Bisection."



[FIG2] An illustration of the Pareto boundary, which is either the complete upper boundary of $\mathcal{G}$ as in (a) or a subset of the upper boundary as in (b). The unattainable utopia point is also shown.

FINDING THE PARETO BOUNDARY BY BISECTION

The single-objective optimization problem in (4) finds the weak Pareto boundary in the direction $\mathbf{v}$ from the origin. This problem can be solved by checking if a series of points, each denoted $\boldsymbol{\mu} = [\mu_1 \dots \mu_M]^T \in \mathbb{R}^M$, belong to the attainable objective set $\mathcal{G}$ or not. This is determined by the membership test

$$\begin{aligned} &\text{find} && \mathbf{x} \in \mathcal{X} \\ &\text{subject to} && g_m(\mathbf{x}^*) = \mu_m. \end{aligned} \tag{S1}$$

The complexity of this feasibility problem is a baseline for other optimization problems that involve the same resource bundle and objective functions—if the membership test is computationally intractable, there is little chance that any meaningful problem formulation is practically solvable. Fortunately, there are many cases when the membership test is efficiently solvable; for example, it is a convex problem in many beamforming design problems for cellular networks [6].

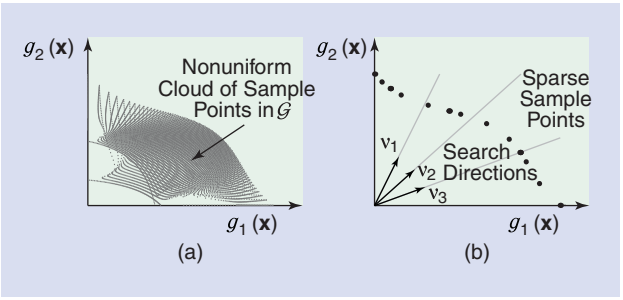
Equipped with a tractable membership test, we can solve (4) by first defining a range $[\lambda_{\min}, \lambda_{\max}]$ of values for λ, such that $\lambda_{\min}\mathbf{v} \in \mathcal{G}$ and $\lambda_{\max}\mathbf{v} \notin \mathcal{G}$. The lower limit can be $\lambda_{\min} = 0$, since the origin is always attainable. The upper limit is selected for the MOOP at hand, for example, by exploiting the utopia point (if it is known) or by relaxing the problem to find other unattainable points. The following algorithm solves (4):

Input: Range $[\lambda_{\min}, \lambda_{\max}]$ and accuracy $\epsilon > 0$
while $\lambda_{\max} - \lambda_{\min} > \epsilon$ **do**
  Make membership test (S1) for $\boldsymbol{\mu} = ((\lambda_{\max} + \lambda_{\min})/2)\mathbf{v}$
  if $\boldsymbol{\mu} \in \mathcal{G}$ **then**
    $\lambda_{\min} \leftarrow (\lambda_{\max} + \lambda_{\min})/2$
  else
    $\lambda_{\max} \leftarrow (\lambda_{\max} + \lambda_{\min})/2$
  end if
end while
Output: Attainable point $\mathbf{a} = \lambda_{\min}\mathbf{v}$.

This is a classical bisection algorithm that cuts the range $[\lambda_{\min}, \lambda_{\max}]$ in half in each iteration [17]. Bisection has fast convergence and the distance between $\mathbf{a}$ and the Pareto boundary is below $\epsilon\|\mathbf{v}\|$ for the given $\epsilon > 0$.

Figure 3 illustrates the two approaches. The first approach gives sample points that provides a sense of the shape of $\mathcal{G}$: Is it convex? What are the numerical ranges? Are the objectives strongly or weakly conflicting? The density of points is nonuniform and it is not guaranteed that any sample point is exactly on the Pareto boundary. In contrast, the second approach gives a sparse set of sample points that are exactly on the Pareto boundary. Each point is found by searching in a certain direction (for example, $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$) from the origin.

By looking at visualizations of the Pareto boundary, such as the ones in Figure 3, the network designer can understand the fundamental properties and tradeoffs between conflicting objectives. Visualization is a powerful tool that supports the network designer in making an informed decision. This is the essence of



[FIG3] An illustration of the two approaches to visualize the objective set $\mathcal{G}$ by computing sample points. (a) Approach 1. (b) Approach 2.

the a posteriori method. Since it is difficult to visualize more than three dimensions at a time, one needs to limit the granularity to a few objectives at a time. This issue can be treated in an iterative fashion where the network designer makes preliminary decisions (for example, regarding the preferred minimal level for different objectives), which replaces the current resource bundle $\mathcal{X}$ with a smaller set $\tilde{\mathcal{X}} \subset \mathcal{X}$. This interactive process continues until the network designer is satisfied—a type of psychological convergence [4].

SOLVING A MOOP BY SCALARIZATION

An alternative way to solve MOOPs in practice is to let the network designer articulate preferences before any computations take place. This is referred to as the *a priori method*. The purpose is to find the operating point $\mathbf{g} \in \mathcal{G}$ that satisfies these preferences as well as possible. In particular, the designer can specify a goal function $f : \mathbb{R}^M \to \mathbb{R}$ that for any conceivable operating point $\mathbf{g}$ (attainable or not) produces a scalar describing how preferable that point is (large value means high preference). The goal function describes a certain subjective tradeoff between the objectives and thus imposes an order on the vectors in the objective set $\mathcal{G}$. Consequently, the MOOP in (1) is converted into the single-objective optimization problem

$$\begin{aligned} &\underset{\mathbf{x}}{\text{maximize}} && f(g_1(\mathbf{x}), g_2(\mathbf{x}), \dots, g_M(\mathbf{x})) \\ &\text{subject to} && \mathbf{x} \in \mathcal{X}. \end{aligned} \tag{5}$$

This conversion is called *scalarization,* and the solution is a weak Pareto boundary point and usually also belongs to the strong Pareto boundary. In contrast to the conventional approach of having a sole performance objective and expressing other potential objectives as constraints, (5) combines the M objectives into a scalar goal function and has no additional constraints. It is indeed possible to impose constraints on the acceptable values for certain objectives also in the scalarization case, but it is not required.

The goal function can take many forms and a variety of classes of functions can be found in the literature; see [3]–[6]. We describe four important goal function classes. The most common goal function might be the weighted sum

$$f_{\text{sum}}(\cdot) = \sum_{m=1}^{M} w_m g_m(\mathbf{x}), \tag{6}$$

where $w_1, \ldots, w_M$ are positive weights that specify the priority of each objective; the priority of the mth objective grows by increasing the corresponding weight w_m. One should be careful when interpreting the relative priorities, because the objectives can have different scales, units, and couplings.

Similarly, one can consider the weighted product

$$f_{\text{product}}(\cdot) = \prod_{m=1}^{M} (g_m(\mathbf{x}))^{w_m}, \tag{7}$$

where the weights are defined as before but act differently. Note that (7) is the (weighted) geometric mean, while (6) is the (weighted) arithmetic mean. Generally speaking, the geometric mean is better at comparing objectives with different numerical ranges, because the relative scaling has no impact.

The weighted Chebyshev formulation, also known as the weighted max-min formulation, played a key role when we computed sample points on the Pareto boundary in the a posteriori method. The weighted Chebyshev goal function is

$$f_{\text{Chebyshev}}(\cdot) = \min_{1 \le m \le M} \frac{g_m(\mathbf{x})}{w_m}. \tag{8}$$

This scalarization is equivalent to (4) if we write it on epigraph form [17] and select the weights $w_1, \ldots, w_M$ as $w_m = v_m$ for all m. Hence, this scalarization searches for the Pareto boundary in the direction $[w_1 \ldots w_M]^T$ from the origin.

Alternatively, the network designer can specify a preferable operating point $\boldsymbol{\upsilon} \in \mathbb{R}^M$ (for example, the utopia point $\boldsymbol{\upsilon} = \mathbf{u}_{\text{utopia}}$). The distance goal function is defined as

$$f_{\text{distance}}(\cdot) = -\| \boldsymbol{\upsilon} - \mathbf{g}(\mathbf{x}) \| \tag{9}$$

and measures the distance from the preferable point in some appropriately selected norm $\| \cdot \|$. The norm $\| \boldsymbol{\upsilon} - \mathbf{g}(\mathbf{x}) \|$ should be small (preferably zero), thus the negative sign in (9) is used to achieve a goal function that is to be maximized.

The final operating point is determined by the choice of goal function. Interestingly, the computational complexity also varies with the goal function; the scalarized problem in (5) may be convex (that is, solvable in polynomial time) for some classes of functions, while other classes give nonconvex problems with exponential complexity—or even worse. For example, [11] proved that transmit beamforming optimization in cellular networks is (quasi)convex for the weighted Chebyshev goal function and strongly NP-hard for most other goal functions. This result has general implications.

PROPERTY 3

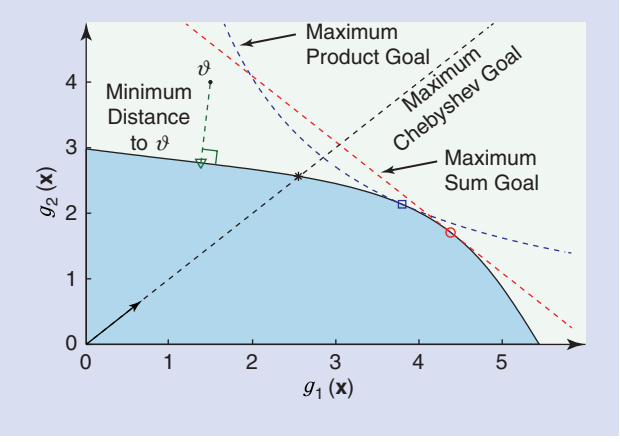
The weighted Chebyshev goal function is the safest choice in terms of computational complexity; if there exists a tractable membership test for the objective set, then it can be solved efficiently as described in "Finding the Pareto Boundary by Bisection."

Since goal functions are inherently subjective, no choice is better than the others in terms of optimality. Property 3 inspired [6] to propose what is known as the *pragmatic approach to resource allocation*: select the weighted Chebyshev goal function (due to its tractable complexity) and exploit the weights to adapt to the needs of the network designer.

The operating points attained by different scalarizations are illustrated in Figure 4, for a scenario where the attainable ranges are different for the two objectives. The goal functions in (6)–(9) are considered for $w_1 = w_2 = 1$. Let f^* denote the optimal function value in (5), which of course is different for each goal function. The optimal operating point with the sum goal function lies on the level curve $f_{\text{sum}}(\cdot) = f^*$, which is the red line in Figure 4. Similarly, $f_{\text{product}}(\cdot) = f^*$ gives the blue parabolic level curve of the product goal function. These level curves touch the objective set $\mathcal{G}$ in unique Pareto boundary points, which are the optimal operating points for the respective scalarized problems. As described earlier, the Chebyshev goal function searches on a line from the origin. For $w_1 = w_2 = 1$, this is the line where the two objectives have equal values. If there is a preferable operating point $\vartheta \notin \mathcal{G}$ as in Figure 4, (5) provides the operating point that minimizes the distance to $\mathcal{G}$ (the Euclidean distance is used in Figure 4).

The function classes in (6)–(8) are parameterized by the weights $\mathbf{w} = [w_1, \ldots, w_M]^T$. Different weight selections give different Pareto optimal points when solving (5). By varying $\mathbf{w}$ over the set $\mathcal{W} = \{\mathbf{w} : w_m \ge 0 \ \forall m, \sum_m w_m = 1\} \in \mathbb{R}^M$ of positive weights that sum up to one, we can attain the whole Pareto boundary or a subset thereof, depending on the function class [4]. Since each scalarization in (5) is a single-objective optimization problem, it is equipped with conventional Karush–Kuhn–Tucker (KKT) optimality conditions [17]. By considering all $\mathbf{w} \in \mathcal{W}$, these can be extended to a joint set of optimality conditions for all points achieved by the function class [5]. These optimality conditions describe the structure of the resource utilizations that achieve the Pareto boundary; for example, it was utilized in [6, Sec. 3.2] to parameterize any efficient transmit beamforming.

Finally, we note that game theory provides an alternative way to select operating points from the Pareto boundary, by specifying



[FIG4] An illustration of the Pareto optimal operating points achieved by scalarization using common goal functions.

the rules of a game instead of a goal function [18]. These techniques are mainly for systems with separate agents/objectives that compete for shared resources, while single-operator networks typically have dedicated resources.

CASE STUDY: DESIGNING MASSIVE MIMO SYSTEMS

We exemplify the usefulness of MOO by a case study. The goal is to visualize tradeoffs between conflicting 5G objectives and describe how the framework can be used to acquire new insights and prove old heuristic observations. In recent years, coordinated multipoint (CoMP) techniques have shown the potential to greatly improve the area rates in cellular networks. This is achieved by deploying antenna arrays at base stations (BSs) and apply a coordinated space division multiple access (SDMA) scheme across the network [6], [19]–[21]. Unfortunately, CoMP is difficult to implement since the coordination signaling is limited [22], the signal processing complexity increases drastically [11], and the performance gains are not robust to the interuser interference caused by having imperfect channel state information (CSI) [20].

The concept of massive MIMO has gained traction since it might eliminate the CoMP issues listed above [23]–[26]. Massive MIMO is based on the idea of deploying large arrays with unconventionally many active antennas at the BSs and serve a much smaller number of users; for example, hundreds of antennas that serve several tens of users. One would imagine that adding more antennas and users into a system would make CoMP even more difficult to implement, but the beauty of massive MIMO is that this is not the case [23]. The excessive number of antennas brings robustness to imperfect CSI, makes low-complexity signal processing close to optimal [24], and allows for simple implicit intercell coordination [25]. Massive MIMO systems are even robust to the distortions caused by hardware imperfections [26].

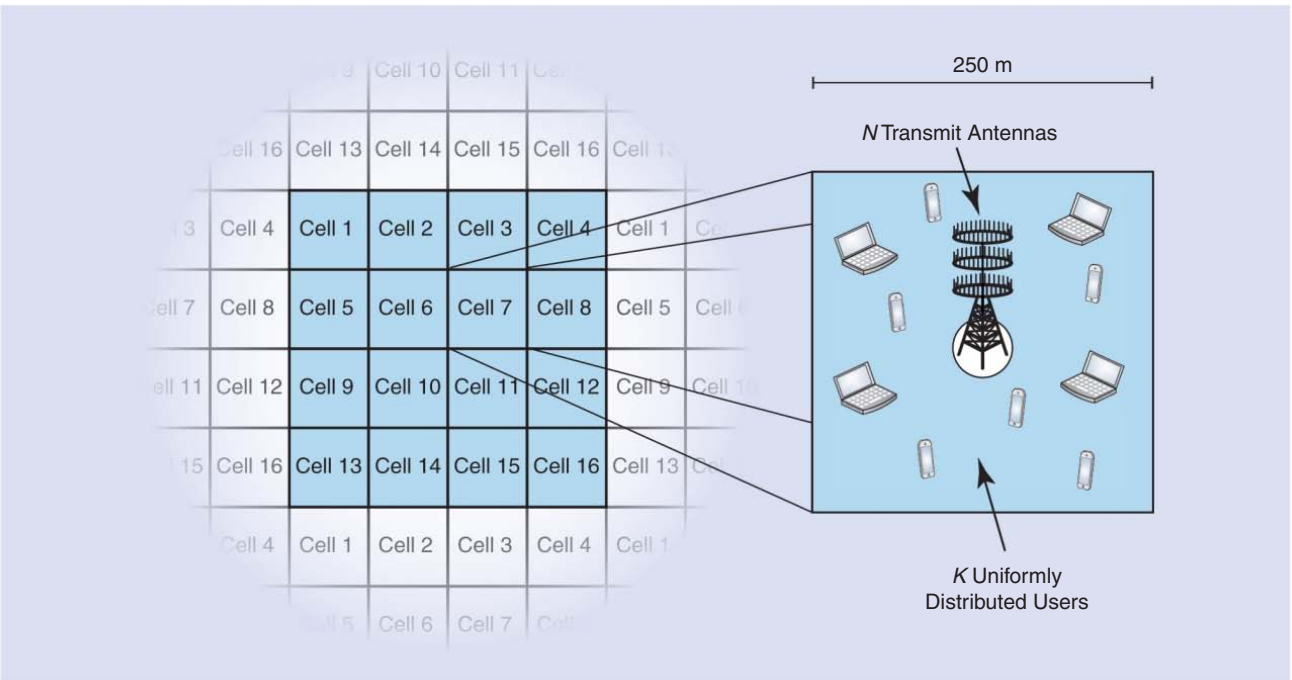
In this case study, we strive to optimize the downlink transmission of a massive MIMO system to balance $M = 3$ conflicting objectives: high average user rates, high average area rates, and high EE. The cellular network that we consider has 16 cells, each consisting of a BS with N antennas and K single-antenna users. The bandwidth is $B = 10$ MHz, the emitted power per BS is denoted P W, and $\sigma^2 = 10^{-13}$ W is the average noise power.

Each cell is a square of size 250×250 m (that is, the area is $A = 0.25^2$ km^2) and we apply classic wrap-around to avoid edge effects; the scenario is shown in Figure 5. The K users are uniformly distributed in the cell, with a minimum distance of 35 m. For a randomly picked user, let λ_{serving} be the channel variance from the serving BS and $P\lambda_{\text{intercell}}$ be the average intercell interference power. We are concerned with average behaviors and define the expectations $\Lambda_1 = \mathbb{E}\{1/\lambda_{\text{serving}}\}$ and $\Lambda_2 = \mathbb{E}\{\lambda_{\text{intercell}}/\lambda_{\text{serving}}\}$ for later use. Using the same 3GPP path loss model as in [16], we get $\Lambda_1 = 1.72 \cdot 10^9$ and $\Lambda_2 = 0.54$.

The optimization/resource variables in this case study are the number of BS antennas N, the number of users K, and the transmit power P per cell. The resource bundle is

$$\mathcal{X} = \left\{ [K\ N\ P]^T : \begin{matrix} 1 \le K \le \frac{N}{2}, \\ 2 \le N \le N_{\max}, \\ 0 \le P \le NP_{\max} \end{matrix} \right\}, \tag{10}$$

where $N_{\max} = 500$ is the maximal number of antennas that can fit at each BS, $P_{\max} = 20$ W is the maximal emitted power per



[FIG5] An illustration of the scenario in the case study: a cellular network with N antennas per BS and K users per cell.

BS antenna, and the constraint $K \leq N/2$ makes sure that we have many more BS antennas than active users.

Next, we define the average user rate and the total power consumption per cell. For simplicity, we assume that each BS has obtained perfect CSI for its users and applies zero-forcing precoding, which is a signal processing technique that cancels out intracell interference by beamforming and adapts the power allocation to guarantee the same rate to each user. Similar to [16], the average user rate can be shown to be

$$R_{\text{average}} = B\left(1 - \frac{K}{\Upsilon}\right)\log_2\left(1 + \frac{\frac{P}{K}(N-K)}{\sigma^2\Lambda_1 + P\Lambda_2}\right), \tag{11}$$

under the assumption that each user knows its useful channel and treats intercell interference as noise. The prelog-factor $(1-K/\Upsilon)$ accounts for the necessary overhead for channel acquisition, and $\Upsilon = 1{,}000$ is the number of channel uses that the channel stays fixed. It is selected as $\Upsilon = B_{\text{coherence}}\tau_{\text{coherence}}$, where $B_{\text{coherence}} = 200$ kHz is the coherence bandwidth and $\tau_{\text{coherence}} = 5$ ms is the coherence time. Looking inside the logarithm of (11), P/K is the average transmit power per user, $N-K$ is the effective array gain, and $\sigma^2\Lambda_1 + P\Lambda_2$ is the average degradation from noise and intercell interference.

Based on the models and the practical numbers in [16], [27], and [28], the total power consumption per cell is given by

$$P_{\text{total}} = \frac{P}{\eta} + NC_N + KC_K + \frac{C_{\text{precoding}}}{L} + C_0, \tag{12}$$

where $\eta = 0.31$ is the efficiency of the power amplifiers at the BS, $C_N = 1$ W is the hardware power consumed per transmit antenna, $C_K = 0.3$ W is the hardware power per user, and $C_0 = 10$ W is the static hardware power. In addition, $C_{\text{precoding}} = 3K^2N(B/T)$ is the floating-point operations per second (flops) required to compute zero-forcing precoding, while $L = 12.8$ Gflops/W is a typical computational efficiency.

We are now ready to define our three objective functions:

$$g_1(\mathbf{x}) = R_{\text{average}} \quad \text{[bit/s/user]} \tag{13}$$

$$g_2(\mathbf{x}) = \frac{K}{A}R_{\text{average}} \quad \text{[bit/s/km}^2\text{]} \tag{14}$$

$$g_3(\mathbf{x}) = \frac{KR_{\text{average}}}{P_{\text{total}}} \quad \text{[bit/J]}, \tag{15}$$

where $\mathbf{x} = [K\,N\,P]^T$ are the optimization/resource variables. The objective $g_1(\mathbf{x})$ is the average user rate, $g_2(\mathbf{x})$ is the average area rate, and $g_3(\mathbf{x})$ is the EE.

DESIGNING MASSIVE MIMO BY MOO FRAMEWORK

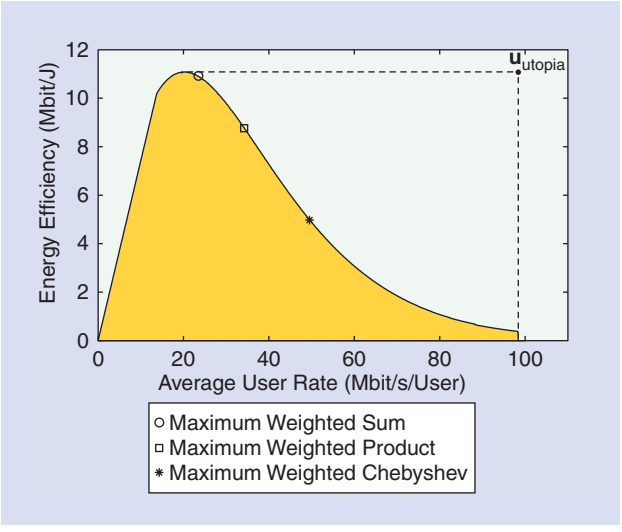
We have now defined a MOOP of the type in (1). The resource bundle is given by (10) and the three objectives are defined in (13)–(15). We now describe how the MOO framework can be used to study tradeoffs between these objectives, with the purpose of deriving new insights and confirming old beliefs.

The tradeoff between the average user rate and the EE is shown in Figure 6. The objective set with respect to these two objectives was generated by the second approach described earlier (that is, searching for the Pareto boundary in different directions). Figure 6 shows that these two objectives are aligned up to the point $g_1 = 20.4$ Mbit/s/user and $g_3 = 11.1$ Mbit/J, where the maximal EE is achieved. The objectives are then conflicting, because the user rates can only be further increased by making drastic sacrifices in the EE.

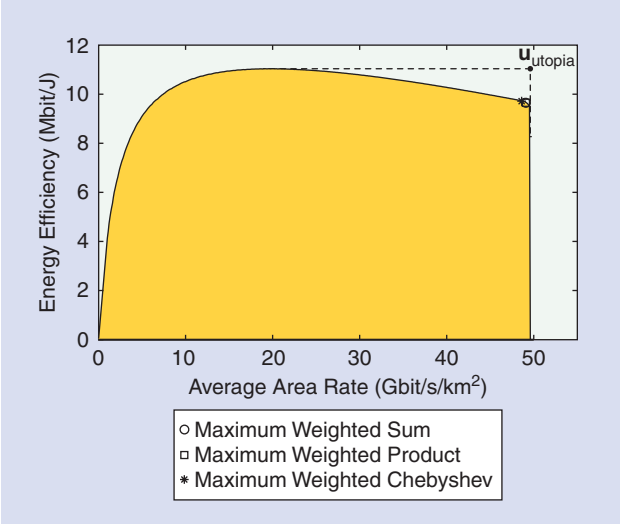
Another tradeoff is illustrated in Figure 7, where the average area rate and the EE are compared. These objectives are also aligned until the EE reaches its maximum value. However, one can increase the area rate beyond this point with only minor losses in EE. By noting that $g_2(\mathbf{x}) = (K/A)g_1(\mathbf{x})$ and comparing with the previous figure, this obviously means that the area rate is improved by transmitting to more users (that is, having a larger K) and not by increasing the rate per user. This conclusion is supported by Figure 8, which shows the three-dimensional objective set with respect to all objectives.

Figure 8 reveals that high area rates are only achievable when the rate per active user is low, which means that we serve many user devices in parallel. In contrast, high rates per user is only achievable by having fewer active users. High EE is possible when the rate per user is small. These different operating points are achieved by different resource utilizations $\mathbf{x} \in \mathcal{X}$; thus, the number of antennas/users are different and the signal processing related to precoding changes. This proves the otherwise heuristic belief that the network architecture must be flexible (for example, in terms of switching off antennas and precoding adaptation) if different operating points should be attainable in different traffic cases.

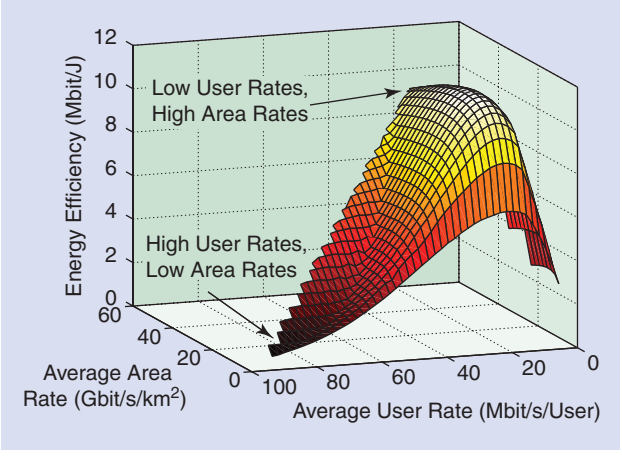
The previous discussion is typical for the a posteriori method; we analyzed the shape of the Pareto boundary and drew conclusions on which operating points that are preferable to us. If we would instead utilize the a priori method, then we need to specify a goal function. This can be done by picking any of the function classes described in (6)–(9) and selecting the corresponding parameters (for example, weights) to describe our subjective goals. To aid us in this process, suppose we know the utopia point $\mathbf{u}_{\text{utopia}}$ [defined in (3)] in advance. This point contains the maximal value for each objective if we would focus completely on it. If the three objectives are equally important to us, it makes sense to normalize their numerical ranges. This is achieved by setting $\mathbf{w} = [(1/u_1)\ (1/u_2)\ (1/u_3)]^T$ in the weighted sum goal, $\mathbf{w} = [1\ \ 1\ \ 1]^T$ in the weight product goal, and $\mathbf{w} = \mathbf{u}_{\text{utopia}}$ in the weighted Chebyshev goal formulation. The corresponding operating points when solving the scalarized problem in (5) are shown in Figures 6 and 7. The shape of the region has a great impact on the spread of the operating points, but different weights still give different operating points (as discussed earlier). The utopia point $\mathbf{u}_{\text{utopia}} = [u_1\,u_2\,u_3]^T$ is also shown in these figures. We observe that it is far outside the attainable objective set in Figure 6, since the two objectives are strongly conflicting. On the contrary, the utopia point is quite close to the objective set in Figure 7, where the conflict is rather mild.



[FIG6] Visualization of the tradeoff between two objectives in the case study: average user rate and EE.



[FIG7] Visualization of the tradeoff between two objectives in the case study: average area rate and EE.



[FIG8] Visualization of the tradeoff between all three objectives in the case study: average user rate, average area rate, and EE.

Finally, we remark that the a posteriori and a priori methods can be combined. The network architecture can, for example, be designed by studying the shape of the attainable objective set and making sure that the network can adapt and achieve different operating points at the Pareto boundary at different times. The system designer can then formulate multiple goal functions that are exploited for efficient real-time network adaptation, based on current traffic load, service requirements, and capability of the user devices.

CONCLUSIONS AND FUTURE DIRECTIONS

The design expectations on 5G wireless networks cannot be properly articulated by a single performance objective. There are many conflicting objectives, such as improving the peak user rates, average area rates, and EE. The network design thus calls for multiobjective optimization, which is rigorous framework for studying and solving design problems with multiple objectives. This article provided a survey on this topic. There is no objectively optimal solution to this type of problems, but there are two main methods to find subjectively optimal solutions that fit the needs of the network designer. The a posterior method computes sample points on the Pareto boundary—the set of tentative operating points where no objective can be improved without degrading another objective. These sample points are used to visualize the Pareto boundary for the network designer, who can then make well-informed design decisions. Alternatively, the network designer can specify a goal function that describes the acceptable tradeoffs between objectives and infers an order on the attainable operating points. One can then maximize this tradeoff by solving a conventional optimization problem and thereby obtain the most suitable Pareto boundary point.

We also provided a case study on network dimensioning of cellular networks that allows for massive MIMO deployment. This example illustrates our vision of how the MOO framework can be utilized to balance conflicting performance objectives when designing future wireless communication networks. While the analytic tools provided by MOO are well established, the applications to communication networks are greatly unexplored. A particular research challenge is to formulate MOOPs with a modeling granularity that allows us to answer fundamental design questions related to how the system can efficiently manage the heterogeneous 5G characteristics described in the introduction of this article. To this end, the models must capture the main practical propagation characteristics, be robust to hardware imperfections and uncertain model parameters, and allow for optimization of the signal processing techniques. All of this is to be done while making the basic optimization operations (for example, the membership test described previously) computationally tractable.

ACKNOWLEDGMENTS

This article has been supported by the International Postdoc grant 2012-228 from the Swedish Research Council and the ERC Starting Grant 305123 MORE (Advanced Mathematical Tools for Complex Network Engineering).

AUTHORS

Emil Björnson (emil.bjornson@liu.se) received the M.S. degree from Lund University, Sweden, in 2007 and the Ph.D. degree in 2011 from KTH Royal Institute of Technology, Sweden. From 2012 to 2014, he was a joint postdoctoral researcher at Supélec, France, and KTH Royal Institute of Technology, Sweden, sponsored by a personal international postdoctoral grant from the Swedish Research Council. He is the lead author of *Optimal Resource Allocation in Coordinated Multi-Cell Systems* and received Best Conference Paper Awards in 2009, 2011, and 2014. He is now a research fellow in the tenure track at Linköping University, Sweden.

Eduard Jorswieck (eduard.jorswieck@tu-dresden.de) received the M.S. and Ph.D. degrees from the Technische Universität Berlin, Germany, in 2000 and 2004, respectively. He was with the Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institut Berlin, from 2000 to 2008. From 2005 to 2008, he was a lecturer at the Technische Universität Berlin. From 2006 to 2008 he was a postdoctoral researcher and assistant professor at the KTH Royal Institute of Technology, Sweden. Since 2008, he has been the head of the Chair of Communications Theory and is a full professor at Dresden University of Technology, Germany. In 2006, he received the IEEE Signal Processing Society Best Paper Award.

Mérouane Debbah (merouane.debbah@supelec.fr) received the M.Sc. degree in 1999 and the Ph.D. degree in 2002 from École Normale Supérieure de Cachan, France. From 1999 to 2002, he worked for Motorola Labs. From 2002 until 2003, he was appointed senior researcher at the Vienna Research Center for Telecommunications, Austria. From 2003 until 2007, he was with the mobile communications department of the Institut Eurecom (Sophia Antipolis, France) as an assistant professor. He is currently a professor at Supélec, France, and holds the position of Alcatel-Lucent Chair on Flexible Radio. He received the 2005 Mario Boella Prize Award, the 2007 GLOBECOM Best Paper Award, the 2009 Wi-Opt Best Paper Award, the 2010 Newcom++ Best Paper Award, as well as the 2007 Valuetools, 2008 Valuetools, and 2009 CrownCom Best Student Paper Awards. He is a WWRF fellow.

Björn Ottersten (bjorn.ottersten@uni.lu) received the M.S. degree from Linköping University, Sweden, in 1986 and the Ph.D. degree in 1989 from Stanford University, California. In 1991 he was appointed professor of signal processing at KTH Royal Institute of Technology, Sweden. During 1996–1997, he was the director of research at ArrayComm Inc, a start-up company based on his patented technology. Currently, he is the director for the Interdisciplinary Centre for Security, Reliability, and Trust at the University of Luxembourg. He coauthored articles that received the IEEE Signal Processing Society Best Paper Award in 1993, 2001, 2006, and 2013, and several IEEE conference papers receiving Best Paper Awards. In 2011, he received the IEEE Signal Processing Society Technical Achievement Award. He is the editor-in-chief of EURASIP's journal, *Signal Processing*. He is a Fellow of the IEEE and EURASIP.

REFERENCES

[1] S. Tombaz, A. Västberg, and J. Zander, "Energy- and cost-efficient ultra-high-capacity wireless access," *IEEE Wireless Commun. Mag.*, vol. 18, no. 5, pp. 18–24, 2011.

[2] L. Zadeh, "Optimality and non-scalar-valued performance criteria," *IEEE Trans. Autom. Control*, vol. 8, no. 1, pp. 59–60, 1963.

[3] R. T. Marler and J. S. Arora, "Survey of multi-objective optimization methods for engineering," *Struct. Multidisc Optim.*, vol. 26, no. 6, pp. 369–395, 2004.

[4] J. Branke, K. Deb, K. Miettinen, and R. Slowinski, Eds., *Multiobjective Optimization: Interactive and Evolutionary Approaches*. New York: Springer, 2008.

[5] R. Bot, S.-M. Grad, and G. Wanka, *Duality in Vector Optimization*. New York: Springer, 2009.

[6] E. Björnson and E. Jorswieck, "Optimal resource allocation in coordinated multi-cell systems," *Found. Trends Commun. Inform. Theory*, vol. 9, nos. 2–3, pp. 113–381, 2013.

[7] A. Osseiran, "The 5G future scenarios identified by METIS—the first step toward a 5G mobile and wireless communications system," press release, Sept. 2013. METIS project. [Online]. Available: https://www.metis2020.com/

[8] F. Kelly, "Charging and rate control for elastic traffic," *Eur. Trans. Telecom.*, vol. 8, no. 1, pp. 33–37, 1997.

[9] D. P. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1439–1451, 2006.

[10] H. Weingarten, Y. Steinberg, and S. Shamai, "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3936–3964, 2006.

[11] Y.-F. Liu, Y.-H. Dai, and Z.-Q. Luo, "Coordinated beamforming for MISO interference channel: Complexity analysis and efficient algorithms," *IEEE Trans. Signal Processing*, vol. 59, no. 3, pp. 1142–1157, 2011.

[12] F. Rashid-Farrokhi, K. J. R. Liu, and L. Tassiulas, "Transmit beamforming and power control for cellular wireless systems," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 8, pp. 1437–1450, 1998.

[13] M. Chiang, P. Hande, T. Lan, and C. W. Tan, "Power control in wireless cellular networks," *Found. Trends Netw.*, vol. 2, no. 4, pp. 355–580, 2008.

[14] Y. Chen, S. Zhang, S. Xu, and G. Y. Li, "Fundamental trade-offs on green wireless networks," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 30–37, 2011.

[15] C. Isheden, Z. Chong, E. Jorswieck, and G. P. Fettweis, "Framework for link-level energy efficiency optimization with informed transmitter," *IEEE Trans. Wireless Commun.*, vol. 11, no. 8, pp. 2946–2957, 2012.

[16] E. Björnson, L. Sanguinetti, J. Hoydis, and M. Debbah, "Optimal design of energy-efficient multi-user MIMO systems: Is massive MIMO the answer?" submitted for publication.

[17] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[18] E. Jorswieck, L. Badia, T. Fahldieck, E. Karipidis, and J. Luo, "Spectrum sharing improves the network efficiency for cellular operators," *IEEE Commun. Mag.*, vol. 52, no. 3, pp. 129–136, 2014.

[19] R. H. Roy and B. Ottersten, "Spatial division multiple access wireless communication systems," European Patent EP0616742, 1994.

[20] D. Gesbert, M. Kountouris, R. W. Heath, C.-B. Chae, and T. Sälzer, "Shifting the MIMO paradigm," *IEEE Signal Processing Mag.*, vol. 24, no. 5, pp. 36–46, 2007.

[21] D. Gesbert, S. Hanly, H. Huang, S. Shamai, O. Simeone, and W. Yu, "Multi-cell MIMO cooperative networks: A new look at interference," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1380–1408, 2010.

[22] P. Marsch and G. Fettweis, "On multicell cooperative transmission in backhaul-constrained cellular systems," *Ann. Telecommun.*, vol. 63, no2. 5–6, pp. 253–269, 2008.

[23] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, 2010.

[24] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Processing Mag.*, vol. 30, no. 1, pp. 40–60, 2013.

[25] J. Hoydis, K. Hosseini, S. ten Brink, and M. Debbah, "Making smart use of excess antennas: Massive MIMO, small cells, and TDD," *Bell Labs Tech. J.*, vol. 18, no. 2, pp. 5–21, 2013.

[26] E. Björnson, J. Hoydis, M. Kountouris, and M. Debbah, "Massive MIMO systems with non-ideal hardware: Energy efficiency, estimation, and capacity limits," *IEEE Trans. Inform. Theory,* to be published.

[27] H. Yang and T. L. Marzetta, "Total energy efficiency of cellular large scale antenna system multiple access mobile networks," in *Proc. OnlineGreenComm*, 2013, pp. 27–32.

[28] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. A. Imran, D. Sabella, M. J. Gonzalez, O. Blume, and A. Fehske, "How much energy is needed to run a wireless network?," *IEEE Wireless Commun. Mag.*, vol. 18, no. 5, pp. 40–49, 2011.

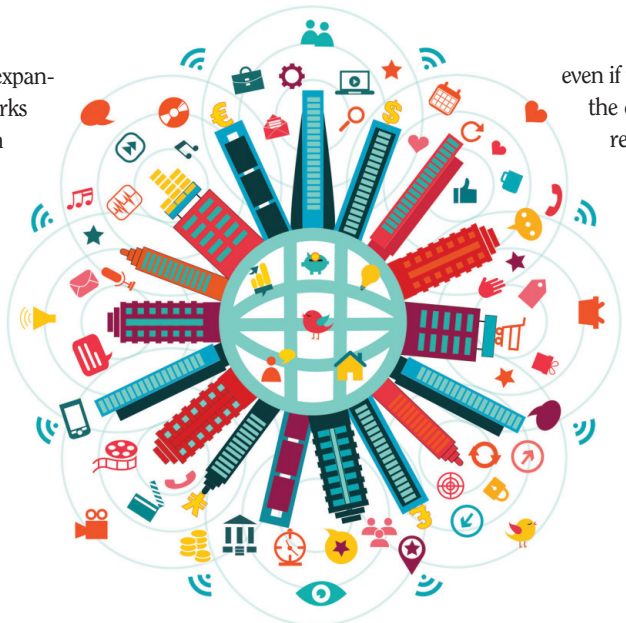
[SP]

Renato L.G. Cavalcante, Slawomir Stańczak, Martin Schubert,
Andreas Eisenblätter, and Ulrich Türke

Toward Energy-Efficient 5G Wireless Communications Technologies

Tools for decoupling the scaling of networks from the growth of operating power

The densification and expansion of wireless networks pose new challenges on energy efficiency. With a drastic increase of infrastructure nodes (e.g. ultradense deployment of small cells), the total energy consumption may easily exceed an acceptable level. While most studies focus on the energy radiated by the antennas, the bigger part of the total energy budget is actually consumed by the hardware (e.g., coolers and circuit energy consumption). The ability to shut down infrastructure nodes (or parts of it) or to adapt the transmission strategy according to the traffic will therefore become an important design aspect of energy-efficient wireless architectures. Network infrastructure should be regarded as a resource that can be occupied or released on demand, and the modeling and optimization of such systems are highly nontrivial problems. In particular, elements of the network infrastructure should be released by taking into account traffic forecasts to avoid losing the required coverage and capacity. However, even if traffic profiles were perfectly known, the determination of the elements to be released is complicated by the potential interference coupling between active elements and the sheer size of the optimization problems in dense networks.



THE 5G REVOLUTION

Digital Object Identifier 10.1109/MSP.2014.2335093
Date of publication: 15 October 2014

INTRODUCTION

Due to the compelling need for broadband mobile access to the Internet, there has been a dramatic growth in demand for wireless access worldwide over the past decade. This growth is expected to continue in the years to come, driven by an increasing interest in various wireless services and novel types of machine-to-machine (M2M) and device-to-device (D2D) communications. The vision is to create the so-called Internet of Things by integrating billions of sensors and actuators into physical objects and connecting them to the network via wireless connections. Requiring no human involvement, such communications may exceed any existing limits on information dissemination, leading to a data explosion of unprecedented magnitude.

This vision can only be brought to reality if we introduce major changes to the way current cellular networks are designed and operated. The need for such changes can be partially justified by the results in the landmark study by Gupta and Kumar [1]. This

1053-5888/14©2014IEEE

study strongly suggests that traditional large-scale networks (i.e., networks of spatially and temporarily independent sources, arbitrarily located source and destination nodes, and arbitrary traffic demands) inevitably face the problem of asymptotically vanishing per-user throughputs whenever the following restrictions hold: 1) nodes are stationary, 2) interference is treated as noise [2], and 3) the network operates without any underlying infrastructure such as the presence of base stations (i.e., the information needs to be carried from node to node in a multihop fashion). As an immediate consequence, to overcome this fundamental limit, we must drop at least one of the restrictions when designing wireless networks to provide additional dimensions for network optimization.

The first restriction of the study in [1] has been lifted in [3], where the authors analyze an infrastructureless network with mobile nodes and interference treated as noise. It has been shown that mobility can stabilize the throughput at the cost of delay. Further studies reveal and characterize a fundamental tradeoff between throughput and delay (for example, [4]). From these studies, we conclude that mobility may be an important ingredient in enhancing the performance of future wireless networks, but, due to strict delay constraints of many wireless applications, it cannot be the ultimate solution to the problem.

By dropping the second restriction, we can make nodes exploit, shape, or reject interference through advanced multiuser transmission and reception techniques. The studies in [5] and [6] have shown that the throughput in large-scale infrastructureless networks can be stabilized by resorting to cooperation and other multiuser communications strategies that are derived from information-theoretic results on broadcast, multiple-access, and relay channels. An additional performance-enhancing approach falling into the class of interference-shaping techniques is interference alignment [7]. However, these interference-mitigating methods alone cannot deliver the promised gains in practice because of, for example, the lack of channel state information and the lack of perfect synchronization in real systems. Therefore, a fixed network infrastructure, which is of vital importance to current networks, is also envisioned to play a crucial role in future systems.

In particular, networks with densely deployed infrastructure nodes are one of the main pillars in the current fifth-generation (5G) discussion to enhance the throughput of cellular networks at relatively low operational costs [8]. The vision is to have small and low-cost base stations to form small cells and to provide Internet access by using short-distance links [9]. This vision is partially motivated by the analysis in [1]; the study in [10] shows that the per-user throughput can be improved significantly even when interference is treated as noise, provided that the density of infrastructure nodes grows sufficiently fast with the number of users.

One of the main challenges that may limit the acceptable density of future networks is the high capital and operational costs. In particular, a large part of operational costs is directly related to the energy consumed for transmission and for operation of the network infrastructure [11]. We argue in this article that future 5G wireless communications technologies need to be energy efficient to reduce the total cost per transmitted bit, thereby providing cost-effective, affordable wireless bandwidth.

ENERGY EFFICIENCY IN WIRELESS NETWORKS

In the literature, there are different notions of energy efficiency, and selecting a suitable definition is a multifaceted problem with profound theoretical and practical implications.

In communication systems with finite energy constraints, it is natural to relate the energy efficiency to the amount of energy that we need to transmit a finite number of bits subject to a given error probability. This fact has led researchers to consider the capacity per unit of energy (or the capacity of finite-energy channels in bits), where the energy of the codewords is kept finite as the code length tends to infinity [12] (see also the discussion in [13, p. 15]). This notion, which measures the maximum rate per unit of energy (bit per second per Joule), is difficult to handle with the framework of classical multiuser information theory. Therefore, information-theoretic studies have typically considered the notion of energy per one bit (power divided by data rate), which is defined as the amount of energy that is required for reliable (i.e., asymptotically error-free) communication of one bit of information at some rate [14]. We emphasize that the two notions are not equivalent because, if the number of bits tends to infinity at some rate, and the energy per bit is fixed, then the total energy used for transmission tends to infinity [13]. In particular, the notion of energy per bit, which commonly only considers the energy radiated by an antenna, is often used to show that recent techniques such as advanced multiuser communication, massive multiple-input, multiple output (MIMO), interference alignment, and network coding are energy-efficient solutions. However, as Example 1 shows, when other sources of energy consumption (such as hardware and signal processing operations) are taken into account, then the energy savings provided by these solutions may not be so clear. For convenience, in the text that follows, we use the general term *radiated energy* to refer to the energy radiated by antennas, and we reserve the term *operating energy* to refer to the remaining sources of energy consumption.

EXAMPLE 1

The study in [15] investigates the impact of infrastructure nodes (base stations, access points, etc.) on the throughput scaling. Inspired by the operation of current practical systems, where the interference is treated as noise and the transmission is considered successful if a given signal-to-interference-plus-noise (SINR) ratio is attained, the authors propose a communication scheme for random networks (i.e., wireless devices placed randomly in an uniform and independent manner, whereas infrastructure nodes are placed arbitrarily in a predefined manner, independent of the placement of the wireless devices) that achieves the throughput scaling of $\Theta(m/n)$, provided that $m(n) \in \omega(\sqrt{(n/\log n)})$ and $m(n) \in O(n/\log n)$. Here, m denotes the number of base stations, and n is the

number of users. It can be shown [16, Deliverable D4.2] that, for this scheme and this scaling of base stations, by choosing $m(n) = (n/\log n)^b$ with $b \in (1/2), 1]$, the radiated energy per information bit diminishes to zero as the number of users tends to infinity. In contrast, the operating energy consumed per information bit $E_b(n)$ increases at the order of $E_b(n) \in \Theta(n(\log n/n)^b)$. ■

Example 1 reveals that, although the transmit energy per bit may vanish as the number of nodes increases in the considered scenario, the operating energy per transmitted information bit increases without bound [in the best case as $\Theta(\log n)$]. Therefore, in highly dense networks (m and n large), the energy consumed by hardware is dominant, so we draw the following important conclusion (Fact 1).

FACT 1

Advanced multiuser communication strategies such as cooperative (relaying) techniques, massive MIMO, interference alignment, and network coding are highly promising approaches to push the performance of wireless networks with respect to throughput, delay, and error probability to orders of magnitude beyond the performance limits of contemporary cellular networks. These technologies can decrease the transmit energy per bit, but alone, they do not reduce significantly the operating energy. On the contrary, they may lead to a significant increase of the operating energy. This is due to the increased number of antennas (and the accompanying hardware), and the energy-expensive and time-consuming signal processing algorithms.

One of the major design principles to reduce the consumption of the radiated energy and the operating energy is to adjust the capacity of the network to the demand. In particular, in delay-sensitive applications, one option is to devise load-dependent algorithms that deactivate network elements in a coordinated manner to provide the desired coverage and throughput performance at any given point in time. The energy-saving capabilities of these algorithms is limited by a fundamental tradeoff between energy efficiency and delay constraints, which, to a large extent, remains an open problem [17]. We can take this idea one step further and consider that the algorithms are also able to choose the most suitable multiuser communication strategy (for example, network coding, MIMO technique, etc.) for the active network elements. Signal processing tools that can be used to develop and to support these load-adaptive schemes are the topic of the remaining sections.

ONE OF THE MAJOR DESIGNS PRINCIPLES TO REDUCE THE CONSUMPTION OF THE RADIATED ENERGY AND THE OPERATING ENERGY IS TO ADJUST THE CAPACITY OF THE NETWORK TO THE DEMAND.

TRAFFIC PATTERNS

In current communication networks, traffic typically follows a roughly periodic pattern; traffic is high during the day and low at night, with some local variations depending on whether we consider industrial, commercial, residential, or rural areas. These spatial and temporal fluctuations create opportunities to save energy by switching off unnecessary network elements. To exploit this approach to its fullest potential, we need reliable forecasts of local traffic, a task that calls for machine-learning algorithms and statistical tools.

In general, the prediction power of learning algorithms improves as we increase the amount of available prior information, and one of the most natural assumptions to use in traffic forecasts is the rough periodicity of the time series. However, especially in future networks, assumptions of this type have to be used carefully for two main reasons. First, in current networks, much of the traffic is generated by users, so the patterns are correlated to those of the human activity. For instance, users are more likely to watch streaming videos during the day or early evening than very late at night. In contrast, in future networks, it is envisioned that M2M and D2D communications will be ubiquitous and, for the devices requesting such services, the time of the day when data is sent may be irrelevant or it may be even desirable to send data when other users are less active (for example, noncritical software updates or backups could be performed while users are sleeping to avoid unnecessary congestion in the network). Second, current studies showing the coarse periodicity of traffic typically consider traffic aggregated over regions of the network, but at a local scale (individual cells of the network) the patterns many not necessarily follow the global trend too closely. This last observation has already been noticed in historical data coming from a real network of a large European city [16, Deliverable D6.2], where the time series of key performance indicators (KPIs) related to data traffic are less regular than those related to voice calls. In Figure 1, we show synthetic voice and data traffic with similar statistical properties to those found in a typical cell of the aforementioned real network. Note that, when data traffic is considered, bursts of traffic are frequently observed in periods when voice traffic is predictably low. The practical implication of these observations is that, especially at a local scale, forecasting algorithms should take into account the type of the service. The clear periodicity of current traffic patterns may not be necessarily present in future systems. In particular, prediction algorithms for data traffic should use statistically robust methods because of the irregularity of the time series, whereas prediction algorithms for voice traffic may be able to safely assume the coarse periodicity observed in Figure 1(a).

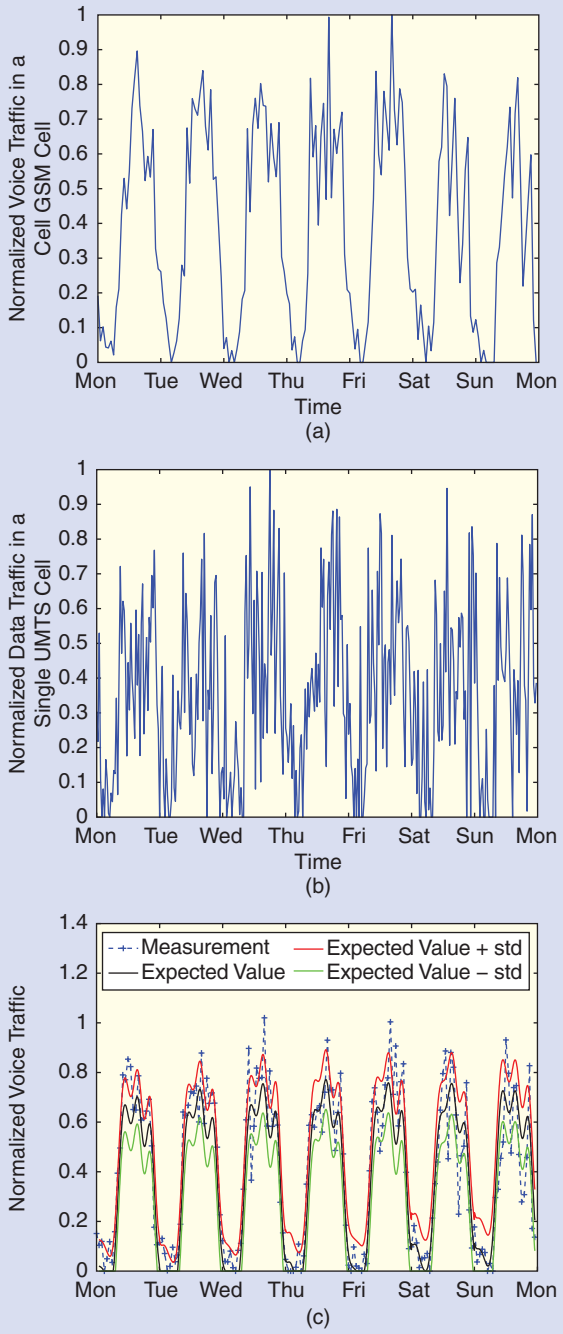
To produce forecasts of time series strongly related to voice calls, we can use approaches similar to those used for electrical load forecasting because electrical load in power grids and voice calls in cellular networks present similar periodicity. This knowledge and other contextual information, such as the presence of holidays and major events, can be easily incorporated into algorithms based on Gaussian processes (GPs) [18] and other kernel-based methods [16,

Deliverable D4.1]. The main idea is to choose a covariance or kernel function based on the observed features of the time series under consideration. In particular, the framework of GPs is a promising approach that enables us to specify those features merely in general terms, and confidence intervals for the predictions are readily available. For example, we assume that the time series in Figure 1(a) has a clear periodic pattern, so we can specify a periodic covariance function and leave the period as a free parameter (hyperparameter) to be selected by maximizing a merit function with a very natural probabilistic interpretation [18, Ch. 5] (for example, the marginal log-likelihood). This example is not particularly interesting because the period can be easily obtained by simply looking at the data, but the approach can be straightforwardly applied to capture more subtle features such as small variations of traffic according to the day of the week. To this end, we can design "atomic" covariance functions, each of which is responsible for capturing one desired feature of the time series. Then we construct the final covariance function by combining the atomic functions with operations that preserve covariance functions, and we compute the hyperparameters by maximizing the marginal log-likelihood. In Figure 1(c), we show results for the prediction of voice traffic obtained by following the general guidelines for the selection of the covariance function outlined in [18, Ch. 5.4.3]. For this result, we use synthetic data, but the same algorithm provides good forecasts for most cells of a real network. However, if the time series contains too many bursts of traffic, such as that depicted in Figure 1(b), the traditional framework of GPs may have bad generalization properties. In such cases, robust statistical tools are required, and we now review possible approaches.

It has been observed that, in the short/medium term (up to a couple of months), samples of traffic-related KPIs for either working days or holidays can be assumed to come from independent and identically distributed (i.i.d.) random variables for most cells of a real network [16, Deliverable D6.2] if they are spaced by multiples of 24 hours (an indication of the validity of this assumption can be obtained with the turning point test [19]). As a result, we can use simple robust tools based on order statistics, such as tolerance intervals [19], to obtain knowledge about upper bounds for traffic at any given hour of the day. More precisely, let $X_{1:n} \leq \ldots \leq X_{n:n}$ be the sorted values of the i.i.d. random variables $X_1, \ldots, X_n$ corresponding to traffic measurements $x_1, \ldots, x_n$ for a given hour of the day. Denote the cumulative distribution function of the random variables by $F(x)$ and the inverse cumulative distribution function by $F^{-1}(p) = \sup\{x \mid F(x) \leq p\}$, where $p \in]0,1[$. Then, for a fixed quantile p, the following holds [20]:

$$P(F(X_{k:n}) > p) = 1 - \sum_{i=k}^{n} \binom{n}{i} p^i \, (1-p)^{n-i},$$

which is an exact value that does not depend on the distribution of the random variables. If this probability is sufficiently low and p sufficiently high, we can configure the cell or base station to serve, at most, traffic value $X_{k:n}$. Similar, strikingly simple, and exact results exist to other types of intervals (for example, prediction intervals), and good agreement between the theoretical



[FIG1] Normalized traffic as a function of time. (a) Synthetic voice traffic in a single cell of a global system for mobile communications (GSM) network. (b) Synthetic data traffic in a single universal mobile telecommunications system (UMTS) cell. (c) Forecast for voice traffic obtained with the framework of Gaussian processes ("std" stands for standard deviation)—three weeks of data are used for training.

and empirical results has been obtained with real network data [16, Deliverable D6.2].

A major limitation of the aforementioned robust approaches is that they are unable to detect trends and to capture correlations between samples from consecutive hours of the day (samples from different hours are analyzed independently). Correlations and trends can be captured by robust machine learning tools such as those described in [21]. To improve further the estimates, we can also try to exploit temporal and spatial correlations among cells. Extensions of this type should consider carefully the computational complexity because of the large number of network elements in future networks. Unfortunately, obtaining data sets for research from real networks is difficult, a fact that has limited the literature on this important topic.

NETWORK INTERFERENCE CALCULUS

Even with good traffic forecasts, identifying the best action to save energy still remains a difficult problem because of the possible interference coupling among active network elements. Therefore, it comes as no surprise that energy-saving algorithms also need at least a rough estimate of the interference patterns; we now turn the attention to some basic results on interference calculus [22], [23], a general mathematical framework that unifies many interference models in wireless systems. The presentation is heavily based on the study in [22], which shows algorithms for power control in code division multiple access systems. From a mathematical perspective, these algorithms solve general fixed-point problems, so they have been used in applications different from that originally envisioned. We expect these algorithms based on interference calculus to play an important role in the analysis of future systems owing to the generality of the framework.

STANDARD INTERFERENCE FUNCTIONS

In the following discussion, $\mathbb{R}_+$ denotes the set of nonnegative reals, $\mathbb{R}_{++}$ is the set of strictly positive reals, and $\mathbf{1}$ is a vector of ones. Inequalities involving vectors should be understood as element-wise inequalities. Interference functions are defined as follows (for convenience, our definitions are slightly different from that originally stated in [22]).

DEFINITION 1

A function $I: \mathbb{R}_+^M \to \mathbb{R}_{++}$ is said to be a standard interference function if the following axioms hold:

1) (Scalability) $\alpha I(\boldsymbol{x}) > I(\alpha \boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}_+^M$ and all $\alpha > 1$.
2) (Monotonicity) $I(\boldsymbol{x}_1) \geq I(\boldsymbol{x}_2)$ if $\boldsymbol{x}_1 \geq \boldsymbol{x}_2$.

Given M standard interference functions $I_i: \mathbb{R}_+^M \to \mathbb{R}_{++}$, $i = 1, \ldots, M$, we call the mapping $\mathcal{J}: \mathbb{R}_+^M \to \mathbb{R}_{++}^M$ with $\mathcal{J}(\boldsymbol{x}) := [I_1(\boldsymbol{x}), \ldots, I_M(\boldsymbol{x})]^T$ a *standard interference mapping* or simply *interference mapping*.

Checking whether a given function is a standard interference function by using the definition is not necessarily easy. Fortunately, the following proposition shows that a large class of functions frequently used to model wireless systems are standard interference functions. We note that the simple result shown below has been explicitly mentioned in the nonpublic report [16, D52] (see also [24] and references therein).

PROPOSITION 1

Concave functions $I: \mathbb{R}_+^M \to \mathbb{R}_{++}$ are standard interference functions.

The usefulness of Proposition 1 lies in the fact that there exist many simple and well-known techniques to identify concave and convex functions [25]. The converse of Proposition 1 does not, in general, hold. However, we can expand the class of functions that can be easily identified with the help of Proposition 1 by using the following operations that preserve interference functions/mappings.

FACT 2

Operations that preserve standard interference functions/mappings are as follows [22], [23]:

1) Standard interference functions are closed under finite addition and multiplication by strictly positive constants. For instance, if $I_1: \mathbb{R}_+^M \to \mathbb{R}_{++}$ and $I_2: \mathbb{R}_+^M \to \mathbb{R}_{++}$ are standard interference functions, then $I'(\boldsymbol{x}) = \alpha_1 I_1(\boldsymbol{x}) + \alpha_2 I_2(\boldsymbol{x})$ for $\alpha_1, \alpha_2 > 0$ is a standard interference function.
2) If $I_i: \mathbb{R}_+^M \to \mathbb{R}_{++}$, $(i \in \{1, \ldots, N\}$, and $N \in \mathbb{N})$ are standard interference functions, then $I'(\boldsymbol{x}) := \min_{i \in \{1,\ldots,N\}} I_i(\boldsymbol{x})$ and $I''(\boldsymbol{x}) := \max_{i \in \{1,\ldots,N\}} I_i(\boldsymbol{x})$ are standard interference functions.
3) Standard interference mappings are closed under finite composition. For example, if $\mathcal{J}_1: \mathbb{R}_+^M \to \mathbb{R}_{++}^M$ and $\mathcal{J}_2: \mathbb{R}_+^M \to \mathbb{R}_{++}^M$ are standard interference mappings, then $\mathcal{J}'(\boldsymbol{x}) = \mathcal{J}_1(\mathcal{J}_2(\boldsymbol{x}))$ is a standard interference mapping.

We are mostly interested in studying fixed points of standard interference mappings, and the following result can be used for this purpose. When interference calculus is used to investigate the performance of communication systems, the fixed points describe, for example, the load or interference experienced by network elements.

FACT 3

Selected properties of standard interference mappings [22]: Let $I_i: \mathbb{R}_+^M \to \mathbb{R}_{++}$ be a standard interference function for every $i \in \{1, \ldots, M\}$, and consider the corresponding mapping $\mathcal{J}: \mathbb{R}_+^M \to \mathbb{R}_{++}^M$ given by $\mathcal{J}(\boldsymbol{x}) := [I_1(\boldsymbol{x}), \ldots, I_M(\boldsymbol{x})]^T$. Then the following holds:

1) If the mapping $\mathcal{J}$ has a fixed point (i.e., $\emptyset \neq \mathrm{Fix}(\mathcal{J}) := \{\boldsymbol{x} \in \mathbb{R}_+^M \mid \boldsymbol{x} = \mathcal{J}(\boldsymbol{x})\}$), then the fixed point is unique.
2) The mapping $\mathcal{J}$ has a fixed point if and only if there exists $\boldsymbol{x}' \in \mathbb{R}_{++}^M$ satisfying $\mathcal{J}(\boldsymbol{x}') \leq \boldsymbol{x}'$.
3) If $\mathcal{J}$ has a fixed point, then the sequence $\{\boldsymbol{x}_n\}_{n \in \mathbb{N}}$ generated by $\boldsymbol{x}_{n+1} = \mathcal{J}(\boldsymbol{x}_n)$ satisfies the following:
   - For an arbitrary vector $\boldsymbol{x}_0 \in \mathbb{R}_+^M$, the sequence $\{\boldsymbol{x}_n\}_{n \in \mathbb{N}}$ converges to the fixed point $\boldsymbol{x}^* \in \mathrm{Fix}(\mathcal{J})$.
   - If $\boldsymbol{x}_0 = \mathbf{0}$, then the sequence $\{\boldsymbol{x}_n\}_{n \in \mathbb{N}}$ is monotonously increasing; i.e., $\boldsymbol{x}_{n+1} \geq \boldsymbol{x}_n$.
   - If $\mathcal{J}(\boldsymbol{x}_0) \leq \boldsymbol{x}_0$, then the sequence $\{\boldsymbol{x}_n\}_{n \in \mathbb{N}}$ is monotonously decreasing; i.e., $\boldsymbol{x}_{n+1} \leq \boldsymbol{x}_n$.

Remark 1

Suppose that, in addition to being an interference mapping, $\mathcal{J}:\mathbb{R}_+^M \to \mathbb{R}_{++}^M$ is also upper bounded; i.e., $\mathcal{J}(\boldsymbol{x}) \leq B\,\mathbf{1}$ for some fixed $B \in \mathbb{R}$ and every $\boldsymbol{x} \in \mathbb{R}_+^M$. In this case, Facts 3.1 and 3.2 guarantee the existence of a unique fixed point $\boldsymbol{x}^*$, and Fact 3.3 suggests the following iterative procedure to compute the fixed point. Produce in parallel two sequences $\{\boldsymbol{x}'_{n+1} = \mathcal{J}(\boldsymbol{x}'_n)\}$ and $\{\boldsymbol{x}''_{n+1} = \mathcal{J}(\boldsymbol{x}''_n)\}$, where $\boldsymbol{x}'_0 = \mathbf{0}$ and $\boldsymbol{x}''_0 = B\,\mathbf{1}$. Fact 3 shows that $\boldsymbol{x}'_n \leq \boldsymbol{x}^* \leq \boldsymbol{x}''_n$ for every $n \in \mathbb{N}$ and that both sequences $\{\boldsymbol{x}'_n\}$ and $\{\boldsymbol{x}''_n\}$ converge to $\boldsymbol{x}^*$. Furthermore, $\boldsymbol{x}'_{n+1} \geq \boldsymbol{x}'_n$ and $\boldsymbol{x}''_{n+1} \leq \boldsymbol{x}''_n$. We can therefore stop the algorithm at iteration n whenever $\|\boldsymbol{x}'_n - \boldsymbol{x}'_n\|_\infty \leq \epsilon$ is satisfied, where $\epsilon > 0$ is the desired precision. The algorithm is guaranteed to terminate with a finite number of iterations. In other words, $\boldsymbol{x}'_n$ and $\boldsymbol{x}''_n$ can serve as lower and upper bounds, respectively, for $\boldsymbol{x}^*$.

We now relate these results to load estimation in multicarrier systems with fast link adaptation (as envisioned in 5G systems). The couplings of the models shown next have been originally studied on a case-by-case basis, but recently their connection to interference calculus has been independently established in [16, Deliverable D5.2] and [24].

LOAD ESTIMATION IN WIRELESS NETWORKS

The model described below is based on the discussion in [26], and this or similar models have been used for various network optimization tasks for many years. Later, we use this model to gain insight onto the challenges associated with the formulation of energy-saving optimization problems.

In more detail, we focus on a cellular radio network with M base stations (or cells) and N test points (an abstract concept to represent demand of users in a given region). We denote the set of base stations by $\mathcal{M} = \{1, 2, \ldots, M\}$ and the set of test points by $\mathcal{N} = \{1, 2, \ldots, N\}$. The quality-of-service (QoS) requirement of each test point $j \in \mathcal{N}$ is represented by a minimum amount of data $d_j \in \mathbb{R}_{++}$ that needs to be sent during a unit of time. We denote by $X \in \{0,1\}^{M \times N}$ the assignment matrix; the component of its ith row and jth column takes the value $x_{i,j} = 1$ if test point j is assigned to base station i or the value $x_{i,j} = 0$ otherwise (we also assume that each base station serves at least one test point). The power gain between base station i and test point j is denoted by $g_{i,j} \in \mathbb{R}_+$. Each base station $i \in \mathcal{M}$ transmits with fixed power spectral density per minimum resource unit (for example, resource blocks in multicarrier systems, time slots in time division multiple access systems, etc.) in scheduling, which we denote by $P_i \in \mathbb{R}_{++}$. The load vector of the base stations is given by $\boldsymbol{\rho} := [\rho_1, \ldots, \rho_M]^T \in \mathbb{R}_+^M$, where $\rho_i \in \mathbb{R}_+$ is the load at base station i. The load ρ_i is defined as the ratio between the number of resource units requested by test points served by base station $i \in \mathcal{M}$ and the number K of resource units available in the system. We can obtain an estimate of $\boldsymbol{\rho}$ by solving the following system of nonlinear equations [26]:

$$\rho_1 = I_1(\boldsymbol{\rho}), \quad \ldots, \quad \rho_M = I_M(\boldsymbol{\rho}), \tag{1}$$

where

$$I_1(\boldsymbol{\rho}) := \sum_{j \in N} \frac{d_j x_{i,j}}{K \omega_{i,j}(\boldsymbol{\rho})}, \tag{2}$$

σ^2 is the noise power, and

$$\omega_{i,j}(\boldsymbol{\rho}) := B \log_2\left(1 + \frac{P_i\, g_{i,j}}{\eta\left(\sum_{l \in \mathcal{M}\setminus\{i\}} P_l\, g_{l,j}\, \rho_l + \sigma^2\right)}\right) \tag{3}$$

is the the spectral efficiency (i.e., the effective bit rate per resource unit) of the link connecting base station i to test point j. In (3), $\omega_{i,j}$ depends on the effective bandwith per resource unit B and the SINR η, which are parameters that are typically fitted from simulations of measurements of the actual spectral efficiency $\omega_{i,j}$ as a function of $\boldsymbol{\rho}$ for a particular network configuration (choice of schedulers, MIMO transmission scheme, etc.); see [27] for additional details. Intuitively, each term in the sum in (2) is the fraction of resource units, relative to the total number of resource units K, that the test point j requests from base station i if data rate d_j is desired.

A positive vector $\boldsymbol{\rho}^* = [\rho_1^*, \ldots, \rho_M^*]^T$ is a solution of the system of nonlinear equations in (1) if and only if $\boldsymbol{\rho}^*$ is a fixed point of $\mathcal{J}(\boldsymbol{\rho}) := [I_1(\boldsymbol{\rho}), \ldots, I_M(\boldsymbol{\rho})]^T$ [i.e., $\boldsymbol{\rho}^* \in \mathrm{Fix}(\mathcal{J})$]. Once a solution is obtained (assuming that it exists and is unique), we can verify whether the network configuration can support the traffic demand by checking whether $\rho_i^* \leq 1$ for every $i \in \{1, \ldots, M\}$; i.e., base stations do not use more resource units than available, in which case we say that the network configuration is feasible. As a result, answering questions regarding uniqueness and existence of a fixed point of $\mathcal{J}$ (and also iterative methods to compute fixed points) is crucial to verify the feasibility of networks. To this end, we can use the result in Proposition 1 and Fact 3. More precisely, we note that $I_i(\boldsymbol{\rho})$ is a standard interference function because it is positive and concave (see Proposition 1). Concavity of I_i follows from simple facts [25]: 1) $f(x) := 1/\log_2(1 + 1/x)$ is a concave function on the domain $\mathbb{R}_{++}$, 2) composition of concave functions with affine transformations preserves concavity, and 3) the set of concave functions is closed under addition and multiplication by strictly positive real numbers. (See [26] for an alternative argument.) By identifying $\mathcal{J}$ as a standard interference mapping, we can now use the known results in Fact 3 to compute fixed points, answer questions regarding existence and uniqueness of fixed points, etc. Note that some previous studies have not identified the functions I_i as standard interference functions. The advantage of working with the framework of interference calculus (in addition to showing that many existing results are direct consequences of Fact 3) becomes clear when we extend the interference coupling model in (1) by using simple operations that preserve interference mappings. By doing so, we can analyze more complex communication systems than those described previously (see also [24] and [16, Deliverable D5.2] for more details on the examples).

EXAMPLE 2

Assume that the spectral efficiency $\omega_{i,j}$ in (3) is upper bounded by $\bar{\omega}$, which is a natural assumption in real systems. In such cases, we can evaluate the feasibility of a network by computing

the fixed point of the mapping $\mathcal{J}'(\boldsymbol{\rho}) := [I_1'(\boldsymbol{\rho}), \ldots, I_M'(\boldsymbol{\rho})]$, where $I_i'(\boldsymbol{\rho}) := \sum_{j \in N} \max\{(d_j\, x_{i,j}/K\omega_{i,j}(\boldsymbol{\rho})), (d_j\, x_{i,j}/K\bar{\omega})\}$. Note that we can easily verify that I_i' is a standard interference function by using Fact 2. ■

EXAMPLE 3

Suppose that the system considered in Example 2 is not feasible; i.e., the interference mapping $\mathcal{J}'$ either does not have a fixed point or, if a fixed point exists, some of its components have value strictly larger than one (in which case at least one of the base stations require more resource units than available in the system). In such cases, the system may still work with overloaded base stations dropping users (i.e., only a fraction of the traffic of the overloaded base stations is served), but the fixed point of $\mathcal{J}'$ is not useful to indicate the load in non-overloaded base stations. To capture this feature of real systems, we can impose limits on the maximum possible load by using the interference mapping $\mathcal{J}''(\boldsymbol{\rho}) := [I_1''(\boldsymbol{\rho}), \ldots, I_M''(\boldsymbol{\rho})]$, where $I_i''(\boldsymbol{\rho}) := \min\{I_i'(\boldsymbol{\rho}), 1\}$. We have already seen that each function I_i' is a standard interference function, so Fact 2 shows that $\mathcal{J}''$ is also a standard interference function. By noticing that $\mathcal{J}''(\boldsymbol{\rho}) \leq \mathbf{1}$ for every $\boldsymbol{\rho} \in \mathbb{R}_+^M$, Fact 3.2 shows that $\mathcal{J}''$ is guaranteed to have a fixed point, which can be computed with the scheme with the nonheuristic stopping criterion in Remark 1. ■

The aforementioned interference coupling models are also useful to highlight limitations of interference calculus, which should be addressed in future extensions of the framework. As discussed in the "Introduction" section, future systems will be composed of combinations of many energy-efficient transmission schemes. In particular, when network elements are cooperative or apply interference-exploiting methods, extending the above models while remaining under the framework of interference calculus is difficult. For example, although the model in (3) can be adjusted to account for some limited advanced communication strategies (the presence of intelligent schedulers, MIMO transmitters, etc.), it is an approximation that can be too crude in future systems where cooperation among network elements will be taken to completely new levels. We refer the reader to [23] for recent advances in the field.

ADVANCED COMMUNICATION TECHNIQUES, SUCH AS MASSIVE MIMO AND INTERFERENCE ALIGNMENT, CAN DECREASE THE TRANSMIT ENERGY PER BIT, BUT ALONE, THEY MAY GREATLY INCREASE THE TOTAL ENERGY CONSUMPTION OF THE NETWORK.

ALGORITHMS FOR ENERGY SAVINGS

We now turn our attention to algorithms that have the objective of switching off as many network elements as possible while satisfying constraints such as coverage and data rate requirements. These algorithms typically use as an input the traffic forecasts and interference models described in previous sections. To avoid notational clutter, we assume, for the moment, that all network elements consume the same amount of energy and that the radiated energy is negligible. In particular, the latter assumption is an acceptable approximation in current cellular systems [11]. All assumptions are later dropped to take into account heterogeneous systems with hardware more energy efficient than that available today.

We associate network elements of a communication system with a vector $\boldsymbol{x} = [x_1, \ldots, x_M]^T \in \{0,1\}^M$, where M is the number of network elements and each variable $x_i \in \{0,1\}$ takes the value one if network element i is active, or the value zero otherwise. By $\mathcal{X}$, we denote the set of configurations $\boldsymbol{x} \in \{0,1\}^M$ satisfying some required constraints (for example, capacity constraints). The mathematical problem that energy-saving algorithms try to solve can be typically described as a variation of the following combinatorial problem:

$$\begin{aligned} &\text{minimize} && \|\boldsymbol{x}\|_0 && (4)\\ &\text{subject to} && \boldsymbol{x} \in \mathcal{X}, \\ & && \boldsymbol{x} \in \{0,1\}^M, && (5)\end{aligned}$$

where $\|\boldsymbol{x}\|_0$ is the function, informally called l_0-norm ($\|\cdot\|_0$ does not satisfy all axioms of norms), which returns the number of nonzero elements of the vector $\boldsymbol{x}$. The above energy problems are typically NP-hard, so we cannot expect to solve them both fast and optimally, especially when considering the densification of future networks, which will lead to problems of huge dimensions. If optimality is desired, we can use a branch and bound algorithm, but solving even fairly small problems often takes a very long time [28]. As a result, with the densification of the networks, recent studies [29]–[32] have focused on fast but suboptimal heuristics, and many of them [28]–[30], [32] build upon theoretically sound methods that aim at solving discrete optimization problems by using the l_1-norm as a proxy of the l_0-norm. The reason for this approximation is that the l_1-norm is a (convex) sparsity promoting norm [25], [28], [33], and convex problems are typically easier to solve than nonconvex problems. (We can also interpret the l_1-norm as the convex envelope of the l_0-norm for an appropriate domain [25].) By also relaxing nonconvex constraints to convex constraints, many efficient optimization techniques become available [25], [32]. In particular, we often replace the discrete constraint in (5) by

$$\boldsymbol{x} \in [0,1]^M. \quad (6)$$

The solution of the resulting convex optimization problem (assuming that $\mathcal{X}$ is also a convex set) can then guide other heuristics to make the final hard decisions on the active set of network elements. This approach is used in, for example, the study in [32], which considers an antenna selection problem in energy limited point-to-point communication systems with constraints given in terms of the required channel capacity.

In the compressive sensing community, problems with the l_0-norm in the objective are increasingly being solved with a method called *reweighted* l_1*-norm* [33], which, in recent years, is finding applications in cellular communication systems [29], [30]. We first review the majorization-minimization (MM) algorithm to then explain these ideas.

DENSE AND ULTRADENSE DEPLOYMENTS OF FUTURE 5G NETWORKS WILL HARDLY BECOME REALITY UNLESS THE SCALING OF BASE STATIONS CAN BE DECOUPLED FROM THE GROWTH OF OPERATING POWER.

THE MM ALGORITHM

The discussion here follows closely that of the studies in [33] and [34]. Suppose that the objective is to minimize a function $f:\mathcal{X}\to\mathbb{R}$, where $\mathcal{X}\subset\mathbb{R}^M$. Unless the optimization problem has a very special structure that can be exploited, such as convexity, finding an optimal solution $\boldsymbol{x}^*$ (provided that one exists) is intractable in general. To devise a suboptimal approach, assume that we are able to construct a function $g:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$, hereafter called a *majorizing function*, satisfying the following properties:

$$f(\boldsymbol{x})\le g(\boldsymbol{x},\boldsymbol{y}),\quad \forall \boldsymbol{x},\boldsymbol{y}\in\mathcal{X} \tag{7}$$

and

$$f(\boldsymbol{x})=g(\boldsymbol{x},\boldsymbol{x}),\quad \forall \boldsymbol{x}\in\mathcal{X}. \tag{8}$$

Then, starting from $\boldsymbol{x}_0\in\mathcal{X}$, the MM algorithm produces a sequence $\{\boldsymbol{x}_n\}\subset\mathcal{X}$ by

$$\boldsymbol{x}_{n+1}\in\arg\min_{\boldsymbol{x}\in\mathcal{X}} g(\boldsymbol{x},\boldsymbol{x}_n). \tag{9}$$

We can verify that the sequence $\{f(\boldsymbol{x}_n)\}$ is monotonously decreasing: $f(\boldsymbol{x}_{n+1})=g(\boldsymbol{x}_{n+1},\boldsymbol{x}_{n+1})\le g(\boldsymbol{x}_{n+1},\boldsymbol{x}_n)\le g(\boldsymbol{x}_n,\boldsymbol{x}_n)=f(\boldsymbol{x}_n)$, where the equalities follow from (8), and the first and second inequalities follow from (7) and (9), respectively. As a result, $f(\boldsymbol{x}_n)\to c\in\mathbb{R}$ for some $c\ge f(\boldsymbol{x}^*)$ as $n\to\infty$, where we assume that $\boldsymbol{x}^*$ is a solution of the original optimization problem. In practice, we stop the algorithm when we observe no progress in the objective value, and we note that convergence of $\{f(\boldsymbol{x}_n)\}$ does not, in general, imply the convergence of the sequence $\{\boldsymbol{x}_n\}$.

The main challenge in applying the MM algorithm is to find a majorizing function g such that the iteration in (9) can be implemented efficiently. Fortunately, in some special cases of practical interest, we can construct a majorizing function easily. For example, if f can be decomposed as $f(\boldsymbol{x})=f_1(\boldsymbol{x})+f_2(\boldsymbol{x})$, where f_1 is a differentiable concave function and f_2 is a convex function, then we can use

$$g(\boldsymbol{x},\boldsymbol{y})=f_1(\boldsymbol{y})+\nabla f_1(\boldsymbol{y})^T(\boldsymbol{x}-\boldsymbol{y})+f_2(\boldsymbol{x}) \tag{10}$$

as the majorizing function, where $\nabla f_1(\boldsymbol{y})$ is the gradient of f_1 at $\boldsymbol{y}$ (if f_1 is not differentiable, we can replace the gradient by an arbitrary subgradient). With this choice, the optimization problem in (9) becomes a convex optimization problem provided that the set $\mathcal{X}$ is convex, hence it can be solved with efficient methods [25], [32]. We can verify the validity of the property in (7) from the first-order characterization of concave functions [25].

Let us turn our attention to the optimization problem in (4) and (5) with the constraint in (5) replaced by that in (6). It is well known that the l_0-norm satisfies the following [33], [34]:

$$\|\boldsymbol{x}\|_0=\lim_{\epsilon\to 0}\sum_{i=1}^{M}\frac{\log(\epsilon+|x_i|)-\log(\epsilon)}{\log(1+\epsilon^{-1})}, \tag{11}$$

which, if each component x_i of the vector $\boldsymbol{x}=[x_1,\ldots,x_M]^T$ is constrained to be nonnegative as in (6), suggests the use of the function $f_\epsilon(\boldsymbol{x})=\sum_{i=1}^{M}(\log(\epsilon+x_i)-\log(\epsilon))/\log(1+\epsilon^{-1})$ with a small design parameter $\epsilon>0$ as a approximation of the l_0-norm (other choices are possible [33]). Furthermore, note that the function $f_\epsilon(\boldsymbol{x})$ is concave, so the MM algorithm can be used to generate a sequence of vectors $\{\boldsymbol{x}_n\}$ with decreasing objective value as discussed earlier.

ENERGY-SAVING OPTIMIZATION PROBLEMS IN CELLULAR NETWORKS

We now exemplify how to apply the aforementioned ideas to energy saving problems in cellular networks. (These problems also highlight the need to study interference coupling in energy efficient networks.) To this end, consider the notation introduced in the previous section on interference calculus, and assume that, for each test point $j\in\mathcal{N}$, we have an estimate of the traffic demand per unit of time, which, as before, we denote by $d_j>0$. This estimate can be obtained with learning algorithms. We assume that the operating energy of each base station (which here can also designate a cell, a radio unit, etc.) $i\in\mathcal{M}$ is given by $c_i>0$, and the energy related to radiation is approximated by a function $f_i:[0,1]\to\mathbb{R}_+$ of the load $\rho_i\in[0,1]$. To detect base stations that can be switched off, we may try to solve the following problem:

$$\min.\ \sum_{i=1}^{M}c_i|\rho_i|_0+\sum_{i=1}^{M}f_i(\rho_i) \tag{12}$$

$$\text{s.t.}\quad \rho_i=\sum_{j=1}^{N}\frac{d_j}{K\omega_{i,j}(\boldsymbol{\rho})}x_{i,j},\qquad i\in\mathcal{M} \tag{13}$$

$$\sum_{i=1}^{M}x_{i,j}=1\qquad j\in\mathcal{N}$$

$$x_{i,j}\in\{0,1\}\qquad i\in\mathcal{M},\ j\in\mathcal{N} \tag{14}$$

$$\rho_i\in[0,1]\qquad i\in\mathcal{M}, \tag{15}$$

where $\{x_{i,j}\}_{i\in\mathcal{M},j\in\mathcal{N}}$ (assignment variables) and $\{\rho_i\}_{i\in\mathcal{M}}$ (load at base stations) are the optimization variables, and $|x|_0$ is the function that takes the value 0 if $x=0$ or the value 1 otherwise.

By solving (12)–(15), base stations with no load (i.e., those with $\rho_i = 0$) can be deactivated. In the problem, we assume that the constraints are feasible. If not, we can add slack variables and an additional l_0 or l_1 penalty norm to the merit function, but, for brevity, we do not consider such extensions here. One of the main complications of the problem in (12)–(15) is the interference coupling appearing in the nonconvex constraint in (13). We have already seen in previous sections that computing interference or load, even with fixed network configurations, is a nontrivial task. Therefore, it is common in the literature to fix the value of the spectral efficiency of the link connecting base station i to test point j to some constant $\tilde{\omega}_{i,j}$, which is obtained by considering the worst-case interference or an average case. Unfortunately, even with this simplification, the resulting problem is difficult to solve because it is a generalization of the standard bin-packing problem, which is NP-hard. To obtain a fast heuristic, we can use the MM algorithm as follows. First, as discussed in the previous section, we relax the discrete constraint in (14) to consider continuous values $x_{i,j} \in [0,1]$. (In particular, this modification is natural when base stations are allowed to serve only a fraction of the traffic requested at test points.) Then, by also using the approximation discussed below (11), we obtain the following optimization problem, which can be efficiently addressed with the MM algorithm if $f_i(\rho_i)$ is a convex or concave function [see the discussion above (10)], and we note that a linear function is often considered a good approximation of the energy related to radiation [11]:

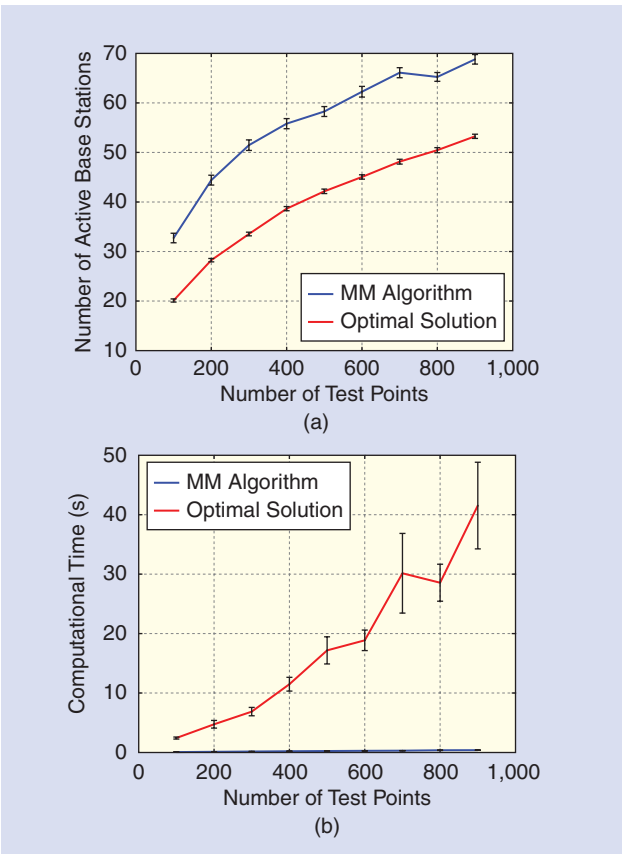
$$\min . \sum_{i=1}^{M} c_i \frac{\log(\epsilon + \rho_i) - \log(\epsilon)}{\log(1+\epsilon^{-1})} + \sum_{i=1}^{M} f_i(\rho_i) \tag{16}$$

$$\text{s.t.} \quad 0 \le \rho_i = \sum_{j=1}^{N} \frac{d_j}{K\tilde{\omega}_{i,j}} x_{i,j} \le 1, \quad i \in \mathcal{M} \tag{17}$$

$$\sum_{i=1}^{M} x_{i,j} = 1, \quad j \in \mathcal{N}$$

$$x_{i,j} \in [0,1] \quad i \in \mathcal{M}, j \in \mathcal{N}. \tag{18}$$

A solution of this modified problem provides a good indication of which base stations to switch off (ρ_i close to zero), a result that can guide other heuristics to make the final hard decisions. Empirical evidence suggests that these algorithms based on convex optimization are competitive, in terms of energy savings, against heuristics that deal with the nonconvex problems directly. We also note that convex optimization techniques can naturally consider the energy related to radiation, and they can exploit the rich structure of the problems to decrease the computational effort (for example, many assignments are impossible owing to the distances involved). We refer the reader to [29] for a comparison of the aforementioned MM-based method against the cell zooming approach in [31]. In Figure 2, we compare the solution obtained with ten iterations of the MM algorithm against the optimal solution of the problem in (12)–(15) with the constraints in (13) replaced by the worst-case constraints in (17). To obtain discrete values for each assignment $x_{i,j}$ after applying the MM algorithm to the problem in (16)–(18), we use the heuristic outlined in [29]. All optimization problems have been solved with CPLEX, a standard commercial solver. The simulated network mimics a long-term evolution system with 100 base stations where the radiated energy is neglected, which is done to simplify the process of obtaining optimal solutions. The bandwidth of each base station is set to 100 MHz, and the data rate requirement of each test point/user is 128 kbit/s. All other parameters are exactly the same as the network simulated in [29]. We observe in Figure 2 that, although we obviously lose optimality by applying the MM algorithm, the time to obtain a network with fairly low energy consumption grows slowly as we increase the number of test points, which stands in sharp contrast to the algorithm that obtains optimal solutions. For a variation of the above ideas to the problem of energy efficiency of small cell networks with best effort users, we refer the reader to [30]. These results indicate that techniques based on convex optimization (which, unlike discrete heuristics, can easily deal with the energy related to radiation) have great potential to scale to very large problems.



[FIG2] A comparison between the optimal solution of the discrete problem with the MM heuristic in a system with 100 base stations. We obtain the error bars corresponding to 95% confidence intervals by estimating the error in the mean from 100 realizations of the simulations. (a) The average number of active base stations as a function of the number test points/users. (b) The average computational time as function of the number of test points/users.

SUMMARY AND OUTLOOK

We showed that many communication schemes aiming at reducing the transmit energy per bit may, in fact, increase the total energy consumption. The analysis and development of wireless communication systems have traditionally considered only the energy radiated by antennas, but neglected the energy required for operating the network. By means of information-theoretic arguments, we showed that the latter cannot be neglected when considering dense deployments of base stations. One way of saving on operating power is to adjust the network to the demand by switching off unnecessary network elements. We devised algorithms to select repeatedly a suitable subset of the network's base station.

ONE WAY OF SAVING ON OPERATING ENERGY IS TO ADJUST THE NETWORK TO THE DEMAND BY SWITCHING OFF UNNECESSARY NETWORK ELEMENTS.

These tools for saving energy need to be refined and extended to consider, for example, mobility, temporal traffic profiles, and the high level of cooperation among network elements of future systems. In particular, the following issues remain open concerning the presented energy-saving optimization approach based on MM:

- Under which conditions does the sequence produced by the MM algorithm converge in this particular application domain? To the best of our knowledge, MM-based algorithms have not been formally shown to converge even when applied to problems in compressive sensing.
- How are realistic load estimates in the optimization problems (not average or worst-case estimates) used, since the spectral efficiency is a function of the assignments $\{x_{i,j}\}_{i\in\mathcal{M},j\in\mathcal{N}}$? Integrating the results on interference calculus into the energy-saving problems may be a direction for future research.
- How are temporal traffic patterns and mobility exploited to save energy? This topic is particularly important for content-centric networking, where we have the additional option of caching content with the intent to save energy by considering current and future channel conditions.
- How are distributed versions of the optimization algorithms implemented? The information exchange among network elements may consume wireless resources, but this fact is not considered in most optimization models.
- How is the flexibility of choosing different communication strategies (for example, cooperative transmission schemes) added while keeping the energy-optimization problems tractable? Adding this level of flexibility complicates the optimization problems. As illustrated earlier, even the simpler task of computing load in the presence of cooperative systems is already difficult.

In summary, our work shows that dense and ultradense deployments of future 5G networks will hardly become reality unless the scaling of base stations can be decoupled from the growth of operating power. Switching off network elements when they are not required is one way to do so.

ACKNOWLEDGMENTS

This work was partially supported by the European Commission within the FP7 Projects GreenNets (grant agreement 286822) and by the German Federal Ministry of Economics and Technology (BMWi) within the project ComGreen (grant 01ME11010). Part of this work has also been performed in the framework of the FP7 project ICT-317669 METIS, which is partly funded by the European Union. We would like to acknowledge the contributions of our METIS colleagues, although the views expressed are ours and do not necessarily represent the project.

AUTHORS

Renato L.G. Cavalcante (renato.cavalcante@hhi.fraunhofer.de) received the electronics engineering degree from the Instituto Tecnológico de Aeronáutica, Brazil, in 2002, and the M.E. and Ph.D. degrees in communications and integrated systems from the Tokyo Institute of Technology, Japan, in 2006 and 2008, respectively. From 2003 to 2008, he was a recipient of the Japanese Government Scholarship (MEXT). He is currently a research fellow with the Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute, Berlin, Germany, and a lecturer at the Technical University of Berlin, Germany. Previously, he held appointments as a research fellow with the University of Southampton, United Kingdom, and as a research associate with the University of Edinburgh, United Kingdom. He received the Excellent Paper Award from the Institute of Electronics, Information, and Communication Engineers of Japan in 2006 and the IEEE Signal Processing Society (Japan Chapter) Student Paper Award in 2008. He also coauthored the study that received the 2012 IEEE International Workshop on Signal Processing Advances in Wireless Communications Best Student Paper Award. His current interests are in signal processing for distributed systems, multiagent systems, and wireless communications.

Slawomir Stańczak (slawomir.stanczak@hhi.fraunhofer.de) received his Dipl.-Ing. and Dr.-Ing. degrees with distinction (summa cum laude) in electrical engineering from the Technical University of Berlin (TU Berlin) in 1998 and 2003, respectively. He has held a habilitation degree (venialegendi) since 2006 and is an associate professor (privatdozent) at TU Berlin. Since 2003, he has led a research group at the Heinrich Hertz Institute. He coauthored two books and more than 130 peer-reviewed journal articles and conference papers. He was the general chair of the 2010 Workshop on Resource Allocation in Wireless Networks. Between 2009 and 2011, he was an associate editor of *European Transactions for Telecommunications*. He has been an associate editor of *IEEE Transactions on Signal Processing* since 2012.

Martin Schubert (martin.schubert@huawei.com) received the diploma and doctoral degrees in electrical engineering from the Technische Universität, Berlin, Germany, in 1998 and 2002, respectively. In 1998, he joined the Heinrich Hertz Institute for Telecommunications, Berlin, as a research assistant. From 2003

to 2012, he was with the Fraunhofer German-Sino Lab for Mobile Communications. He was lecturer at the Technical University of Berlin (2003–2012) and Technical University of Munich (2012–2013). He was a corecipient of the 2007 Johann–Philipp–Reis Award, and he coauthored the 2007 Best Paper Award of the Association of German Electrical Engineers (VDE), and he was a coauthor of the 2007 Best Paper Award of the IEEE Signal Processing Society. From 2009 to 2013, he was an associate editor of *IEEE Transactions of Signal Processing*. Since 2013, he has been a principal researcher at Huawei's European Research Center in Munich, Germany, where he is engaged in research on 5G wireless systems.

Andreas Eisenblätter (eisenblaetter@atesio.de) received a Dipl-Ing. from the Berufsakademie Stuttgart in 1990, an M.Sc. degree in mathematics from the University of Heidelberg in 1995, and a Ph.D. degree in mathematics from Berlin University of Technology, Germany in 2001. He is a managing director at atesio GmbH, Berlin, Germany, a company providing consulting services and software solutions for network infrastructure optimization to the telecommunication industry. He is responsible for all of atesio's activities in the domain of wireless communications. He has been participating in several national and international research projects such as the European Union FP7 projects Momentum, Socrates, GreenNets, and Semafour. He is a (co)author of numerous papers and has contributed to several books. Between 1995 and 2010, he was a researcher at the Zuse Institute Berlin, and he was a founding member of Matheon in 2002.

Ulrich Türke (tuerke@atesio.de) received the M.S. degree in electrical engineering from Aachen University of Technology in 2000. From 2001 to 2007, he worked for Siemens AG on planning and optimization methods for mobile radio networks as well as radio resource management algorithms. In 2007, he received a Ph.D. degree from the University of Bremen. In 2007 he joined atesio GmbH, where he continues to work on analysis, planning, and optimization methods for global system for mobile communications, universal mobile telecommunications system, and long-term evolution networks. He participates in the European Union FP7 projects Momentum, Socrates, GreenNets, and Semafour.

REFERENCES

[1] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Trans. Inform. Theory*, vol. 46, no. 2, pp. 388–404, Mar. 2000.

[2] S. Stańczak, M. Wiczanowski, and H. Boche, *Fundamentals of Resource Allocation in Wireless Networks*, 2nd ed., ser. Foundations in Signal Processing, Communications and Networking, W. Utschick, H. Boche, and R. Mathar, Eds. Berlin, Heidelberg: Springer, 2009.

[3] M. Grossglauser and D. Tse, "Mobility increases the capacity of ad hoc wireless networks," *IEEE/ACM Trans. Networking*, vol. 10, no. 4, pp. 477–486, Aug. 2002.

[4] M. J. Neely and E. Modiano, "Capacity and delay tradeoffs for ad hoc mobile networks," *IEEE Trans. Inform. Theory*, vol. 51, no. 6, pp. 1917–1937, June 2005.

[5] L.-L. Xie and P. R. Kumar, "A network information theory for wireless communication: Scaling laws and optimal operation," *IEEE Trans. Inform. Theory*, vol. 50, no. 5, pp. 748–767, May 2004.

[6] A. Ozgur, O. Leveque, and D. Tse, "Hierarchical cooperation achieves optimal capacity scaling in ad hoc networks," *IEEE Trans. Inform. Theory*, vol. 53, no. 10, pp. 3549–3572, Oct. 2007.

[7] V. Cadambe and S. Jafar, "Interference alignment and the degrees of freedom for the k user interference channel," *IEEE Trans. Inform. Theory*, vol. 54, no. 8, pp. 3425–3441, Oct. 2008.

[8] J. Hoydis, M. Kobayashi, and M. Debbah, "Green small-cell networks," *IEEE Veh. Technol. Mag.*, vol. 6, no. 1, pp. 37–43, Mar. 2011.

[9] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka, H. Tullberg, M. A. Uusitalo, B. Timus, and M. Fallgren, "Scenarios for the 5G mobile and wireless communications: The vision of the METIS project," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 26–35, May 2014.

[10] B. Liu, Z. Liu, and D. Towsley, "On the capacity of hybrid wireless networks," in *Proc. IEEE INFOCOM 2003*, Mar. 2003, vol. 2, pp. 1543–1552.

[11] L. M. Correia, D. Zeller, O. Blume, D. Ferling, Y. Jading, I. Gódor, G. Auer, and L. Van Der Perre, "Challenges and enabling technologies for energy aware mobile radio networks," *IEEE Commun. Mag.*, vol. 48, no. 11, pp. 66–72, 2010.

[12] R. G. Gallager, "Energy limited channels: Coding, multi-access and spread spectrum," in *Proc. Conf. Information Sciences and Systems (CISS)*, Mar. 1988, p. 372.

[13] A. Goldsmith and S. Wicker, "Design challenges for energy-constrained ad-hoc wireless networks," *IEEE Wireless Commun. Mag.*, vol. 9, pp. 8–27, Aug. 2002.

[14] S. Verdu, "Recent results on the capacity of wideband channels in the low-power regime," *IEEE Wireless Commun. Mag.*, vol. 9, pp. 40–45, Aug. 2002.

[15] A. Zemlianov and G. de Veciana, "Capacity of ad hoc wireless networks with infrastructure support," *IEEE J. Select. Areas Commun.*, vol. 23, no. 3, pp. 657–667, Mar. 2005.

[16] R. L. G. Cavalcante, E. Pollakis, S. Stańczak, F. Penna, and J. Bühler, "GreenNets deliverables," GreenNets Project, Tech. Rep. FP7.SME.2011.1, 2013.

[17] Y. Chen, S. Zhang, S. Xu, and G. Li, "Fundamental trade-offs on green wireless networks," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 30–37, 2011.

[18] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press, 2006.

[19] J.-Y. Le Boudec, "Performance evaluation of computer and communication systems," EPFL Press: Lausanne, Switzerland, 2010.

[20] B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja, *A First Course in Order Statistics*. Philadelphia, PA: SIAM, 2008.

[21] I. Takeuchi, Q. V. Le, T. D. Sears, and A. J. Smola, "Nonparametric quantile estimation," *J. Mach. Learn. Res.*, vol. 7, pp. 1231–1264, July 2006.

[22] R. D. Yates, "A framework for uplink power control in cellular radio systems," *IEEE J. Select. Areas Commun.*, vol. 13, no. 7, pp. 1341–1348, Sept. 1995.

[23] M. Schubert and H. Boche, *Interference Calculus—A General Framework for Interference Management and Network Utility Optimization*. Berlin: Springer, 2012.

[24] A. Fehske, H. Klessig, J. Voigt, and G. Fettweis, "Concurrent load-aware adjustment of user association and antenna tilts in self-organizing radio networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 5, pp. 1974–1988, June 2013.

[25] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2006.

[26] I. Siomina and D. Yuan, "Analysis of cell load coupling for LTE network planning and optimization," *IEEE Trans. Wireless Commun.*, vol. 11, no. 6, pp. 2287–2297, June 2012.

[27] P. Mogensen, W. Na, I. Z. Kovacs, F. Frederiksen, A. Pokhariyal, K. I. Pedersen, T. Kolding, K. Hugl, and M. Kuusela, "LTE capacity compared to the Shannon bound," in *Proc. 65th IEEE Vehicular Technology Conf., VTC Spring 2007*, Apr., pp. 1234–1238.

[28] S. Joshi and S. Boyd, "Sensor selection via convex optimization," *IEEE Trans. Signal Processing*, vol. 57, no. 2, pp. 451–462, 2009.

[29] E. Pollakis, R. L. G. Cavalcante, and S. Stanczak, "Base station selection for energy efficient network operation with the majorization-minimization algorithm," in *Proc. 2012 IEEE 13th Int. Workshop Signal Processing Advances Wireless Communications* (*SPAWC*), June 2012, pp. 219–223.

[30] L. Su, C. Yang, Z. Xu, and A. F. Molisch, "Energy-efficient downlink transmission with base station closing in small cell networks," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Processing* (*ICASSP*), May 2013, pp. 4784–4788.

[31] Z. Niu, Y. Wu, J. Gong, and Z. Yang, "Cell zooming for cost-efficient green cellular networks," *IEEE Commun. Mag.*, vol. 48, no. 11, pp. 74–79, Nov. 2010.

[32] I. Yamada, M. Yukawa, and M. Yamagishi, "Minimizing the Moreau envelope of nonsmooth convex functions over the fixed point set of certain quasi-nonexpansive mappings," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, H. Bauschke, R. Burachick, P. L. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz, Eds. New York: Springer-Verlag, pp. 345–390, 2011.

[33] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted l_1 minimization," *J. Fourier Anal. Appl.*, vol. 14, no. 5, pp. 877–905, Dec. 2008.

[34] B. K. Sriperumbudur, D. A. Torres, and G. R. G. Lackriet, "A majorization-minimization approach to the sparse generalized eigenvalue problem," *Mach. Learn.*, vol. 85, no. 1–2, pp. 3–39, Oct. 2011.

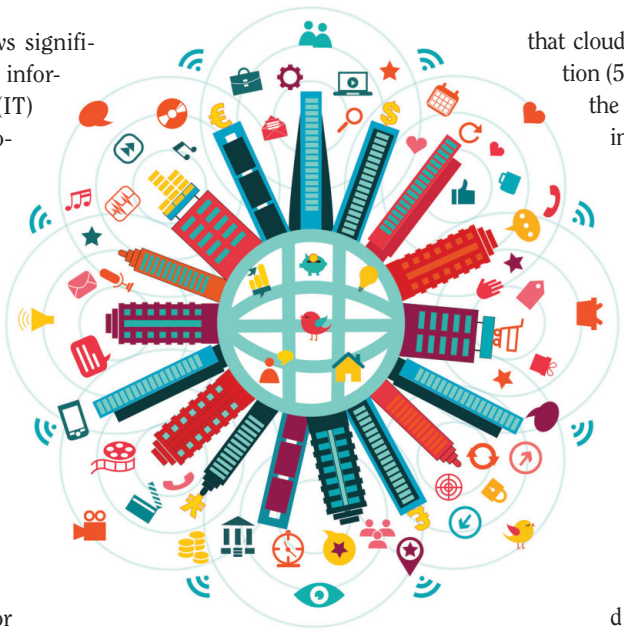
[SP]

Dirk Wübben, Peter Rost, Jens Bartelt, Massinissa Lalam, Valentin Savin, Matteo Gorgoglione, Armin Dekorsy, and Gerhard Fettweis

Benefits and Impact of Cloud Computing on 5G Signal Processing

Flexible centralization through cloud-RAN

Cloud computing draws significant attention in the information technology (IT) community as it provides ubiquitous on-demand access to a shared pool of configurable computing resources with minimum management effort. It gains also more impact on the communication technology (CT) community and is currently discussed as an enabler for flexible, cost-efficient and more powerful mobile network implementations. Although centralized baseband pools are already investigated for the radio access network (RAN) to allow for efficient resource usage and advanced multicell algorithms, these technologies still require dedicated hardware and do not offer the same characteristics as cloud-computing platforms, i.e., on-demand provisioning, virtualization, resource pooling, elasticity, service metering, and multitenancy. However, these properties of cloud computing are key enablers for future mobile communication systems characterized by an ultra-dense deployment of radio access points (RAPs) leading to severe multicell interference in combination with a significant increase of the number of access nodes and huge fluctuations of the rate requirements over time. In this article, we will explore the benefits that cloud computing offers for fifth-generation (5G) mobile networks and investigate the implications on the signal processing algorithms.



THE 5G REVOLUTION

©ISTOCKPHOTO.COM/ZONADEARTE

INTRODUCTION

The evolution toward 5G mobile networks is characterized by an exponential growth of traffic. This growth is caused by an increased number of user terminals, richer Internet content, more frequent usage of Internet-capable devices, and by more powerful devices with larger screens. This implies also the need for more scaling possibilities in mobile networks to handle spatially and temporally fluctuating traffic patterns, terminals with different quality requirements, and more diverse services. Current mobile networks are not able to support this diversity efficiently but are designed for peak-provisioning and typical Internet traffic.

The use of very dense, low-power, small-cell networks with very high spatial reuse is a promising way to allow for handling future data rate demands [1], [2]. Small cells exploit two fundamental effects. First, the distance between the RAP and terminals is reduced, which increases the line-of-sight probability and reduces the path loss. Second, the spectrum is used more efficiently because each RAP uses the same spectrum. Small cells complement existing macrocellular deployments that are required to provide coverage for fast-moving users and in areas with low

Digital Object Identifier 10.1109/MSP.2014.2334952
Date of publication: 15 October 2014

user density. In Third-Generation Partnership Project (3GPP) long-term evolution (LTE), small cells draw significant attention on both the physical and higher layer [3], [4], where impacts on the RAN protocol and system architecture are discussed.

As networks become denser, intercell interference increases and interference scenarios become more complex due to multitier interference. Furthermore, the higher the deployment density, the higher the chance that a RAP will carry no or only low traffic-load due to spatial and temporal traffic fluctuations. Currently, 15–20% of all sites carry about 50% of the total traffic [5]. Centralized processing permits to selectively turn RAPs on and off to address the spatiotemporal traffic fluctuations. In addition, it allows for efficient interference avoidance and cancellation algorithms across multiple cells as well as joint detection algorithms. Centralized RAN (C-RAN) recently attracted attention as one possible way to efficiently centralize RAN processing [6]. In C-RAN, remote radio heads (RRHs) are connected through optical fiber links to a data center where all baseband processing is performed [7], [8]. Thus, by pooling baseband processing in baseband units (BBUs), centralization gains are achieved. However, BBUs are based on specialized hardware platforms utilizing digital signal processors (DSPs) [9]. As a long-term goal, it is beneficial to deploy cloud-computing platforms running on general-purpose hardware, leading to a cloud-RAN system as outlined subsequently in this article.

Only fiber links are capable of supporting the necessary data rates between the RRH and the BBU. This constitutes the main drawback of C-RAN, i.e., it requires very high data rate links to the central BBU. In [8], the authors report a required backhaul (BH) transmission rate of 10 Gbit/s for time-domain LTE (TD-LTE) with eight receive antennas and 20-MHz bandwidth. Due to the use of optical fiber, C-RAN deployments are less flexible as only spots with existing fiber access may be chosen or fiber access must be deployed, which is very cost-intense. Future mobile networks will deploy heterogeneous BH solutions that are optimized for different scenarios. This mix of BH characteristics will also imply a mix of more C-RAN solutions that require high-capacity BH and more decentralized solutions compatible with BH solutions that introduce high latency and stronger throughput constraints [10].

The RAN as a Service (RANaaS) concept is introduced in [11]. It addresses the deficiencies of C-RAN to allow for a centralization over heterogeneous BH. The main characteristics of RANaaS are the flexible assignment of RAN functionality between the RAPs and the central processor, the deployment of commodity hardware at the central processor, and the tight integration of RAN, BH network, and central processor. In this article, we focus on the challenges and benefits of implementing signal processing algorithms on a cloud-computing platform. Hence, in the following, we refer to the concept of centralization toward commodity cloud-computing platforms as cloud-RAN. More details on the architecture design of the underlying 5G mobile network as well as fundamental concepts from medium access control (MAC) and network layer of the cloud-RAN concept are given in [11]. Further challenges in 5G mobile networks, which are beyond the scope of this article, are introduced in [2] and [12], among others. However, cloud-RAN will foster approaches currently under discussion for 5G such as massive multiple-input, multiple-output (MIMO) and multiple radio access technologies.

FLEXIBLE CENTRALIZATION THROUGH CLOUD-RAN

FLEXIBLE ASSIGNMENT OF RAN FUNCTIONALITY

A flexible assignment of RAN functionality can consider both the cloud-platform resource availability and the small-cell BH characteristics. In addition, cloud-computing platforms allow for the scalability that is required to cope with temporal and spatial traffic fluctuations in mobile networks. This scalability is a fundamental requirement to improve the utilization of mobile networks and to allow for an economically and ecologically sustainable operation of mobile networks.

Cloud-RAN is a disruptive technology in many ways and imposes new challenges on the signal processing in 5G mobile networks. Most importantly, it will exploit standard processor technology [general-purpose processors (GPPs)] to execute RAN functionality. By contrast, currently discussed C-RAN technology considers a baseband pooling approach where a large number of DSPs are provided at a central entity [8], [9]. Although this allows for resource sharing, C-RAN still uses specialized and expensive hardware and software. Hence, it is misleading to consider C-RAN as an example of cloud computing according to the IT definition by the U.S. National Institute of Standards and Technology [13].

Cloud-RAN will further foster scalable algorithms that are designed for cloud-computing environments and leverage massive parallelism. This implies that algorithms should not be simply ported to cloud-computing platforms but rather redesigned to gain from the available computing resources. Cloud-RAN allows for the deployment of algorithms that scale with the need for cooperation and coordination among the individual cells, i.e., depending on the traffic demand and user density, RAPs may be differently grouped or different algorithms may be deployed. In the following sections, this article provides more detailed examples for algorithms that benefit from an application to cloud-computing platforms.

To enable cloud-RAN, it is necessary to have a system architecture that provides the required interfaces without disruptive changes to an existing deployment. This architecture has been introduced in [11]. It does not imply changes to existing interfaces but introduces the concept of a virtual eNodeB (veNB). A veNB is composed of one or more RAPs, a cloud-computing platform, and the necessary BH links between these nodes. It maintains the same interfaces as a 3GPP LTE eNodeB (eNB) to maximize backward-compatibility. This system architecture requires 1) that the functionality at the eNB can be decomposed into reassignable functions and 2) that each function can be assigned either to the central processor or local RAPs. Furthermore, a tight integration of RAN, BH, and central processor is required, e.g., through joint coding as introduced in [14].

OPPORTUNITIES OF CLOUD-RAN

Cloud computing offers the ability of computational load balancing to RANs. This allows for spending more computational

efforts on critical operations, e.g., in the case of interference scenarios or difficult channel conditions. In these scenarios, more advanced and computationally intense algorithms may be needed and could be executed in a cloud environment. By contrast, traditional implementations are hard real-time systems. Hence, a certain task such as decoding or scheduling is always executed within the same time window.

A flexible assignment of functionality will also allow for shaping the signaling load on the BH connection. For instance, in the case of high-capacity, low-latency BH, the central processor may process directly in-phase/quadrature (I/Q) samples. In the case of higher latency and lower bandwidth on the BH, the central processor may only perform upper-layer functionality. This will require changes to the operation of the BH and the signal processing platform, and it may require changes to the RAN protocol stack.

Cloud-RAN will open the door for many new applications in 5G. It offers the possibility of using signal processing software dedicated to a special purpose based on the actual service. It reflects the diversity of services, use cases, and deployments through flexibility and scalability of the signal processing platform. In addition, it may even take into account the complexity and abilities of terminals during the processing of signals. Finally, cloud-RAN avoids the typical vendor lock-in as in current deployments that follow a similar development observed in the mobile core network, which may be implemented on cloud-platforms [15].

The flexible centralization of RAN functionality will impact the operation of the 3GPP LTE RAN protocol stack and may even be limited by dependencies within the protocol stack. Table 1 provides an overview of promising functions of the 3GPP LTE radio protocol stack, which may be considered for a partial centralization. In general, the lower we place the functional split within the protocol stack, the higher the overhead and the more stringent BH requirements are. Centralizing functionality on the physical layer (PHY) allows for computational diversity that depends directly on the number of users per RAP. Due to temporal and spatial fluctuations, the computational load can be balanced across cells. Central processing also allows for implementing multicell algorithms to avoid or exploit interference, e.g., intercell interference coordination and cooperative multipoint processing [16].

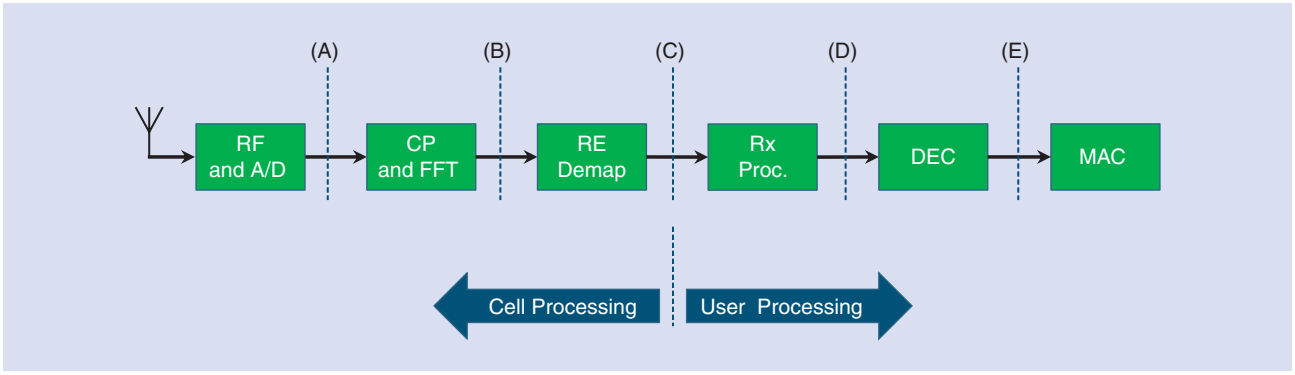
FUNCTIONAL SPLIT

In this subsection, we introduce several functional split options that determine the execution of processing in the RAP or in the cloud-platform and directly influence the required BH data rate. The discussion is focused on the uplink (UL) since its processing load dominates the downlink (DL) processing. Detailed investigations of such splits have also been conducted in [17], but here we focus more on the opportunities of a flexible split. By relying on GPPs as opposed to dedicated hardware as used in the C-RAN concept, and through extensive use of function virtualization, the envisioned architecture allows us to adapt the functional split flexibly in time (e.g., according to traffic demand) and location (e.g., depending on the density of the deployment). Figure 1 illustrates the principle LTE signal processing chain of an UL receiver and different options of placing a functional split. Notice that similar shifts are also possible for DL processing as considered, e.g., in the context of precoding for massive MIMO systems in [18].

Subsequently, we discuss these split options and give numerical results on the required BH data rates per link between one RAP and the cloud-platform for a simple configuration as specified in Table 2.

[TABLE 1] THE BENEFITS AND SIGNAL PROCESSING CHALLENGES FOR THE CENTRALIZATION OF SELECTED 3GPP LTE RADIO PROTOCOL FUNCTIONALITY ON THE PHY AND LOWER MAC LAYER.

CENTRALIZED FUNCTIONALITY	CENTRALIZED REQUIREMENTS	CENTRALIZATION BENEFITS	CHALLENGES FOR SIGNAL PROCESSING
DETECTION AND FEC-DECODING	■ DEPENDS ON CONTROL OVERHEAD IN UL ■ LATENCY REQ. DEPENDS ON TIMING REQ. IN DL ■ STRONG RELIABILITY	■ COOPERATIVE RECEIVER (RX) ■ COMPUTATIONAL DIVERSITY	■ PREDETECTION AT RAP TO REDUCE BH OVERHEAD ■ OPTIMAL QUANTIZATION OF SIGNALS AND EXCHANGE OVER BH
FEC-ENCODING AND MODULATION AND PRECODING	■ DEPENDS ON CONTROL OVERHEAD IN DL ■ STRONG RELIABILITY	■ COOPERATIVE TRANSMITTER (TX) ■ ADVANCED PRECODING ■ COMPUTATIONAL DIVERSITY	■ SEPARATE PRECODING DECISION AND EXECUTION AT RAP AND CENTRAL PROCESSOR ■ OPTIMAL QUANTIZATION OF SIGNALS AND EXCHANGE OVER BH
LINK RELIABILITY PROTOCOLS (E.G., HARQ)	■ DEPENDS ON ENTITY THAT PERFORMS RETRANSMISSION DECISION	■ SIMPLIFIED CENTRALIZATION OF SCHEDULING AND DECODING	■ PREDEFINED TIMING OF (N)ACK MESSAGES ■ SEPARATION OF RETRANSMISSION DECISION AND PACKET COMBINING ■ STRONG INTERACTION WITH OTHER FUNCTIONS, E.G., SCHEDULER, EN-/DECODER
SCHEDULING AND INTERCELL RRM	■ FLEXIBLE REQUIREMENTS	■ MULTICELL GAINS ■ COMPUTATIONALLY EXPENSIVE ALGORITHMS ■ GAINS DEPEND ON BH QUALITY	■ SCALABLE LATENCY REQUIREMENTS MUST BE SUPPORTED ■ INTERCELL INTERFERENCE COORDINATION (ICIC) BASED ON CHANGING QUALITY OF CHANNEL STATE INFORMATION ■ CHANGING COMPUTATIONAL COMPLEXITY



[FIG1] The functional split between RAPs and the cloud-platform for UL transmission.

I/Q FORWARDING (A)

By immediately forwarding the time-domain receive signals that have been downconverted to the baseband and analog-to-digital (AD) converted (indicated by block RF/AD), the complete receive frame including the cyclic prefix (CP) has to be transmitted over the BH link to the cloud-platform. This approach is usually referred to as *radio-over-fiber* (*RoF*) and is used in the common public radio interface (CPRI) standard [19]. The main benefit of this split is that almost no digital processing devices are required at the RAPs, potentially making them very small and cheap. If a flexible split varying over time is envisioned, the processing devices would have to be available at the RAPs anyway, nullifying this benefit. Also, the required BH data rate for I/Q forwarding is comparatively high and given as

$$D_{\mathrm{BH}}^{A} = N_{\mathrm{O}} \cdot f_{\mathrm{S}} \cdot 2 \cdot N_{\mathrm{Q}} \cdot N_{\mathrm{R}} = 2 \cdot 30.72\ \mathrm{MHz} \cdot 2 \cdot 10\ \mathrm{bit} \cdot 2 = 2.46\ \mathrm{Gbit/s}. \quad (1)$$

SUBFRAME FORWARDING (B)

By removing the CP and transforming the Rx signal to frequency-domain using fast Fourier transformation (FFT), guard subcarriers can be removed (block CP/FFT). Since the number of guard subcarriers in LTE is $\approx 40\%$, this decreases the required BH data rate significantly.

$$D_{\mathrm{BH}}^{B} = N_{\mathrm{Sc}} \cdot T_{\mathrm{S}}^{-1} \cdot 2 \cdot N_{\mathrm{Q}} \cdot N_{\mathrm{R}} = 1{,}200 \cdot (66\ \mu\mathrm{s})^{-1} \cdot 2 \cdot 10\ \mathrm{bit} \cdot 2 = 720\ \mathrm{Mbit/s}. \quad (2)$$

[TABLE 2] EXEMPLARY TRANSMISSION PARAMETERS FOR CALCULATING THE IMPACT OF FUNCTIONAL SPLIT CHOICES ON THE BH DATA RATE.

PARAMETER	SYMBOL	VALUE
BANDWIDTH	B	20 MHz
SAMPLING FREQUENCY	f_{S}	30.72 MHz
OVERSAMPLING FACTOR	N_{O}	2
NUMBER OF USED SUBCARRIERS	N_{Sc}	1,200
SYMBOL DURATION	T_{S}	66.6 µs
QUANTIZATION/SOFT BITS PER I/Q	N_{Q}	10
RX ANTENNAS	N_{R}	2
SPECTRAL EFFICIENCY	S	3 bit/cu
ASSUMED RB UTILIZATION	η	50%

As an FFT can be implemented on dedicated hardware very efficiently, the implementation in the RAP is worthwhile compared to the split option I/Q forwarding (A). As the per-cell based processing does not depend on the actual load of the RAP, load balancing gains can be only achieved if RAPs are completely turned off.

RX DATA FORWARDING (C)

If only a part of the resource elements (REs) are actually utilized by the user equipment (UE) in a cell, only these REs remain after RE demapping (block RE Demap) and have to be forwarded to the cloud-platform. The required BH data rate is directly given by the fraction of utilized RE and thus, the subsequent splits can profit from load balancing gains.

$$D_{\mathrm{BH}}^{C} = D_{\mathrm{BH}}^{B} \cdot \eta = 720\ \mathrm{Mbit/s} \cdot 0.5 = 360\ \mathrm{Mbit/s}. \quad (3)$$

To allow for a joint processing of received signals from multiple RAPs, it has to be ensured that only REs of UE not considered for joint processing are removed, even if they are not (primarily) associated with the current RAP.

SOFT-BIT FORWARDING (D)

The receive processing (block Rx Proc) per user consists of equalization in frequency domain, inverse discrete Fourier transformation (IDFT), MIMO receive processing, and demapping. In a MIMO scheme utilizing receiver diversity, the signals of multiple antennas are combined during channel equalization, thus removing the dependency on the number of receive antennas. This results in a reduced BH load of $D_{\mathrm{BH}}^{D} = D_{\mathrm{BH}}^{C}/N_{\mathrm{R}} = 180$ Mbit/s. In contrast, for spatial multiplexing with N_{S} layers per UE, the BH would correspond to $D_{\mathrm{BH}}^{D} = D_{\mathrm{BH}}^{C} \cdot N_{\mathrm{S}}/N_{\mathrm{R}}$. By this split, only joint decoding of soft bits forwarded by several RAPs is possible in the cloud-platform. Also note that usually the number of soft bits per symbol would depend on the modulation scheme (e.g., three soft-bits per information bit), and thus N_{Q} and the BH data rate would depend directly on the modulation order, which in turn depends on the access channel quality due to radio resource management (RRM).

MAC (E)

During forward error correction (FEC) decoding (block DEC), data bits are recovered from the received symbols and redundant

bits are removed, resulting in the pure MAC payload at the decoder output. The resulting BH data rate depends largely on the used modulation and coding scheme (MCS), which is reflected here by the exemplary spectral efficiency $S = 3$ bit/cu.

$$D_{\mathrm{BH}}^{E} = N_{\mathrm{Sc}} \cdot T_{\mathrm{S}}^{-1} \cdot \eta \cdot S = 1{,}200 \cdot (66\,\mu\mathrm{s})^{-1} \cdot 0.5 \cdot 3\ \mathrm{bit/cu} = 27\ \mathrm{Mbit/s}. \quad (4)$$

FEC decoding is a complex task that is commonly performed on dedicated hardware and hence a centralized decoding on GPPs has not been considered in C-RAN. However, as outlined later in this article, recent results show that it can be performed on GPPs. On the other hand, performing decoding in the RAPs according to the split option MAC (E) terminates the possibility for joint PHY-layer processing in the cloud-platform and only cooperation on higher layers, e.g., joint scheduling, remains possible. As PHY-layer cooperation mainly revolves around interference mitigation, this option is beneficial in scenarios were RAPs are well separated, e.g., for indoor deployments or in narrow street canyons.

Obviously, the required BH data rate and the required processing power in the cloud decreases significantly when the functional split is shifted to the higher PHY processing layers or even to the MAC. However, this is traded off with lower centralization gains in terms of spectral efficiency and computational load balancing. The advantage of a flexible split is that we can reap the benefits of both extremes: load balancing for low traffic situations and high spectral efficiency by cooperative processing for high traffic. Since current BH standards like CPRI only support a very specific functional split, new and more flexible standards will have to be defined to enable cloud-RAN architectures.

The huge BH bandwidth requirements of functional shifts on the lower PHY layers also shows that improved and optimized BH technologies are required. While technologies offering sufficient bandwidth are already available [10], a joint design of radio access and BH links should be also considered to use the deployed capacity as efficiently as possible. Additionally, to further limit the BH rate between the RAPs and the cloud-platform, cooperative processing strategies could be used to directly exploit lower-layer interaction between RAPs. This would allow the use of heterogeneous BH technologies to interconnect the RAPs and implement joint distributed detection techniques as depicted in Figure 2 and discussed in the next section.

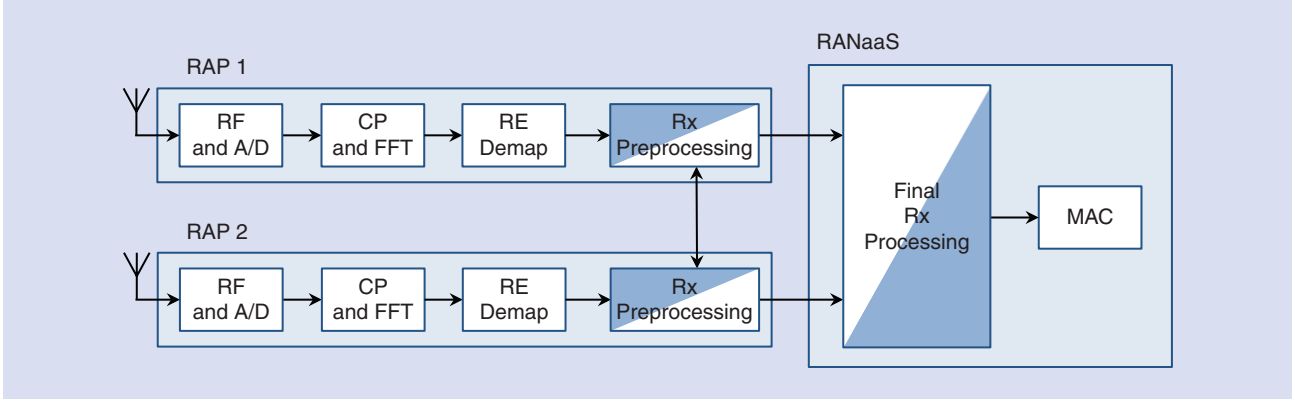
SIGNAL PROCESSING IN THE CLOUD

The difficulty of implementing RAN functionality in a cloud-platform lies in the tight constraints caused by the 3GPP LTE protocol stack. This implies that individual tasks need to finish within a predefined time window. Figure 3 shows relevant parts of the 3GPP LTE protocol stack and two exemplary functional splits that correspond to options (C) and (D) in Figure 1. In the following, we discuss the benefits and challenges of a cloud implementation of three representative parts of the signal processing chain.

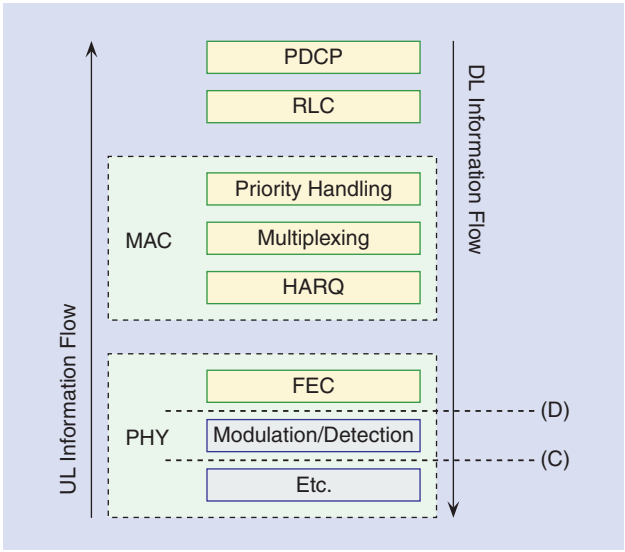
HYBRID AUTOMATIC REPEAT REQUEST

Among all the timers defined in LTE, the one associated to the acknowledgment (ACK) of a UL physical frame at the MAC layer is the most critical one. The reception status of any frame sent through the air interface needs to be fed back to the transmitter, to proceed to the transmission of a new frame ACK or to attempt a retransmission negative ACK (NACK). This hybrid automatic repeat-request (HARQ) operation is performed at the MAC level, after all the physical processing of a codeword is done (detection, demodulation, and FEC decoding). In LTE, each frame sent at subframe n needs to be acknowledged (ACK or NACK) at subframe $n+4$ in both UL and DL directions, a subframe lasting 1 ms [20]. Hence, the overall receive process has to finish in 3 ms to stay compliant with the 3GPP LTE HARQ timing. This timing includes the processing at the RAPs of the physical blocks located before the split (see Figure 3 and both functional split options therein), the processing at the cloud-platform of the physical blocks located after the split and the round-time trip through the BH. However, some algorithms such as turbo-decoders underly a computational jitter which implies that the decoding time may vary. Hence, it may happen that packets are retransmitted even though they would have been decoded with more computational resources, i.e., either more time or more parallel processors. This computational jitter also adds up to the overall delay that needs to be considered.

To relax the timing constraint for the receive processing, we may adapt the HARQ process. The authors in [17] suggest for



[FIG2] The cooperative Rx preprocessing among RAPs with final Rx processing in the central processor.



[FIG3] **The LTE protocol stack and exemplary functional splits.**

example to suspend the HARQ process until the end of the receive processing. In the case that the receive processing is not finished in time, an ACK is sent after 3 ms to meet the timing requirements while receive processing is continued. If, at the end, successful decoding is not possible, a NACK is sent. As the UE does not immediately drop out a package when receiving an ACK to cope with transmission errors on the feedback channel, a retransmission of the particular packet can be scheduled later. However, this approach halves the achievable UE peak rate [17]. This drawback can be avoided by a preliminary HARQ process, where the initial feedback message is determined by estimating the decoding success based on the quality of the received signals (e.g., using models from link level simulations [21]). If correct decoding is likely, a preliminary ACK is sent to the UE, otherwise a preliminary NACK. Again, the standard techniques capturing feedback errors automatically handle erroneous preliminary feedback messages. This approach relaxes the timing constraints for the receive processing chain. It separates the most complex processing parts and the most latency-critical parts but still allows for high data rates depending on the reliability of preliminary ACK/NACK.

FORWARD ERROR CORRECTION

The tight requirement of finishing the overall detection within 3 ms poses a significant challenge for executing FEC decoding within the cloud-platform due to its high complexity. Usually, FEC decoders are implemented in specialized hardware, such as application-specific integrated circuit (ASIC) designs or field-programmable gate array (FPGA) implementations [22]. However, the introduction of many-core architectures opens new perspectives for massively parallel implementations. To meet stringent requirements on data rates, cloud-based FEC decoders will need to fully exploit the available parallelism of a cloud-computing platform. In this context, low-density parity check (LDPC) [23] and turbo codes [24] are two promising candidates because both allow for accommodating various degrees of parallelization.

From a high-level perspective, two main approaches can be used to exploit parallelism in multicore platforms. The first approach parallelizes the decoder itself through decomposition of the decoding algorithms into multiple threads that run in parallel. Second, multiple codewords may be decoded in parallel. The first approach decreases the latency per codeword but introduces more synchronization overhead across different threads. By contrast, the second approach uses less synchronization objects and therefore increases the parallelization gain. However, it may introduce a higher latency per codeword compared to the first approach.

For very high throughput applications, LDPC codes are known to compare favorably against turbo codes because LDPC decoding allows for a higher degree of parallelism [25], [26]. Hence, LDPC codes are suitable for the first approach of decoder parallelization. However, software-based parallel LDPC decoders barely achieve throughputs of a few tens of Mbit/s, as reported in [27] for graphical processing units (GPUs), or in [28] for the signal processing on-demand architecture (SODA). In both cases, the main reason is the need for synchronization across different threads to access shared objects that results in scalability issues [27].

By contrast, parallelizing multiple codewords eliminates the need for synchronizing objects. This results in better scalability properties and the throughput of the multicodeword decoder is known to increase almost linearly with the number of cores [29]. Furthermore, it allows for different codes, algorithms, and configurations running in parallel. Multicodeword LDPC decoders have been reported to achieve throughputs up to 80 Mbit/s on the IBM CELL Broadband Engine [27], [30], with 24–96 codewords decoded in parallel. Recently, central processing unit (CPU) and GPU implementations of multicodeword turbo decoders have also been reported in [31] with a peak throughput from 55 Mbit/s to 122 Mbit/s, as the number of decoding iterations decreases from eight to four.

Figure 4 shows experimental results for spectral efficiency and required computational complexity of an 3GPP LTE UL decoder. To obtain these results, the turbo-decoder has been implemented on a default VMWare ESXi server with Ubuntu Linux host operating system, GNU C++ compiler, and codeword multithreading to account for the virtualization overhead. We measured the required CPU time to decode one codeword and determined the average CPU time within the 90% confidence interval.

Figure 4(a) shows the achievable spectral efficiency for a given signal-to-noise ratio (SNR) (additive white Gaussian noise, no fading). We illustrate the results for two cases: maximum throughput (high number of iterations possible) and low complexity (number of iterations limited to two). Reducing the complexity of the decoding process results in a performance penalty of 1–2 dB. In Figure 4(b), we show the required computational resources for a 10-MHz 3GPP LTE system. The required complexity strongly depends upon the SNR. First, it increases linearly with the number of information bits, which implies a logarithmic increase of complexity in SNR. Second, the complexity increases with the number of iterations that are necessary to decode a codeword. As shown in [32], the complexity increases superlinearly with decreasing SNR (in decibels) for a fixed MCS. In Figure 4(b),

markers show the SNR where the next higher MCS has been chosen. We notice at each of these markers an increase of the computational demand, which is then quickly decreasing in SNR.

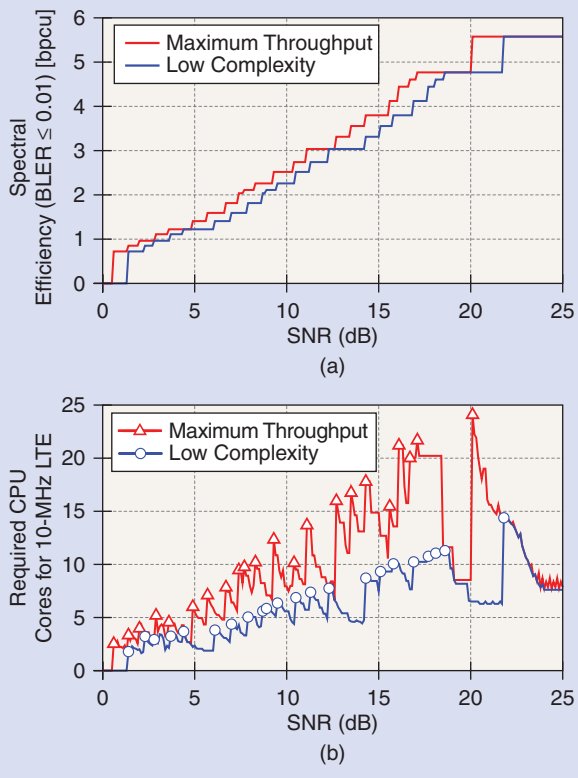
Apparently, this strongly varying computational demand allows for the exploitation of multiuser computational diversity at the centralized processor. For instance, the central processor can perform computational load balancing across multiple users to reduce the ratio of peak-to-average computational efforts. Furthermore, the central processor can actively shape the computational demand by selecting MCS to satisfy a computational constraint, e.g., in the case of a traffic burst the computational requirements may significantly increase and may exceed the available resources if MCSs are chosen based on maximum throughput. Finally, the computational load can be actively shaped by adjusting the number of quantization bits N_Q used for forwarding the Rx signals from the RAP to the cloud-platform over the BH. Figure 5 shows the tradeoff between number of turbo iterations and quantization bits N_Q for different modulation schemes at a target bit error rate (BER) of 10^{-4}. Obviously, the decoding latency can significantly been reduced by increasing the number of quantization bits N_Q on the cost of a higher BH transmission rate.

MULTIUSER DETECTION

Consider again the functional split option, Rx Data Forwarding (C) in Figure 3. In this case, I/Q samples are forwarded over high-capacity BH links to the central processor that performs joint multiuser detection (MUD) using the Rx signals of several RAPs. The joint processing of many RAPs implements a virtual MIMO architecture and the huge computational power offered by the cloud-platform allows for aggressive RRM across the RAPs. However, due to the heterogeneous nature of BH networks, it is also beneficial to use a mix of local processing at RAPs, cooperative processing among RAPs, and central processing in the cloud-platform. Promising techniques that are adaptable to changing BH and radio access parameters are, among others, multipoint turbo detection (MPTD) and in-network processing (INP).

The underlying idea of MPTD [33] is to schedule (edge) users attached to different RAPs on the same resource. Then, a joint detection of these users through a turbo processing approach is performed [21], [34]. Such processing could be done either centrally on the cloud-platform or locally in each RAP. If it is fully centralized, MPTD benefits from high degree of spatial diversity due to the different locations of the involved RAPs. Due to this spatial diversity increase, the centralization gain can be quite significant compared to a classical distributed detection.

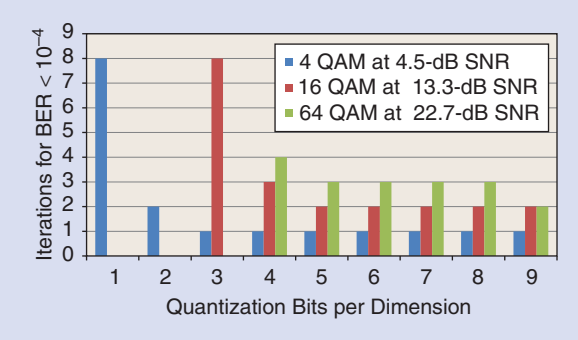
This split of functionality may offer significant centralization gains compared to distributed detection methods. This is illustrated in Figure 6 for an UL scenario with $N_{UE} = 2$ users each equipped with $N_T = 1$ transmit antenna. Both users interfere with each other at $N_{RAP} = 2$ RAPs each equipped with $N_R = 2$ receive antennas. We assume the worst case of identical path-losses. In addition, these results consider Rayleigh channel fading and LTE-compliant MCSs [20]. Figure 6 shows that at a frame error rate (FER) of 0.01 a centralization



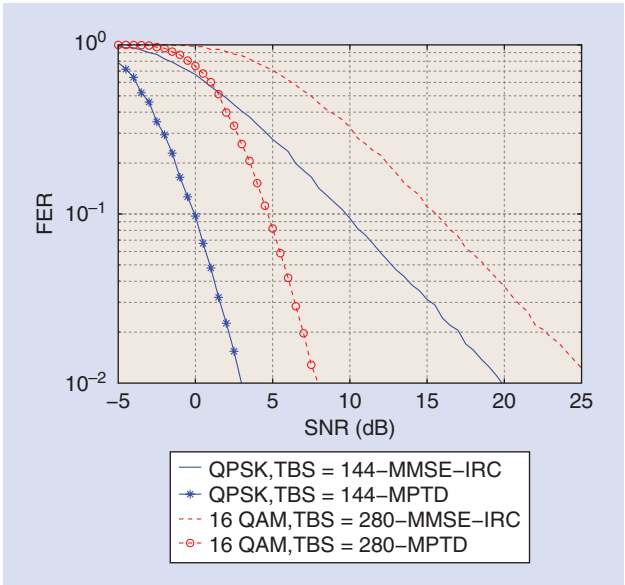
[FIG4] Throughput and computational complexity results for turbo-decoding using an out-of-the-box cloud-computing platform and 3GPP LTE MCSs.(a) Spectral efficiency. (b) Required CPU cores.

gain of about 17 dB can be achieved by MPTD compared to a linear minimum mean square error (MMSE) filter with interference rejection combining (IRC) [35].

An alternative approach that faces the joint MUD problem from an optimization perspective is INP. It allows for the solution of general estimation problems in a distributed, decentralized way within a network. The special class of consensus-based algorithms achieves this by iteratively reaching consensus of the estimates among the processing nodes [36], [37]. The adaptation of INP for an iterative distributed MUD has recently been presented in [38]. Due to its generic structure, INP can also be implemented with



[FIG5] The number of turbo iterations required for BER $< 10^{-4}$ versus number of quantization bits N_Q per I/Q dimension.

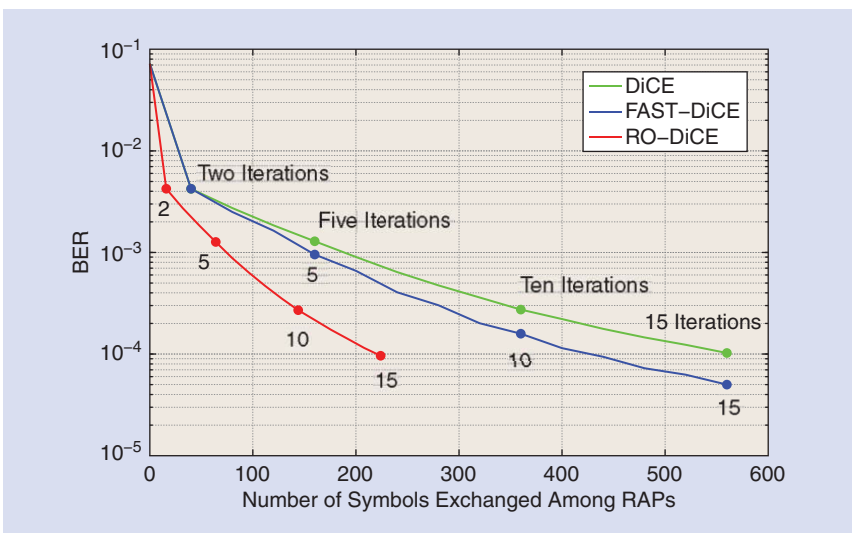


[FIG6] The FER over SNR for MPTD (centralized) and MMSE-IRC (distributed).

the desired mix of local, cooperative, and central processing within the distributed architecture in Figure 2 allowing for shifting the BH traffic flexibly within the network.

By combining for each RE the N_T transmit signals of all N_{UE} users into the signal vector $\boldsymbol{x}$, the receive signal vector at RAP j is given by $\boldsymbol{y}_j = \boldsymbol{H}_j\boldsymbol{x} + \boldsymbol{n}_j$ with $\boldsymbol{H}_j$ denoting the effective channel matrix and $\boldsymbol{n}_j$ representing the additive noise vector. In case of a fully centralized solution, the receive signals $\boldsymbol{y}_j$ of all N_{RAP} RAPs have to be forwarded to the central processing node and can be collected into the receive signal vector $\boldsymbol{y} = [\boldsymbol{y}_1^H \cdots \boldsymbol{y}_{N_{RAP}}^H]^H = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}$, where $\boldsymbol{H}$ and $\boldsymbol{n}$ denote the stacked channel matrix and the stacked noise vector. The solution of a centralized least squares (LS) problem

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmin}} \| \boldsymbol{y} - \boldsymbol{H}\boldsymbol{x} \|^2 \tag{5}$$



[FIG7] The BER over BH overhead at SNR of 10 dB.

is given by $\hat{\boldsymbol{x}} = \boldsymbol{H}^+\boldsymbol{y}$ with the Moore–Penrose pseudo-inverse $\boldsymbol{H}^+ = (\boldsymbol{H}^H\boldsymbol{H})^{-1}\boldsymbol{H}^H$. For a distributed calculation of this central solution, local estimates $\tilde{\boldsymbol{x}}_j$ per node j are introduced to reformulate the LS problem by a set of local optimization problems

$$\hat{\boldsymbol{x}}_j = \underset{\{\boldsymbol{x}_j | j \in \mathcal{J}\}}{\operatorname{argmin}} \sum_{j=1}^{N_{RAP}} \| \boldsymbol{y}_j - \boldsymbol{H}_j\boldsymbol{x}_j \|^2 \tag{6a}$$

$$\text{s.t.} \quad \boldsymbol{x}_j = \boldsymbol{x}_i \quad \forall j \in \mathcal{J}, i \in \mathcal{N}_j, \tag{6b}$$

where $\mathcal{J}$ denotes the set of all RAPs and $\mathcal{N}_j$ the set of all RAPs connected with RAP j. The consensus constraint (6b) directly couples estimates of neighboring nodes guaranteeing that the estimates of all nodes converge to the central LS solution (5). In [37], the distributed consensus-based estimation (DiCE) algorithm has been introduced, which allows for parallel processing across the involved RAPs. Furthermore, the required information exchange is reduced by the reduced-overhead-DiCE (RO-DiCE) [39] approach and the fast-DiCE implementation improves the convergence speed [40].

Figure 7 shows the BER for uncoded binary phase shift keying transmission with $N_{UE} = 2$ users with $N_T = 2$ transmit antennas to $N_{RAP} = 4$ RAPs with $N_R = 4$ over Rayleigh-fading channels and fully connected mesh network of RAPs. It further compares the three different DiCE implementations for a fixed SNR of 10 dB versus the number of signals exchanged among the RAPs. Obviously, with an increasing number of iterations the BER performance improves at the cost of an increased communication overhead. In particular, the Fast-DiCE approach allows for a faster convergence and the RO-DiCE reduces the overhead by 60% at the same BER.

MUD imposes new challenges on signal processing within a cloud-computing environment. Among other challenges, synchronization needs to be maintained and taken into account. Furthermore, the data exchange between virtual machines needs to be orchestrated to allow for low delays during the MUD process. Scalability and resource pooling are two major advantages of cloud computing. This requires a hypervisor that takes into account requirements and constraints from the RAN functionality and distributes the work load accordingly, e.g., resources per virtual machine, assignment of users to virtual machines, mapping of communication clusters to virtual machines, and massive parallelization across multiple virtual machines and possible different hardware racks.

CONCLUSIONS

This article discussed benefits and challenges that may be implied by cloud-computing platforms on signal processing algorithms. The novel RANaaS concept was introduced, which realizes cloud technologies in 5G mobile networks and allows for a flexible functional split between RAPs and the centralized cloud-platform. This allows

for centralization benefits, but also introduces challenges due to the strict timing constraints imposed by the 3GPP LTE protocol stack. These challenges were identified and enabling technologies were discussed.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007–2013) under grant agreement number 317941. We would like to acknowledge our iJOIN colleagues' contributions, although the views expressed in this article are our own and do not necessarily represent the project.

AUTHORS

Dirk Wübben (wuebben@ant.uni-bremen.de) received the Dipl.-Ing. (FH) degree in electrical engineering from the University of Applied Science Münster, Germany, in 1998, and the Dipl.-Ing. (Uni) degree and the Dr.-Ing. degree in electrical engineering from the University of Bremen, Germany, in 2000 and 2005, respectively. In 2001, he joined the Department of Communications Engineering, University of Bremen, Germany, where he is currently a senior researcher and lecturer. His research interests include wireless communications, signal processing for multiple antenna systems, cooperative communication systems, and channel coding. He has published more than 80 papers in international journals and conference proceedings. He is a board member of the Germany Chapter of the IEEE Information Theory Society and an editorial board member of Elsevier's journal, *Physical Communication*.

Peter Rost (peter.rost@ieee.org) received his Ph.D. degree from Technische Universität Dresden, Germany, in 2009, and his M.Sc. degree from the University of Stuttgart, Germany, in 2005. From 1999 to 2002 he was with the Fraunhofer Institute for Beam and Material Technologies, Dresden, Germany, and from 2002 to 2005 he was with IBM Deutschland Entwicklung GmbH, Böblingen, Germany. In June 2005 he joined the Vodafone chair of Prof. Gerhard Fettweis at Technische Universität Dresden and focused on different aspects of relaying in the context of mobile communications systems. Since April 2010, he has been a member of the Mobile and Wireless Networks group at NEC Laboratories Europe, where he is working as a senior researcher in business unit projects, 3GPP RAN2 as active delegate, and the European Union's Seventh Framework Programme projects FLAVIA and iJOIN, which he currently manages as technical manager. He was the Technical Program Committee chair at the Spring 2013 IEEE Vehicular Technology Conference and a member of the IEEE Communications Society GLOBECOM/ICC Technical Committee (GITC) and *IEEE Transactions on Wireless Communications* Executive Editorial Committee. He has published more than 30 scientific publications and authored 18 patents and patent applications.

Jens Bartelt (Jens.Bartelt@tu-dresden.de) received his Dipl.-Ing. (M.S.E.E.) degree from Technische Universität Dresden, Germany, in 2012. In 2011–2012 he worked as an intern for Rohde & Schwarz in Munich. Since 2013, he has been a research associate at the Vodafone Chair Mobile Communications Systems at Technische Universität Dresden, Germany, working toward his Ph.D. degree. His research interests include cloud-based mobile networks, millimeter-wave communication, and channel coding.

Massinissa Lalam (massinissa.lalam@sagemcom.com) received his M.S. degree from the Institut National Polytechnique de Grenoble (ENSIMAG, INPG, France) in 2002. He received his Ph.D degree from the Ecole Nationale Supérieure de Télécommunications de Bretagne (Telecom Bretagne, France) in 2006. In 2007, he took a postdoctoral position with Telecom Bretagne, and from 2008 to mid-2009 he was with Orange Labs (France). He is now with Sagemcom (France) where he works on wireless (Wi-Fi) and cellular (third generation/long-term evolution) technologies. His expertise includes link- and system-level performance evaluation, heterogeneous network, network modeling, and radio resource management. He is the author or coauthor of over 15 international publications.

Valentin Savin (valentin.savin@cea.fr) received his M.S. degree in mathematics from École Normale Supérieure de Lyon, France, in 1997 and his Ph.D. degree in mathematics from J. Fourier Institute, Grenoble, France, in 2001. He also holds an M.S. degree in cryptography, security, and coding theory from the University of Grenoble 1. Since 2005, he has been with the Digital Communications Laboratory of CEA-LETI, working on the analysis and design of binary and nonbinary low-density parity check codes for physical- and upper-layer applications. He has published more than 40 papers in international journals and conference proceedings, holds six patents, and is currently participating in or coordinating several French and European Union Seventh Framework Programme research projects in information and communications technology.

Matteo Gorgolione (Matteo.Gorgoglione@cea.fr) received his M.Sc. degree in telecommunications engineering in 2009 from the Polytechnic University of Turin (Italy) and his Ph.D. degree in information and communication sciences in 2012 from the University of Cergy-Pontoise (France). His Ph.D. dissertation was conducted in collaboration with the Digital Communications Laboratory of CEA-LETI, Grenoble (France). He is currently with CEA-LETI as a postdoctoral research fellow. His research interests are mainly related to the design of binary and nonbinary low-density parity check codes for cooperative communications.

Armin Dekorsy (dekorsy@ant.uni-bremen.de) received his Dipl.-Ing. (FH) (B.Sc.) degree from Fachhochschule Konstanz, Germany, Dipl.-Ing. (M.Sc.) degree from the University of Paderborn, Germany, and Ph.D. degree from the University of Bremen, Germany, all in communications engineering. From 2000 to 2007, he worked as research engineer at Deutsche Telekom AG and as a distinguished member of technical staff at Bell Labs Europe, Lucent Technologies. In 2007 he joined Qualcomm GmbH as a European research coordinator conducting Qualcomms' internal and external European research projects like ARTIST4G, BeFemto, and WINNER+. He has held the chair position of the Department of Communications Engineering, University of Bremen, since April 2010. His current research interests include resource management, transceiver design and digital

signal processing for wireless communications systems in health care, automation, and mobile communications. He is a member of the Information Technology Society (ITG) expert committee "Information and System Theory" of the Association for Electrical, Electronic, and Information Technologies (VDE) and the IEEE Communications Society and IEEE Signal Processing Society.

Gerhard Fettweis (fettweis@ifn.et.tu-dresden.de) earned his Ph.D. degree from RWTH Aachen in 1990. After one year at IBM Research in San Jose, California, he moved to TCSI Inc., Berkeley, California. Since 1994, he has been the Vodafone chair professor at TU Dresden, Germany, where currently 20 companies from Asia/Europe/United States sponsor his research on wireless transmission and chip design. He coordinates two DFG centers, the Center for Advancing Electronics Dresden (cfAED) and Highly Adaptive Energy-Efficient Computing (HAEC), at Technische Universität Dresden, Germany. He is an IEEE Fellow, member of the German Academy of Science and Engineering (acatech), has an honorary doctorate from Tampere University of Technology, and has received multiple awards. He has helped organize IEEE conferences, was the Technical Program Committee chair of the 2009 IEEE International Conference on Communications as well as the 2012 IEEE Technology Time Machine, and general chair of the Spring 2013 IEEE Vehicular Technology Conference.

REFERENCES

[1] M. Dohler, R. Heath, A. Lozano, C. Papadias, and R. A. Valenzuela, "Is the PHY layer dead?" *IEEE Commun. Mag.*, vol. 49, no. 4, pp. 159–165, Apr. 2011.

[2] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. T. Sukhavasi, C. Patel, and S. Geirhofer, "Network densification: The dominant theme for wireless evolution into 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 82–89, Feb. 2014.

[3] 3GPP, "Small cell enhancements for E-UTRA and E-UTRAN—Physical layer aspects," Tech. Rep. TR.36.872, Sept. 2013.

[4] 3GPP, "Small cell enhancements for E-UTRA and E-UTRAN—Higher layer aspects," Tech. Rep. TR.36.842, May 2013.

[5] H. Guan, T. Kolding, and P. Merz, "Discovery of Cloud-RAN," in *Proc. Cloud-RAN Workshop*, Beijing, China, Apr. 2010.

[6] NGMN, "Suggestions on Potential Solutions to C-RAN by NGMN Alliance," Tech. Rep., Jan. 2013.

[7] D. Wake, A. Nkansah, and N. J. Gomes, "Radio over fiber link design for next generation wireless systems," *IEEE/OSA J. Lightwave Technol.*, vol. 28, no. 16, pp. 2456–2464, Aug. 2010.

[8] K. Chen, C. Cui, Y. Huang, and B. Huang, "C-RAN: A green RAN framework," in *Green Communications: Theoretical Fundamentals, Algorithms and Applications*, J. Wu, S. Rangan, and H. Zhang, Eds. Boca Raton, FL: CRC Press, pp. 279–304, 2013.

[9] G. Li, S. Zhang, X. Yang, F. Liao, T. Ngai, S. Zhang, and K. Chen, "Architecture of GPP based, scalable, large-scale C-RAN BBU pool," in *Proc. Int. Workshop Cloud Base-Station Large-Scale Cooperative Communications, IEEE GLOBECOM 2012 Workshops*, Anaheim, CA, Dec., pp. 267–272.

[10] J. Bartelt, G. Fettweis, D. Wübben, M. Boldi, and B. Melis, "Heterogeneous backhaul for cloud-based mobile networks," in *Proc. 1st Int. Workshop Cloud Technologies Energy Efficiency Mobile Communication Networks (CLEEN 2013), IEEE VTC2013-Fall Workshops*, Las Vegas, NV, Sept.

[11] P. Rost, C. J. Bernardos, A. De Domenico, M. Di Girolamo, M. Lalam, A. Maeder, D. Sabella, and D. Wübben, "Cloud technologies for flexible 5G radio access networks," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 68–76, May 2014.

[12] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[13] P. Mell and T. Grance, "The NIST definition of cloud computing," Tech. Rep., National Inst. Standards Technology, Special Publication 800-145, Sept. 2011.

[14] J. Bartelt and G. Fettweis, "Radio-over-radio: I/Q-stream backhauling for cloud-based networks via millimeter wave links," in *Proc. 1st Int. Workshop Cloud Technologies Energy Efficiency Mobile Communication Networks, IEEE GLOBECOM 2013 Workshops (IWCPM 2013)*, Atlanta, GA, Dec., pp. 778–783.

[15] J. Kempf, B. Johansson, S. Pettersson, H. Luning, and T. Nilsson, "Moving the mobile evolved packet core to the cloud," in *Proc. IEEE Int. Conf. Wireless Mobile Computing, Networking, Communications*, Barcelona, Spain, Oct. 2012, pp. 784–791.

[16] R. Irmer, H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H.-P. Mayer, L. Thiele, and V. Jungnickel, "Coordinated multipoint: Concepts, performance, and field trial results," *IEEE Commun. Mag.*, vol. 49, no. 2, pp. 102–111, Feb. 2011.

[17] U. Dötsch, M. Doll, H. P. Mayer, F. Schaich, J. Segel, and P. Sehier, "Quantitative analysis of split base station processing and determination of advantageous architectures for LTE," *Bell Labs Tech. J.*, vol. 18, no. 1, pp. 105–128, May 2013.

[18] S. Park, C.-B. Chae, and S. Bahk, "Before/after precoded massive MIMO in cloud radio access networks," *J. Commun. Netw.*, vol. 15, no. 4, pp. 398–406, Aug. 2013.

[19] CPRI. (2013, Aug.). Common public radio interface (CPRI); Interface specification (V6.0). Tech. Rep. [Online]. Available: http://www.cpri.info/

[20] 3GPP, "Radio access network; evolved universal terrestrial radio access; (E-UTRA); physical channels and modulation (Release 10)," Tech. Rep. TS.36.211, Dec. 2012.

[21] R. Visoz, A.O. Berthet, and M. Lalam, "Semi-analytical performance prediction methods for iterative MMSE-IC multiuser MIMO joint decoding," *IEEE Trans. Commun.*, vol. 58, no. 9, pp. 2576–2589, Sept. 2011.

[22] F. Kienle, N. Wehn, and H. Meyr, "On complexity, energy- and implementation-efficiency of channel decoders," *IEEE Trans. Commun.*, vol. 59, no. 12, pp. 3301–3310, Dec. 2011.

[23] R. G. Gallager, *Low-Density Parity Check Codes* (Research Monograph Series). Cambridge, U.K.: MIT Press, 1963.

[24] C. Berrou and A. Glavieux, "Near optimum error correcting coding and decoding: Turbo-codes," *IEEE Trans. Commun.*, vol. 44, no. 10, pp. 1261–1271, Oct. 1996.

[25] M. M. Mansour and N. R. Shanbhag, "High-throughput LDPC decoders," *IEEE Trans. Very Large Scale Integration (VLSI) Syst.*, vol. 11, no. 6, pp. 976–996, Dec. 2003.

[26] C. Zhang, Z. Wang, J. Sha, L. Li, and J. Lin, "Flexible LDPC decoder design for multigigabit-per-second applications," *IEEE Trans. Circuits Syst. I*, vol. 57, no. 1, pp. 116–124, Jan. 2010.

[27] G. Falcao, L. Sousa, and V. Silva, "Massively LDPC decoding on multicore architectures," *IEEE Trans. Parallel Distrib. Syst.*, vol. 22, no. 2, pp. 309–322, Feb. 2011.

[28] S. Seo, T. Mudge, Y. Zhu, and C. Chakrabarti, "Design and analysis of LDPC decoders for software defined radio," in *Proc. IEEE Workshop Signal Processing Systems*, Shanghai, China, Oct. 2007, pp. 201–215.

[29] A. Diavastos, P. Petrides, G. Falcao, and P. Trancoso, "LDPC decoding on the Intel SCC," in *Proc. IEEE 20th Euromicro Int. Conf. Parallel, Distributed Network-Based Processing (PDP)*, Garching, Germany, Feb. 2012, pp. 57–65.

[30] G. Falcao, V. Silva, L. Sousa, and J. Marinho, "High coded data rate and multicodeword WiMAX LDPC decoding on Cell/BE," *Electron. Lett.*, vol. 44, no. 24, pp. 1415–1417, Feb. 2008.

[31] M. Wu, G. Wang, B. Yin, C. Studer, and J. R. Cavallaro, "HSPA+/LTE-A turbo decoder on GPU and multicore CPU," in *Proc. 47th IEEE Asilomar Conf. Signals, Systems, Computers (ASILOMAR)*, Pacific Grove, CA, Nov. 2013, pp. 824–828.

[32] P. Grover, K. A. Woyach, and A. Sahai, "Towards a communication-theoretic understanding of system-level power consumption," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 8, pp. 1744–1755, Sept. 2011.

[33] iJOIN. (2013, Nov.). State-of-the-art of and promising candidates for PHY layer approaches on access and backhaul network. Tech. Rep. [Online]. Available: www.ict-ijoin.eu/wp-content/uploads/2014/01/D2.1.pdf

[34] G. Caire and R. Müller, "The optimal received power distribution of IC-based iterative multiuser joint decoders," in *Proc. 39th Annu. Allerton Conf. Communications, Control Computing*, Monticello, IL, Oct. 2001.

[35] R4-132017, "WF on NAICS receiver terminology," 3GPP TSG-RAN WG4# 66bis, Apr. 2013.

[36] H. Zhu, A. Cano, and G. B. Giannakis, "Distributed consensus-based demodulation: Algorithms and error analysis," *IEEE Trans. Wireless Commun.*, vol. 9, no. 6, pp. 2044–2054, June 2010.

[37] H. Paul, J. Fliege, and A. Dekorsy, "In-network-processing: Distributed consensus-based linear estimation," *IEEE Commun. Lett.*, vol. 17, no. 1, pp. 59–62, Jan. 2013.

[38] H. Paul, B.-S. Shin, D. Wübben, and A. Dekorsy, "In-network-processing for small cell cooperation in dense networks," in *Proc. 1st Int. Workshop Cloud Technologies Energy Efficiency Mobile Communication Networks, IEEE VTC2013-Fall Workshops (CLEEN 2013)*, Las Vegas, NV, Sept.

[39] B.-S. Shin, H. Paul, D. Wübben, and A. Dekorsy, "Reduced overhead distributed consensus-based estimation algorithm," in *Proc. 1st Int. Workshop Cloud Technologies Energy Efficiency Mobile Communication Networks, IEEE GLOBECOM Workshops 2013 (IWCPM 2013)*, Atlanta, GA, Dec., pp. 784–789.

[40] G. Xu, H. Paul, D. Wübben, and A. Dekorsy, "Fast distributed consensus-based estimation for cooperative wireless sensor networks," in *Proc. 18th Int. ITG Workshop Smart Antennas (WSA 2014)*, Erlangen, Germany, Mar.

[SP]

Sergio Barbarossa, Stefania Sardellitti, and Paolo Di Lorenzo

Communicating While Computing

Distributed mobile cloud computing over 5G heterogeneous networks

Current estimates of mobile data traffic in the years to come foresee a 1,000 × increase of mobile data traffic in 2020 with respect to 2010, or, equivalently, a doubling of mobile data traffic every year. This unprecedented growth demands a significant increase of wireless network capacity. Even if the current evolution of fourth-generation (4G) systems and, in particular, the advancements of the long-term evolution (LTE) standardization process foresees a significant capacity improvement with respect to third-generation (3G) systems, the European Telecommunications Standards Institute (ETSI) has established a roadmap toward the fifth-generation (5G) system, with the aim of deploying a commercial system by the year 2020 [1]. The European Project named "Mobile and Wireless Communications Enablers for the 2020 Information Society" (METIS), launched in 2012, represents one of the first international and large-scale research projects on fifth generation (5G) [2]. In parallel with this unparalleled growth of data traffic, our everyday life experience shows an increasing habit to run a plethora of applications specifically devised for mobile devices, (smartphones, tablets, laptops)for entertainment, health care, business, social networking, traveling, news, etc. However, the spectacular growth in wireless traffic generated by this lifestyle is not matched with a parallel improvement on mobile handsets' batteries, whose lifetime is not improving at the same pace [3]. This determines a widening gap between the energy required to run sophisticated applications and the energy available on the mobile handset. A possible way to overcome this obstacle is to enable the mobile devices, whenever possible and convenient, to offload their most energy-consuming tasks to nearby fixed servers. This strategy has been studied for a long time and is reported in the literature under different names, such as *cyberforaging* [4] or *computation offloading* [5], [6]. In recent years, a strong impulse to computation offloading has come through cloud computing (CC), which enables the users to utilize resources on demand. The resources made available by a cloud service provider are: 1) infrastructures, such as network devices, storage, servers, etc., 2) platforms, such as operating systems, offering an integrated environment for developing and testing custom applications, and 3) software, in the form of application programs. These three kinds of services are labeled, respectively, as *infrastructure as a service*, *platform as a service*, and *software as a service*. In particular, one of the key features of CC is virtualization, which makes it possible to run multiple operating systems and multiple applications over the same machine (or set of machines), while guaranteeing isolation and protection of the programs and their data. Through virtualization, the number of virtual machines (VMs) can scale on demand, thus improving the overall system computational

THE 5G REVOLUTION

©ISTOCKPHOTO.COM/ZONADEARTE

Digital Object Identifier 10.1109/MSP.2014.2334709
Date of publication: 15 October 2014

efficiency. Mobile CC (MCC) is a specific case of CC where the user accesses the cloud services through a mobile handset [5]. The major limitations of today's MCC are the energy consumption associated to the radio access and the latency experienced in reaching the cloud provider through a wide area network (WAN). Mobile users located at the edge of macrocellular networks are particularly disadvantaged in terms of power consumption and, furthermore, it is very difficult to control latency over a WAN. As pointed out in [7]–[9], humans are acutely sensitive to delay and jitter: as latency increases, interactive response suffers. Since the interaction times foreseen in 5G systems, in particular in the so-called tactile Internet [10], are quite small (in the order of milliseconds), a strict latency control must be somehow incorporated in near future MCC. Meeting this constraint requires a deep rethinking of the overall service chain, from the physical layer up to virtualization.

Within this framework, the goal of this article is to review first a series of offloading mechanisms and then to provide a mathematical formulation of the computation offloading problem aimed at optimizing the communication and computation resources jointly, posing a strict attention to latency and energy constraints. Wherever possible, we try to emphasize those features of 5G systems that can help meet the strict latency constraints while keeping the energy consumption at a minimum level. Signal processing can play a significant role in mobile cloud computing from different perspectives: from a rigorous mathematical formulation and efficient solution of the resource allocation problem to the development of applications specifically built to take full advantage of computation offloading, etc.

COMPUTATION OFFLOADING FOR MOBILE CLOUD COMPUTING

MCC can bring the following advantages:

- prolong battery lifetime, by offloading energy-consuming tasks from the mobile handset to the cloud
- enable mobile devices to run sophisticated applications and provide significantly higher data storage capabilities
- improve reliability, since data can be stored and backed up from the mobile device to a set of reliable fixed devices specifically designed for storage purposes.

These advantages come on top of the usual advantages of CC, specifically making resources, either storage or applications, available on demand without the need for the user to own sophisticated devices or software tools.

Offloading strategies may be classified in different ways, depending on what aspects are identified as most relevant. From the point of view of the protocols used to handle the exchange of data between mobile device and server, three classes are identified [5]: client-server communication, virtualization, and mobile agents. (A mobile agent is a program able to migrate across a network carrying its own code and execution state.) The client–server protocol requires the services to be preinstalled in the participating devices. Examples of this class are: Spectra [11], Chroma [12], and Cuckoo [13]. The second class of methods requires the instantiation of VMs on the server. Virtualization ensures a relatively secure execution as a VM encapsulates and separates the guest software from the host software. Examples of methods based on virtualization are Mobile Assistance Using Infrastructure (MAUI) [7], CloneCloud [14], and MobiCloud [15]. Finally, methods based on mobile agents use a mobile approach to partition and distribute jobs and are more suitable for disconnected operations typical of wireless access. An example of this class is Scavenger [16].

Offloading a computation does not necessarily imply transferring all the program execution to the remote server. Typically, a program is first subdivided into modules, some of which need to be run on the mobile handset, such as, for instance, all modules controlling the input/output peripherals. For the rest of the modules, a decision has to be taken on what is more appropriate to offload. Offloading can be either static or dynamic. Static offloading means that the program partitioning is given before execution, and the decision about what modules to transfer is taken, once for all, at the beginning of the execution. Examples of static offloading are Spectra [11], [17], [18] and Chroma [12]. In contrast, in dynamic offloading, the decision on whether and what to offload is taken at run-time based on current conditions. Dynamic offloading is, in principle, more efficient than static offloading; however, it induces more overhead on the system relating to latency, profiling, and run-time decision making. Examples of dynamic offloading are [19]–[22].

Offloading typically requires code partitioning [7], [23], [24], aimed to decide which parts of the code should run locally and which parts should be offloaded, depending on contextual parameters, such as computational intensity of each module, size of the program state to be exchanged to transfer the execution from one site to the other, battery level, delay constraints, channel state, and so on. MAUI [7] is an example of an offloading method aimed at selecting what program modules to offload to minimize energy consumption at the mobile terminal. The approach is based on the so-called call graph representation of a program. A call graph is a representation that models the relations between the modules (procedures) of a computer program in the form of a directed graph $\mathcal{G} = (V, E)$, where each vertex $v \in V$ represents a procedure in the call stack, and each directed edge $e = (u, v)$ represents the invocation of procedure v from procedure u. The call graph includes also auxiliary information concerning, for instance, the number of instructions within each module and the amount of data exchanged among modules. For nonrecursive languages with reasonable assumptions on the program structure [25], the call graph is a directed, acyclic graph. Given the call graph of the application, MAUI collects information about energy consumption and data transfer requirements and solves an integer linear program to determine which modules are more suitable for being offloaded. In this way, MAUI does not offload the whole application, but only the most energy-consuming modules. The critical aspect is the prediction of energy consumption. MAUI saves information about past offloaded methods and uses online profiling to create an energy consumption model. When new offloading requests are received, MAUI uses history data to predict the execution

time of the task. Of course, the effectiveness of this offloading scheme depends on the accuracy achievable in the prediction of energy consumption and time execution. ThinkAir is an alternative strategy supporting method-level offloading to a smartphone clone executing in the cloud [26]. ThinkAir uses three profilers at the mobile side, monitoring: the energy consumption of the device hardware; the program parameters, such as execution time, acquired memory, number of instructions; and communication related parameters, such as bandwidth, connectivity, and delay.

Besides computational aspects, there are two major issues about offloading associated with radio access: power consumption and latency. These are indeed two of the major bottlenecks in the deployment of an effective MCC in current cellular networks. In macrocellular systems, the power spent from mobile users, especially those located in the edge of the cell, may be significant. In some cases, this large transmit power may nullify all potential benefits in terms of energy saving. A possible way to reduce this power consumption is to bring computational resources closer to the mobile user. This idea was put forward in [8], where the concept of a cloudlet was introduced. In such a case, the mobile handset offloads its workload to a local cloudlet consisting of a set of multicore computers connected to the remote cloud server. The storage and computational capabilities of the cloudlet are much smaller than those available at the cloud server, but, at the same time, installing a cloudlet is much less expensive than installing a cloud server. The main advantage of this solution is scalability—the powerful cloud resources are used only when really necessary, otherwise computation is offloaded to a cloudlet. The radio access to the cloudlet could be through Wi-Fi. The idea of bringing cloud services closer to the mobile users has been further pushed in the current European Union project named "Distributed Computing, Storage, and Radio Resource Allocation over Cooperative Femtocells" (TROPIC) [30], where it was proposed to endow small-cell base stations with additional, albeit limited, cloud functionalities. The new base stations are denoted, in LTE terminology, small-cell cloud enhanced e Node B (SCceNB). In this way, a mobile user is able to find a radio access point within a short distance, enabling it to access cloud functionalities. The scenario is depicted in Figure 1, where the SCceNB's are interconnected through the so-called femtoclouds, i.e., small clouds with intermediate storage and computation capabilities. The femtoclouds manage the allocation of VMs to the users accessing through the associated base stations. The femtoclouds are then interconnected with each other and to the cloud provider. Whenever the users' request can be met by the local femtocloud, everything is performed locally. Otherwise, the SCceNB may ask the intervention of the cloud server through high-capacity wired links. In this way, both radio and computational resources are brought closer to the user, thus improving scalability in both radio and computation aspects. The radio access is based on LTE, which yields some advantages over Wi-Fi: 1) it provides a single technology solution for offloading, with no need to switch from 3G/4G to Wi-Fi and vice versa, and 2) it provides QoS guarantees.

Bringing resources closer to the user improves not only power consumption at the terminal side but also the other major issue, latency. More specifically, the latency ℓ contains three terms:

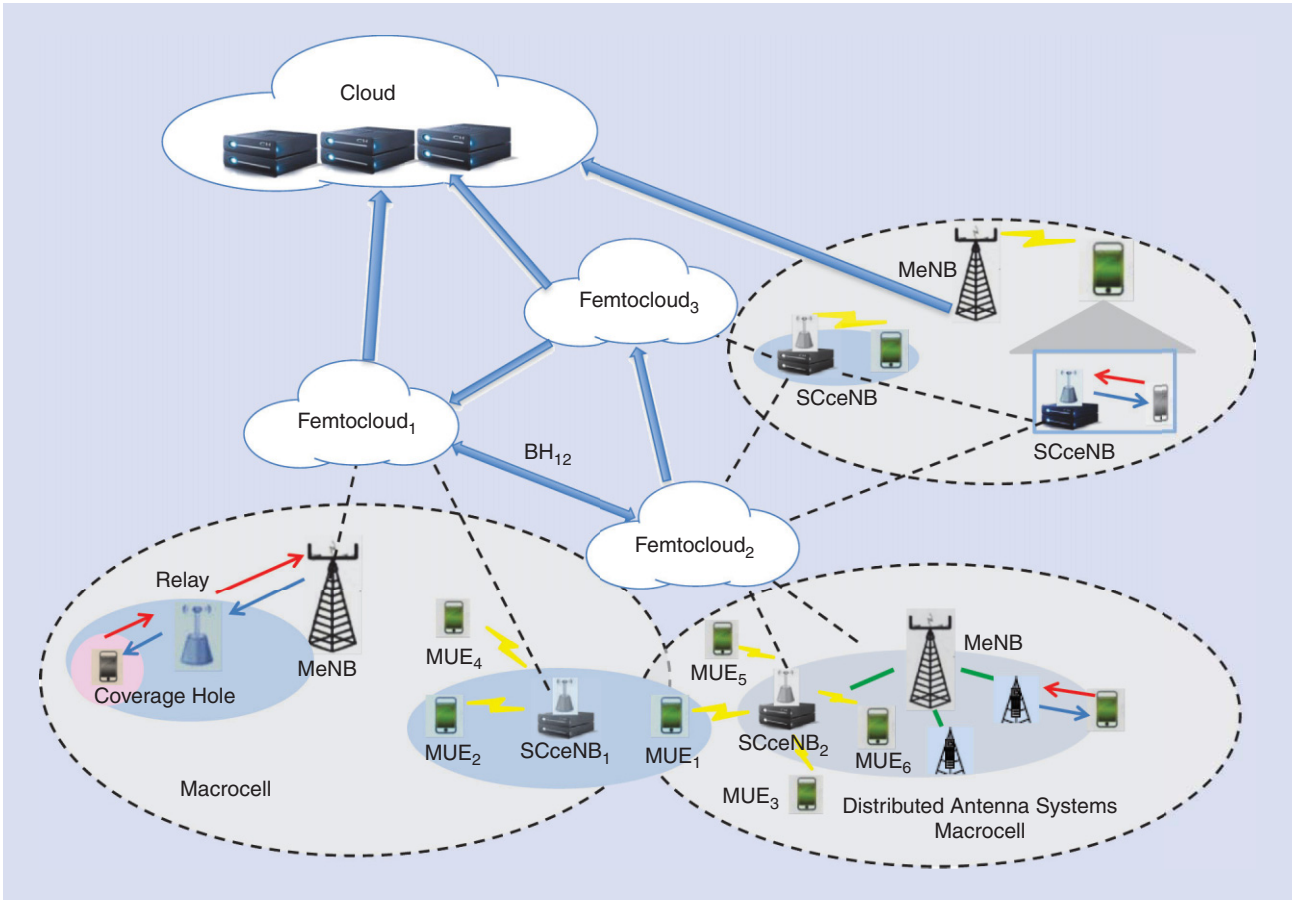
$$\ell = \Delta_T + \Delta_{\text{exe}}^{\text{rem}} + \Delta_R, \tag{1}$$

where Δ_T is the time needed to send the information necessary to transfer the program execution from the mobile device to the cloud, $\Delta_{\text{exe}}^{\text{rem}}$ is the time necessary to run the program at the remote side (cloud), and Δ_R is the time necessary to the cloud to send the result back to the mobile unit. More specifically, the time Δ_T includes the time for the mobile unit to reach the radio access point plus the time for this information to travel from the access point to the cloud server through the backhaul. This second term depends on the technology involved in the backhaul: the latency over a fiber optic cable may be negligible, but the latency over an asymmetric digital subscriber line (ADSL) link may not, depending on traffic. The term $\Delta_{\text{exe}}^{\text{rem}}$ depends on the computational load and on how many computational resources [e.g., VM, central processing unit (CPU) cycles, etc.] are assigned to the user. This depends on the server capabilities, but also on the number of users asking for the service. Equation (1), in its simplicity, shows how, in MCC, the issues related to communication and computation are strictly related to each other. In fact, the effectiveness of an offloading scheme depends on both the radio access and computational aspects. Very schematically, every application, or part of it, is characterized by an input, a number of instructions to be executed, and an output. To offload a computation, it is necessary to transfer the program state (input and state variables) from the mobile user to the server (cloud). Clearly, the applications more suitable for offloading are those characterized by a limited input (or state) size and a high number of instructions to be executed. An example of this class of applications is a chess game. On the contrary, offloading may not be convenient for those applications where it is necessary to transmit a large set of data and the computational load is not so heavy. These qualitative statements will be corroborated by quantitative evaluations in the ensuing sections. But before delving into the mathematical formulation, it is worth outlining how the 5G revolution can have an impact on MCC.

MOBILE CLOUD COMPUTING IN 5G SYSTEMS

Even if 5G is still undefined, some key features have already been identified; see, e.g., [1]. A nonexhaustive list of some of the key features, as relevant to computation offloading, is reported below:

- (*Ultra*)*dense cell deployment:* The deployment of small-cell networks is already part of 4G evolution, but network "densification" is going to play a major role in 5G systems [29]. Small-cell networks include femtocell networks, specifically devised to cover indoor environments or outdoor cells served by small base stations placed on lamp posts or on the facades of a building. Small-cell networks, covering areas with a radius in the order of a few tens of meters, are going to coexist with conventional macrocell networks. With respect to MCC, a dense deployment of small-cell base stations carries two advantages: 1) it reduces the transmit power necessary for



[FIG1] The distributed cloud scenario.

computation offloading and the latency over the wireless access channel and 2) it increases the probability for the mobile handset to find an access point within a short range; furthermore, if some small-cell base stations are endowed with additional cloud functionalities, scalability improves along both radio and computational resources.

■ *Millimeter-wave links:* The usage of wideband links, with very high capacity and directivity, provides an effective way for the radio access points to forward the users' offloading requests to the cloud with reduced latency, thus overcoming the limitations of today's ADSL backhaul links, often used in femtocell networks; furthermore, these high-capacity links facilitate cooperation among small-cell base stations at both radio and computing levels.

■ *Massive multiple-input, multiple-output (MIMO):* MIMO transceivers improve spectral efficiency, thus reducing the time necessary to transfer the program execution from the mobile site to the cloud; furthermore, the usage of extensive beamforming allows an efficient management of intercell interference through adaptive null steering.

■ *Multicell cooperation:* Since computation offloading involves both communication and computation aspects, the cooperation among cells is fundamental to distribute radio access and computing requests in the most effective way; in particular, cooperation may occur at the radio level to reduce interference and, at the application level to implement a distributed cloud capability [28].

■ *Cognitive radio:* The incorporation of cognitive radio capabilities helps to improve the overall system efficiency; the specific novelty brought by cognitive techniques in the MCC context is that the cognition activity can involve both radio aspects as well as learning and adaptation of appropriate energy consumption models, which play a basic role in devising the most appropriate offloading strategy across a network of computing resources; furthermore, exploiting the cooperation among base stations, it is possible to implement collaborative sensing techniques to augment the learning capabilities of the individual access points, in terms of channel sensing and energy consumption models.

■ *Quality of experience (QoE) versus quality of service (QoS):* A design driven by a user's QoE implies a system approach that does not consider radio or networking aspects separated from the application requirements; in this sense, this change of perspective matches perfectly with the synergic approach that is going to be most effective for MCC.

OPTIMAL ALLOCATION OF RADIO RESOURCES IN A SINGLE-USER SCENARIO

For simplicity, we start with a single user case before moving to the multiser/multicell scenario. We consider alternative optimization criteria to have a set of strategies to tackle alternative application requirements and user needs.

MINIMUM TRANSMIT ENERGY UNDER COMPUTATIONAL CONSTRAINT

The first criterion we analyze is the minimization of the energy consumption at the mobile side under a computational rate constraint, besides the usual power budget constraint. We assume here that the computation has to take place within a time window of T seconds. The mobile decides to carry out the computations locally or to offload computations to the cloud depending on which strategy requires less energy consumption at the mobile side. In case of offloading, the user needs to send all necessary data (input, program state, etc.) to the cloud. This involves a time τ. The other system parameters are: f_{loc} is the local computational rate (CPU cycles/s), f_S is the server computational rate (CPU cycles/s), B is the bandwidth of the link from MUE to SCceNB, p_{proc} is the power spent for local processing, N is the dimension (number of bits) of the program state, and λ is the computational rate (CPU cycles/s) required to run the application while meeting the user's requests.

The degrees of freedom are the following: $d \in \{0, 1\}$ is the decision variable, set to 0 if the program is executed at the mobile side, or 1 otherwise; $p \in [0, P_T]$ is the power spent for transmitting the program state from mobile to server; $\tau \in [0, T]$ is the duration of the interval necessary for transmitting the program state to the server, necessary to enable the program execution transfer.

Offloading takes place if the energy $p\tau$ consumed by the terminal device for offloading is less than the energy $p_{\text{proc}}T$ needed for carrying out the computations locally. In case of offloading, the effective computing rate, taking into account the time necessary to transfer the program execution to the cloud, is $\tilde{f} = (1 - \tau/T) f_S$. The decision variable d is explicitly related to p and τ through

$$d = u(p_{\text{proc}}T - p\tau), \tag{2}$$

where $u(\cdot)$ is the unit step function. Assuming adaptive modulation, the time τ necessary to send the N bits encoding the program state across a channel of bandwidth B is related to p as

$$\tau = \frac{N}{B\log(1 + \alpha p)} \tag{3}$$

with $\alpha = |h|^2/\Gamma(\text{BER})\sigma_n^2$, where h is the channel coefficient, $\Gamma(\text{BER}) = -(2\log(5\,\text{BER})/3)$ is the signal-to-noise ration (SNR) margin introduced to meet a target BER, and σ_n^2 is the receiver noise power. Note that the gap factor $\Gamma(\text{BER})$ is valid under the assumption $\Gamma(\text{BER}) > 0$, i.e., BER $< 1/5$. Exploiting the relations among the free variables p, τ, and d, the objective function can be expressed in terms of the single variable τ and the optimization problem can be formulated as:

$$\min_{\tau} \min\left[p_{\text{proc}}T, \frac{\tau}{\alpha}(2^{N/B\tau} - 1)\right] \tag{4}$$

subject to (s.t.)

$$C.1\ \frac{N}{B\log(1 + \alpha P_T)} \le \tau \le T$$

$$C.2\ f_{\text{loc}} + [f_S(1 - \tau/T) - f_{\text{loc}}] \cdot u(p_{\text{proc}}T - p\tau) \ge \lambda.$$

After some algebraic manipulations, it is possible to show that this problem is feasible if $f_S \ge \lambda$ and the (equivalent) channel coefficient α exceeds a minimum value, i.e., $\alpha > \alpha_{\min}$, where

$$\alpha_{\min} = \max\left[\frac{1}{P_T}, \frac{(1 - \frac{\lambda}{f_S})}{p_{\text{proc}}}\right](2^{\frac{N}{BT(1-\lambda/f_S)}} - 1). \tag{5}$$

This last condition states, in closed form, that offloading can take place only if the channel is sufficiently good, as expected. The interesting point is that the minimum channel value is dictated by parameters related to both radio and computational parameters. The previous problem is convex and, if the feasible set is nonempty, the optimal transmit power can be expressed in closed form as

$$p = \frac{1}{\alpha}(2^{\frac{N}{BT(1-\lambda/f_S)}} - 1). \tag{6}$$

Clearly, since the wireless channel is random, there is a nonnull probability that offloading takes place or not, depending on both channel status and computational requests. The consequence is that the real computing rate experienced over a fading channel is going to depend on the channel statistics and typically will be lower than the rate achievable under ideal channel conditions. Interestingly, building on previous expressions, we can derive the equivalent computing rate λ^*, to be asked by the mobile device to ensure the desired rate λ, in the presence of fading, provided that the fading statistics are known. More specifically, the probability of offloading is simply

$$\mathcal{P} := \text{Prob}\{\alpha > \alpha_{\min}\} = 1 - D_A(\alpha_{\min}), \tag{7}$$

where $D_A(\alpha)$ denotes the cumulative distribution function of α and $\alpha_{\min}$ is given in (5), with λ^* instead of λ. The expected value of the rate is then

$$f_{\text{ave}} = (1 - \mathcal{P}(\lambda^*))f_{\text{loc}} + \mathcal{P}(\lambda^*)\lambda^*, \tag{8}$$

where we made explicit the dependence of $\mathcal{P}$ from λ^* (through $\alpha_{\min}$). Imposing this average rate to be equal to the target rate λ, we end up with a nonlinear equation in λ^*. This equation admits a unique solution. As a numerical example, Figure 2(a) shows λ^* as a function of the distance between mobile user and base station for different antenna configurations. The parameters of the simulation are $f_S = 10^{10}$, $f_{\text{loc}} = 10^7$, $p_{\text{proc}} = 0.1$, $P_T = 0.1$, $N = 5 \times 10^3$, $B = 2 \times 10^6$, $T = 10^{-2}$, and $\lambda = 10^8$. The channels are generated as statistically independent Rayleigh fading channels. As expected, at short distances, λ^* tends to coincide

with λ, because channel attenuation is negligible and offloading occurs most of the times. However, as the distance exceeds a certain value, λ^* starts increasing considerably to compensate for the missing opportunities to offload. In parallel, Figure 2(b) shows the average energy spent for offloading as a function of the distance between mobile user and base station, for different antenna configurations. As expected, the energy increases at larger distance until reaching a constant value dictated by the energy required to run the application locally. It is interesting to see how MIMO transceivers yield a larger saving and then, ultimately, widen the area over which offloading is beneficial.

MINIMUM TRANSMIT POWER UNDER DELAY CONSTRAINT

In this section, we assume all radio equipment to be equipped with multiple antennas and the action available at the mobile user to select the precoding matrix to satisfy some optimality criterion for offloading. We use the following symbols: n_T and n_R are the number of antennas at the transmit and receive sides, respectively; L is the maximum latency limit; P_T is the maximum transmit power budget; w is the number of CPU cycles to be executed; N is the number of bits to be transmitted in case of offloading; $\mathcal{E}_{\text{loc}}$ is the energy spent to run the program at the mobile side; $\mathbf{Q}$ is the covariance matrix of the transmitted symbols; $\boldsymbol{H}$ is the $n_R \times n_T$ channel matrix between mobile user and base station; and $\boldsymbol{R}_n$ is the disturbance (interference plus noise) covariance matrix.

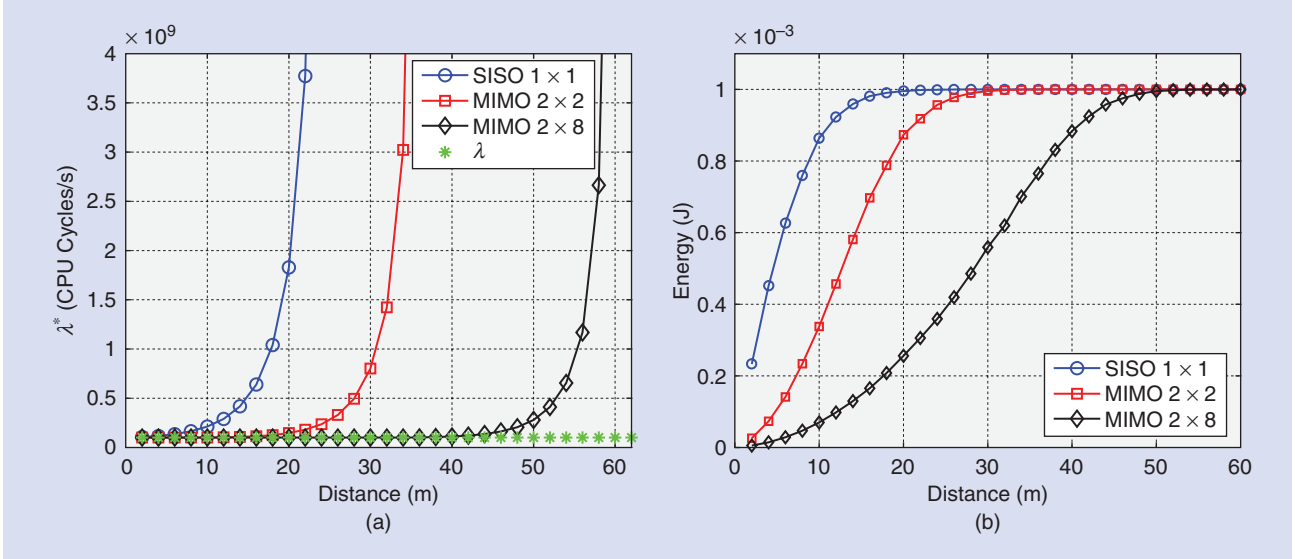
The mobile user offloads its computations if the energy spent for offloading is less than the energy necessary for processing the data locally. The goal of the optimization is then to find the optimal precoding matrix (equivalently, the covariance matrix of the transmitted symbols) to minimize power consumption, s.t. the following constraints: 1) latency constraint, 2) energy for offloading less than energy to be spent for local processing, and 3) transmit power less than available power budget. The problem can be formulated as:

$$\min_{\mathbf{Q}\succeq\mathbf{0}} \ \text{trace}(\mathbf{Q}) \quad \text{s.t.} \quad \left.\begin{array}{l} \text{i)} \ \dfrac{N}{B\log_2\det(\mathbf{I}+\mathbf{H}\mathbf{Q}\mathbf{H}^H\mathbf{R}_n^{-1})}+\dfrac{w}{f_S}+\Delta_R \le L \\ \text{ii)} \ \dfrac{\text{tr}(\mathbf{Q})N}{B\log_2\det(\mathbf{I}+\mathbf{H}\mathbf{Q}\mathbf{H}^H\mathbf{R}_n^{-1})} \le \mathcal{E}_{\text{loc}} \\ \text{iii)} \ \text{tr}(\mathbf{Q}) \le P_T, \quad \mathbf{Q}\succeq\mathbf{0} \end{array}\right\} \triangleq \mathcal{X} \qquad (\mathcal{P}.1) \tag{9}$$

where the three line constraints reflect the constraint list mentioned above; $N/B\log_2\det(\mathbf{I}+\mathbf{H}\mathbf{Q}\mathbf{H}^H\mathbf{R}_n^{-1})$ is the time necessary to transmit N bits over a bandwidth B using optimal coding. The symbol $\mathcal{X}$ denotes the feasible set: If $\mathcal{X}$ is empty, offloading is not convenient or impossible to carry out within the user's requirements and processing is performed at the mobile device; if $\mathcal{X}$ is nonempty, the previous problem is convex, offloading takes place and the optimal precoding matrix can be expressed in closed form, as proved in [31]. In particular, denoting with $\mathbf{H}^H\mathbf{R}_n^{-1}\mathbf{H}=\mathbf{U}\mathbf{D}\mathbf{U}^H$ the eigendecomposition of the composite channel matrix, weighted with the inverse of the disturbance covariance matrix $\mathbf{R}_n$, the optimal covariance matrix $\mathbf{Q}$ is given by

$$\mathbf{Q}=\mathbf{U}(\alpha\mathbf{I}-\mathbf{D}^{-1})^+\mathbf{U}^H, \tag{10}$$

where $\alpha=(\beta\mathcal{E}_{\text{loc}}+\lambda(L-w/f_S-\Delta_R))/(\mu+\beta c+1)$ is a positive constant, a function of the three Lagrangian multipliers β, λ, and μ is associated with the three constraints in (9) with $c=N/B$. Interestingly, the solution (10) has the well known "water-filling"



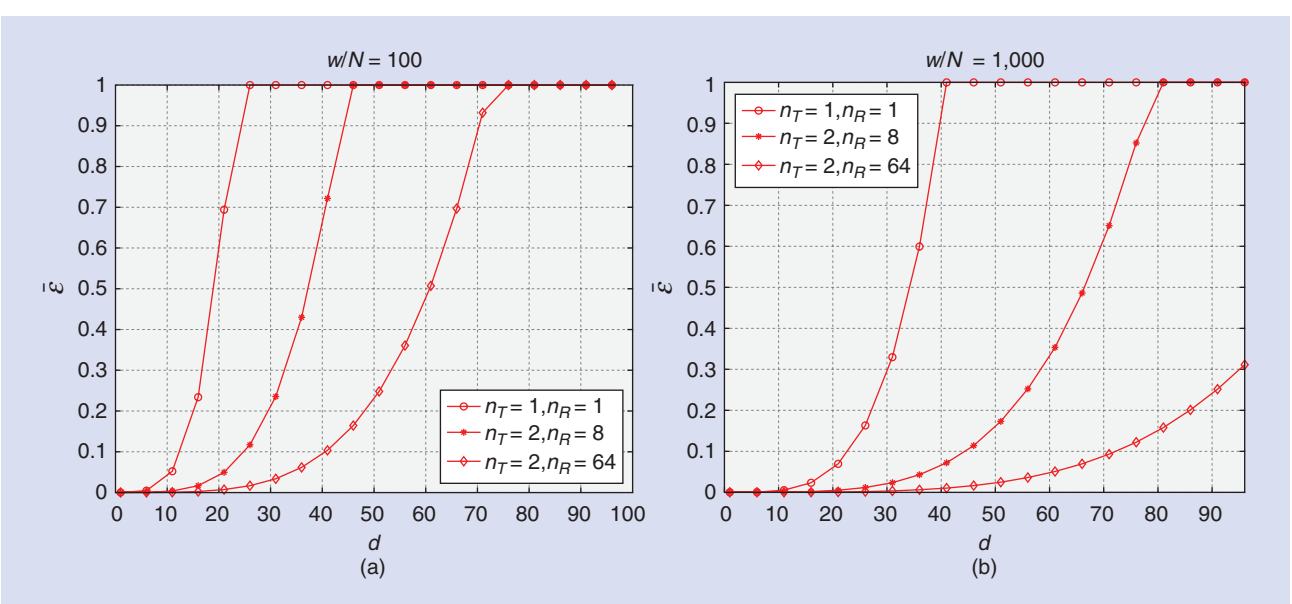
[FIG2] (a) The equivalent rate λ^* versus distance between mobile user and access point for different communication strategies. (b) The average energy spent for offloading versus distance between mobile user and access point for different communication strategies.

form but with a water level depending on parameters specifying the computational features, e.g., a number of CPU cycles, size of program state, energy necessary for local processing. As a numerical example, in Figure 3, we report the average energy spent with or without offloading (normalized to the energy to be spent locally) as a function of the distance between a mobile user and access point for different MIMO configurations and different applications. The average has been evaluated over 100 independent realizations of Rayleigh fading channels. The difference between the applications is captured by the ratio w/N between the number w of CPU cycles to be executed to run the program and the number of bits N to be transmitted to transfer the program execution. The maximum level in each figure is the energy spent for local processing at the mobile side (all energy values are normalized with respect to this value, so that it is easier to read offloading gains in percentage). From Figure 3, we can observe how offloading is advantageous at short distances and, depending on the distance, it can also bring substantial savings. Furthermore, it is also evident how MIMO transceivers enlarge the area over which offloading is advantageous. Comparing (a) and (b), we can observe that, as expected, offloading is more advantageous for program modules characterized by a higher ratio w/N, i.e., a higher computational load (number of CPU cycles), for a number of bits to be exchanged. These curves are simple examples of how MCC can benefit from a dense deployment of base stations and from massive MIMO, as foreseen in 5G, because having proximity access facilitates offloading and then enables a higher energy saving at the mobile terminal.

JOINT OPTIMIZATION OF PROGRAM PARTITIONING AND RADIO RESOURCE ALLOCATION

So far, we have considered a single module to offload. In this section, we consider a more structured program, represented as a call graph, as in MAUI, and we illustrate an approach that optimizes code partitioning and radio resource allocation jointly. By code partitioning, we mean the decision about which modules are to be offloaded and which ones are to be executed locally. The difference with respect to MAUI is that, whereas in MAUI the energy consumption and latencies are supposed to be given (or estimated), here we optimize jointly across code partitioning and radio resource allocation. In our computation offloading framework, we label each vertex $v \in V$ of the call graph with the energy E_v^l it takes to execute the procedure locally, and with the overall number of instructions w_v (CPU cycles), of which the procedure is composed. At the same time, each edge $e = (u, v)$ is characterized by a label describing the number of bits $N_{u,v}$ representing the size of the program state that needs to be exchanged to transfer the execution from node u to node v. In general, some procedures cannot be offloaded, such as the program modules controlling the user interface or the interaction with input/output devices. The set of procedures that must be executed locally is denoted by V_l. Intuitively speaking, the modules more amenable for offloading are the ones requiring intensive computations and limited exchange of data to transfer the execution from one site to the other. Our goal now is to make this intuition the result of an optimization procedure. To this end, we formulate the offloading decision problem jointly with the selection of the transmit power and the constellation size used for transmitting the program state necessary to transfer the execution from the mobile handset to the cloud or vice versa. The objective is to minimize the energy consumption at the mobile site, under power budget and latency constraints.

Let us indicate with I_v the indicator variable, which is equal to one, if the procedure v of the call graph is executed remotely, or zero, if it is executed locally. To incorporate the fact that the program initiates and terminates at the mobile site, we introduce



[FIG3] The average energy spent for offloading versus distance between mobile device and radio access point for different antenna configurations and applications: (a) $w/N = 100$ and (b) $w/N = 1{,}000$.

two auxiliary vertices, namely the initial and terminating vertices, whose indicator variables are set to zero by default. The addition of these two nodes gives rise to an extended edge set $\mathcal{E}_e$ that comprises all the edges of the original call graph, plus the edges from the initiating node to the call graph, and the edges from the call graph to the terminating node. We also denote by $p_{u,v}$ the power spent to transmit the program state between the procedures u and v. To decide which modules of the call graph should be executed remotely, we need to solve the following optimization problem [27]:

$$(\mathcal{P}.2) \quad \min_{I,p} \sum_{v\in V}(1-I_v)\cdot E_v^l + \sum_{(u,v)\in\mathcal{E}_e}[J_{u,v}(p_{u,v})I_v + \varepsilon_{u,v}I_u - (J_{u,v}(p_{u,v})+\varepsilon_{u,v})I_uI_v] \tag{11}$$

s.t.

$$\sum_{v\in V}[(1-I_v)T_v^l + I_vT_v^r] + \sum_{(u,v)\in\mathcal{E}_e}[D_{u,v}(p_{u,v})I_v + \gamma_{u,v}I_u - (D_{u,v}(p_{u,v})+\gamma_{u,v})I_uI_v] \le L \tag{12}$$

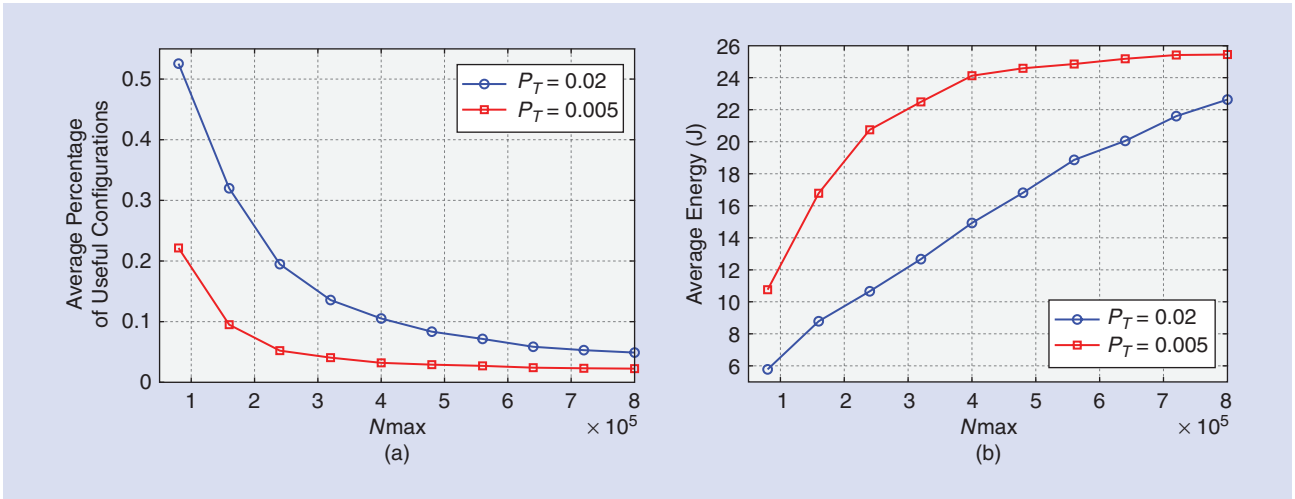
$$I_v\in\{0,1\},\ I_v=0,\ \forall v\in V_l,\ 0\le p_{u,v}\le P_T\ \forall(u,v)\in\mathcal{E}_e, \tag{13}$$

where $I=\{I_v\}_{v\in V}$, $p=\{p_{u,v}\}_{(u,v)\in\mathcal{E}}\in\mathbb{R}^{\mathrm{card}(\mathcal{E}_e)}$, with $\mathrm{card}(\mathcal{E}_e)$ denoting the cardinality of set $\mathcal{E}_e$, T_v^l, and T_v^r are the time it takes to execute the program module v locally or remotely, respectively. The objective function in (11) represents the total energy spent by the MUE for executing the application. In particular, the first term in (11) is the sum (over all the vertices of the call graph) of the energies spent for executing the procedures locally, whereas the second term is the sum (over the edges of the extended call graph) of the energies spent to transfer the execution from the MUE to the SCceNB. The quantity $J_{u,v}(p_{u,v})$ represents the energy necessary to transmit the $N_{u,v}$ bits encoding the program state from the MUE to the SCceNB, whereas $\varepsilon_{u,v}$ is the cost for the MUE to decode the $N_{u,v}$ bits of the program state transmitted back by the SCceNB. The cost $\varepsilon_{u,v}$ is not a function of the MUE's transmitted power and it depends only on the size of the program state $N_{u,v}$. The specific form of the objective function in (11) has been derived so that there is an energy cost associated to offloading only if the procedures u and v are executed at different locations, i.e., $I_u\neq I_v$. More specifically, if $I_u=0$ and $I_v=1$, the energy cost is equal to the energy $J_{u,v}(p_{u,v})$ needed to transmit the program state $N_{u,v}$ from the MUE to the SCceNB, whereas, if $I_u=1$ and $I_v=0$, the cost is equal to the energy $\varepsilon_{u,v}$ needed by the MUE to decode the $N_{u,v}$ bits of the program state transmitted by the SCceNB. Now, assuming an adaptive modulation scheme that selects the QAM constellation size as a function of channel conditions and computational requirements, the minimum time $D_{u,v}(p_{u,v})$ necessary to transmit $N_{u,v}$ bits of duration T_b over an additive white Gaussian noise (AWGN) channel is as in (3), with $p=p_{u,v}$. The energy $J_{u,v}(p_{u,v})$, associated to the transfer of a program state of size $N_{u,v}$, is simply $J_{u,v}(p_{u,v})=p_{u,v}D_{u,v}(p_{u,v})$. The constraint in (12) is a latency constraint, and it contains two summations: the first summation includes the time to run the local modules plus the time to run the offloaded modules; the second summation is the overall delay resulting from transferring the program state from one site (e.g., the MUE) to the other (e.g., the cloud). The constant L represents the maximum latency, dictated by the application. The quantity $\gamma_{u,v}$ is the time needed by the MUE to decode the $N_{u,v}$ bits of the program state transmitted back by the SCceNB. From (12), we note that no delay occurs if the two procedures u and v are both executed in the same location, i.e., $I_u=I_v$. Furthermore, if $I_u=0$ and $I_v=1$, the delay is equal to $D_{u,v}(p_{u,v})$, whereas, if $I_u=1$ and $I_v=0$, the delay is equal to $\gamma_{u,v}$. The constraint in (13) specifies that the variables I_v are binary and that for all procedures contained in the set V_l, which is the set of procedures that are to be executed locally, $I_v=0$. The last constraint, in (13), is the power budget constraint on the maximum transmit power P_T.

Clearly, this optimization procedure is rather complex. Since the state variables I_u are integer, problem $(\mathcal{P}.2)$ is inherently a mixed nonlinear integer programming problem, which might be very complicated to solve. However, a series of simplifications are possible to reduce the complexity of the overall algorithm. An important simplification comes from observing that, for any set of integer values I_v, the remaining optimization over the power coefficients is a convex problem [27]. Furthermore, it is possible to derive closed-form expressions that allow us to check the feasibility of the convex optimization problem for any fixed graph configuration. This feasibility check enables us to discard a priori all graph configurations that cannot be offloaded, for any transmission power satisfying the budget constraint P_T in $(\mathcal{P}.2)$. This check may considerably reduce the complexity of the search. These statements are corroborated by the numerical results reported in Figure 4(a), where we report the average number of call graph configurations worth of offloading (the ones passing the feasibility check) and the total energy consumption (including the energy spent for local processing and the energy spent for offloading) versus the maximum size $N\max$ of the program state to be transferred. The results have been averaged over 1,000 call graph realizations, where the graph has six modules and the number of bits to be transferred for each module are generated as a uniform random variable in the interval $[0, N\max]$. From Figure 4(a), we can see how, increasing the transmit power P_T, there are more feasible configurations because offloading is more likely to occur. At the same time, Figure 4(b) shows that the energy consumption is smaller for the higher transmit power, because offloading occurs more frequently (and then less energy is consumed for local processing). This curve shows an interesting tradeoff between energy consumption and complexity.

OPTIMAL ALLOCATION OF COMMUNICATION/ COMPUTATION RESOURCES IN A MULTIUSER SCENARIO

We consider now the more challenging scenario composed of a set of N_c clouds, N_b small-cell access points (eNBs), and K mobile users. The goal is to find the optimal strategy to assign each user to a base station and to a cloud, to minimize the overall energy



[FIG4] (a) The average percentage of useful call graph configurations versus maximum program state size for different mobile transmit power. (b) The average energy spent for offloading versus maximum program state size for different mobile transmit power.

consumption, under latency constraints. The degrees of freedom are the precoding matrix $\mathbf{Q}_k \in \mathbb{C}^{n_T \times n_T}$ for each user, assuming MIMO transceivers with n_T transmit antennas, the number f_{mk} of CPU cycles/second for running the application of user k over the mth cloud, and the assignment of each user to a base station and then to a cloud. The optimal assignment is performed by selecting the binary values $a_{nmk} \in \{0,1\}$ for $n = 1,\ldots,N_b$, $m = 1,\ldots,N_c$, $k = 1,\ldots,K$. For each k, $a_{nmk} = 1$ if user k accesses the network through the nth BS and it is then served by the mth cloud; all other values are set to zero. In principle, a user could be served by multiple base stations, as in cooperative communications, and by multiple clouds. However, this scenario would make the overall computation management much more complicated. The objective is the minimization of a weighted sum of the energies spent by each mobile terminal: $\mathcal{E}_{\text{tot}} := \sum_{k=1}^{K} c_k \mathcal{E}_k(\mathbf{Q}, \mathbf{a}_k)$, with $\mathcal{E}_k(\mathbf{Q}, \mathbf{a}_k) = \text{trace}(\mathbf{Q}_k) \sum_{n=1}^{N_b} \sum_{m=1}^{N_c} a_{nmk} \Delta_{nk}^t(\mathbf{Q})$, where $\mathbf{Q} \triangleq (\mathbf{Q}_k)_{k=1}^K$ is the set of all covariance matrices. The coefficients c_k are positive parameters that could be varied dynamically to enforce some sort of fairness among the users. The overall latency experienced by the kth MUE for accessing the network through the base station n and being served by the cloud m is now

$$\Delta_{nmk} = \Delta_{nk}^t + \frac{w_k}{f_{mk}} + T_{rx}^{nmk} + T_{Bmn}, \tag{14}$$

where Δ_{nk}^t is the time needed to send the information necessary to transfer the program execution from the kth MUE to the nth base stations, w_k/f_{mk} is the time necessary to execute w_k CPU cycles at the mth server, and T_{rx}^{nmk} is the time necessary for the server to send the results back to the kth MUE. The new term with respect to (1) is the delay T_{Bmn} over the backhaul used to transfer the program state from the nth base station to the mth cloud. The transmission delay Δ_{nk}^t is

$$\Delta_{nk}^t(\mathbf{Q}) = \frac{N_k}{B \log_2 \det\left(\mathbf{I} + \mathbf{H}_k^n \mathbf{Q}_k \mathbf{H}_k^{nH} \tilde{\mathbf{R}}_{nk}(\mathbf{Q}_{-k})^{-1}\right)}, \tag{15}$$

where $\mathbf{H}_l^n$ is the channel matrix between the lth user and the nth base station and the covariance matrix

$$\tilde{\mathbf{R}}_{nk}(\mathbf{Q}_{-k}) = N_0 \mathbf{I} + \sum_{l=1, l\neq k}^{K} \mathbf{H}_l^n \mathbf{Q}_l \mathbf{H}_l^{nH}$$

now contains noise plus multiuser interference.

Formally, the optimization problem can be formulated as:

$$\begin{aligned}
&\min_{\mathbf{Q},\mathbf{f},\mathbf{a}} \sum_{k=1}^{K} c_k \mathcal{E}_k(\mathbf{Q}, \mathbf{a}_k), \text{ subject to:} \\
&\text{i) } a_{nmk}\left(\frac{N_k}{B\log_2 \det\left(\mathbf{I}+\mathbf{H}_k^n \mathbf{Q}_k \mathbf{H}_k^{nH} \tilde{\mathbf{R}}_{nk}(\mathbf{Q}_{-k})^{-1}\right)} + \frac{w_k}{f_{mk}} + T_{Bmn}\right) \\
&\qquad \le L_k, \quad \forall k, n, m \\
&\text{ii) } \text{tr}(\mathbf{Q}_k) \le P_T, \quad \mathbf{Q} \succeq 0, \forall k = 1,\ldots,K, \\
&\text{iii) } f_{mk} \ge 0, \forall m,k \quad \sum_{k=1}^{K} \sum_{n=1}^{N_b} a_{nmk} f_{mk} \le F_m \quad \forall m = 1,\ldots,N_c \\
&\text{iv) } \sum_{n=1}^{N_b} \sum_{m=1}^{N_c} a_{nmk} = 1, a_{nmk} \in \{0,1\}, \quad \forall k, n, m,
\end{aligned} \tag{$\mathcal{P}$.3}$$

(16)

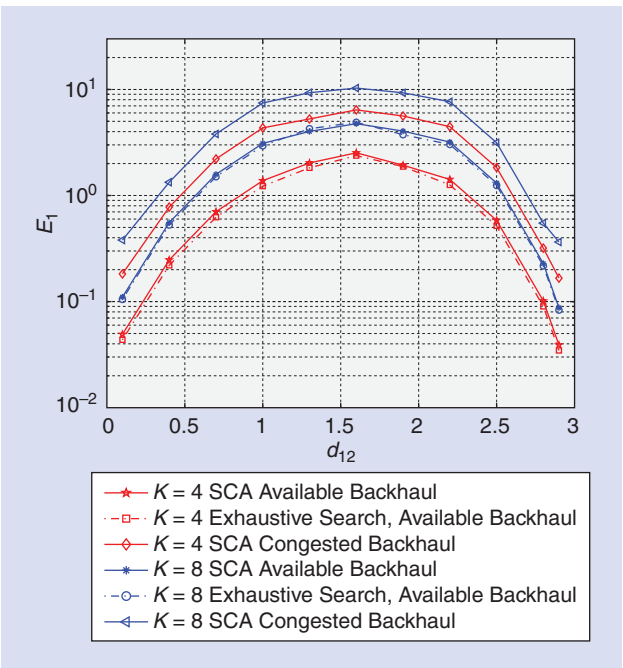
where the constraints are: 1) the overall latency for each user k must be lower than the maximum tolerable value L_k, 2) the total power spent by each user must be lower than its total power budget, 3) the sum of the computational rates assigned by each server cannot exceed the server computational capability F_m, and 4) each mobile user should be served by one couple base station-cloud. To drive the solution toward the situation where each user is served by a single base station and a single cloud, we enforce the constraint $\sum_{n=1}^{N_b} \sum_{m=1}^{N_c} a_{nmk} = 1$ for each k. For simplicity, we have incorporated the term T_{rx}^{nmk} in the latency limit L_k.

MOBILITY MANAGEMENT

As a particular case of the above formulation, we can handle user mobility as follows. Consider, with reference to Figure 1, a single user case, i.e., $K = 1$, where a user (MUE 1) moves from a

position near SCceNB 1 to a position close to SCceNB 2. While moving, a conventional cellular system would perform a base station handover to associate that user to the best station, i.e., the station providing the higher SNR. However, if we consider the allocation of radio and computing resources jointly, we need to consider also a cloud handover. This means that if we switch the radio access from SCceNB 1 to SCceNB 2 but we keep the VM running over SCceNB 1, to avoid VM migration, we need to take into account the delay for sending data over the backhaul from SCceNB 2 to SCceNB 1. Alternatively, we might consider switching both radio access point and serving cloud, but in such a case, we need to migrate the VM. The solution of this problem can be achieved as a particular case of problem $\mathcal{P}.3$, with $K = 1$.

Unfortunately, problem $(\mathcal{P}.3)$ is inherently combinatorial and then NP hard. This makes its exact solution hard to achieve even for moderate values of N_c, N_b, and K. To overcome this obstacle, we assume the coefficients a_{nmk} to be real variables belonging to the interval $[0, 1]$. Then, we adopt a successive convex approximation (SCA) approach to solve problem $(\mathcal{P}.3)$ as a sequence of strongly convex problems converging to a local solution of the original problem (see, e.g., [31] for details). As an example, we consider the scenario sketched in Figure 1, where MUE 1 moves from SCceNB 1 toward SCceNB 2. While moving, the optimal assignment and resource allocation is periodically recomputed as a solution of problem $(\mathcal{P}.3)$. The optimization involves $N_b = 4$ base stations, $N_c = 4$ clouds and $K = 4$ or 8 mobile users. In Figure 5 we compare the average energy consumption obtained by solving the relaxed form of problem $(\mathcal{P}.3)$ in the case where the backhaul between the clouds is congested or only lightly loaded. We also report, as a benchmark, the optimal results achievable with the exhaustive search. It is remarkable to see how our relaxed algorithm gets very close to the optimal exhaustive search solution. Also, we can see the advantage resulting from having a good backhaul between the two base stations.



[FIG5] Energy consumption versus distance d_{12}.

CONCLUSIONS AND FURTHER DEVELOPMENTS

In this article, we have proposed a system perspective of next 5G systems centered on the need to empower energy-hungry mobile terminals with computation offloading capabilities via proximity radio access through small-cell base stations endowed with cloud functionalities. We showed how the optimal resource allocation involves a joint allocation of radio and computation resources, within a fully cross-layer approach. The proposed solution has an impact on several aspects related to the new radio interface, signaling strategies and overall network management. First of all, in case of pervasive computation offloading, the traffic statistics over the uplink and downlink channels are going to change significantly with respect to the current situation, where there is a clear preponderance of traffic on the downlink channel. This change will call for a different (possibly dynamic) partitioning between the capacity associated to the uplink and downlink channels. Furthermore, since the proposed distributed cloud approach requires a (possibly) intensive exchange of data among base stations (in case of base station/cloud handover), the traffic over the wired backhaul linking the base station may become a critical issue. This problem call for the inclusion of high-capacity wireless links between small-cell base stations, e.g., using millimeter-wave direct links, already considered for 5G. We showed how computations can be distributed among a pool of clouds, but assuming only one cloud is active at each time. In principle, the approach can be extended to the parallel computing case, but this would raise extra complexity issues. We also considered a centralized solution. A distributed solution could be reached through the usual primal/dual decomposition methods. Of course this will require a limited exchange of data among the nodes concurring to reach the solution. The wide cross-layer approach proposed in this article of course does not come without a price—the need for higher signaling among applications and physical layers. However the results can justify this extra complexity. Finally, in this article, we assumed many parameters to be known, such as channel state, computation load, traffic over the backhaul, latency, etc. In practice, it would be useful to incorporate suitable learning mechanisms to predict the evolution of most of them. In summary, we believe that in the scenario depicted in this work, signal processing can play a key role in several aspects.

ACKNOWLEDGMENTS

This work was funded by the European Community 7th Framework Programme Project ICT-TROPIC, under grant 318784.

AUTHORS

Sergio Barbarossa (sergio@infocom.uniroma1.it) is a full professor at the University of Rome "La Sapienza." He has held various visiting positions at the Environmental Institute of Michigan (1988), the University of Virginia (1995 and 1997), and the University of Minnesota (1999). He received the 2010 EURASIP

Technical Achievements Award and the 2000 IEEE Best Paper Award. He is an IEEE Distinguished Lecturer and an IEEE Fellow. He is a member of the editorial board of *IEEE Signal Processing Magazine*. He is involved in international projects on heterogeneous networks, cloud computing, and radar remote sensing. His current research interests include distributed optimization, bioinspired signal processing, and self-organizing networks.

Stefania Sardellitti (stefania.sardellitti@uniroma1.it) received the M.Sc. degree in electronic engineering from the University of Rome "La Sapienza," Italy, in 1998 and the Ph.D. degree in electrical and information engineering from the University of Cassino, Italy, in 2005. She is currently a research assistant in the Department of Information Engineering, Electronics, and Telecommunications, University of Rome. She has participated in the European project WINSOC (on wireless sensor networks) and in the European project FREEDOM (on femtocell networks). She is currently involved in the European project TROPIC on distributed computing, storage, and radio resource allocation over cooperative femtocells. Her research interests are in the area of statistical signal processing, mobile cloud computing, femtocell networks, and wireless sensor networks.

Paolo Di Lorenzo (dilorenzo@infocom.uniroma1.it) received the M.Sc. and Ph.D. degrees in electrical engineering from the University of Rome "Sapienza," Italy, in 2008 and 2012, respectively, where he is currently a postdoctoral researcher in the Department of Information, Electronics, and Telecommunications. He has participated in the European projects Freedom (on femtocell networks) and SIMTISYS (on moving target detection through satellite constellations). He is currently involved in the European project TROPIC (on distributed computing, storage, and radio resource allocation over cooperative femtocells). His research interests are in bioinspired signal processing, adaptation and learning over networks, mobile cloud computing, and synthetic aperture radar processing. He received Best Student Paper Awards at the 2010 IEEE International Workshop on Signal Processing Advances for Wireless Communications, the 2011 European Signal Processing Conference, and 2011 International Workshop on Computational Advances in Multisensor Adaptive Processing. He also received the 2012 Gruppo Telecomunicazioni e Tecnologie dell'Informazione Best Doctoral Thesis Award.

REFERENCES

[1] (2013, Nov.). ETSI summit on future mobile and standards for 5G. [Online]. Available: http://www.3gpp.org/news-events/conferences/1515-etsi-summit-on-future-mobile-and

[2] European Project METIS. [Online]. Available: https://www.metis2020.com

[3] M. R. Palacin, "Recent advances in rechargeable battery materials: A chemists perspective," *Chem. Soc. Rev.*, vol. 38, no. 9, pp. 2565–2575, 2009.

[4] M. Sharifi, S. Kafaie, and O. Kashefi, "A survey and taxonomy of cyber foraging of mobile devices," *IEEE Commun. Surveys Tutorials*, vol. 14, no. 4, pp. 1232–1243, 2012.

[5] N. Fernando, S. W. Loke, and W. Rahayu "Mobile cloud computing: A survey," *Future Gen. Comput. Syst.*, vol. 29, pp. 84–106, Jan. 2013.

[6] K. Kumar, J. Liu, Y.-H. Lu, and B. Bhargava, "A survey of computation offloading for mobile systems," *Mobile Networks Applicat.*, vol. 18, no. 1, pp. 129–140, Feb. 2013.

[7] E. Cuervo, A. Balasubramanian, D. Cho, A. Wolman, S. Saroiu, R. Chandra, and P. Bahl, "MAUI: Making smartphones last longer with code offload," in *Proc. ACM Int. Conf. Mobile Systems, Applications, and Services*, San Francisco, CA, 15–18 June 2010, pp. 49–62.

[8] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies, "The case for VM-based cloudlets in mobile computing," *IEEE Pervasive Comput.*, vol. 8, no. 4, pp. 14–23, Oct.–Dec. 2009.

[9] M. V. Barbera, S. Kosta, A. Mei, and J. Stefa, "To offload or not to offload? The bandwidth and energy costs of mobile cloud computing," in *Proc. INFOCOM 2013*, Apr. 2013, Turin, Italy, pp. 1285–1293.

[10] G. Fettweiss, "A 5G wireless communication vision," *Microwave J.*, vol. 55, no. 12, pp. 24–36, Dec. 2012.

[11] J. Flinn, S. Park, and M. Satyanarayanan, "Balancing performance, energy, and quality in pervasive computing," in *Proc. 22nd IEEE Int. Conf. Distributed Computing Systems*, 2002, pp. 217–226.

[12] R. Balan, M. Satyanarayanan, S. Park, and T. Okoshi, "Tactics-based remote execution for mobile computing," in *Proc. 1st Int. Conf. Mobile Systems, Applications and Services*, 2003, pp. 273–286.

[13] R. Kemp, N. Palmer, T. Kielmann, and H. Bal, "Cuckoo: A computation offloading framework for smartphones," in *Proc. 2nd Int. Conf. Mobile Computing, Applications, and Services (MobiCASE)*, pp. 59–79, Oct. 2010.

[14] B.-G. Chun and P. Maniatis, "Augmented smartphone applications through clone cloud execution," in *Proc. HotOS*, Monte Verita, Switzerland, p. 8, May 2009.

[15] D. Huang, X. Zhang, M. Kang, and J. Luo, "Mobicloud: Building secure cloud framework for mobile computing and communication," in *Proc. 5th IEEE Int. Symp. Service Oriented System Eng. (SOSE)*, June 2010, pp. 27–34.

[16] M. D. Kristensen and N. O. Bouvin, "Scheduling and development support in the scavenger cyber foraging system," *Pervasive Mobile Comput.*, vol. 1, no. 6, pp. 677–692, 2010.

[17] J. Flinn, D. Narayanan, and M. Satyanarayanan, "Self-tuned remote execution for pervasive computing," in *Proc. 8th IEEE Workshop Hot Topics in Operating Systems*, Schloss Elmau, Germany, 2001, pp. 61–66.

[18] K. Kumar and Y.-H. Lu, "Cloud computing for mobile users: Can offloading computation save energy?" *Computer*, vol. 43, no. 4, pp. 51–56, Apr. 2010.

[19] X. Gu, K. Nahrstedt, A. Messer, I. Greenberg, and D. Milojicic, "Adaptive offloading for pervasive computing," *IEEE Pervasive Comput.*, vol. 3, no. 3, pp. 66–73, 2004.

[20] S. Ou, K. Yang, and Q. Zhang, "An efficient runtime offloading approach for pervasive services," in *Proc. IEEE Wireless Communication and Networking Conf. (WCNC2006)*, Las Vegas, NV, 2006, pp. 2229–2234.

[21] D. Huang, P. Wang, and D. Niyato, "A dynamic offloading algorithm for mobile computing," *IEEE Trans. Wireless Commun.*, vol. 11, no. 6, pp. 1991–1995, June 2012.

[22] G. Chen, B. T. Kang, M. Kandermir, N. Vijaykrishnan, M. J. Irwin, and R. Chandranouli, "Studying energy trade offs in offloading computation/compilation in Java-enabled mobile devices," *IEEE Trans. Parallel Distrib. Syst.*, vol. 15, no. 9, pp. 795–809, Sept. 2004.

[23] Y. Wen, W. Zhang, and H. Luo, "Energy-optimal mobile application execution: Taming resource-poor mobile devices with cloud clones," *Proc. IEEE Int. Conf. Computer Communications 2012*, Mar. pp. 2716–2720.

[24] Z. Li, C. Wang, and R. Xu, "Computation offloading to save energy on handheld devices: A partition scheme," in *Proc. Int. Conf. Compilers, Architecture, and Synthesis for Embedded Systems*, Atlanta, GA, 2001, pp. 238–246.

[25] B. G. Ryder, "Constructing the call graph of a program," *IEEE Trans. Softw. Eng.*, vol. 5, no. 3, pp. 216–226, 1979.

[26] S. Kosta, A. Aucinas, P. Hui, R. Mortier, and X. Zhang, "ThinkAir: Dynamic resource allocation and parallel execution in the cloud for mobile code offloading," *Proc. INFOCOM 2012*, Mar., pp. 945–953.

[27] P. Di Lorenzo, S. Barbarossa, and S. Sardellitti, "Joint optimization of radio resources and code partitioning in mobile cloud computing," submitted for publication.

[28] R. Kaewpuang, D. Niyato, P. Wang, and E. Hossain, "A framework for cooperative resource management in mobile cloud computing," *IEEE J. Select. Areas Commun.*, vol. 31, pp. 2685–2700, Dec. 2013.

[29] I. Hwang, B. Song, and S. Soliman, "A holistic view on hyper-dense heterogeneous and small cell networks," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 20–27, June 2013.

[30] FP7 European Project. (2012). Distributed computing, storage and radio resource allocation over cooperative femtocells (TROPIC). [Online]. Available: http://www.ict-tropic.eu/

[31] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint optimization of radio and computational resources for mobile cloud computing in a multicell environment," submitted for publication.

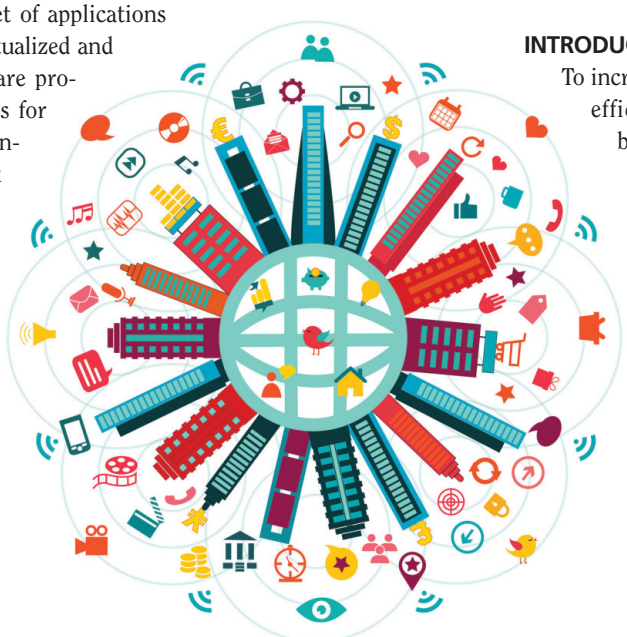
[SP]

Hadi Baligh, Mingyi Hong, Wei-Cheng Liao, Zhi-Quan Luo,
Meisam Razaviyayn, Maziar Sanjabi, and Ruoyu Sun

Cross-Layer Provision of Future Cellular Networks

A WMMSE-based approach

To cope with the growing demand for wireless data and to extend service coverage, future fifth-generation (5G) networks will increasingly rely on the use of low-power nodes to support massive connectivity in a diverse set of applications and services [1]. To this end, virtualized and mass-scale cloud architectures are proposed as promising technologies for 5G in which all the nodes are connected via a backhaul network and managed centrally by such cloud centers. The significant computing power made available by the cloud technologies has enabled the implementation of sophisticated signal processing algorithms, especially by way of parallel processing, for both interference management and network provision. The latter two are among the major signal processing tasks for 5G due to an increased level of frequency sharing, node density, interference, and network congestion. This article outlines several theoretical and practical aspects of joint interference management and network provisioning for future 5G networks. A cross-layer optimization framework is proposed for joint user admission, user-base station (BS) association, power control, user grouping, transceiver design, as well as routing and flow control. We show that many of these cross-layer tasks can be treated in a unified way and implemented in a parallel manner using an efficient algorithmic framework called *weighted minimum mean squared error* (*WMMSE*). Some recent developments in this area are highlighted and future research directions are identified.



THE 5G REVOLUTION

INTRODUCTION

To increase network capacity and spectral efficiency, network operators have been adding many low-power micro/pico/femto BSs, relays and Wi-Fi access points (APs) to the network, thereby reducing the signal transmission distances. This has resulted in the so-called heterogeneous network (HetNet) architecture; see [2] and the references therein. The HetNet architecture naturally replaces the traditional single hop access mode between the high-power BS and its users by a wireless mesh network consisting of a large number of densely deployed wireless APs with either wireline or wireless backhaul support; see Figure 1. A natural extension to this concept would be virtualized radio access (VRA) and mass-scale cloud architectures, which are proposed as promising technologies for 5G [1]. In VRA, all nodes are connected via a backhaul network and managed centrally by cloud centers. Such a multihop network must be self-organized and fast adaptive to the changes of traffic demands caused by users joining, leaving, or requesting new data packets at any time. To maximize the performance of such networks, the multiuser interference, which is a major performance limiting factor, should be astutely managed through advanced signal processing techniques.

Digital Object Identifier 10.1109/MSP.2014.2335237
Date of publication: 15 October 2014

1053-5888/14©2014IEEE

In addition, traffic engineering within the entire radio access network (RAN), including the backhaul network, should be jointly optimized with resource allocation across the wireless links.

In this article, we present a cross-layer optimization framework for joint resource allocation and provision of future 5G networks. Our approach integrates several important cross-layer techniques: 1) advanced signal processing techniques for physical-layer (PHY) interference management; 2) medium access control (MAC)-layer algorithms to handle user scheduling, BS assignment and BS clustering; and 3) network layer solutions such as software defined networking (SDN) [1], [3], [4] to manage the network services and to control individual flows. Such cross-layer optimization is more challenging than the traditional cellular network optimization. Routing and scheduling of user traffic in the presence of multiuser interference is a difficult problem by itself. It is made more complicated by practical issues such as backhaul capacity limitations, channel state information (CSI) overhead, and distributed implementation of the algorithms. Our goal is to highlight some recent advances in this area, illustrate the potential of signal processing in the provision of future 5G networks, and identify some future research directions.

The concept of cross-layer optimization is not new, and such a task is known to be computationally complex. However, with the significant computing power at cloud centers of future 5G networks, we believe the time has come to consider the use of advanced signal processing algorithms for network provision. In fact, with the network densification and Cloud-RAN architecture, it is crucial to jointly provision the RAN and backhaul networks, addressing such issues as users-BS assignments, interference management, and traffic engineering in a coordinated manner. Our proposed cross-layer approach addresses many of these important issues in a coherent algorithmic framework based on a WMMSE technique [5]–[7].

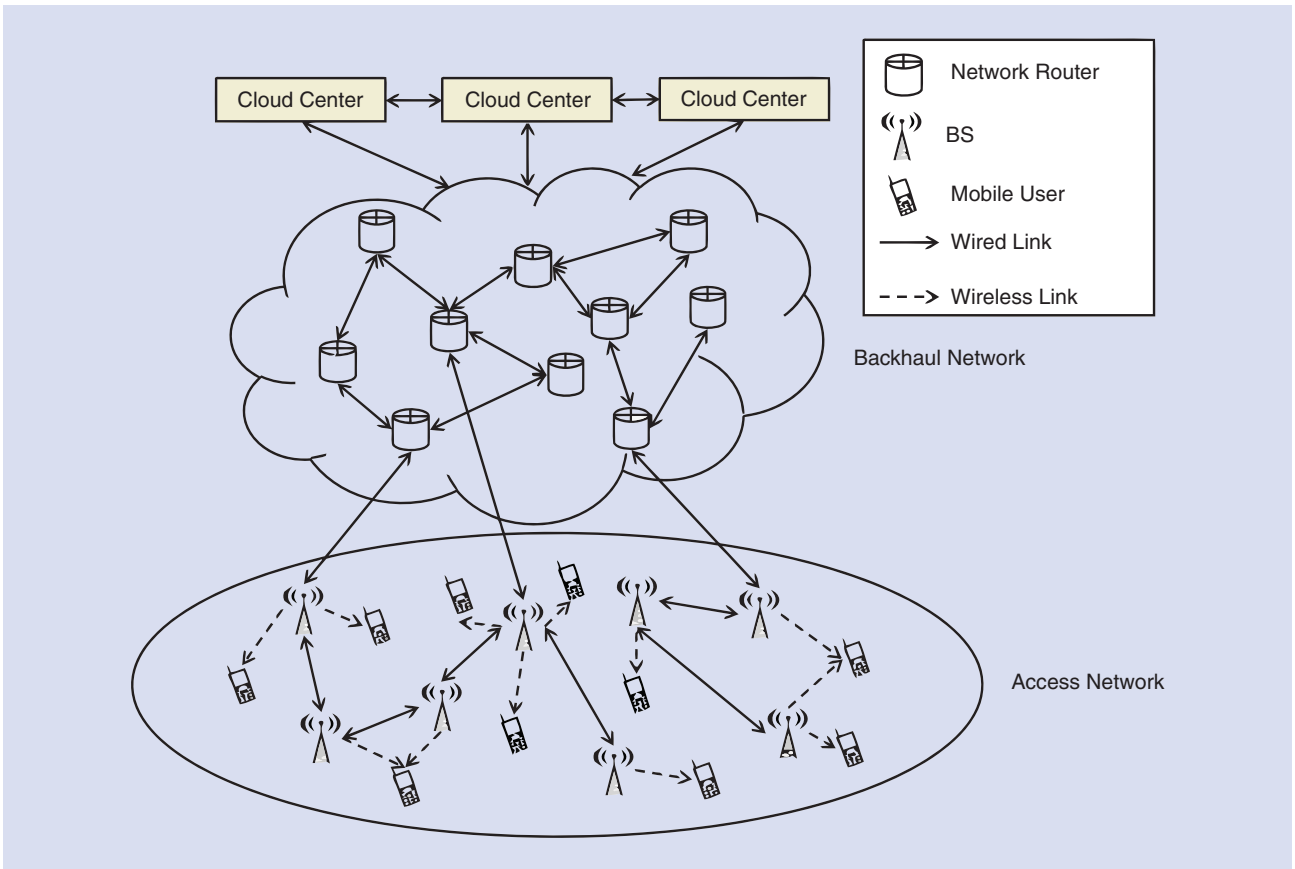
A CROSS-LAYER FORMULATION

A GENERAL MODEL

For simplicity, we restrict our discussion to the downlink direction in which the traffic flows from the core network to the users through the wired/wireless backhaul network and the RAN.

Let the backhaul network be composed of a set $\mathcal{N}$ of routers and a set $\mathcal{L}^{\mathrm{w}}$ of wired links, whose main purpose is to route the traffic toward the RAN. Each backhaul link $l \in \mathcal{L}^{\mathrm{w}}$ has a fixed capacity, denoted as $\bar{C}^l$. The RAN consists of a set $\mathcal{Q}$ of BSs and a set $\mathcal{I}$ of users. Assume that each BS/AP is equipped with M transmit antennas, and that the total bandwidth is divided into F frequency tones.

Each user i has N receive antennas and achieves an instantaneous rate of $R_i[t]$ for a given time instance t. Such a rate is a function of a collection of system parameters, defined across



[FIG1] An illustration of the SDN-RAN.

various network layers. These parameters include the transmit strategies of the BSs (e.g., precoders, power control mechanism), the cooperation strategies among the BSs [e.g., the coordinated multipoint (CoMP) scheme], the routing decisions within the backhaul, and so on. The network provision in this context amounts to achieving certain overall performance by optimizing, possibly in different time-scales, all the system parameters. Let $U(\cdot)$ denote a utility function that measures the system-level performance. We are interested in solving the following cross-layer system utility maximization problem:

$$\begin{aligned} &\text{maximize} && U(\{R_i[t]\}_{i\in\mathcal{I}}) \\ &\text{subject to} && \text{Per BS resource constraints} \\ &&& \text{Per data flow QoS constraints} \\ &&& \text{Network structure constraints} \\ &&& \text{Per node flow conservation constraints.} \end{aligned} \tag{1}$$

At this stage, the problem is described in a very generic form. Below we illustrate how each component of the problem can be instantiated in practice.

A large family of utility functions is the "α-fair" utility functions, which is given by

$$U(\{R_i[t]\}_{i\in\mathcal{I}}) = \sum_{i\in\mathcal{I}} \frac{(R_i[t])^{1-\alpha}}{1-\alpha}. \tag{2}$$

Different choices of the parameter α give different priorities to user fairness and overall system performance. For example, when $\alpha = 0$, we obtain the sum rate utility: $U(\{R_i[t]\}_{i\in\mathcal{I}}) = \sum_{i\in\mathcal{I}} R_i[t]$. When $\alpha = 1$, we obtain the proportional fair utility $U(\{R_i[t]\}_{i\in\mathcal{I}}) = \sum_{i\in\mathcal{I}} \log(R_i[t])$. Popular choices of α and their corresponding utility functions can be found in [8] and the references therein. Furthermore, when certain parameters in the system (e.g., the channel realizations) are random, we can choose the average throughput $\sum_{i\in\mathcal{I}} \mathbb{E}[R_i[t]]$ as the system utility. In general, the system utilities expressed as a function of the users' rates are typically concave, but are nonconcave in terms of the system parameters (e.g., power, routing variables) due to multiuser interference [9].

The first type of constraints in (1) is related to the BSs' resources or their individual transmit strategies. Suppose a BS q uses a transmit precoder $\mathbf{V}_i^q[f,t] \in \mathbb{C}^{M\times N}$ to transmit to user i on tone f at time t. Then we have the following per BS transmit power constraint

$$\sum_{f=1}^{F} \sum_{i\in\mathcal{I}} \mathrm{Tr}[\mathbf{V}_i^q[f,t](\mathbf{V}_i^q[f,t])^H] \le \overline{P}^q, \tag{3}$$

where $\overline{P}^q$ is BS q's power budget. Other constraints in this category include per group of antenna constraints, user scheduling constraints, or the zero forcing constraints.

The second type of constraints in (1) is related to the quality of service (QoS) required by the users. One such requirement is given by $R_i[t] \ge \gamma_i[t], \forall\, i \in \mathcal{I}$, where $\gamma_i[t]$ is the predefined minimum rate requested by user i at time t. Alternatively, one can require that the outage probability is bounded above, i.e., $\mathbb{P}(R_i[t] \le \gamma_i) \le \delta_i, \forall\, i \in \mathcal{I}$.

The third type of constraints is closely related to the densification of the BS site deployment in 5G systems. New architectures such as the HetNet offer unprecedented flexibility in terms of how the access network can be formed. For example, we can cluster a subset of BSs to cooperatively serve a user i and optimize the cluster membership to yield the best performance. To further elaborate on such network structure constraint, let us define a set of binary variables $\{z_i^q\}$ to represent the BS-user association ($z_i^q = 1$ means BS q serves user i, and zero otherwise). The set of serving BSs for user i is then given by $\mathcal{S}_i := \{q \mid q \in \mathcal{Q}, z_i^q = 1\}$, and the set of active users for BS q is $\mathcal{I}^q := \{i \mid i \in \mathcal{I}, z_i^q = 1\}$. Then constraining on $|\mathcal{S}_i|$ controls the size of each cooperative BS cluster. Other structural constraints include the ones that dictate that, at a given time, only a subset of BSs in $\mathcal{Q}$ are activated. Such constraints are useful to keep the operational costs of the access network under control.

The last type of constraint has to do with the control of data flow in the backhaul network. Without loss of generality, suppose that each user i requests a single data flow. Then we use $s(i)$ and $d(i)$ to denote a source-destination pair for a given data flow, where $s(i)$ typically represents a packet data gateway, and $d(i)$ is some user device. Let us use $\mathcal{L}^{\mathrm{wl}}$ to denote the set of wireless links in the system. That is, $\mathcal{L}^{\mathrm{wl}} \triangleq \{(s,d,f) \mid s \in \mathcal{Q}, d \in \mathcal{I}, f = 1\cdots, F\}$ with (s,d,f) being the wireless link from BS s to user d on tone f. Denote the rate for user i on link l as $R_i^{(l)}$. Then the following flow conservation constraint must hold true:

$$\sum_{l\in\mathrm{In}(v)} R_i^{(l)} + 1_{\{s(i)\}}(v)R_i = \sum_{l\in\mathrm{Out}(v)} R_i^{(l)} + 1_{\{d(i)\}}(v)R_i, \forall i_k \in \mathcal{I},\ \forall v \in \mathcal{V}, \tag{4}$$

where $\mathrm{In}(v)$ and $\mathrm{Out}(v)$ denote the set of links going into and coming out of a node v, respectively; $1_{\mathcal{A}}(x)$ denotes the indicator function for a set $\mathcal{A}$, i.e., $1_{\mathcal{A}}(x) = 1$ if $x \in \mathcal{A}$ and $1_{\mathcal{A}}(x) = 0$, otherwise.

Finally, we give a concrete example showing how the users' rates are dependent on various system parameters introduced so far. Let $\mathbf{H}_i^q[f,t] \in \mathbb{C}^{N\times M}$ denote the channel matrix between BS q and user i on tone f at slot t. Let σ_i denote the noise power at user i. Further define a binary variable $\alpha_i[f,t] \in \{0,1\}$ to signify whether user i is served on channel f at slot t. Then user i's rate $R_i[t]$, under perfect CSI, can be expressed as (we use natural logarithm throughout this article)

$$R_i[t] = \sum_{f=1}^{F} \alpha_i[t,f] \log\det\Big(\mathbf{I} + \sum_{q\in\mathcal{S}_i} \mathbf{H}_i^q[t,f]\mathbf{V}_i^q[t,f] \times \sum_{q\in\mathcal{S}_i} (\mathbf{V}_i^q[t,f])^H (\mathbf{H}_i^q[t,f])^H \mathbf{C}_i^{-1}[t,f]\Big) \tag{5}$$

where $\mathbf{C}_i[t,f] \in \mathbb{C}^{N\times N}$ is the interference matrix for user i

$$\mathbf{C}_i[t,f] = \sigma_i^2\mathbf{I} + \sum_{(q,j):j\ne i} \mathbf{H}_i^q[t,f]\mathbf{V}_j^q[t,f](\mathbf{V}_j^q[t,f])^H(\mathbf{H}_i^q[t,f])^H.$$

THE CHALLENGES AND THE OVERVIEW OF SOLUTIONS

The generic problem (1) is huge in size because of the increasingly large number of access nodes in 5G systems, as well as the fact that it covers quite a few aspects of the system design. It would be very challenging, if not impossible, to optimize (1) directly in its most general form. In practice, system parameters are usually optimized over different time scales. For example, network structure typically changes at a slower rate compared with the transmit/receive beamformer. Therefore, at any given time instance, a reduced version of (1) is solved, maybe with a much smaller problem dimension.

This article presents a few important cross-layer design problems formulated as special cases of (1). We discuss existing approaches for these problems, and advocate a unifying framework based on the WMMSE method. We demonstrate that the WMMSE [5]–[7], known as one of the state-of-the-art methods for PHY-layer precoder design, can be generalized in multiple ways to deal with cross-layer designs in 5G networks. More importantly, the resulting schemes are often amendable to efficient parallel implementation, which suits ideally to the cloud-based software-defined architecture of the 5G networks.

IN FACT, WITH THE NETWORK DENSIFICATION AND CLOUD-RAN ARCHITECTURE, IT IS CRUCIAL TO JOINTLY PROVISION THE RAN AND BACKHAUL NETWORKS, ADDRESSING SUCH ISSUES AS USERS-BS ASSIGNMENTS, INTERFERENCE MANAGEMENT, AND TRAFFIC ENGINEERING IN A COORDINATED MANNER.

CROSS-LAYER NETWORK MANAGEMENT

From a global perspective, the densification and heterogeneity of the access network give rise to new design issues across different layers of the network. In fact, the traditional boundary between different layers of the network has been blurred. For example, the possibility of using multiple closely located BSs to jointly serve one user impacts the traffic routing strategy in the network layer, which further affects the ways that the scheduling in the MAC layer and interference management in the physical layer are performed. In this section, we address various issues in the joint design across the PHY, MAC, and network layer. We take a bottom-up approach and show how PHY-layer signal processing techniques can be generalized to deal with MAC layer problems such as scheduling and BS management/assignment, or network layer problems such as traffic management and congestion control. In particular, the well-known WMMSE algorithm [5]–[7] is presented as a unifying framework for such a purpose.

PHY-LAYER INTERFERENCE MANAGEMENT

Let us first consider the simple classical problem of precoder design in a multiuser wireless network, where only the PHY-layer transmit and receive strategies are the design variables. Assume the users have been properly scheduled in different time/frequency slots and we only consider the problem in one time/frequency slot. Moreover, we assume that the user-BS association has been determined and each user is served by only one BS, i.e., the network is an interfering broadcast channel (IBC). The utility maximization problem in an IBC can be formulated as

$$\begin{aligned} &\text{maximize} \quad U(\{R_i\}_{i\in\mathcal{I}}) \\ &\text{subject to} \quad (5),\ \sum_{i\in\mathcal{I}^q} \mathrm{Tr}\big[\mathbf{V}_i^q(\mathbf{V}_i^q)^H\big] \le \overline{P}^q,\ \ \forall q. \end{aligned} \tag{6}$$

Except for some special cases [e.g., maximizing the minimum rate in the multiple-input, single output (MISO) or single-input, multiple output (SIMO) interference channel (IC)], (6) is nonconvex and NP-hard (see [9] and the references therein). A few nonconvex instances of (6), such as the sum rate maximization in MISO or SIMO IC, can be solved by global optimization methods (see [10] and the references therein). However, these algorithms are usually too complex to be used for large-scale networks. Moreover, it is not clear whether and how they can be extended to more general network settings such as MIMO IBC, or networks with flexible BS association and limited backhaul availability. In general, finding the global maximum for (6) is NP-hard and is difficult even for reasonably sized networks. Consequently, much research effort has been devoted to designing efficient algorithms that produce high-quality suboptimal solutions to (6); see [9] and [11] for recent surveys. In this subsection, we briefly review the WMMSE algorithm as it forms the basis of the cross-layer optimization approach presented in this article.

The WMMSE algorithm was first proposed in [5] for the multiple-input, multiple-output (MIMO) broadcast channel and later extended to MIMO IC, where each user transmits a single data stream [6]. It was extended significantly to MIMO IC and MIMO IBC/I-MAC where each user transmits multiple data streams and to a large family of utility functions, including the sum rate utility [7]. The main idea is to transform the original sum utility maximization problem to a WMMSE minimization problem where the weights are adaptively updated. We illustrate this idea for the sum rate maximization in MIMO IC where each user only transmits a single data stream. Suppose the transmit beamformer of BS i is $\mathbf{v}^i \in \mathbb{C}^{M\times 1}$ [i.e., $\mathbf{V}_i^i$ in the formulation (6)] and the receive beamformer of BS i is $\boldsymbol{u}_i \in \mathbb{C}^{N\times 1}$, which has unit norm. The SINR of user i is given by $\gamma_i = \big(|\boldsymbol{u}_i^H \mathbf{H}_i^i \mathbf{v}^i|^2\big)/\big(\sum_{j\neq i}|\boldsymbol{u}_i^H \mathbf{H}_i^j \mathbf{v}^j|^2 + \sigma_i^2\big)$, and the rate of user i is given by $R_i = \log(1+\gamma_i)$. With these notations, the sum rate maximization problem becomes

$$\begin{aligned} &\underset{\{\boldsymbol{u}_i,\mathbf{v}^i\}}{\text{maximize}} \quad \sum_i R_i, \\ &\text{subject to} \quad \|\mathbf{v}^i\|^2 \le \overline{P}^i,\ \ \forall i. \end{aligned} \tag{7}$$

The mean squared error (MSE) of user i is given as $e_i = |\boldsymbol{u}_i^H \mathbf{H}_i^i \mathbf{v}^i - 1|^2 + \sum_{j\neq i}|\boldsymbol{u}_i^H \mathbf{H}_i^j \mathbf{v}^j|^2 + \sigma_i^2$. There is a well-known relationship between SINR and MMSE: MMSE = 1/(1 + SINR). More precisely, for any given $\{\mathbf{v}^i\}$, we have

$$1+\max_{u_i}\gamma_i=\max_{u_i}\frac{1}{e_i}. \tag{8}$$

Consequently, the sum rate maximization problem (7) is equivalent to the following problem [7]:

$$\begin{aligned}&\underset{\{u_i,\upsilon^i\}}{\text{minimize}} \quad \sum_i \log(e_i),\\ &\text{subject to} \quad \|\upsilon^i\|^2\le \overline{P}^i, \quad \forall i.\end{aligned} \tag{9}$$

One can further prove that (9) is equivalent to the following problem (in the sense that there is a one-to-one correspondence between the stationary points of the two problems)

$$\begin{aligned}&\underset{\{u_i,\upsilon^i,w_i\}}{\text{minimize}} \quad \sum_i (w_i e_i-\log(w_i))\\ &\text{subject to} \quad \|\upsilon^i\|^2\le \overline{P}^i, \quad \forall i.\end{aligned} \tag{10}$$

The optimization problem (10) is an adaptively weighted MMSE problem and can be solved by alternate optimization. Specifically, e_i is a convex quadratic function over $\{\boldsymbol{u}_i\}$ and $\{\boldsymbol{\upsilon}^i\}$ respectively, thus $\{\boldsymbol{u}_i\}$ and $\{\boldsymbol{\upsilon}^i\}$ can be updated in closed forms (with an additional bisection step for updating $\{\boldsymbol{\upsilon}^i\}$); the optimal w_i is given by $w_i=1/e_i$. The convergence of the WMMSE algorithm can be derived from the classical theory of the alternate optimization [7]. The details of the WMMSE algorithm for sum rate maximization are given in Algorithm 1. We emphasize that the $\boldsymbol{\upsilon}$ subproblem can be solved very efficiently because of the equivalent transformation from (7) to (10), which is the key technique of the WMMSE algorithm. As will be seen in subsequent sections, the alternate optimization framework and the simplification of the $\boldsymbol{\upsilon}$ subproblem are crucial to generalize the WMMSE algorithm to solve cross-layer design problems.

Algorithm 1 The WMMSE algorithm for sum rate maximization in MIMO IC (single beam case).

Initialize $\boldsymbol{\upsilon}^i, \boldsymbol{u}_i, w_i$ randomly.

repeat

- $\boldsymbol{u}_i \leftarrow \left(\sum_{j=1}^K \mathbf{H}_i^j \boldsymbol{\upsilon}^j (\boldsymbol{\upsilon}^j)^H (\mathbf{H}_i^j)^H+\sigma_i^2 I\right)^{-1}\mathbf{H}_i^i\boldsymbol{\upsilon}^i, \forall i;$
- $w_i \leftarrow (1-\boldsymbol{u}_i^H\mathbf{H}_i^i\boldsymbol{\upsilon}^i)^{-1}, \forall i;$
- $\boldsymbol{\upsilon}^i \leftarrow \left(\sum_{j=1}^K w_j (\mathbf{H}_j^i)^H \boldsymbol{u}_j\boldsymbol{u}_j^H\mathbf{H}_j^i+\mu_i^* I\right)^{-1}\mathbf{H}_i^i\boldsymbol{u}_i w_i, \forall i,$

where μ_i^* is computed by bisection such that $\|\boldsymbol{\upsilon}^i\|^2\le\overline{P}^i$.

until Convergence

Assuming all computation is executed at cloud centers, the WMMSE algorithm can be easily implemented in a parallel fashion. Let $\boldsymbol{u},\boldsymbol{\upsilon},\boldsymbol{w}$ denote the sets of variables $\{\boldsymbol{u}_i\},\{\boldsymbol{\upsilon}^i\},\{w_i\}$, respectively. Fixing the two sets of variables $\boldsymbol{u}$ and $\boldsymbol{w}$, the objective function of (10) is decomposable across all BSs with respect to $\boldsymbol{\upsilon}$. As a result, the update of each $\boldsymbol{\upsilon}^i$ only depends on $\boldsymbol{u}$ and $\boldsymbol{w}$ and does not depend on other $\boldsymbol{\upsilon}^j, \forall j\neq i$. All $\boldsymbol{\upsilon}^i$ can be updated in a parallel fashion. Similarly, the variables $\boldsymbol{u}_i$ and w_i can also be updated in parallel.

The WMMSE algorithm can be interpreted as an inexact alternate optimization approach to solve (9), where $\{\boldsymbol{u}_k\}$ and $\{\boldsymbol{\upsilon}^k\}$ are updated iteratively. The receive beamformers $\{\boldsymbol{u}_k\}$ are updated to MMSE receivers, which exactly minimize $\sum_k \log(e_k)$. The transmit precoders $\{\boldsymbol{\upsilon}^k\}$ are updated to minimize $\sum_j (\partial \log(e_j))/(\partial e_j)e_j$, a "linear" approximation of $\sum_j \log(e_j)$, where $(\partial \log(e_j))/(\partial e_j)=1/e_j$ [i.e., the optimal w_j for (10)] is computed based on the previous iterates. With this interpretation, the convergence of the WMMSE algorithm can also be derived from the result in [12].

We summarize the advantages of the WMMSE algorithm as follows. First, the alternate optimization framework and the simplicity of each subproblem makes the WMMSE algorithm easily extendable to cross-layer design. Additionally, no stepsize tuning is needed. Second, the WMMSE algorithm exploits the special structure of the rate function, thus converging faster than general nonlinear optimization methods such as the gradient descent algorithm. Third, since at each step the subproblem can be decomposed across each user, it is amenable to parallel and distributed implementation.

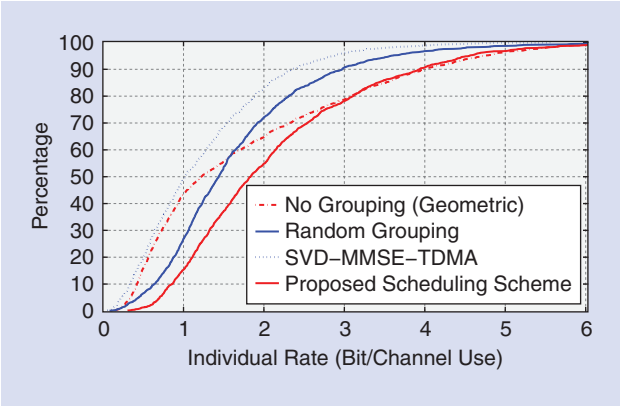
Let us briefly compare the WMMSE algorithm with the parallel successive convex approximation (SCA) algorithm [13], which is designed for general nonconvex problems. The latter algorithm, when applied to the sum utility maximization problem, can be viewed as the parallel version for the MIMO case of the sequential interference pricing (IP) algorithm proposed in [14]. The main advantage of the SCA compared with the IP algorithm is that by properly introducing appropriate stepsize control, it allows parallel updates and the convergence of the algorithm is guaranteed (note that in [14], the IP algorithm is designed for SISO and MISO ICs, and there is no convergence proof for the parallel IP algorithms). Similar to the WMMSE algorithm, the parallel SCA exploits the structure of the rate function, can be implemented in parallel, and has been extended to solve some cross-layer design problems [15]. Compared with the WMMSE, parallel SCA aims to design transmit covariance matrices rather than the beamformer, so it cannot be directly used when the number of transmit antennas are larger than the receive antennas, or there are restrictions on the number of transmitted data streams. Since parallel SCA shows good performance in certain problem setups, it will be interesting to see whether it can be generalized to other cross-layer design problems.

JOINT PHY-MAC-LAYER OPTIMIZATIONS: SCHEDULING AND RESOURCE ALLOCATION

Once we know how to deal with PHY-layer precoder design using WMMSE, we move up one layer and show that it is in fact easy to integrate various tasks from the MAC layer, such as user scheduling across different time slots/frequency tones, BS assignment, and BS clustering. The resulting PHY-MAC joint designs are again stepsize free, decompose nicely across the network nodes, and are thus easily parallelizable.

BEAMFORMING AND SCHEDULING

For a HetNet, user scheduling and beamforming are two effective techniques to mitigate multiuser interference. Joint user scheduling and beamforming has been investigated in



[FIG2] The rate CDF of different methods [16].

different works, including [16] and [17]. The basic problem is to design linear transmit/receive beamformers and schedule the users across a fixed set of time slots/frequency tones in a way that maximizes a given system utility. For simplicity of presentation, let there be one frequency tone and T time slots. We assume that the BS assignment is fixed and each user is served by only one BS, i.e., $|\mathcal{S}_i| = 1, \forall i \in \mathcal{I}$. For notational simplicity, we only present the algorithm for the simple IC model, although the algorithm and the analysis can be generalized to the IBC [16]. Let us define a binary variable $\alpha_i[t] \in \{0,1\}$ to indicate if user i is served in time slot t or not. We assume no joint decoding across different time slots and therefore the rate of each user is the summation of the rates over different time slots/frequency tones. Notice that the time slots here are much longer than the symbol length and typically of the size of the block used for channel coding. Making this assumption, the joint beamforming and scheduling problem can be formulated as

$$\begin{aligned} \underset{\mathbf{V},\alpha}{\text{maximize}} \quad & U(\{R_i\}_{i\in\mathcal{I}}) \\ \text{subject to} \quad & \mathrm{Tr}[\mathbf{V}^i[t](\mathbf{V}^i[t])^H] \le \overline{P}^i, \forall i,t \\ & R_i = \sum_t \alpha_i[t] \log\det\Big(\mathbf{I} + \mathbf{H}_i^i\mathbf{V}^i[t](\mathbf{V}^i[t])^H(\mathbf{H}_i^i)^H \times \\ & \Big(\sigma_i^2\mathbf{I} + \sum_{j\neq i}\mathbf{H}_i^j\mathbf{V}^j[t](\mathbf{V}^j[t])^H(\mathbf{H}_i^j)^H\Big)^{-1}\Big) \\ & \alpha_i[t] \in \{0,1\}, \forall i,t. \end{aligned} \tag{11}$$

It can be shown that without losing optimality, the scheduling variables $\alpha_i[t]$ can be initially set to 1. Furthermore, local linearization of the utility function around the current rate values results in similar problem as in the sum rate maximization problem. Using this observation, an algorithm similar to WMMSE (Algorithm 2) can be used to solve the optimization problem (11); see [16] for more details. After solving the above optimization problem, we can update the scheduling variables $\alpha_i[t]$ using the obtained precoders.

Figure 2 illustrates the performance gain of the joint scheduling and beamforming optimization in a 19-hexagonal wraparound cell layout. There are 285 users served by 57 BSs in the network. The other simulation details can be found in [16]. The figure shows the achieved rate cumulative distribution function (CDF) of different approaches: The "no scheduling" corresponds to the WMMSE algorithm with no scheduling; the "random scheduling" curve represents a random scheduler combined with the WMMSE algorithm; while in the "SVD-MMSE-TDMA" approach, each BS serves its own users in a time-division multiple access approach (TDMA) fashion by using the singular value decomposition (SVD) precoder and the users deploy MMSE receivers. As shown in Figure 2, the joint beamforming and scheduling can significantly improve the system throughput as well as the user fairness.

Algorithm 2 The joint scheduling and beamforming algorithm.

initialize $\mathbf{V}^i[t]$'s randomly
repeat
- $\mathbf{U}_i[t] \leftarrow \Big(\sum_j \mathbf{H}_i^j\mathbf{V}^j[t](\mathbf{V}^j)^H(\mathbf{H}_i^j)^H + \sigma_i^2\mathbf{I}\Big)^{-1} \times \mathbf{H}_i^i\mathbf{V}^i[t], \ \forall i,t$
- Update $\mathbf{W}_i[t]$ by calculating the local gradient of the utility function; see [16] for details
- $\mathbf{V}^i[t] \leftarrow \Big(\sum_j (\mathbf{H}_j^i)^H\mathbf{U}_j[t]\mathbf{W}_j[t](\mathbf{U}_j[t])^H\mathbf{H}_j^i + \mu_i^*\mathbf{I}\Big)^{-1} \times (\mathbf{H}_i^i)^H\mathbf{U}_i[t]\mathbf{W}_i[t], \ \forall i,t$

until convergence
if $\|\mathbf{V}^i[t]\| > \epsilon$ then $\alpha_i[t] \leftarrow 1$; otherwise $\alpha_i[t] \leftarrow 0, \forall i,t$

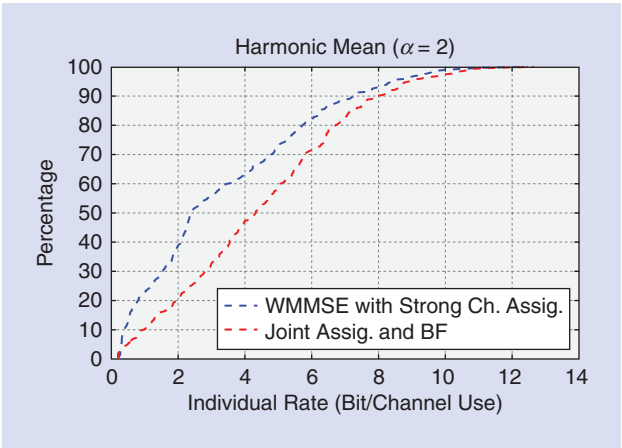
BEAMFORMING AND BS ASSIGNMENT

The densification of the BS deployment gives the users the flexibility of choosing its serving BS from potentially a large pool of nearby BSs. Traditionally, BS assignment is made on the basis of signal strength (or the distances between the BSs and a given user). However, such a greedy scheme often leads to suboptimal solutions and causes unfairness among users, especially when the network is congested [18]. It may be more beneficial to assign users to a different BS when the closest BS is congested [18]. Such freedom in assignment, if properly exploited, can result in substantial gains both in terms of network throughput and user fairness [8], [19].

We show below how to integrate BS assignment with beamforming in the WMMSE framework. Consider the following problem over both the BS assignment and beamforming variables $\{z_i^q, \mathbf{V}_i^q\}_{i,q}$:

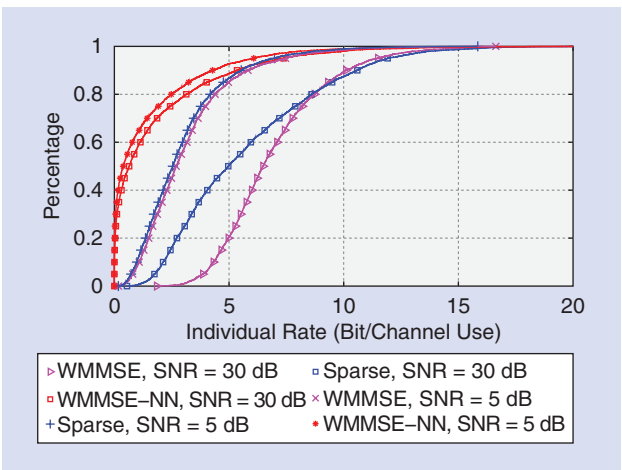
$$\begin{aligned} \underset{z,\mathbf{V}}{\text{maximize}} \quad & U(\{R_i\}_{i\in\mathcal{I}}) \\ \text{subject to} \quad & \sum_q z_i^q \le 1, \quad z_i^q \in \{0,1\}, \forall i,q \\ & \sum_{i\in\mathcal{I}^q} \mathrm{Tr}(\mathbf{V}_i^q(\mathbf{V}_i^q)^H) \le \overline{P}^q \\ & R_i = \sum_q z_i^q \log\det\Big(\mathbf{I} + \mathbf{H}_i^q\mathbf{V}_i^q(\mathbf{V}_i^q)^H(\mathbf{H}_i^q)^H\Big(\sigma_i^2\mathbf{I} \\ & \qquad + \sum_{(j,p)\neq(i,q)}\mathbf{H}_i^p\mathbf{V}_j^p(\mathbf{V}_j^p)^H(\mathbf{H}_i^p)^H\Big)^{-1}\Big). \end{aligned} \tag{12}$$

If the optimization variable $\{z_i^q\}$ is fixed, then the above optimization problem reduces to (6). The extra binary constraint in (12) makes the problem difficult to solve. One way to



[FIG3] Rate CDF achieved by joint BS assignment and beamforming [8].

alleviate this difficulty is to relax it to $0 \le z_i^q \le 1$. It is not hard to see that the relaxation is tight when the objective is linearized with respect to the rate of the users. On the other hand, when the association variables are fixed and the objective is linearized with respect to the rate of the users, the problem is similar to the beamforming for sum rate maximization [cf. (7)], which can be handled using the WMMSE algorithm. To update the association variables, we use the gradient projection method. Combining the gradient projection step for updating the association variables and the WMMSE algorithm for updating the beamformers, we can solve the above problem to a stationary point; see [8] for more details. Figure 3 illustrates the performance gain achieved by the joint precoder design and BS assignment. For a benchmark, we have chosen the greedy BS assignment method, i.e., first each user is assigned to the BS with the strongest channel and then the WMMSE is used to optimize the beamformers. Figure 3 shows that the joint beamforming and BS assignment achieves significantly higher throughput and fairness. For other relaxations of the BS assignment problem, we refer to [18].



[FIG4] The rate CDF achieved by joint BS clustering and beamforming [21].

BS CLUSTERING FOR CoMP

When there are a large number of BSs available, coordinated transmission and reception is shown to be very effective in improving the overall spectrum efficiency [20]. Two popular approaches are coordinated beamforming and joint processing. In the first approach, a user's data resides only at its serving BS and the beamformers are jointly optimized among the coordinated BSs to suppress multiuser interference. The second approach is joint transmission where the user data signals are shared among the cooperating BSs, resulting in a single virtual BS with a large set of antennas. In this way, the HetNet is reduced to a large MIMO broadcast channel. However, full cooperation among all BSs may require significant overhead across the backhaul. In practice, it is desirable to limit the amount of coordination by grouping the BSs into cooperating clusters of small sizes, within which joint processing is performed. In this way, users' data signals are shared only within clusters, thus reducing the overall backhaul signaling overhead. Many recent works have developed various BS clustering strategies for such purposes where clusters are formed either greedily or by an exhaustive search procedure. Once the clusters are formed, various approaches can be used to design beamforming strategies for each BS.

An important cross-layer design issue in this context is how to form BS clusters in conjunction with beamforming and BS coordination so as to strike the best tradeoff among system throughput performance and signaling overhead. To formally state the problem, let us define $\mathcal{Q}_i$ to be a subset of BSs that can potentially cooperate to serve a particular user i, e.g., the set of BSs deployed in the same cell as user i. The goal of clustering is to reduce the system overhead by setting the less beneficial precoders to zero. Such sparse precoder structure can be imposed by solving the following problem, where a group Lasso regularizer is introduced in the objective to promote sparsity among potential precoders [21]:

$$\begin{aligned} \underset{\{\mathbf{V}_i^q\}}{\text{maximize}} \quad & U(\{R_i\}_{i\in\mathcal{I}}) - \lambda \sum_{i\in\mathcal{I}} \sum_{q\in\mathcal{Q}_i} \|\mathbf{V}_i^q\|_F \\ \text{subject to} \quad & (5),\ \sum_{i\in\mathcal{I}^q} \mathrm{Tr}(\mathbf{V}_i^q(\mathbf{V}_i^q)^H) \le \overline{P}^q, \forall q. \end{aligned} \tag{13}$$

Once again, the WMMSE algorithm is naturally suited to solve this problem. The key here is to recognize that after performing the equivalent transformation [cf. (10)], the $\mathbf{u}$, $\mathbf{w}$ subproblems are exactly the same as before (with closed-form solutions), and the $\mathbf{v}$ subproblem becomes a quadratic problem penalized by a group-LASSO regularizer, which is still easy to solve. By again using the alternate optimization approach, (13) can be solved to a stationary point; see [21] for more details. It is worth noticing the tradeoff in the choice of parameter λ: larger values of λ result in smaller size clusters, but lower system throughput. Based on the numerical experiments, a simple rule for selecting the value of λ is suggested in [21]. Figure 4 shows the performance gain of the sparse-WMMSE algorithm. This numerical experiment is run over a network with 20 BSs and 40 users in the system; the number of transmit (respectively receive)

antenna is four (respectively 2). The algorithm is evaluated in the low- and high-signal-to-noise (SNR) regime; see [21] for more details. For comparison, we have also simulated the performance of two other approaches: WMMSE and NN-WMMSE. In WMMSE, full cooperation is considered among the BSs of each cell and the WMMSE algorithm is used to optimize the precoders. Clearly, this corresponds to one single cluster that requires the most system overhead. In contrast, the NN-WMMSE only picks the BS with the strongest channel to serve each user. Thus, the NN-WMMSE method corresponds to greedy BS assignment followed by optimal beamforming.

JOINT PHY-NETWORK LAYER OPTIMIZATION: RESOURCE ALLOCATION AND TRAFFIC ENGINEERING

In light of growing traffic demand and large network size, backhaul links may be capacity limited [1], [3], [4]. Therefore traffic engineering within the multi-hop backhaul network (e.g., traffic routing) must be considered together with resource allocation in the radio air interface (e.g., precoder design and scheduling).

AN IMPORTANT CROSS-LAYER DESIGN ISSUE IN THIS CONTEXT IS HOW TO FORM BS CLUSTERS IN CONJUNCTION WITH BEAMFORMING AND BS COORDINATION SO AS TO STRIKE THE BEST TRADEOFF AMONG SYSTEM THROUGHPUT PERFORMANCE AND SIGNALING OVERHEAD.

One interesting approach to such joint optimization problem, which is gaining support from both academia and industry, is to manage the entire network by a few cloud centers. The shift of the computation tasks from a large number of heterogeneous BSs to a few cloud centers is attractive to the operators, as it allows for an effective and energy efficient way of managing the entire network. This new architecture is called Cloud-RAN or SDN-RAN [1], [3], [4].

There are two main methods in the literature to deal with the per-link capacity constraints in the backhaul. The first method allows the cloud center to compute a joint precoding strategy for all the BSs and then compress the precoded messages before sending to the BSs via the backhaul; see [22] and the references therein. The limited backhaul capacity determines the level of compression needed for each data stream. To illustrate the idea, let us consider a joint downlink compression and sum-rate maximization problem in the Cloud-RAN. Suppose that each BS q is connected to the cloud center via a dedicated line with capacity $\bar{C}^q$. Let $\mathbf{V}^q \triangleq \{\mathbf{V}_i^q\}_{i\in\mathcal{I}}$ denote the precoder used by BS q to serve all the users and let $\mathbf{H}_i \triangleq \{\mathbf{H}_i^q\}_{q\in\mathcal{Q}}$ denote the channel for user i. Define $\mathbf{V}_i$ similarly. Let $\Omega_{q,p} \in \mathbb{C}^{M\times M}$ denote the correlation matrix between the quantization noises of BS q and p. Let Ω be the compression covariance with $\Omega_{q,p}$ as its (q,p)th block component. By using the precoding and compression scheme described in [22], user i's rate can be expressed as

$$R_i(\mathbf{V},\Omega) = \log|\mathbf{I}+\mathbf{H}_i(\mathbf{V}\mathbf{V}^H+\Omega)\mathbf{H}_i^H| - |\mathbf{I}+\mathbf{H}_i(\sum_{j\neq i}\mathbf{V}_j\mathbf{V}_j^H+\Omega)\mathbf{H}_i^H|. \quad (14)$$

Then the problem can be formulated as

$$\underset{\mathbf{V},\Omega\succeq 0}{\text{maximize}} \quad \sum_{i\in\mathcal{I}} R_i(\mathbf{V},\Omega) \quad (15a)$$

$$\text{subject to} \quad \sum_{q\in\mathcal{S}} \log\frac{|\mathbf{V}^q\mathbf{V}^q+\Omega_{q,q}|}{|\Omega_{q,q}|} \le \sum_{q\in\mathcal{S}}\bar{C}^q, \forall \mathcal{S}\subseteq\mathcal{Q}, \quad (15b)$$

$$\mathrm{Tr}[\mathbf{V}^q(\mathbf{V}^q)^H+\Omega_{q,q}] \le \bar{P}^q, \forall q\in\mathcal{Q}, \quad (15c)$$

where the constraints in (15b) ensure that the messages can be reliably transferred to the BSs through the backhaul. With this formulation, the compression covariance as well as transmit precoders can be jointly designed using optimization algorithms such as SCA [12], [13]. However, this line of work usually assumes that there is a single-hop direct connection between the BSs and the cloud centers, and that the routing of the traffic within the backhaul has been predetermined.

The second approach combines the traditional utility-based traffic engineering with the resource allocation in the access network. The idea is to optimize the routing scheme in the backhaul and the power allocation/precoding strategy in the access network together. Different from the precoding-compression scheme, this approach allows for flexible backhaul structure, and the resulting algorithm can be implemented in parallel among the cloud centers. To illustrate the main ideas, let us consider an instance of such joint optimization problem for a single time slot (thus ignoring the time indices):

$$\text{maximize} \quad U(\{R_i\}_{i\in\mathcal{I}}) \quad (16a)$$

$$\text{subject to} \quad R_i^{(l)} \ge 0,\ i\in\mathcal{I},\ \forall l\in\mathcal{L}, \text{and (3)} \quad (16b)$$

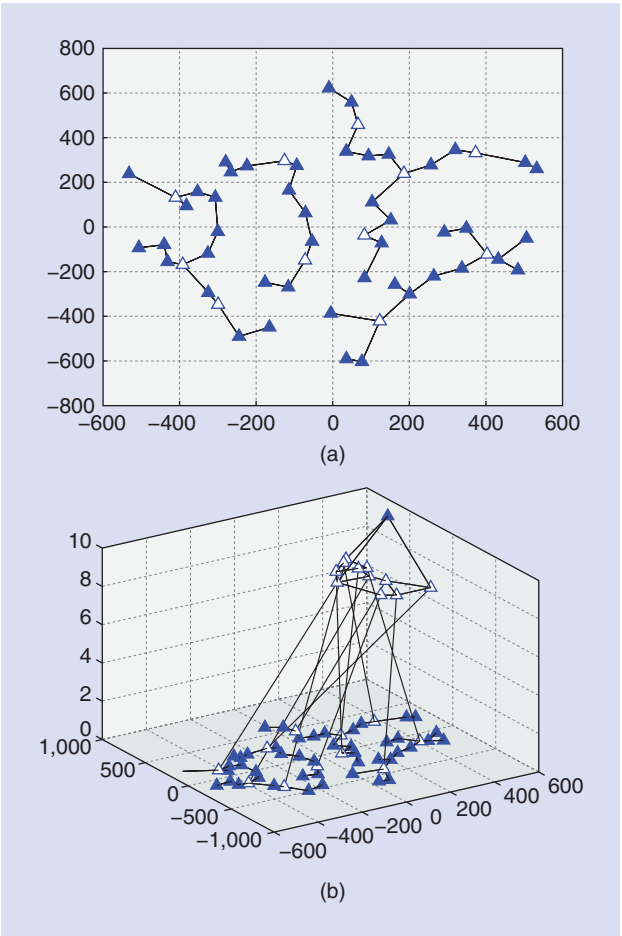
$$\sum_{i\in\mathcal{I}} R_i^{(l)} \le \bar{C}^l, \forall\, l\in\mathcal{L}^{\mathrm{w}}, \text{and (4)} \quad (16c)$$

$$\sum_{i\in\mathcal{I}} R_i^{(l)} \le \bar{R}^l, \forall\, l\in\mathcal{L}^{\mathrm{wl}}, \quad (16d)$$

where $\bar{R}^l$ represents the capacity of a wireless link l, which is a function of transmit strategies of all BSs interfering link l. Compared with the problems studied in the previous sections, the joint design in (2) is more difficult because of the large number of additional variables $\{R_i^{(l)}\}$ and constraints (16b) and (16c), which together describe how the traffic is routed within the backhaul.

Variants of the joint optimization problem (2) have been studied extensively in the framework of cross-layer optimization; see [23]. However, most of the existing methods assume that the air interface is interference free, leading to a tractable convex problem [23]. As argued throughout this article, such interference-free transmission is unrealistic in future RANs.

Recently, [24] proposed solving the joint routing and interference management problem by combining the WMMSE and



[FIG5] **The considered network topology. (a) The locations and the connectivity of all the BSs. (b) The connections between BSs and routers, which are displayed in the upper part of the graph [24].**

the powerful alternating direction method of multipliers (ADMMs). The ADMM algorithm is designed for structured convex optimization problem with linearly coupled constraints. The idea in [24] is quite simple: use WMMSE to take care of the interference management in the RAN, and apply ADMM to handle the traffic engineering in the backhaul. To highlight ideas, we adopt the max-min utility function and consider the single antenna case (generalization to MIMO case and other utilities is nontrivial, but possible). Transmit and receive beamformers now become scalars, denoted using lowercase letters. Similar to the WMMSE algorithm, we first transform (16) into an equivalent but easily manageable form. To this end, introduce two sets of additional optimization variables $\mathbf{u} \triangleq \{u_l \mid l \in \mathcal{L}^{\text{wl}}\}$ (the receive beamformers) and $\mathbf{w} \triangleq \{w_l \mid l \in \mathcal{L}^{\text{wl}}\}$ (the weights), and explore the well-known rate-MSE relationship for each wireless link $l = (s, d, f) \in \mathcal{L}^{\text{wl}}$:

$$\overline{R}^l = \max_{u_l, w_l} c_{1,l} + c_{2,l} v_d^s[f] - \sum_{n=(s',d',f')} c_{3,\text{ln}} \mid v_{d'}^{s'}[f] \mid^2. \quad (17)$$

Here $(c_{1,l}, c_{2,l}, c_{3,\text{ln}})$ are functions of $\mathbf{u}$ and $\mathbf{w}$ given by $c_{1,l} = 1 + \log(w_l) - w_l(1 + \sigma_d^2 \mid u_l \mid^2)$, $c_{2,l} = 2w_l \text{Re}\{u_l^* h_d^s[f]\}$, and $c_{3,\text{ln}} = w_l \mid u_l \mid^2 \mid h_d^{s'}[f] \mid^2$. The difficult wireless channel rate constraints (16d) are then replaced with the following rate-MSE constraint

$$\sum_{i \in \mathcal{I}} R_i^{(l)} \le c_{1,l} + c_{2,l} v_d^s[f] - \sum_{n=(s',d',f')} c_{3,\text{ln}} \mid v_{d'}^{s'}[f] \mid^2, \forall l \in \mathcal{L}^{\text{wl}}. \quad (18)$$

This transformation is motivated by the following facts: 1) any $\{\mathbf{v}, \mathbf{R}\}$ satisfies (18) must also satisfy (16d); 2) constraint (16d) is replaced by (18), so the modified version of (2), is convex with respect to any single variable $\mathbf{w}$, $\mathbf{u}$, and $\{\mathbf{v}, \mathbf{R}\}$. Therefore, the alternating optimization technique is again applied to solve the transformed problem; see the N-MaxMin algorithm outlined in Algorithm 3. As shown in [24], the iterates generated by this algorithm converge to a stationary solution of (2).

Algorithm 3 The N-MaxMin algorithm for minimum rate maximization in SDN-RAN (SISO case).

Initialize feasible $\mathbf{u}$, $\mathbf{w}$, $\mathbf{v}$, and $\mathbf{R}$ randomly
repeat
- $\forall\, l = (s, d, f) \in \mathcal{L}^{\text{wl}}$
  $u_l \leftarrow \left(\sum_{n=(s',d',f')} h_d^{s'}[f] \mid^2 \mid v_{d'}^{s'}[f] \mid^2 + \sigma_d^2 \right)^{-1} h_d^s[f] v_d^s[f]$
  $w_l \leftarrow (1 - (h_d^s[f] v_d^s[f])^* u_l)^{-1}$
- Update $\{\mathbf{v}, \mathbf{R}\}$ via inner ADMM algorithm; cf. [24]

until Convergence

In the N-MaxMin, solving $\mathbf{u}$ and $\mathbf{w}$ is again easy and in closed form. What is different from the sum-rate WMMSE algorithm so far is that the update for $\{\mathbf{v}, \mathbf{R}\}$ is not in closed form. This step requires solving a huge convex optimization subproblem, which couples all the nodes in the system. To make the entire algorithm scalable and suitable for distributed implementation, the ADMM is then used to decompose the huge problem for $\{\mathbf{v}, \mathbf{R}\}$ into a collection of easy subproblems. The trick here is to decouple the flow conservation constraints (4) and the rate-MSE constraints (18) by introducing a few extra variables. We refer to [24] for a detailed description of the algorithm. We mention that all the steps of the resulting ADMM iterations are separable over the nodes/links of the network, and each of them can be updated in (semi)closed-form. More importantly, such separability structure makes the N-MaxMin algorithm implementable in parallel by a few cloud centers, each handling a subset of nodes/flows.

NUMERICAL EXAMPLES

Let us illustrate the performance of the N-MaxMin algorithm using a few examples. The topology and the connectivity of the testing network are shown in Figure 5. The source (destination) node of each commodity is randomly selected from network routers (mobile users). For the backhaul links of this network, a fixed capacity is assumed in the range between 2.88 Mnats/s to 1.44 Gnats/s, and they are the same in both directions. The number of subchannels is $K = 3$ and each subchannel has 1-MHz bandwidth, and the wireless links follow the distribution $CN(0, (200/\text{dist})^3)$, where "dist" is the distance between a BS and mobile user.

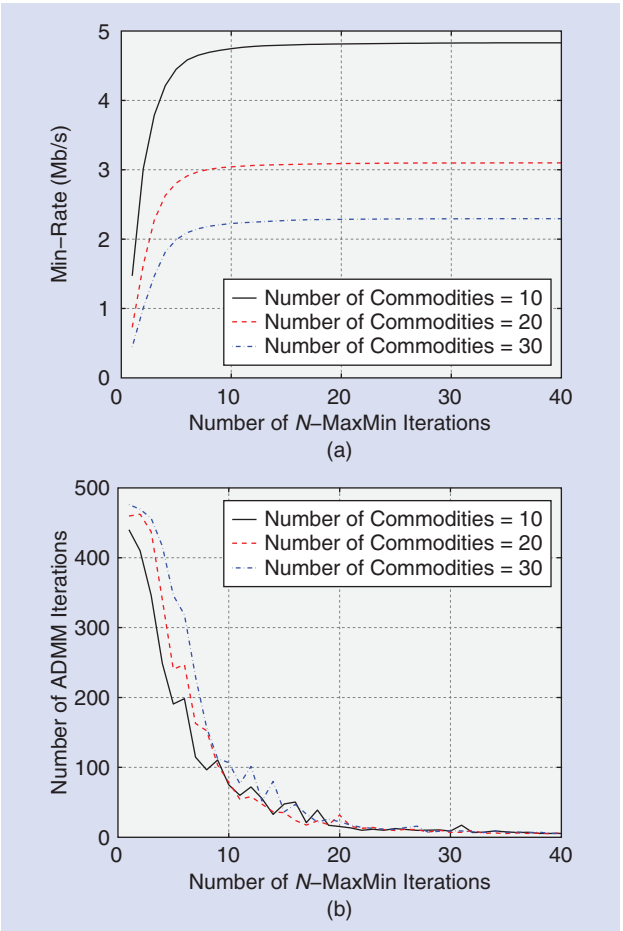
To accelerate the initial steps, a maximum of 500 inner iterations are allowed for the first five outer iterations. Figure 6 shows that the min-rate converges at about the tenth outer iteration when the number of commodities is up to 30, while fewer than 500 inner iterations are needed per outer iteration. Moreover, after the tenth outer iteration, the number of inner ADMM iterations drops below 100.

Next we demonstrate how parallel implementation can speed up the algorithm considerably. To illustrate the benefit of parallelization, we consider a larger network that is derived by merging two identical networks shown in Figure 5(a). For simplicity, we removed all the wireless links, so this reduces (2) to a network flow problem (a very large linear program). Here, the N-MaxMin algorithm is implemented by the Open Message Passing Interface (MPI) package with nine computation cores, and is compared with Gurobi [a state-of-the-art commercial linear program (LP) solver] in terms of efficiency. In Table 1, we observe that parallel implementation leads to more than fivefold improvement in computation time, compared with its sequential counterpart. We also note that when the problem size increases to 8×10^4 variables and over 10^5 constraints, Gurobi becomes slower than the proposed algorithm (implemented in parallel). Thus, the proposed ADMM algorithm scales well with problem size.

PRACTICAL CONSIDERATIONS: OVERHEAD REDUCTION

In the previous sections, the channel states are assumed known. In practice, the direct channel state is estimated using pilot messages. Performing only a few channel estimations might lead to effects such as channel aging, resulting in imperfect channel estimation. For future cellular networks, allowing simultaneous use of frequency slots leads to interference. Hence, in addition to the direct channels, the states of the interfering links also need to be estimated. Moreover, all these channel estimations should be fed back to the transmitters. With the large number of cochannel transmissions, estimating the states of all interfering links is infeasible. A major challenge to implement the cross-layer management algorithms (such as the ones presented in the previous sections) is the overhead associated with channel estimation. In this section, we discuss some of the possible ways to partially relax the CSI at the transmitter side (CSIT).

For wireless links, the channel state can be modeled as a random variable with a known distribution. The distributions can be calibrated using experimental data and are usually defined by long-term physical parameters such as path-loss coefficients, propagation distance, and channel training time/power. The first approach to CSIT mismatch is the use of robust optimization techniques; see, e.g., [25]. These methods are mostly designed for the worst-case scenarios and, therefore, due to their nature, are suboptimal when the worst cases happen with small probability. On one hand, these methods can give guarantees on the QoS constraint with high probability (measured based on the channel distribution). On the other hand, as they need to work with outage probability type of constraints, these methods are typically



[FIG6] The min-rate performance and the required number of iterations for the proposed N-MaxMin algorithm. (a) plots the obtained min-rate versus the iterations of N-MaxMin. (b) plots the required number of inner ADMM iterations versus the iteration for the outer N-MaxMin algorithm [24].

rather complex compared to their nonrobust counterparts. In addition, they have a limited scope and cannot deal with realistic channel distributions due to analytical intractability. In fact, most of them use a Gaussian channel distribution with estimated channels as the means for the Gaussian distributions. Therefore, utilizing these methods still requires estimating the channel for all the links in the network (including all the interfering links), albeit the channel estimation need not be very accurate. For a large HetNet, this approach is still not practical since estimating

[TABLE 1] THE COMPARISON OF COMPUTATION TIME USED BY DIFFERENT IMPLEMENTATIONS OF THE ADMM ALGORITHM FOR THE ROUTING PROBLEM.

NUMBER OF COMMODITIES	50	100	200	300
NUMBER OF VARIABLES ($\times 10^4$)	1.4	2.9	5.8	8.7
NUMBER OF CONSTRAINTS ($\times 10^4$)	2.1	4.2	8.4	13
SEQUENTIAL UPDATE (s)	1.04	2.03	4.73	8.53
PARALLEL UPDATE (s)	0.20	0.37	0.75	1.10
GUROBI (s)	0.20	0.64	1.65	2.51

all the channel states requires excessive training overhead.

An alternative approach is to maximize the expected performance using a stochastic optimization framework that requires only the statistical channel knowledge rather than the full instantaneous CSI. This method is proposed in [26], where the goal is to maximize the sum of expected utilities of the users. To illustrate this, we can consider a similar formulation as (7) in an IC, but using an averaged sum rate as the objective (extensions to IBC or CoMP is possible)

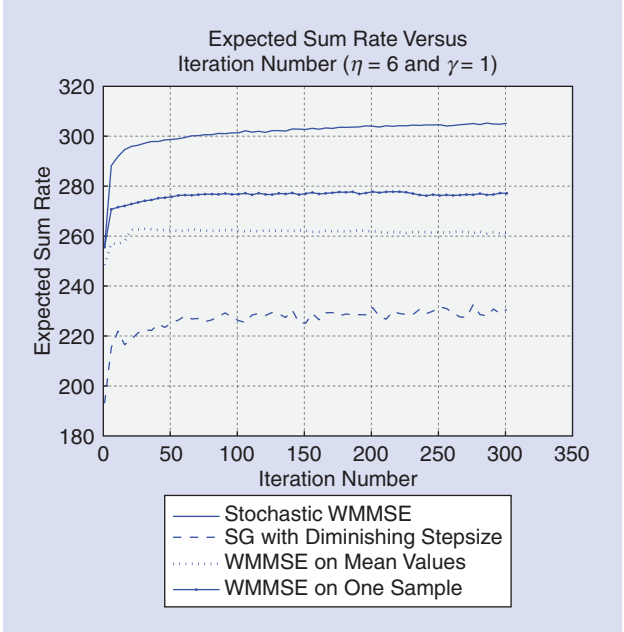
$$\begin{aligned} &\underset{\mathbf{v}^i,\mathbf{u}_i}{\text{maximize}} \quad \sum_i \mathbb{E}\,[R_i] \\ &\text{subject to} \quad \|\mathbf{v}^i\| \le \overline{P}^i, \forall i. \end{aligned} \tag{19}$$

The algorithm can be viewed as a generalization of the WMMSE algorithm (Algorithm 1) to the stochastic setting. The algorithm proceeds by drawing/receiving a sample of the channels in each step using the known channel distributions and then applying a step of the WMMSE to an average of the utilities defined by the current, as well as the previous, channel samples. The details of the algorithm are presented in Algorithm 4. It can be seen that the steps of the algorithm are similar to Algorithm 1, with the only difference being in the use of matrices $\mathbf{A}$ and $\mathbf{B}$, which serve to accumulate the information of all the previous channel samples thus far. Note that in stochastic WMMSE, transmit precoders $\mathbf{v}$ are adapted to channels statistics known at transmitters, while receive beamformers are adapted optimally to real-time channels. Therefore, the transmit precoder designed by the stochastic WMMSE algorithm is less sensitive to the changes in the real-time channel. Furthermore, in practice, it is easy to incorporate the changes in statistical models of the channels into the algorithm by using a forgetting factor that reduces the effect of the past channel realizations on the choice of the current precoder.

WHEN THERE ARE A LARGE NUMBER OF BSs AVAILABLE, COORDINATED TRANSMISSION AND RECEPTION IS SHOWN TO BE VERY EFFECTIVE IN IMPROVING THE OVERALL SPECTRUM EFFICIENCY. TWO POPULAR APPROACHES ARE COORDINATED BEAMFORMING AND JOINT PROCESSING.

Algorithm 4 Stochastic WMMSE algorithm.

Initialize $\mathbf{v}$ randomly such that $\|\mathbf{v}^i\|^2 = \overline{P}^i, \forall\, i$
repeat
  Obtain the new channel estimate/realization $\mathbf{H}$
  $\mathbf{u}_i \leftarrow \left(\sum_j \mathbf{H}_i^j \mathbf{v}^j (\mathbf{v}^j)^H (\mathbf{H}_i^j)^H + \sigma_i^2 \mathbf{I}\right)^{-1} \mathbf{H}_i^i \mathbf{v}^i, \forall\, i$
  $w_i \leftarrow (\mathbf{I} - \mathbf{u}_i^H \mathbf{H}_i^i \mathbf{v}^i)^{-1}, \forall\, i$
  $\mathbf{A}_i \leftarrow \mathbf{A}_i + \sum_j w_j (\mathbf{H}_j^i)^H \mathbf{u}_j \mathbf{u}_j^H \mathbf{H}_j^i, \forall\, i$
  $\mathbf{B}_i \leftarrow \mathbf{B}_i + (\mathbf{H}_i^i)^H \mathbf{u}_i w_i, \forall\, i$
  $\mathbf{v}^i \leftarrow (\mathbf{A}_i + \mu_i^* \mathbf{I})^{-1} \mathbf{B}_i, \forall\, i,$
  where μ_i^* is computed by bisection to ensure that $\|\mathbf{v}^i\|^2 \le \overline{P}^i$.
until convergence



[FIG7] The expected sum rate versus iteration number. We set $\eta = 6$, $\gamma = 1$ and consequently only 3% of the channel matrices are estimated, while the rest are generated by their path-loss coefficients plus Rayleigh fading. The SNR is set SNR = 15 (dB) [26].

We emphasize that the stochastic WMMSE algorithm converges to a stationary solution for any channel distributions (and not just Gaussian). In practice, each user can estimate the direct channel and as well as a few (say, one to three) interfering channels. The estimated channels can be modeled as Gaussian random variables whose variances may be specified by the training SNR. For the rest of the interfering channels, we can use standard path-loss statistics plus scattering to model them. Estimating path loss, which is slow varying, is much easier and requires significantly less communication overhead than estimating the fast changing channels. Note that the stochastic WMMSE is computationally as cheap as its nonstochastic counterpart of the section "PHY-Layer Interference Management." Moreover, it can be generalized to handle joint clustering and beamforming for CoMP transmission. Such generalization can be done easily by adding a sparsity promoting penalty term to the objective in (19).

To evaluate the effectiveness of the stochastic WMMSE algorithm, we consider a network of 19 cells, each with three separate sectors, for a total of $K = 57$ BSs. Each BS is equipped with four antennas and serves the users (all equipped with two antennas) in its own sector. The path loss and the power budget of the transmitters are generated using the Third-Generation Partnership Project (3GPP) (TR 36.814) evaluation methodology [27]. We assume that

partial channel state information is available for some of the links. In particular, each user estimates only its direct channel in addition to the interfering links whose powers are at most η (dB) below its direct channel power (η is a simulation parameter). For these estimated channels, we assume a channel estimation error model in the form of $\hat{h} = h + z$, where h is the actual channel; $\hat{h}$ is the estimated channel, and z is the estimation error. Given an MMSE channel estimate $\hat{h}$, we can determine the distribution of h as $\mathcal{CN}(\hat{h}, (\sigma_l^2)/(1+\gamma\text{SNR}))$, where γ is the effective SNR coefficient depending on the system parameters (e.g., the number of pilot symbols used for channel estimation) and σ_l is the path loss. Moreover, for the channels that are not estimated, we assume the availability of estimates of the path loss σ_l and use them to construct statistical models (Rayleigh fading is considered on top of the path loss).

A MAJOR CHALLENGE TO IMPLEMENT THE CROSS-LAYER MANAGEMENT ALGORITHMS IS THE OVERHEAD ASSOCIATED WITH CHANNEL ESTIMATION.

We compare the performance of four different algorithms:

- one-sample WMMSE
- mean WMMSE
- stochastic gradient
- stochastic WMMSE.

In the one-sample WMMSE and the mean WMMSE, we apply the WMMSE algorithm [7] to one realization of all channels and to the mean channel matrices, respectively. For the stochastic gradient method, we use a diminishing stepsize rule to the ergodic sum rate maximization problem. For more numerical experiments and the details of the simulations, we refer interested readers to [26]. Figure 7 shows our simulation results when each user only estimates the strongest 3% of its channels, while the others are generated synthetically according to the channel distributions. The expected sum rate in each iteration is approximated in this figure by a Monte-Carlo averaging over 500 independent channel realizations. As can be seen from Figure 8, the stochastic WMMSE algorithm significantly outperforms the rest of the algorithms.

CONCLUDING REMARKS

A key feature of future wireless networks is the large number of BSs, each with varying capability and connected to a backhaul network to serve the user needs. How to best manage such networks to meet the users' high-speed, high-mobility requirement is a major challenge. In this article, we outlined a cross-layer approach, one that leverages resource allocation in the physical layer, user scheduling in the access layer, and traffic engineering in the network layer to maximize an overall system utility. Optimization and signal processing are the key elements in this approach.

While we have highlighted a few recent developments in this area, much remains to be done, especially in the directions of high-performance and low energy footprint network provision algorithms that can be implemented efficiently in a parallel manner. In particular, it will be important to include the emerging concept of "green communication" [28] into the algorithm design. Energy efficiency can be indirectly achieved by, e.g., MAC layer mechanisms similar to the BS clustering in the section "BS Clustering for CoMP." That is, by dynamically shutting down a few BSs that are underutilized, the overall energy consumption can be greatly reduced; see, e.g., the recent work [29]. Alternatively, PHY-layer approaches can be used to directly maximize the energy efficiency measured by the throughput per unit of energy consumption; see [30] and the references therein. It will be interesting to see how these separate approaches can be integrated in a cross-layer design framework. Another important direction is to evaluate and extend the algorithms proposed so far using realistic network-layer protocols, traffic patterns, and network models. Performance metrics such as throughput, delay, traffic queue stability, and achieved QoS need to be carefully evaluated to validate the benefits brought by the cross-layer design.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation, grant CCF-1216858.

AUTHORS

Hadi Baligh (hadi.baligh@huawei.com) received his B.Sc. degree in electrical engineering from Isfahan University of Technology, Iran, in 1996, and his M.Sc. degree from Sharif University of Technology, Tehran, Iran, in 1999. He received his Ph.D. degree from the University of Waterloo, Ontario, Canada, in 2006. From 1999 to 2001, he was with Basamadnegar, in Tehran, Iran, and from 2006 to 2009, he was with Wireless Technology Labs in Nortel, Ottawa, Ontario, Canada. In 2009, he joined Huawei Canada in Ottawa, where he is currently a senior engineer. His research interests include interference management, multiple-input, multiple-output systems, cooperative communication, and channel coding.

Mingyi Hong (mhong@umn.edu) received his B.E. degree in communications engineering from Zhejiang University, China, in 2005, his M.S. degree in electrical engineering from Stony Brook University in 2007, and Ph.D. degree in systems engineering from the University of Virginia in 2011. He was a research assistant professor with the Department of Electrical and Computer Engineering, University of Minnesota, Twin Cities. He is currently an assistant professor with the Department of Industrial and Manufacturing Systems Engineering, Iowa State University, Ames. His research interests include signal processing, wireless communications, large-scale optimization and its applications in compressive sensing, complex networks, and high-dimensional data analysis.

Wei-Cheng Liao (liaox146@umn.edu) received his B.S. degree in electrical engineering and the M.S. degree in communications engineering from the National Tsing Hua University, Hsinchu, Taiwan, in 2008 and 2010, respectively. He is now working toward his Ph.D. degree in the Department of Electrical and Computer Engineering, University of Minnesota.

His research interests are in wireless communications, optimization theory, and signal processing.

Zhi-Quan Luo (luozq@umn.edu) received his B.Sc. degree in applied mathematics in 1984 from Peking University, Beijing, China. He received a Ph.D. degree in operations research in 1989 from the Massachusetts Institute of Technology. From 1989 to 2003, he held a faculty position with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, Canada, where he eventually became the department head and held a Canada Research Chair in information processing. Since 2003, he has been with the Department of Electrical and Computer Engineering at the University of Minnesota, Twin Cities as a full professor and holds an endowed ADC chair in digital technology. His research interests include optimization algorithms, signal processing, and digital communication.

Meisam Razaviyayn (meisam@umn.edu) received his B.Sc. degree in electrical engineering from Isfahan University of Technology, Iran, in 2008, his M.S. degrees in mathematics and electrical engineering from the University of Minnesota in 2013, and his Ph.D. degree in electrical engineering from the University of Minnesota in 2014. He is currently a visiting research assistant at Daniel J. Epstein Industrial and Systems Engineering at the University of Southern California. During the summer of 2010, he worked as a research intern with the Wireless Research Group of Huawei Technologies North America, Chicago. His research interests include wireless communication and its computational issues, large-scale optimization, machine learning, and statistical signal processing.

Maziar Sanjabi (maz@umn.edu) received the B.Sc. degree from Isfahan University of Technology, Iran, in 2008. He is currently working toward his Ph.D. degree in electrical engineering at the University of Minnesota.

Ruoyu Sun (sunxx394@umn.edu) is working toward his Ph.D. degree in the Department of Electrical and Computer Engineering, University of Minnesota, Twin Cities. He received the B.S. degree in mathematics from Peking University, Beijing, China, in 2009. He worked as a research intern with the Wireless Group at Huawei North America in 2011 and 2012. His research interests lie in the areas of convex optimization, wireless communications, and low-dimensional data recovery.

REFERENCES

[1] Huawei Technologies Inc., "5G: A technology vision," 2013, white paper.

[2] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A survey on 3GPP heterogeneous networks," *IEEE Wireless Commun.*, vol. 18, no. 3, pp. 10–21, 2011.

[3] Open Networking Foundation, "Software-defined networking: The new norm for networks," 2012, white paper.

[4] China Mobile, "C-RAN: The road towards green ran," Oct. 2011, white paper, ver. 2.5.

[5] S. S. Christensen, R. Agarwal, E. D. Carvalho, and J. M. Cioffi, "Weighted sum-rate maximization using weighted MMSE for MIMO-BC beamforming design," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 4792–4799, 2008.

[6] D. A. Schmidt, C. Shi, R. A. Berry, M. L. Honig, and W. Utschick, "Minimum mean squared error interference alignment," in *Proc. 2009 Conf. Record of the 43rd Asilomar Conf. Signals, Systems and Computers*, 2009, pp. 1106–1110.

[7] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Processing*, vol. 59, no. 9, pp. 4331–4340, 2011.

[8] M. Sanjabi, M. Razaviyayn, and Z.-Q. Luo, "Optimal joint base station assignment and downlink beamforming for heterogeneous networks," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2012, pp. 2821–2824.

[9] M. Hong and Z.-Q. Luo, "Signal processing and optimal resource allocation for the interference channel," in *Academic Press Library in Signal Processing*. New York: Academic, 2013.

[10] L. Zheng and C. W. Tan, "Maximizing sum rates in cognitive radio networks: Convex relaxation and global optimization algorithms," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 3, pp. 667, 2014.

[11] E. Bjornson and E. Jorswieck, "Optimal resource allocation in coordinated multi-cell systems," *Found. Trends Commun. Inform. Theory*, vol. 9, no. 1–2, pp. 113–381, Jan. 2013.

[12] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. Optim.*, vol. 23, no. 2, pp. 1126–1153.

[13] G. Scutari, F. Facchinei, P. Song, D. P. Palomar, and J.-S. Pang, "Decomposition by partial linearization: Parallel optimization of multi-agent systems," *IEEE Trans. Signal Processing*, vol. 63, no. 3, pp. 641–656, 2014.

[14] D. A. Schmidt, C. Shi, R. A. Berry, M. L. Honig, and W. Utschick, "Distributed resource allocation schemes," *IEEE Signal Processing Mag.*, vol. 26, no. 5, pp. 53–63, 2009.

[15] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint cell selection and radio resource allocation in MIMO small cell networks via successive convex approximation," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, May 2014, pp. 850–854.

[16] M. Razaviyayn, H. Baligh, A. Callard, and Z.-Q. Luo, "Joint user grouping and transceiver design in a MIMO interfering broadcast channel," *IEEE Trans. Signal Processing*, vol. 62, no. 1, pp. 85–94, 2013.

[17] W. Yu, T. Kwon, and C. Shin, "Multicell coordination via joint scheduling, beamforming and power spectrum adaptation," in *Proc. INFOCOM*, 2011, pp. 2570–2578.

[18] R. Stridh, M. Bengtsson, and B. Ottersten, "System evaluation of optimal downlink beamforming with congestion control in wireless communication," *IEEE Trans. Wireless Commun.*, vol. 5, no. 4, pp. 743–751, 2006.

[19] E. Bjornson, M. Kountouris, and M. Debbah, "Massive MIMO and small cells: Improving energy efficiency by optimal soft-cell coordination," in *Proc. 2013 20th Int. Conf. Telecommunications (ICT)*, 2013, pp. 1–5.

[20] G. J. Foschini, K. Karakayali, and R. A. Valenzuela, "Coordinating multiple antenna cellular networks to achieve enormous spectral efficiency," *IEE Proc. Commun.*, vol. 153, no. 4, pp. 548–555, 2006.

[21] M. Hong, R. Sun, H. Baligh, and Z.-Q. Luo, "Joint base station clustering and beamformer design for partial coordinated transmission in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 226–240, 2013.

[22] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai (Shitz), "Joint precoding and multivariate backhaul compression for the downlink of cloud radio access networks," *IEEE Trans. Signal Processing*, vol. 61, no. 22, pp. 5646–5658, Nov. 2013.

[23] M. Chiang, S. Low, A. R. Calderbank, and J. C. Doyle, "Layering as optimization decompositioon: A mathematical theory of network architectures," *Proc. IEEE*, vol. 95, no. 1, pp. 255–312, Jan. 2007.

[24] W.-C. Liao, M. Hong, H. Farmanbar, X. Li, Z.-Q. Luo, and H. Zhang, "Min flow rate maximization for backhaul constrained heterogeneous wireless network," *IEEE J. Sel. Areas Commun.*, to be published. arXiv:1312.5345

[25] M. Shenouda and T. N. Davidson, "On the design of linear transceivers for multiuser systems with channel uncertainty," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 6, pp. 1015–1024, Aug. 2008.

[26] M. Razaviyayn, M. Sanjabi, and Z.-Q. Luo, "A stochastic successive minimization method for nonsmooth nonconvex optimization with applications to transceiver design in wireless communication networks," to be published. arXiv:1307.4457.

[27] 3GPP TR 36.814. [Online]. Available: http://www.3gpp.org/ftp/specs/archive/36_series/36.814/

[28] Z. Hasan, H. Boostanimehr, and V. K. Bhargava, "Green cellular networks: A survey, some research issues and challenges," *IEEE Commun. Surv. Tutorials*, vol. 13, no. 4, pp. 524–540, 2011.

[29] W.-C. Liao, M. Hong, Y.-F. Lui, and Z.-Q. Luo, "Base station activation and linear transceiver design for optimal resource management in heterogeneous networks," *IEEE Trans. Signal Processing*, vol. 62, no. 15, pp. 3939–3952, Aug. 2014.

[30] S. He, Y. Huang, L. Yang, and B. Ottersten, "Coordinated multicell multiuser precoding for maximizing weighted sum energy efficiency," *IEEE Trans. Signal Processing*, vol. 62, no. 3, pp. 741–751, Mar. 2014.

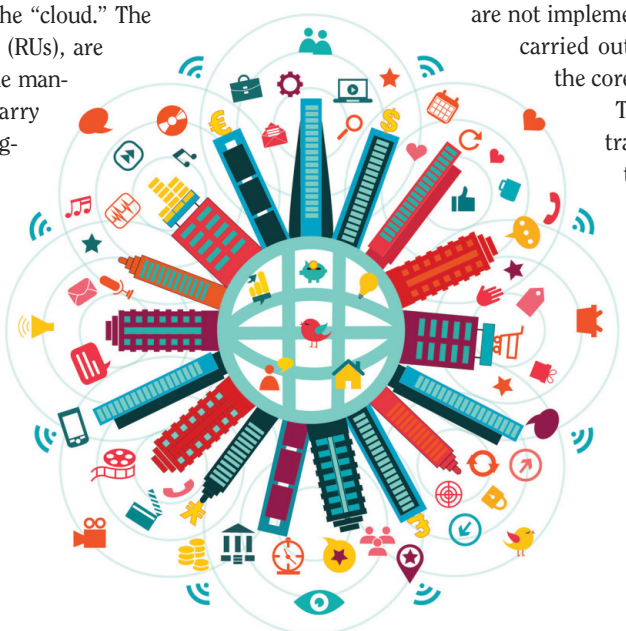
[SP]

Seok-Hwan Park, Osvaldo Simeone, Onur Sahin, and Shlomo Shamai (Shitz)

Fronthaul Compression for Cloud Radio Access Networks

Signal processing advances inspired by network information theory

Cloud radio access networks (C-RANs) provide a novel architecture for next-generation wireless cellular systems whereby the baseband processing is migrated from the base stations (BSs) to a control unit (CU) in the "cloud." The BSs, which operate as radio units (RUs), are connected via fronthaul links to the managing CU. The fronthaul links carry information about the baseband signals—in the uplink from the RUs to the CU and vice versa in the downlink—in the form of quantized in-phase and quadrature (IQ) samples. Due to the large bit rate produced by the quantized IQ signals, compression prior to transmission on the fronthaul links is deemed to be of critical importance and is receiving considerable attention. This article provides a survey of the work in this area with emphasis on advanced signal processing solutions based on network information theoretic concepts. Analysis and numerical results illustrate the considerable performance gains to be expected for standard cellular models.



THE 5G REVOLUTION

INTRODUCTION

C-RANs provide a promising architecture for next-generation wireless cellular systems that is based on the separation of distributed RUs and centralized information processing nodes [1], [2]. Unlike current cellular systems, in C-RANs, the functionalities needed to process the baseband complex or IQ envelopes of the radio signals received/transmitted by the RUs are not implemented at the RUs; instead, they are carried out remotely within the "cloud" of the core network.

To this end, the baseband signals are transferred between the cloud and the RUs on a network of fronthaul links. As an example, Figure 1 illustrates the uplink of a C-RAN in a heterogeneous cellular network with RUs consisting of macro-BSs and pico-BSs and a multihop fronthaul topology between the RUs and the cloud (see, e.g., [3]). Note that on the used nomenclature, fronthaul links are often distinguished from backhaul links in that they have more stringent requirements on latency and synchronization to enable baseband processing in the cloud [3].

The centralization of information processing made possible by C-RANs enables interference management at the geographical scale covered by the distributed RUs (see, e.g., [4]). In fact, C-RANs provide an effective means to implement network multiple-input, multiple-output (MIMO) [5], [6] in heterogeneous wireless networks via the joint processing of the baseband signals at a CU, also known as baseband unit, in the cloud.

As discussed, the key feature of C-RANs is the use of a fronthaul network for the transfer of baseband information to

Digital Object Identifier 10.1109/MSP.2014.2330031
Date of publication: 15 October 2014

and from the cloud. Current solutions, which are the object of various standardization efforts [3], prescribe the use of conventional scalar quantizers for this purpose. However, with this approach, fronthaul capacity limitations are known to impose a formidable bottleneck to the system performance.

EXAMPLE

Consider an RU consisting of an long-term evolution (LTE) macro-BS that serves three cell sectors with five carriers and two receive antennas. As summarized in [7], it can be calculated that, using standard scalar quantization techniques with 15 bits/baseband IQ sample, the throughput required on the fronthaul links exceeds even the 10 Gbit/s provided by standard fiber optics links. The problem is even more pronounced for smaller RUs, e.g., pico-BSs or home-BSs, that, while operating with fewer antennas, channels, and sectors, are typically connected to fronthaul links of lower capacity, such as DSL-based wireline or millimeter-wave channels. ■

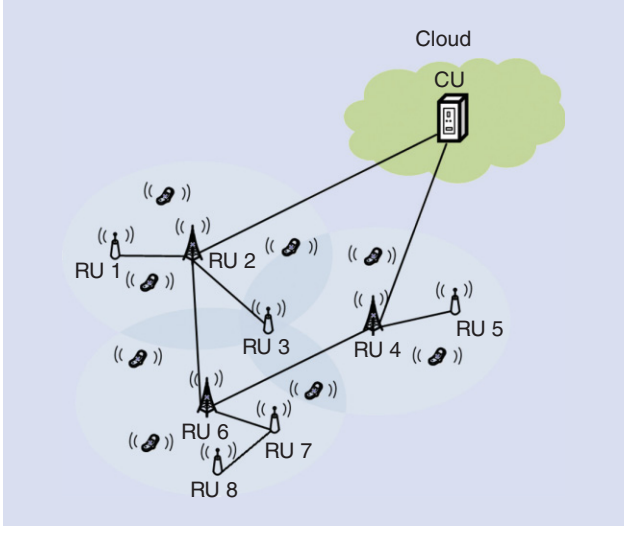
To alleviate the performance bottleneck identified above, recent efforts have targeted the design of more advanced fronthaul compression schemes. These schemes are based on point-to-point compression algorithms (see, e.g., [1], [7], and [8]). However, as is well known from network information theory, point-to-point techniques fail to achieve the optimal performance in the context of even the simplest networks, such as star, or single-hop, topologies [9].

Motivated by the previous discussion, this article aims at providing a survey of the work in the area of fronthaul compression with emphasis on advanced signal processing solutions based on network information theoretic concepts. Specifically, the main ideas that are brought to bear from network information theory are:

1) *Multiterminal compression*: In contrast to point-to-point compression, multiterminal compression allows for the joint processing of the compressed IQ samples of different RUs at the CU. Specifically, in the uplink, joint decompression enables the CU to leverage the correlation among the signals received by neighboring RUs. The key technique that makes this possible is distributed compression or Wyner–Ziv coding [10]. Instead, in the downlink, joint compression allows the CU to correlate the quantization noises of the baseband signals transmitted by neighboring RUs. This can be done via the information-theoretic technique of multivariate compression [9, Ch. 9].

2) *Structured coding*: Point-to-point and multiterminal compression employ unstructured quantization codebooks that are designed independently of the channel codebooks used for transmission on the wireless channels. As a conceptually different alternative, structured codes that are matched to the channel codebooks may instead be used. This leads to new strategies for C-RANs based on the framework of compute-and-forward [11].

In the following, we review point-to-point/multiterminal fronthaul compression and structured coding for the uplink and downlink of a C-RAN. Throughout, we provide numerical results to illustrate the key concepts. We also provide simulation results over standard cellular models to substantiate the gains that are expected from the implementation of multiterminal fronthaul compression in real-world systems. See "Information Theoretic Measures" for a brief review of the standard information theoretic notations used in the article.



[FIG1] The uplink of a C-RAN with a multihop fronthaul topology between the RUs and the cloud, which contains the CU. The solid lines represent the fronthaul links.

UPLINK

SYSTEM MODEL

In a C-RAN, the RUs are partitioned into clusters, such that all RUs within a cluster are managed by a single CU. Within the area covered by a given cluster, there are N_U multiantenna user equipment (UE) and N_R multiantenna RUs. In the uplink, the UE transmits wirelessly to the RUs. In turn, the RUs compress the received baseband signals and transmit the compressed signals on the fronthaul network toward the managing CU.

INFORMATION THEORETIC MEASURES

Throughout the article, we adopt standard information-theoretic definitions for the mutual information $I(X;Y)$, conditional mutual information $I(X;Y|Z)$, differential entropy $h(X)$ and conditional differential entropy $h(X|Y)$ [9]. For jointly complex Gaussian variables $(\mathbf{x},\mathbf{y}) \sim \mathcal{CN}(0, \Omega_{\mathbf{x},\mathbf{y}})$, we define the conditional covariance matrix as $\Omega_{\mathbf{x}|\mathbf{y}} = \mathbb{E}[\mathbf{x}\mathbf{x}^\dagger|\mathbf{y}] = \Omega_{\mathbf{x}} - \Omega_{\mathbf{x},\mathbf{y}} \Omega_{\mathbf{y}}^{-1} \Omega_{\mathbf{x},\mathbf{y}}^\dagger$, where $\Omega_{\mathbf{x}} = \mathbb{E}[\mathbf{x}\mathbf{x}^\dagger]$, $\Omega_{\mathbf{y}} = \mathbb{E}[\mathbf{y}\mathbf{y}^\dagger]$, $\Omega_{\mathbf{x},\mathbf{y}} = \mathbb{E}[\mathbf{x}\mathbf{y}^\dagger]$ and the operation $(\cdot)^\dagger$ denotes the Hermitian transpose of a matrix or vector. Then, for joint complex Gaussian vectors $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$, the quantities $I(\mathbf{x};\mathbf{y})$ and $I(\mathbf{x};\mathbf{y}|\mathbf{z})$ are computed as $I(\mathbf{x};\mathbf{y}) = h(\mathbf{x}) - h(\mathbf{x}|\mathbf{y}) = \text{logdet}(\Omega_{\mathbf{x}}) - \text{logdet}(\Omega_{\mathbf{x}|\mathbf{y}})$ and $I(\mathbf{x};\mathbf{y}|\mathbf{z}) = h(\mathbf{x}|\mathbf{z}) - h(\mathbf{x}|\mathbf{y},\mathbf{z}) = \text{logdet}(\Omega_{\mathbf{x}|\mathbf{z}}) - \text{logdet}(\Omega_{\mathbf{x}|\mathbf{y},\mathbf{z}})$, respectively.

The fronthaul network connecting the RUs to the CU may have a single-hop topology, in which all RUs are directly connected to the CU or, more generally, a multihop topology. We first concentrate on the single-hop topology and then discuss the multihop case. An example of a single-hop C-RAN is the network shown in Figure 1 when restricted to RU 2 and RU 4.

Assuming flat-fading channels, the discrete-time pulse-matched baseband or IQ signal $\mathbf{y}_i^{\text{ul}}$ received by the ith RU at any given time sample can be written using the standard linear model

$$\mathbf{y}_i^{\text{ul}} = \mathbf{H}_i^{\text{ul}}\mathbf{x}^{\text{ul}} + \mathbf{z}_i^{\text{ul}}, \tag{1}$$

where $\mathbf{H}_i^{\text{ul}}$ represents the channel matrix from all the UE in the cluster toward the ith RU; $\mathbf{x}^{\text{ul}}$ is the vector of IQ symbols transmitted by all the UE in the cluster; and $\mathbf{z}_i^{\text{ul}} \sim \mathcal{CN}(\mathbf{0}, \Omega_{\mathbf{z}_i^{\text{ul}}})$ models thermal noise and the interference arising from the other clusters. The signals $\mathbf{x}^{\text{ul}}$ transmitted by the UE are assumed to be jointly complex Gaussian and independent across the UE. This corresponds to assuming standard point-to-point channel codes at the UE (see, e.g., [9, Ch. 3]). The channel matrices are assumed to be fixed and to remain constant during a coding block, which is of size n samples. Note that in (1) and in the following, we do not denote explicitly the dependence of the signals on the sample index to simplify the notation.

In the single-hop topology under study, each RU i is connected to the CU via a fronthaul link of capacity C_i bits/s/Hz. The fronthaul capacity is normalized to the bandwidth of the uplink channel. This implies that for any uplink coding block of n symbols, nC_i bits can be transmitted on the ith fronthaul link.

REMARK 1

Model (1), which is typically used in related literature, assumes implicitly that the RUs perform time and frequency synchronization locally. In fact, signal (1) is free of frequency drift and is sampled at the baud rate. It is noted that, if time synchronization is not carried out at the RUs, then the RUs need to oversample the baseband signals prior to transferring them on the fronthaul links. This is, for instance, prescribed in the CPRI standard [3].

REMARK 2

Following Remark 1, while model (1) assumes that time and frequency synchronization is done locally, the optimal allocation of layer 1 functionalities, such as synchronization and channel estimation, between the RUs and the CU is a subject of ongoing investigations (see, e.g., [12] for a related discussion).

POINT-TO-POINT FRONTHAUL COMPRESSION

In baseline C-RAN systems, each ith RU uses conventional point-to-point compression strategies to process the n samples of the received IQ signal $\mathbf{y}_i^{\text{ul}}$, as illustrated in Figure 2.

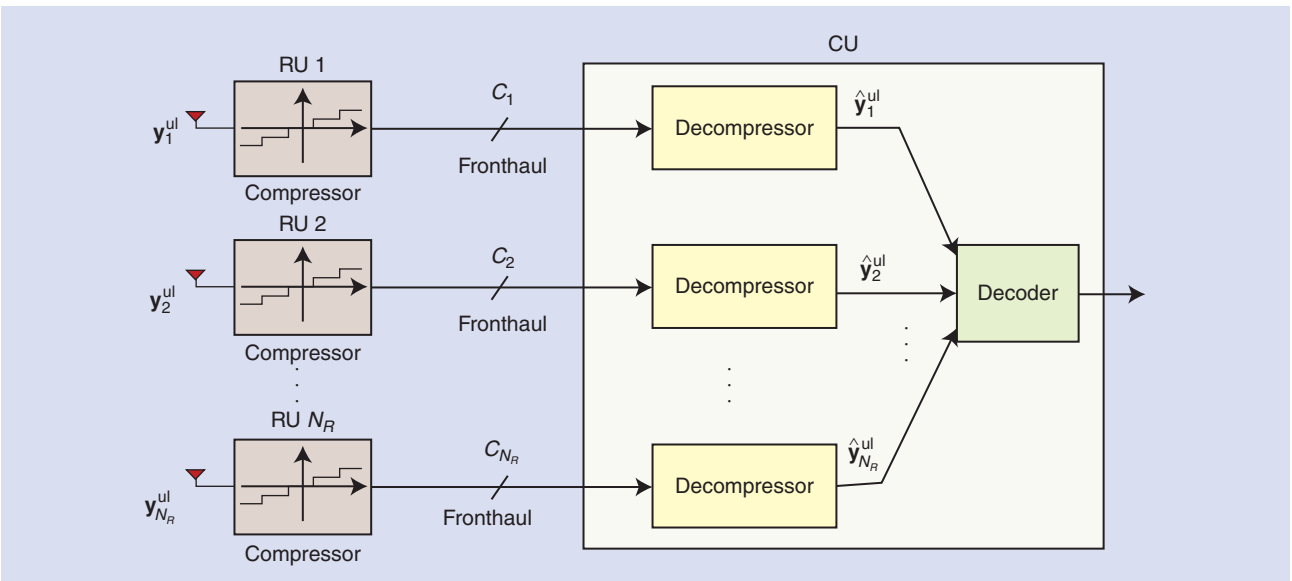
POINT-TO-POINT COMPRESSION

As a result of compression, each ith RU produces a binary string of (at most) nC_i bits, which allows the corresponding decompressor at the CU to identify the quantized signal within the quantization codebook. The quantized signal consists of n samples $\hat{\mathbf{y}}_i^{\text{ul}}$, and is selected by the ith RU from a quantization codebook of 2^{nC_i} codewords (see, e.g., [13]). The example of a scalar quantizer ($n = 1$) at each RU is illustrated in Figure 2 for either the I or the Q component of the IQ sample.

KEY INFORMATION-THEORETIC RESULTS

A standard way of modeling the relationship between the received baseband signal $\mathbf{y}_i^{\text{ul}}$ and its compressed version $\hat{\mathbf{y}}_i^{\text{ul}}$ at RU i is to follow information-theoretic considerations (see, e.g., [9, Ch. 3]) and adopt the Gaussian "test channel"

$$\hat{\mathbf{y}}_i^{\text{ul}} = \mathbf{y}_i^{\text{ul}} + \mathbf{q}_i^{\text{ul}}, \tag{2}$$



[FIG2] Point-to-point fronthaul compression for the uplink of C-RANs.

where the quantization noise $\mathbf{q}_i^{\text{ul}}$ is independent of the signal $\mathbf{y}_i^{\text{ul}}$ and distributed as $\mathbf{q}_i^{\text{ul}} \sim \mathcal{CN}(\mathbf{0}, \Omega_i^{\text{ul}})$. The quantization noise statistics are thus defined by the covariance matrix Ω_i^{ul}. Connecting the information-theoretic viewpoint with classical vector quantization, the covariance matrix Ω_i^{ul} can be thought of defining the shape of the quantization regions of the compressor.

Information theory provides analytical conditions that relate the quantization noise statistics Ω_i^{ul} to the size of the quantization codebooks, and hence to the fronthaul capacity C_i, needed to satisfy the condition (2). More precisely, under these conditions (and for n large enough), the theory guarantees that a quantization codebook exists that contains a codeword of n samples $\hat{\mathbf{y}}_i^{\text{ul}}$ for any input sequence of n samples $\mathbf{y}_i^{\text{ul}}$, such that the joint empirical statistic of the two sequences is "close" to the joint distribution implied by (2) [9, Ch. 3].

Specifically, a standard result in information theory states that, if the fronthaul capacity C_i satisfies the condition

$$I(\mathbf{y}_i^{\text{ul}}; \hat{\mathbf{y}}_i^{\text{ul}}) \leq C_i, \tag{3}$$

where the mutual information is calculated using (2), then it is possible to design a compression strategy that realizes the given quantization error covariance matrix Ω_i^{ul} in the sense discussed above (see, e.g., [9, Ch. 3]). At an intuitive level, condition (3) says that a "smaller" covariance matrix Ω_i^{ul}, and hence a larger mutual information $I(\mathbf{y}_i^{\text{ul}}; \hat{\mathbf{y}}_i^{\text{ul}})$, calls for a larger required fronthaul capacity C_i.

SYSTEM DESIGN

Assuming that the condition (3) is satisfied for all RUs, the quantized IQ signals $\hat{\mathbf{y}}_1^{\text{ul}}, \ldots, \hat{\mathbf{y}}_{N_R}^{\text{ul}}$ are successfully recovered at the CU. The CU then performs joint decoding of the messages sent by all UE, which are encoded in the signals $\mathbf{x}^{\text{ul}}$. As a result, the uplink sum-rate

$$R_{\text{sum}}^{\text{ul}} = I(\mathbf{x}^{\text{ul}}; \hat{\mathbf{y}}_1^{\text{ul}}, \ldots, \hat{\mathbf{y}}_{N_R}^{\text{ul}}), \tag{4}$$

where the mutual information can be calculated from (1) and (2), is achievable (see, e.g., [9, Ch. 4]). Note that individual rates could also be similarly calculated using standard results on the capacity region of multiple access channels, and so could rates achievable with suboptimal decoding strategies such as treating interference as noise (see [9, Ch. 4]).

The sum-rate (4) depends on the compression strategies used by the RUs through the covariance matrices Ω_i^{ul}, $i = 1, \ldots, N_R$. The sum-rate can then be maximized with respect to these matrices to identify the optimal compression strategies to be used at the RUs. The nonconvex problem of maximizing the sum-rate under the fronthaul constraints (3), for $i = 1, \ldots, N_R$, over the matrices Ω_i^{ul}, $i = 1, \ldots, N_R$, falls in the category of difference-of-convex problems and can be tackled by using the so-called majorization minimization (MM) algorithm [14].

REMARK 3

To compress its received signal $\mathbf{y}_i^{\text{ul}}$, each RU i must only be informed about the quantization codebook to be used. Furthermore, the achievability of the sum-rate (4) hinges on the assumption that the CU is aware of the channel matrices of all the active UE. Each ith RU may estimate the channel matrix $\mathbf{H}_i^{\text{ul}}$ based on standard uplink training and then forward the estimated matrix to the CU on the fronthaul links. The CU can then optimize the compression strategies as discussed above and inform accordingly the RUs. We refer to [15] and [16] for an analysis of the overhead associated with the transfer of channel state information on the fronthaul links for ergodic fading channels.

DISTRIBUTED FRONTHAUL COMPRESSION

As seen in Figure 2, with standard point-to-point compression, compression and decompression across different RUs take place in parallel. This separate processing across the RUs neglects the key fact that the baseband signals $\mathbf{y}_i^{\text{ul}}$ in (1) are correlated across the RU index i, since they are noisy observations of the same transmitted signals $\mathbf{x}^{\text{ul}}$. Based on this fact, the joint processing of the signals received on the fronthaul links at the CU via distributed compression is expected to be advantageous, as first proposed in [17].

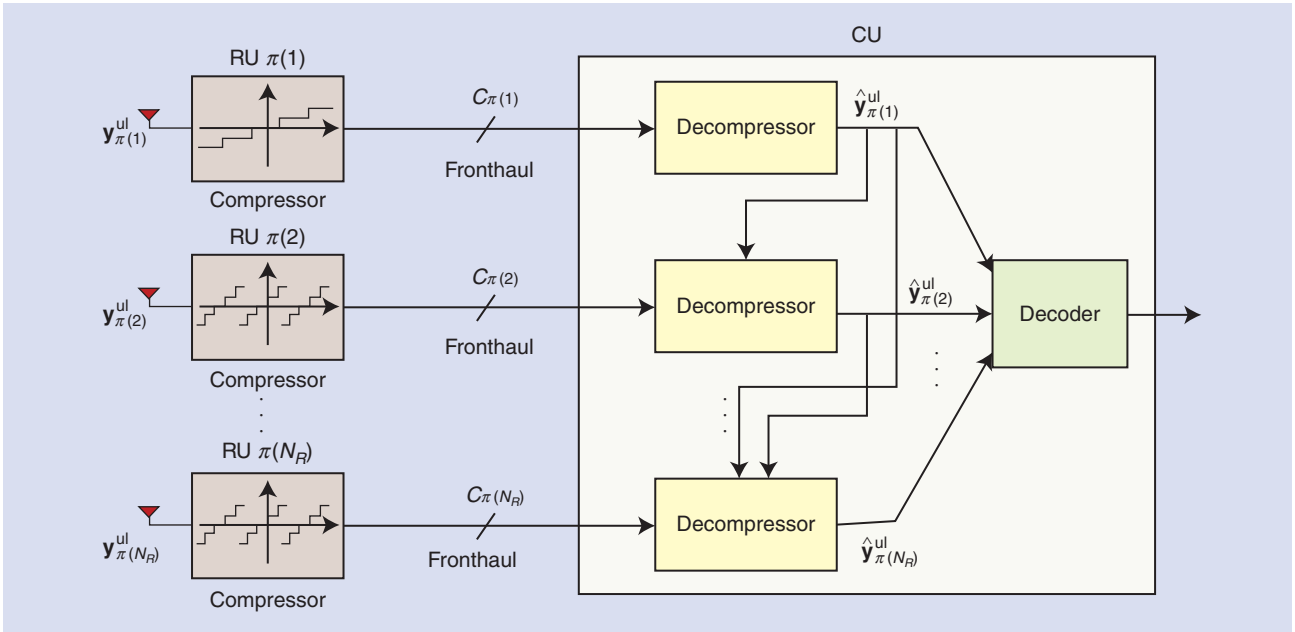
DISTRIBUTED COMPRESSION

To explain distributed compression, here we concentrate on the practical implementation that uses sequential decompression (see [9, Ch. 10] and also [18] and [19]). To this end, we fix an ordering π on the RU indices $\{1, \ldots, N_R\}$. As shown in Figure 3, the CU first decompresses the signal $\hat{\mathbf{y}}_{\pi(1)}^{\text{ul}}$, then $\hat{\mathbf{y}}_{\pi(2)}^{\text{ul}}$, and so on until $\hat{\mathbf{y}}_{\pi(N_R)}^{\text{ul}}$. Therefore, when decompressing $\hat{\mathbf{y}}_{\pi(i)}^{\text{ul}}$, the CU has already retrieved the signals $\{\hat{\mathbf{y}}_{\pi(1)}^{\text{ul}}, \ldots \hat{\mathbf{y}}_{\pi(i-1)}^{\text{ul}}\}$, which are correlated with the signal of interest $\hat{\mathbf{y}}_{\pi(i)}^{\text{ul}}$.

Wyner–Ziv compression offers the information-theoretically optimal approach to leverage side information available at the decompressor to improve the quality of the description $\hat{\mathbf{y}}_{\pi(i)}^{\text{ul}}$. Specifically, Wyner–Ziv compression enables the compressor to use a finer quantizer and hence to obtain a better description $\hat{\mathbf{y}}_{\pi(i)}^{\text{ul}}$, as compared to conventional point-to-point compression, for the same fronthaul capacity $C_{\pi(i)}$.

The approach works as follows. Since a finer quantizer has more codewords than the $2^{nC_{\pi(i)}}$ binary strings that can be supported on the fronthaul link, Wyner–Ziv compression associates the same binary string of $nC_{\pi(i)}$ bits to a subset of codewords. This is in contrast to point-to-point compression in which a distinct binary string is associated with each codeword in the quantization codebook. This subset is known as *bin*, and the *binning* step can be, in practice, realized by using a coset of linear codes or hashing (see, e.g., [10]). Therefore, the complexity of compression is not significantly increased as compared to the point-to-point approach. An example is shown in Figure 3, where RUs $\pi(i)$ with $i > 1$ use a scalar quantizer ($n = 1$) that assigns the same quantization level to multiple regions of the real line (for the I and Q components).

When using Wyner–Ziv compression, the decompressor is thus faced with the problem of having to distinguish among all codewords $\hat{\mathbf{y}}_{\pi(i)}^{\text{ul}}$ that belong to the bin indexed by the binary string received on the fronthaul link. As long as the bins are not



[FIG3] Multiterminal fronthaul compression for the uplink of C-RANs.

too large, this can be done by leveraging the available correlated side information $\{\hat{\mathbf{y}}^{\mathrm{ul}}_{\pi(1)}, \ldots, \hat{\mathbf{y}}^{\mathrm{ul}}_{\pi(i-1)}\}$. In fact, because of their statistical dependence, the real codeword $\hat{\mathbf{y}}^{\mathrm{ul}}_{\pi(i)}$ is expected to be "closer" to the side information sequences. This detection step can be in practice performed by using channel decoding algorithms such as message passing or trellis search (see, e.g., [10]).

KEY INFORMATION-THEORETIC RESULTS

A classical information-theoretic result states that, using Wyner–Ziv compression, a given quantization error matrix $\mathbf{\Omega}^{\mathrm{ul}}_{\pi(i)}$ in (2) is attainable if the fronthaul capacity $C_{\pi(i)}$ satisfies the inequality

$$I(\mathbf{y}^{\mathrm{ul}}_{\pi(i)}; \hat{\mathbf{y}}^{\mathrm{ul}}_{\pi(i)} \,|\, \hat{\mathbf{y}}^{\mathrm{ul}}_{\pi(1)}, \ldots, \hat{\mathbf{y}}^{\mathrm{ul}}_{\pi(i-1)}) \leq C_{\pi(i)}. \tag{5}$$

It is observed that, by standard properties of the mutual information [9, Ch. 2], the constraint (5) imposed on the quantization covariances $\mathbf{\Omega}^{\mathrm{ul}}_{\pi(i)}$ is weaker than the constraint (3) corresponding to point-to-point compression. Specifically, the gap between the two mutual information quantities on the left-hand sides of (3) and (5) increases as the correlation between the useful signal $\hat{\mathbf{y}}^{\mathrm{ul}}_{\pi(i)}$ and the side information $\{\hat{\mathbf{y}}^{\mathrm{ul}}_{\pi(1)}, \ldots, \hat{\mathbf{y}}^{\mathrm{ul}}_{\pi(i-1)}\}$ grows and vanishes if signal and side information are independent.

SYSTEM DESIGN

We are now interested in maximizing the achievable sum-rate (4) with respect to the quantization noise covariances $\mathbf{\Omega}^{\mathrm{ul}}_i$, for $i = 1, \ldots, N_R$ under the fronthaul constraints (5) imposed by distributed compression for a fixed decompression order π. This order can be also optimized upon, as further discussed in Remark 4.

The optimization problem at hand is generally challenging. In [18, Sec. III], a (suboptimal) block-coordinate optimization approach was proposed that leverages a key result in [20]. Accordingly, one optimizes the covariance matrices following the same order π that is employed for decompression. In particular, at the ith step, for fixed (already optimized upon) covariances $\mathbf{\Omega}^{\mathrm{ul}}_{\pi(1)}, \ldots, \mathbf{\Omega}^{\mathrm{ul}}_{\pi(i-1)}$, the covariance $\mathbf{\Omega}^{\mathrm{ul}}_{\pi(i)}$ is obtained by solving the following problem:

$$\begin{aligned} &\underset{\mathbf{\Omega}^{\mathrm{ul}}_{\pi(i)} \succeq 0}{\text{maximize}}\; I(\mathbf{x}^{\mathrm{ul}}; \hat{\mathbf{y}}^{\mathrm{ul}}_{\pi(i)} \,|\, \hat{\mathbf{y}}^{\mathrm{ul}}_{\pi(1)}, \ldots, \hat{\mathbf{y}}^{\mathrm{ul}}_{\pi(i-1)}) \\ &\text{s.t.}\; I(\mathbf{y}^{\mathrm{ul}}_{\pi(i)}; \hat{\mathbf{y}}^{\mathrm{ul}}_{\pi(i)} \,|\, \hat{\mathbf{y}}^{\mathrm{ul}}_{\pi(1)}, \ldots, \hat{\mathbf{y}}^{\mathrm{ul}}_{\pi(i-1)}) \leq C_{\pi(i)}. \end{aligned} \tag{6}$$

In (6), by the chain rule of mutual information (see [9, Ch. 2]), the objective function measures the sum-rate increase obtained by transmitting the signal $\hat{\mathbf{y}}^{\mathrm{ul}}_{\pi(i)}$ to the CU that already has the knowledge of the signals $\{\hat{\mathbf{y}}^{\mathrm{ul}}_{\pi(1)}, \ldots, \hat{\mathbf{y}}^{\mathrm{ul}}_{\pi(i-1)}\}$.

It was shown in [20] that an optimal covariance $\mathbf{\Omega}^{\mathrm{ul}}_{\pi(i)}$ of this problem is given as

$$\mathbf{\Omega}^{\mathrm{ul}}_{\pi(i)} = \mathbf{U}_{\pi(i)} \mathbf{D}_{\pi(i)} \mathbf{U}^{\dagger}_{\pi(i)}, \tag{7}$$

where $\mathbf{U}_{\pi(i)}$ is a unitary matrix whose columns are the orthonormal eigenvectors of the covariance matrix of the signal $\mathbf{y}^{\mathrm{ul}}_{\pi(i)}$ conditioned on the signals $\{\hat{\mathbf{y}}^{\mathrm{ul}}_{\pi(1)}, \ldots, \hat{\mathbf{y}}^{\mathrm{ul}}_{\pi(i-1)}\}$, and $\mathbf{D}_{\pi(i)}$ is a diagonal matrix whose diagonal elements are obtained following a procedure similar to conventional reverse waterfilling [20, Th. 1].

The compression strategy described by the test channel (2) with the derived covariance matrix $\mathbf{\Omega}^{\mathrm{ul}}_{\pi(i)}$ in (7) can be implemented at the RU $\pi(i)$ using classical transform coding [13] as discussed in [20, Sec. III-A]. Accordingly, the RU $\pi(i)$ first applies the linear preprocessing matrix $\mathbf{U}^{\dagger}_{\pi(i)} \mathbf{D}^{-1/2}_{\pi(i)}$ to the received signal vector $\mathbf{y}^{\mathrm{ul}}_{\pi(i)}$ and then independently compresses the resulting signal streams using a Gaussian test channel with

noise of unit variance. It can be proved that multiplication by the unitary transform $\mathbf{U}_{\pi(i)}^{\dagger}$, also referred to as conditional Karhunen–Loeve transform (KLT) [21], decorrelates the received signal streams when conditioned on the side information signals $\{\hat{y}_{\pi(1)}^{\mathrm{ul}}, \ldots, \hat{y}_{\pi(i-1)}^{\mathrm{ul}}\}$.

REMARK 4

The decompression order π generally affects the achievable performance and should be optimized upon. A choice that is generally sensible, and close to optimal, is that of decompressing first the signals coming from macro-BSs and then those from pico- or femto-BSs in their vicinity. The rationale for this approach is that macro-BSs tend to have a larger fronthaul capacity and hence their decompressed signals provide relevant side information for the signals coming from smaller cells, which are typically connected with lower capacity fronthaul links.

REMARK 5

In the previous discussion, it was assumed that the CU first decompresses the signals and then decodes the messages of the UE based on the decompressed signals. The performance may be improved by performing joint decompression and decoding at the cost of an increased computational complexity [17].

COMPUTE-AND-FORWARD

In the schemes discussed so far, the quantization codebooks used by the RUs are designed separately from the channel codebooks used by the UE for transmission in the uplink. A conceptually different approach was instead proposed in [11] based on the principle of compute-and-forward. Accordingly, the same codebook is used both for channel encoding at all the UE and for quantization at the RUs.

The approach proposed in [11] selects a (nested) lattice code. Lattice codes have the property that the weighted-sum—more precisely the modulo-sum with respect to the coarse lattice—of two codewords is also a codeword, as long as the weights are integer numbers. In the scheme of [11], each RU then decodes an appropriate (modulo-)sum, with integer weights, of the codewords transmitted by the UE. The bit stream sent on the fronthaul link identifies the decoded codeword within the lattice code. The idea is that, upon receiving a sufficient number of linear combinations of codewords from the RUs, the CU can invert the resulting linear system and recover the transmitted codewords.

The key potential advantage of the compute-and-forward strategy is that no quantization noise is introduced by the CU due to the fact that the channel and quantization codebooks are matched. On the flip side, the baseband signal (1) received at each RU is a sum with noninteger weights of the codewords transmitted by the UE. Therefore, the difference between the decoded integer combination of codewords and the actual noninteger combination of codewords resulting from the channel affects decoding at each RU as an additional noise term.

NUMERICAL EXAMPLE

We now discuss the performance of point-to-point compression and of more advanced strategies in the context of a specific example. We focus on a standard three-cell circulant Wyner model (see, e.g., [6]), where each cell contains a single-antenna UE and a single-antenna RU, and intercell interference takes place only between adjacent cells (the first and third cell are considered to be adjacent). This implies that the received signal (1) is given as $y_i^{\mathrm{ul}} = x_i^{\mathrm{ul}} + g x_{[i-1]_3}^{\mathrm{ul}} + g x_{[i+1]_3}^{\mathrm{ul}} + z_i^{\mathrm{ul}}$, where x_j^{ul} is the signal sent by the UE in cell j and $[\cdot]_3$ represents the modulo-3 operation. The intercell channel gain is equal to $g = 0.4$. Moreover, every RU has the same fronthaul capacity of 3 bits/s/Hz.

Figure 4 plots the achievable per-cell sum-rate for point-to-point compression, distributed compression, and compute-and-forward versus the transmitted UE power P, which can be taken as a measure of signal-to-noise ratio (SNR). For reference, we also show the per-cell sum-rate achievable with single-cell processing, whereby each RU decodes the signal of the in-cell UE by treating all other UE signals as noise, and the cut-set upper bound [6]. It can be seen that the performance advantage of distributed compression over point-to-point compression increases as the SNR grows larger. This is because the correlation of the received signals in (1) at the RUs becomes more pronounced as the SNR increases. As for compute-and-forward, its performance at low SNR coincides with single-cell processing, as the RUs tend to decode trivial combinations consisting only of the signals of the local UE. On the other hand, compute-and-forward outperforms all the other schemes as the SNR increases, i.e., in the regime where the fronthaul capacity is the main performance bottleneck. Further discussion can be found in the "Downlink" section.

MULTIHOP FRONTHAUL TOPOLOGY

In this subsection, we study the case in which the fronthaul network has a general multihop topology. As an example, in Figure 1, RU 6 communicates to the CU via a two-hop fronthaul connection that passes through RU 2 and RU 4. Note that each RU may have multiple incoming and outgoing fronthaul links.

ROUTING

To convey the quantized IQ samples from the RUs to the CU through multiple hops, each RU must decide on the information to be transmitted on each outgoing fronthaul link based on the information received on the incoming fronthaul links. A first option is to use routing: the bits received on the incoming links are simply forwarded on the outgoing links without any additional processing. This approach requires the optimization of standard flow variables that define the allocation of fronthaul capacity to the different bit streams. The problem is formulated and addressed via the MM algorithm in [22].

IN-NETWORK PROCESSING

Routing may be highly inefficient in the presence of a dense deployment of RUs. In fact, in this case, an RU may be connected to a large number of nearby RUs, all of which receive

correlated baseband signals. In this case, it is wasteful of the fronthaul capacity to merely forward all the bit streams received from the connected RUs. Instead, it is possible to combine the correlated baseband signals at the RU to reduce redundancy. We refer to this processing of incoming signals as *in-network processing*.

To allow for in-network processing, the RU at hand must first decompress the received bit streams from the connected RUs to recover the baseband signals. The decompressed baseband signals are then linearly processed, along with the IQ signal received locally by the RU. After in-network processing, the obtained signals must be recompressed before they can be sent on the outgoing fronthaul links. The effect of the resulting quantization noise must thus be counterbalanced by the advantages of in-network processing to make the strategy preferable to routing. The optimal design of in-network processing is addressed in [22] using the MM algorithm.

NUMERICAL EXAMPLE

We now compare the sum-rates achievable with routing and with in-network processing for the uplink of a C-RAN with a two-hop fronthaul network. Specifically, there are N RUs in the first layer and two RUs in the second layer, all receiving in the uplink. The RUs in the first layer do not have direct fronthaul links to the CU, while the RUs in the second layer do. Half of the RUs in the first layer is connected to one RU in the second layer, and half to the other RU in the second layer. We assume that all fronthaul links have capacity equal to 2–4 bits/s/Hz and all channel matrices have identically and independently distributed (i.i.d.) complex Gaussian entries with unit power (Rayleigh fading). Figure 5 shows the average sum-rate versus the number N of RUs in layer 1 with $N_U = 4$ UE and average received per-antenna SNR of 20 dB at all RUs. It is observed that the performance gain of in-network processing over routing becomes more pronounced as the number N of RUs in the first layer increases. This suggests that, as the density of the RUs' deployment increases, it is desirable for each RU in layer 2 to perform in-network processing of the signals received from layer 1.

DOWNLINK

SYSTEM MODEL

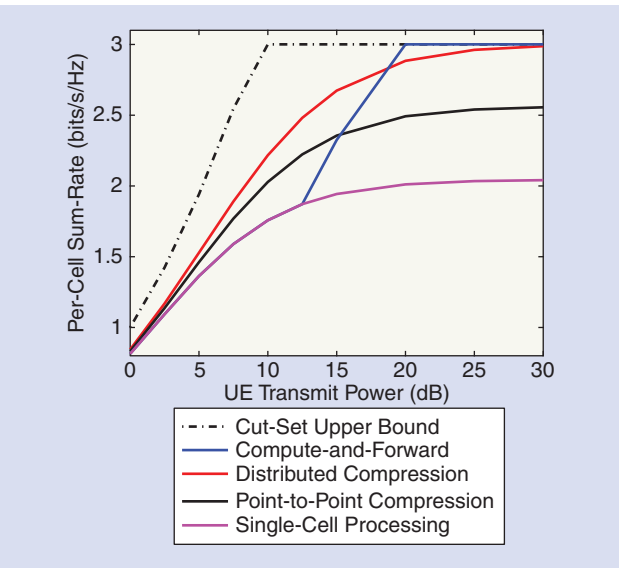
In the downlink, the CU that manages a given cluster processes the information messages of the UE within the cluster by performing channel coding and precoding on behalf of the RUs. As seen in Figure 6, the precoded baseband signals are then compressed by the CU, which finally forwards the compressed IQ signals to the RUs on the fronthaul links. Each RU decompresses the signal received on the fronthaul link (by looking up the corresponding quantization codebook), performs pulse shaping, upconverts the resulting signal, and transmits it to the UE on the wireless downlink channel. Note that we concentrate here on a single-hop fronthaul topology. The multihop case can be addressed following the analysis for the uplink, but this is not further detailed here and is left as an interesting future work.

Similar to the uplink, assuming flat-fading channels, each UE k in the cluster under study receives a discrete-time baseband signal given as

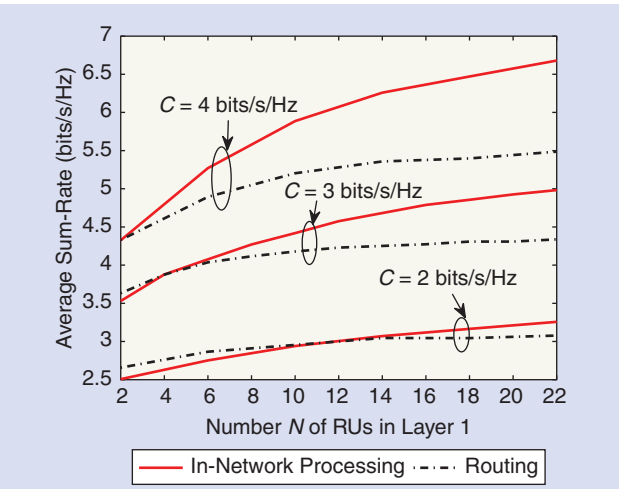
$$\mathbf{y}_k^{\mathrm{dl}} = \mathbf{H}_k^{\mathrm{dl}}\mathbf{x}^{\mathrm{dl}} + \mathbf{z}_k^{\mathrm{dl}}, \tag{8}$$

where $\mathbf{x}^{\mathrm{dl}}$ is the aggregate baseband signal vector transmitted by all the RUs in the cluster; the additive noise $\mathbf{z}_k^{\mathrm{dl}} \sim \mathcal{CN}(\mathbf{0}, \Omega_{\mathbf{z}_k^{\mathrm{dl}}})$ accounts for thermal noise and interference from the other clusters; and the matrix $\mathbf{H}_k^{\mathrm{dl}}$ denotes the channel response matrix from all the RUs in the cluster toward UE k.

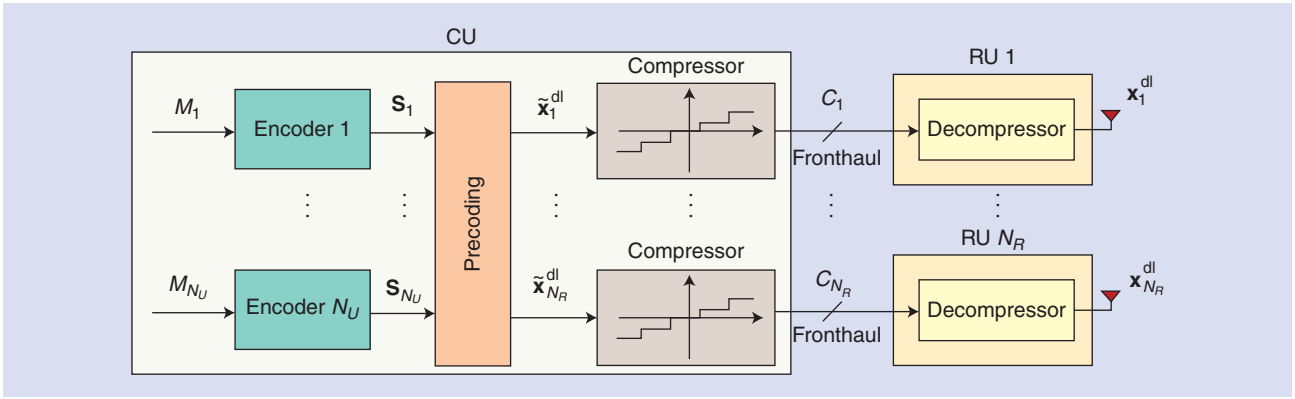
As mentioned, the transmitted signals $\mathbf{x}^{\mathrm{dl}}$ are quantized versions of the baseband signals produced by the CU that manages



[FIG4] Per-cell uplink sum-rate versus the transmitted UE power P for the circulant Wyner model with $C = 3$ bits/s/Hz dB and intercell channel gain equal to $g = 0.4$.



[FIG5] Average uplink sum-rate versus the number of RUs in layer 1 with $N_U = 4$ UEs, average received per-antenna SNR of 20 dB and fronthaul capacity of 2–4 bits/s/Hz.



[FIG6] Point-to-point fronthaul compression for the downlink of C-RANs.

the cluster. As shown in Figure 6, to obtain $\mathbf{x}^{\mathrm{dl}}$, the CU first performs channel encoding separately for different UE. This produces the IQ samples $\mathbf{s} = [\mathbf{s}_1; \ldots; \mathbf{s}_{N_U}]$, with $\mathbf{s}_k$ representing the signal intended for UE k. The CU then performs linear precoding of the channel-encoded baseband signals $\mathbf{s}$. We observe that non-linear precoding via "dirty-paper" coding can also be considered with minor modifications. The precoded IQ signals $\tilde{\mathbf{x}}^{\mathrm{dl}}$ produced by the CU can be written as

$$\tilde{\mathbf{x}}^{\mathrm{dl}} = [\tilde{\mathbf{x}}_1^{\mathrm{dl}}; \ldots; \tilde{\mathbf{x}}_{N_R}^{\mathrm{dl}}] = \mathbf{A}\mathbf{s}, \tag{9}$$

where $\tilde{\mathbf{x}}_i^{\mathrm{dl}}$ is the precoded signal intended for transmission by RU i and $\mathbf{A} = [\mathbf{A}_1, \ldots, \mathbf{A}_{N_U}]$ is the precoding matrix with the submatrix $\mathbf{A}_k$ multiplied to the signal $\mathbf{s}_k$. The compression of the signals $\tilde{\mathbf{x}}_i^{\mathrm{dl}}$, with $i = 1, \ldots, N_R$ to produce $\mathbf{x}^{\mathrm{dl}}$ is discussed next.

POINT-TO-POINT FRONTHAUL COMPRESSION

Similar to the uplink, in the conventional C-RAN implementation, the CU compresses separately the precoded IQ signals $\tilde{\mathbf{x}}_i^{\mathrm{dl}}$ intended for transmission by different RUs i using point-to-point compression, as shown in Figure 6. The index describing the compressed signal $\mathbf{x}_i^{\mathrm{dl}}$ is sent to the ith RU via the corresponding fronthaul link of capacity C_i. Using compression with a Gaussian test channel, the compressed signal $\mathbf{x}_i^{\mathrm{dl}}$ is given as

$$\mathbf{x}_i^{\mathrm{dl}} = \tilde{\mathbf{x}}_i^{\mathrm{dl}} + \mathbf{q}_i^{\mathrm{dl}}, \tag{10}$$

where the compression noise $\mathbf{q}_i^{\mathrm{dl}}$ is distributed as $\mathbf{q}_i^{\mathrm{dl}} \sim \mathcal{CN}(\mathbf{0}, \Omega_i^{\mathrm{dl}})$. The quantization noises are independent across the RU index i due to the separate compression of the RUs' IQ signals.

Using the information theoretic results reviewed in the previous section, the quantization error matrix Ω_i^{dl} can be realized if the fronthaul link capacity C_i satisfies the inequality

$$I(\tilde{\mathbf{x}}_i^{\mathrm{dl}}; \mathbf{x}_i^{\mathrm{dl}}) \leq C_i. \tag{11}$$

Moreover, assuming that each kth UE treats the signals intended for other UE as noise, the information rate

$$R_k^{\mathrm{dl}} = I(\mathbf{s}_k; \mathbf{y}_k^{\mathrm{dl}}) \tag{12}$$

can be achieved for UE k. The optimization of the weighted-sum-rate $R_{\mathrm{sum}}^{\mathrm{dl}} = \sum_{k=1}^{N_U} w_k R_k^{\mathrm{dl}}$ subject to per-RU power constraints and to the constraints (11), for $i = 1, \ldots, N_R$, with respect to the variables $\mathbf{A}$ and Ω_i^{dl} for $i = 1, \ldots, N_R$ was tackled in [23, Sec. V-C] by using the MM algorithm.

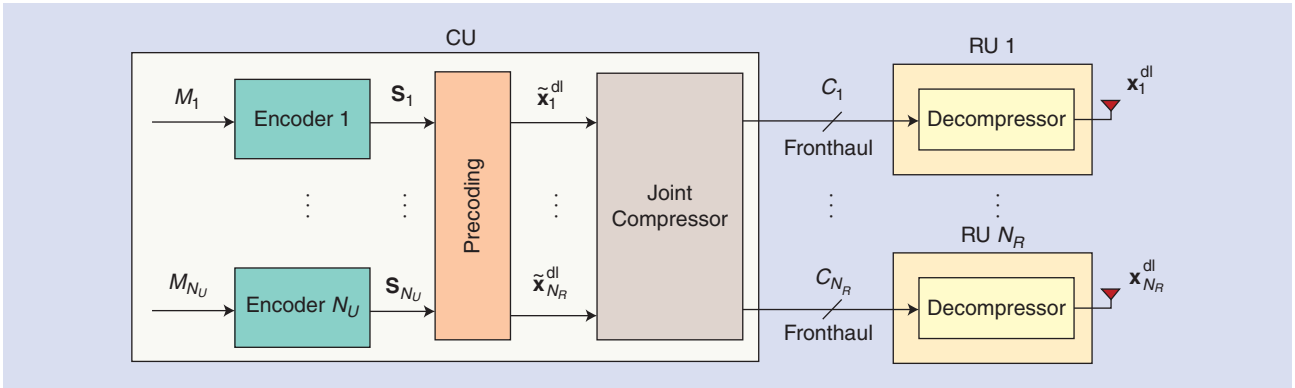
MULTIVARIATE FRONTHAUL COMPRESSION

We now investigate possible improvements to point-to-point compression based on multiterminal compression principles. Our starting observation is that point-to-point compression yields quantization errors that are independent across the RUs. In contrast, multivariate compression [9, Ch. 7] allows correlated quantization noises to be produced, at the expense of a joint, rather than separate, compression of the baseband signals $\tilde{\mathbf{x}}_i^{\mathrm{dl}}$ for $i = 1, \ldots, N_R$ at the CU.

MULTIVARIATE COMPRESSION

The block diagram of the CU and RUs in a cluster operating with multivariate fronthaul compression is shown in Figure 7. As for the conventional point-to-point case of Figure 6, the CU performs channel encoding separately for each UE and applies precoding, hence obtaining the baseband signals $\tilde{\mathbf{x}}_i^{\mathrm{dl}}$ in (9) for $i = 1, \ldots, N_R$. However, unlike point-to-point compression, the signals $\tilde{\mathbf{x}}_i^{\mathrm{dl}}$ are jointly compressed to select the quantized signals $\mathbf{x}_i^{\mathrm{dl}}$ from the corresponding quantization codebooks $i = 1, \ldots, N_R$.

Before providing some details on multivariate compression, we observe that correlating the quantization noises can be beneficial to control the effect of the additive quantization noises on the reception of the UE. To see this, assume that the quantization noise vector $\mathbf{q}^{\mathrm{dl}} = [\mathbf{q}_1^{\mathrm{dl}}; \ldots; \mathbf{q}_{N_R}^{\mathrm{dl}}]$ in (10) are distributed as $\mathcal{CN}(\mathbf{0}, \Omega^{\mathrm{dl}})$, where the covariance matrix Ω^{dl} is a block matrix whose (i, j)th block $\Omega_{i,j}^{\mathrm{dl}} = \mathbb{E}[\mathbf{q}_i^{\mathrm{dl}} \mathbf{q}_j^{\mathrm{dl}\dagger}]$ defines the correlation between the quantization noises of RU i and RU j. By using (8) and (10), the effective noise at the kth UE is given by $\mathbf{z}_k^{\mathrm{dl}} + \mathbf{H}_k^{\mathrm{dl}} \mathbf{q}^{\mathrm{dl}}$. The covariance matrix of the effective noise is then given as $\Omega_{z_k^{\mathrm{dl}}} + \mathbf{H}_k^{\mathrm{dl}} \Omega^{\mathrm{dl}} \mathbf{H}_k^{\mathrm{dl}}$, and can hence be controlled by designing the quantization error covariance matrix Ω^{dl}. As a result, one can reduce the impact of the effective noise on the reception of the useful signal and enhance the achievable rates (12).



[FIG7] Multiterminal fronthaul compression for the downlink of C-RANs.

With reference to vector quantization concepts, one can think of the matrix Ω^{dl} as defining the shape of the quantization regions in the space of the baseband signals of all RUs. Specifically, while point-to-point compression leads to regions that are merely the Cartesian product of the quantization regions of the separate quantizers, multivariate compression allows for more general shapes.

KEY INFORMATION-THEORETIC RESULTS

The multivariate compression lemma in [9, Ch. 9] provides sufficient conditions on the fronthaul capacities under which a given joint quantization error matrix Ω^{dl} can be realized. It is recalled that, if the error matrix Ω^{dl} is block diagonal, i.e., if the submatrices $\Omega_{i,j}^{dl}$ have all zero entries for $i \neq j$, then the conditions at hand reduce to (11) for $i = 1, \ldots, N_R$. Instead, for a general covariance matrix Ω^{dl}, the multivariate compression lemma requires that the following inequality

$$\sum_{i \in \mathcal{S}} h(\mathbf{x}_i^{dl}) - h(\mathbf{x}_{\mathcal{S}}^{dl} \,|\, \tilde{\mathbf{x}}^{dl}) \leq \sum_{i \in \mathcal{S}} C_i \tag{13}$$

be satisfied for all subsets $\mathcal{S} \subseteq \{1, \ldots, N_R\}$. Using standard properties of the mutual information, it can be seen that if Ω^{dl} is block diagonal, then the system of conditions (13) for all subsets $\mathcal{S}$ is equivalent to the system of inequalities (11) for $i = 1, \ldots, N_R$. Otherwise, the inequalities (13) provide more stringent constraints on the fronthaul capacities than (11). The optimization over the precoding matrix $\mathbf{A}$ and the compression noise covariance Ω^{dl} was tackled by using the MM algorithm in [23].

REMARK 6

Similar to Figure 3 for the uplink, multivariate compression can be implemented using a sequential architecture, whereby the baseband signals of different RUs are sequentially, rather than jointly, compressed [23, Sec. IV-D].

COMPUTE-AND-FORWARD

Similar to the "Uplink" section, we now observe that the schemes discussed so far for the downlink employ quantization codebooks that are designed separately from the channel codebooks used for encoding the messages of the UE. An alternative approach, which is dual to the one studied for the uplink, leverages instead the same (nested) lattice code for both channel coding and quantization.

Specifically, according to the approach introduced in [24], the CU employs the same lattice code to perform channel encoding for all UE. Then, it performs precoding using only integer (modulo-)sum operations. In this fashion, the resulting precoded baseband signals are still codewords of the same lattice code. Finally, the CU transmits on the fronthaul links directly the index of the obtained precoded codewords.

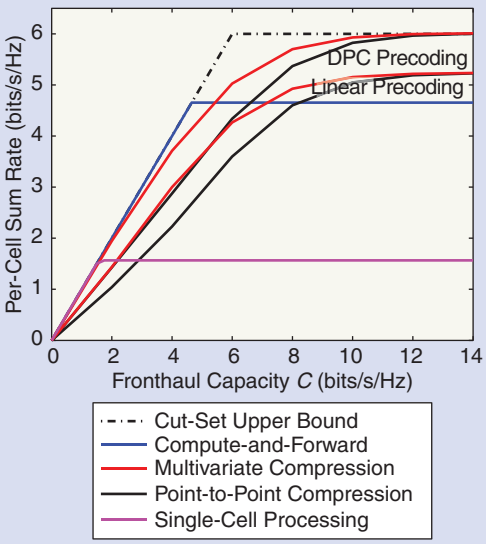
The scheme at hand has similar advantages and disadvantages as compared to its counterpart for the uplink. Specifically, while not adding any quantization noise, it is limited by the integrality constraints on the coefficients of the precoding matrix.

NUMERICAL EXAMPLE

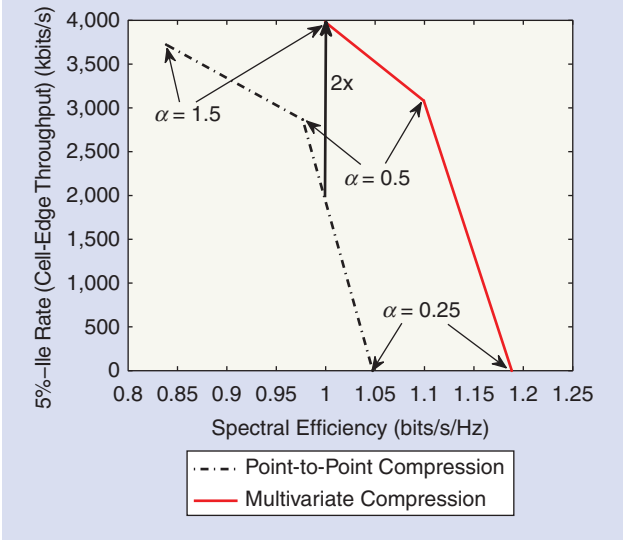
We consider here the same three-cell circulant Wyner model used in Figure 4 for the uplink, where the intercell channel gain is equal to $g = 0.5$, and every RU uses the same transmit power of 20 dB and has the same fronthaul capacity C. Figure 8 shows the per-cell sum-rate of point-to-point compression and multivariate compression, as applied to both linear precoding and "dirty paper" nonlinear precoding [25], and also of compute-and-forward. For reference, we also show the cut-set upper bound and the performance with single-cell processing, whereby each RU transmits only the signal of the in-cell UE. It is observed that multivariate compression significantly outperforms point-to-point compression for both linear precoding and "dirty paper" nonlinear precoding. Moreover, compute-and-forward is the most effective strategy in the regime of moderate fronthaul capacity C in which the limitations imposed by integer precoding are not dominant. In contrast, for sufficiently large fronthaul capacity C, both compression-based schemes attain the upper bound, while compute-and-forward is limited by the mentioned integrality constraints.

PERFORMANCE EVALUATION

This section provides a performance evaluation of the discussed fronthaul compression techniques using the standard cellular topology and channel models of [26]. We focus on the performance of the macrocell located at the center of a



[FIG8] Per-cell downlink sum-rate versus the fronthaul capacity C for the circulant Wyner model with transmitted UE power $P = 20$ dB and intercell channel gain equal to $g = 0.5$.



[FIG9] Cell-edge throughput, i.e., fifth percentile rate, versus the average per-UE spectral efficiency for various fairness constants α in the downlink of a C-RAN with $K = 4$ UEs, $(C_{\text{macro}}, C_{\text{pico}}) = (3, 1)$ bits/s/Hz, $T = 5$ and $\beta = 0.5$.

two-dimensional 19-cell hexagonal cellular layout. In each macrocell, there are K randomly and uniformly located UE; a macro-BS with three sectorized antennas placed in the center; and a single randomly and uniformly located single-antenna pico-BS. A single-hop fronthaul topology is assumed, where each macrocell is a cluster served by a CU that is connected directly to the macro-BS and the pico-BS in the macrocell. Specifically, the fronthaul links to each macro-BS antenna and to each pico-BS have capacities C_{macro} and C_{pico}, respectively. All interference signals from other macrocells are treated as independent noise signals. The system parameters are as indicated in [26]. We focus here on the downlink, but comparable results were observed also for the uplink [27].

We adopt the conventional metric of cell-edge throughput versus the average per-UE spectral efficiency (see, e.g., [8, Fig. 5]). This is obtained by running a proportional fairness scheduler on a sequence of T time-slots with independent fading realizations, and by then evaluating the cell-edge throughput as the fifth percentile rate and the average spectral efficiency as the average sum-rate normalized by the number of UE. We recall that the proportional fairness scheduler maximizes at each time-slot the weighted-sum-rate $R_{\text{sum}}^{\text{fair}} = \sum_{k=1}^{N_U} R_k^{\text{dl}} / \bar{R}_k^{\alpha}$, with $\alpha \geq 0$ being a fairness constant, R_k^{dl} in (12), and $\bar{R}_k$ being the average data rate accrued by UE k so far. After each time-slot, the rate $\bar{R}_k$ is updated as $\bar{R}_k \leftarrow \beta \bar{R}_k + (1-\beta) R_k^{\text{dl}}$ where $\beta \in [0, 1]$ is a forgetting factor. Increasing α leads to a more fair rate allocation among the UE.

Figure 9 plots the cell-edge throughput versus the average spectral efficiency for $K = 4$ UE, $(C_{\text{macro}}, C_{\text{pico}}) = (3, 1)$ bits/s/Hz, $T = 5$ and $\beta = 0.5$. The curve is obtained by varying the fairness constant α in the utility function $R_{\text{sum}}^{\text{fair}}$. It is observed that spectral efficiencies larger than 1.05 bits/s/Hz are not achievable with point-to-point compression, while they can be obtained with multivariate compression. Moreover, it is seen that multivariate compression provides $2\times$ gain in terms of cell-edge throughput for a spectral efficiency of 1 bits/s/Hz.

CONCLUSIONS AND OUTLOOK

The design of C-RANs poses a host of new research challenges to the signal processing community. One key problem is that of devising effective compression algorithms for the fronthaul links connecting the RUs with the CU that resides within the "cloud" of the operator's core network. As reviewed in this article, the performance of conventional point-to-point compression strategies can be substantially improved by leveraging techniques inspired by network information theory. Most notably, we have emphasized the potential gains of multiterminal compression—distributed compression for the uplink and multivariate compression for the downlink—and of structured coding via compute-and-forward. Among the many open issues, we mention here the investigation of structured coding schemes that are robust to nonintegrality limitations (see [11] and [24]); the performance analysis for limited frame lengths; the optimal allocation of layer-1 functionalities between the RUs and the CU [12]; the study of the impact of the fronthaul latency on higher-layer performance metrics; and the analysis of the interplay of the considered techniques with multiuser scheduling and limited-feedback channel state information.

AUTHORS

Seok-Hwan Park (seokhwan81@gmail.com) received the B.Sc. and Ph.D. degrees in electrical engineering from Korea University, Seoul, South Korea in 2005 and 2011, respectively. From 2012 to 2014, he was a postdoctoral research associate with the Center for Wireless Communication and Signal Processing Research, New Jersey Institute of Technology, Newark. Since March 2014, he has

been with Samsung Electronics, Suwon, South Korea, as a senior engineer. His research interests include communication, information, and optimization theories with applications to various multiple-input, multiple-output wireless systems. He received the Best Paper Award at the 2006 Asia-Pacific Conference on Communications and an Excellent Paper Award at the IEEE Student Paper Contest in 2006.

Osvaldo Simeone (osvaldo.simeone@njit.edu) received the M.Sc. degree (with honors) and the Ph.D. degree in information engineering from Politecnico di Milano, Italy, in 2001 and 2005, respectively. He is currently with the Center for Wireless Communications and Signal Processing Research, New Jersey Institute of Technology, Newark, where he is an associate professor. His research interests include wireless communications, information theory, data compression, and machine learning. He is a corecipient of Best Paper Awards of the 2007 IEEE International Workshop on Signal Processing Advances in Wireless Communications and 2007 IEEE Conference on Wireless Rural and Emergency Communications. He is an editor of *IEEE Transactions on Information Theory*.

Onur Sahin (Onur.Sahin@interdigital.com) received the B.S. degree in electrical and electronics engineering from Middle East Technical University, Ankara, Turkey, in 2003 and the M.Sc. and Ph.D. degrees in electrical engineering from the Polytechnic Institute of New York University, Brooklyn, in 2005 and 2009, respectively. Since November 2009, he has been with the Advance Air Interface Group, InterDigital Inc., Melville, New York, as a staff engineer. He conducts research on the cross-layer design of the next-generation cellular and Wi-Fi systems such as ultradense networks, long-term evolution advanced, and beyond. He is the author of more than 20 publications and 12 international patent applications on next-generation wireless communication techniques and system design. His recent research interest includes small cell networks with the utilization of millimeter-wave (60 GHz) link technology to achieve multi-Gigabit/s throughput experience. He received the 2012 InterDigital Innovation Award.

Shlomo Shamai (Shitz) (sshlomo@ee.technion.ac.il) is with the Department of Electrical Engineering, Technion, Israel Institute of Technology, where he is now the William Fondiller Technion Distinguished Professor. He is an IEEE Fellow, a member of the Israeli Academy of Sciences and Humanities, and a foreign member of the U.S. National Academy of Engineering. He received the 2011 Claude E. Shannon Award and the 2014 Rothschild Prize in Mathematics/Computer Sciences and Engineering. He received the URSI 1999 van der Pol Gold Medal, the 2000 IEEE Donald G. Fink Prize Paper Award, the 2003 and 2004 Joint IEEE Information Theory Society and IEEE Communications Society Paper Award, the 2007 IEEE Information Theory Society Paper Award, the 2009 European Commission FP7, the 2014 EURASIP Best Paper Award, and the 2010 Thomson Reuters International Excellence in Scientific Research Award. He served on the executive editorial board of *IEEE Transactions on Information Theory*.

REFERENCES

[1] J. Segel and M. Weldon, "Lightradio portfolio—Technical overview," Technology White Paper 1, Alcatel-Lucent.

[2] China Mobile, "C-RAN: The road towards green RAN," White Paper, ver. 2.5, China Mobile Research Institute, Oct. 2011.

[3] Ericsson AB, Huawei Technologies, NEC Corporation, Alcatel Lucent and Nokia Siemens Networks, "Common public radio interface (CPRI); interface specification," CPRI specification v5.0, Sept. 2011.

[4] Z. Ding and H. V. Poor, "The use of spatially random base stations in cloud radio access networks," *IEEE Signal Processing Lett.*, vol. 20, no. 11, pp. 1138–1141, Nov. 2013.

[5] G. J. Foschini, K. Karakayali, and R. Valenzuela, "Coordinating multiple antenna cellular networks to achieve enormous spectral efficiency," *IEE Proc. Commun.*, vol. 153, no. 4, pp. 548–555, Aug. 2006.

[6] O. Simeone, N. Levy, A. Sanderovich, O. Somekh, B. M. Zaidel, H. V. Poor, and S. Shamai (Shitz), "Cooperative wireless cellular systems: an information-theoretic view," *Found. Trends Commun. Inform. Theory*, vol. 8, no. 1–2, pp. 1–177, 2012.

[7] Integrated Device Technology, Inc., "Front-haul compression for emerging C-RAN and small cell networks," Apr. 2013.

[8] R. Irmer, H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H.-P. Mayer, L. Thiele, and V. Jungnickel, "Coordinated multipoint: Concepts, performance, and field trial results," *IEEE Commun. Mag.*, vol. 49, no. 2, pp. 102–111, Feb. 2011.

[9] A. E. Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[10] Z. Xiong, A. Liveris, and S. Cheng, "Distributed source coding for sensor networks," *IEEE Signal Processing Mag.*, vol. 21, no. 5, pp. 80–94, Sept. 2004.

[11] B. Nazer and M. Gastpar, "Compute-and-forward: Harnessing interference through structured codes," *IEEE Trans. Inform. Theory*, vol. 57, no. 10, pp. 6463–6486, Oct. 2011.

[12] C. J. Bernardos, A. De Domenice, J. Ortin, P. Rost, and D. Wubben, "Challenges of designing jointly the backhaul and radio access network in a cloud-based mobile network," in *Proc. Future Networks Mobile Summit 2013*, Lisbon, Portugal, pp. 1–10.

[13] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Norwell, MA: Kluwer, 1992.

[14] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *Amer. Statist.*, vol. 58, no. 1, pp. 30–37, 2004.

[15] J. Hoydis, M. Kobayashi, and M. Debbah, "Optimal channel training in uplink network MIMO systems," *IEEE Trans. Signal Processing*, vol. 59, no. 6, pp. 2824–2833, June 2011.

[16] J. Kang, O. Simeone, J. Kang, and S. Shamai (Shitz), "Joint signal and channel state information compression for the backhaul of uplink network MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1555–1567, Mar. 2014.

[17] A. Sanderovich, O. Somekh, H. V. Poor, and S. Shamai (Shitz), "Uplink macro diversity of limited backhaul cellular network," *IEEE Trans. Inform. Theory*, vol. 55, no. 8, pp. 3457–3478, Aug. 2009.

[18] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai (Shitz), "Robust and efficient distributed compression for cloud radio access networks," *IEEE Trans. Veh. Technol.*, vol. 62, no. 2, pp. 692–703, Feb. 2013.

[19] L. Zhou and W. Yu, "Uplink multicell processing with limited backhaul via per-base-station successive interference cancellation," *IEEE J. Select. Areas Commun.*, vol. 31, no. 10, pp. 1981–1993, Oct. 2013.

[20] A. del Coso and S. Simoens, "Distributed compression for MIMO coordinated networks with a backhaul constraint," *IEEE Trans. Wireless Commun.*, vol. 8, no. 9, pp. 4698–4709, Sept. 2009.

[21] M. Gastpar, P. L. Dragotti, and M. Vetterli, "The distributed Karhunen-Loeve transform," *IEEE Trans. Inform. Theory*, vol. 52, no. 12, pp. 5177–5196, Dec. 2006.

[22] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai (Shitz), "Multi-hop backhaul compression for the uplink of cloud radio access networks," arXiv:1312.7135. [Online]. Available: http://arxiv.org/abs/1312.7135

[23] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai (Shitz), "Joint precoding and multivariate backhaul compression for the downlink of cloud radio access networks," *IEEE Trans. Signal Processing*, vol. 61, no. 22, pp. 5646–5658, Nov. 2013.

[24] S.-N. Hong and G. Caire, "Compute-and-forward strategies for cooperative distributed antenna systems," *IEEE Trans. Inform. Theory*, vol. 59, no. 9, pp. 5227–5243, Sept. 2013.

[25] O. Simeone, O. Somekh, H. V. Poor, and S. Shamai (Shitz), "Downlink multicell processing with limited backhaul capacity," *EURASIP J. Adv. Signal Process.*, vol. 2009, 2009.

[26] 3GPP, TR 136.931 ver. 9.0.0 Rel. 9, May 2011.

[27] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai (Shitz), "Performance evaluation of multiterminal backhaul compression for cloud radio access networks," in *Proc. CISS*, Princeton, NJ, Mar. 19–21, 2014.

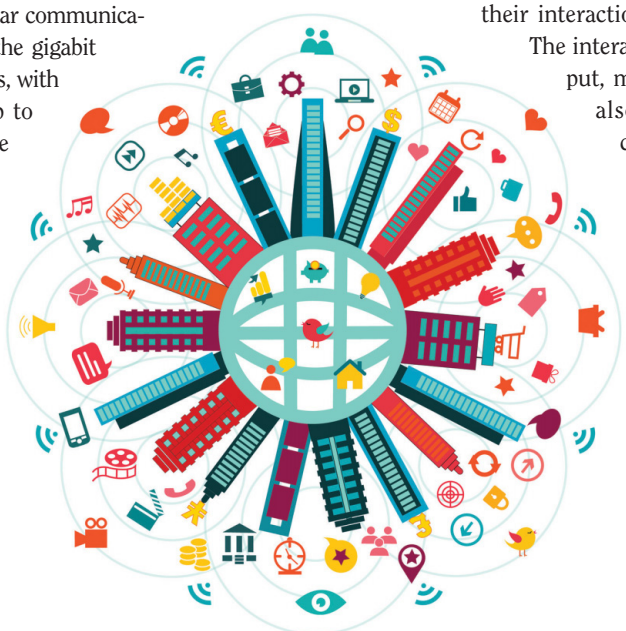
[SP]

Paolo Banelli, Stefano Buzzi, Giulio Colavolpe, Andrea Modenini, Fredrik Rusek, and Alessandro Ugolini

Modulation Formats and Waveforms for 5G Networks: Who Will Be the Heir of OFDM?

An overview of alternative modulation schemes for improved spectral efficiency

Fifth-generation (5G) cellular communications promise to deliver the gigabit experience to mobile users, with a capacity increase of up to three orders of magnitude with respect to current long-term evolution (LTE) systems. There is widespread agreement that such an ambitious goal will be realized through a combination of innovative techniques involving different network layers. At the physical layer, the orthogonal frequency division multiplexing (OFDM) modulation format, along with its multiple-access strategy orthogonal frequency division multiple access (OFDMA), is not taken for granted, and several alternatives promising larger values of spectral efficiency are being considered. This article provides a review of some modulation formats suited for 5G, enriched by a comparative analysis of their performance in a cellular environment, and by a discussion on their interactions with specific 5G ingredients. The interaction with a massive multiple-input, multiple-output (MIMO) system is also discussed by employing real channel measurements.



THE 5G REVOLUTION

©ISTOCKPHOTO.COM/ZONADEARTE

INTRODUCTION

OFDM and OFDMA are the modulation technique and the multiple access strategy adopted in LTE fourth-genereation (4G) cellular network standards, respectively [1]. OFDM and OFDMA succeeded code division multiple access (CDMA), employed in third-generation (3G) networks for several reasons, such as the ease of implementation of both transmitter and receiver thanks to the use of fast Fourier transform (FFT) and inverse FFT (IFFT) blocks; the ability to counteract multipath distortion, the orthogonality of subcarriers which eliminates intercell interference; the possibility of adapting the transmitted power and the modulation cardinality; and the ease of integration with multiantenna hardware, both at the transmitter and receiver.

Digital Object Identifier 10.1109/MSP.2014.2337391
Date of publication: 15 October 2014

1053-5888/14©2014IEEE

Nonetheless, despite such a pool of positive properties, OFDM/OFDMA are not exempt of defects, and their adoption in the forthcoming generation of wireless networks is not taken for granted. Indeed, the spectral efficiency of OFDM is limited by the need of a cyclic prefix (CP) and by its large sidelobes (which require some null guard tones at the spectrum edges), OFDM signals may exhibit large peak-to-average-power ratio values [2], and the impossibility of having strict frequency synchronization among subcarriers makes OFDM and OFDMA not really orthogonal techniques. In particular, synchronization is a key issue in the uplink of a cellular network wherein different mobile terminals transmit separately [3], and, also, in the downlink when base station coordination is used [4], [5]. For instance, with regard to the spectral efficiency loss of sidelobes and the CP, in an LTE system operating at 10 MHz bandwidth, only 9 MHz of the band is used. In addition, the loss of the CP is around 7%, so the accumulated loss totals at 16%. These drawbacks, which invalidate many of the above-mentioned OFDM/OFDMA advantages, form the basis of an open and intense debate on what the modulation format and multiple access strategy should be in next-generation cellular networks. Fifth-generation cellular systems will feature several innovative strategies with respect to existing LTE systems, including, among others, extensive adoption of small cells, use of millimeter (mm)-wave communications for short-range links, large-scale antenna arrays installed on macro base stations, cloud-based radio access network, and, possibly, opportunistic exploitation of spectrum holes through a cognitive approach [6]. All of these strategies will be impacted by the modulation format used at the physical layer. At the same time, 5G cellular networks will have more stringent requirements than LTE in terms of latency, energy efficiency, and data rates, which again are impacted by the adopted modulation scheme. This article provides a review of some of the most credited alternatives to OFDM, performs a critical mutual comparison in terms of spectral efficiency, and discusses their possible interactions with the cited technologies and requirements of 5G networks. The focus of the article is on linear modulations and, after a quick review of OFDM, the emphasis is shifted on filter bank multicarrier (FBMC), time-frequency packing, and single-carrier modulations (SCMs). Particularly, we will focus on spectral efficiency employing quite a general signal processing framework coupled with an information theoretic approach, which permits evaluating the practical information rate associated with a specific signal format. The aim indeed is far away to be exhaustive with respect to all the possible implementation issues and scenarios, still highlighting possible research directions and approaches that deserve to be further investigated. The article is also enriched by a specific section on massive MIMO systems, and by a performance study based on real channel measurements from a massive MIMO testbed from Lund University in Sweden.

SYSTEM MODEL

For all the modulation formats considered in this work, the complex baseband equivalent of the transmitted signal, say $x(t)$, can be expressed as

$$x(t) = \sqrt{PT_s} \sum_{\ell=-G}^{G} s_\ell(t - \ell T_s), \tag{1}$$

where P is the signal power, T_s is the symbol period, $2G+1$ is the number of temporal slots spanned by each data packet, and the waveform $s_\ell(t)$ is the complex baseband equivalent of the waveform associated to the ℓth temporal slot [7], [8], and is written as

$$s_\ell(t) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} d_{k,\ell} p(t) e^{j2\pi k \frac{\delta_f \delta_t}{T_s}(t+\ell T_s)}. \tag{2}$$

In (2), N is the number of subcarriers, $d_{k,\ell}$ is the transmitted symbol associated with the (k,ℓ)th resource element (i.e., kth subcarrier and ℓth symbol interval), $p(t)$ is the underlying shaping pulse, and δ_t and δ_f are two dimensionless constants that rule the actual time and frequency spacing among the transmitted symbols $d_{k,\ell}$. In particular, letting T be a reference symbol time used for normalization and defined as $T = T_t/\delta_t$, it is seen that symbols $d_{k,\ell}$ are spaced in time by $T_s = \delta_t T$ and in frequency by $\delta_f \delta_t / T_s = \delta_f / T$. Note that letting $\delta_f = \delta_t = 1$, we obtain the usual orthogonality-preserving frequency spacing $1/T$ that holds for OFDM systems, while the dimensionless product $\delta_f \delta_t$ can be interpreted as a measure of how much symbols are packed with respect to the classical OFDM choice [8]–[10]. Combining (1) and (2) we also obtain

$$x(t) = \sqrt{\frac{PT_s}{N}} \sum_{k=0}^{N-1} \sum_{\ell=-G}^{G} d_{k,\ell} p(t - \ell T_s) e^{j2\pi k \frac{\delta_f \delta_t}{T_s} t}. \tag{3}$$

Note also that the shaping pulse $p(t)$ has no restrictions in its (practically finite) time duration, but it is assumed to be of unit energy, i.e., $\| p(t) \|^2 = 1$. Moreover, as specified later, variables $\{d_{k,\ell}\}$, the transmitted symbols, are not necessarily equal to pure modulation symbols as they may include some form of signal processing, that, for instance, allows us to consider other (staggered) lattice structures. The pure data symbols are denoted by $\{a_{k,\ell}\}$, which we assume to be of unit average power, i.e., $E\left[|a_{k,\ell}|^2\right] = 1$.

It is easy to show that the above signal model is representative of several modulation formats.

- *OFDM*: Classical OFDM systems assume $p(t)$ as a rectangular pulse of duration $T_s = T + T_{cp}$, where T_{cp} is the CP duration. Consequently, $\delta_t = 1 + T_{cp}/T$ and $\delta_f = 1$ to grant orthogonality on the useful symbol duration T. Note that the transmitted symbols $d_{k,\ell}$ at the edge bands can be set to zero to limit out-of-band emissions. We also recall here that OFDM is the modulation format used in the downlink of current LTE systems, whereas, for the uplink, an OFDM variant known as single-carrier FDMA (SC-FDMA) is adopted, to limit the peak-to-average power ratio (PAPR) [1].
- *FBMC*: FBMC is an OFDM-like modulation format wherein subcarriers are passed through filters that suppress signals' sidelobes, making them eventually strictly bandlimited. The transmitter and receiver may still be implemented through FFT/IFFT blocks or polyphase filter structures [8], [9], and bandlimitedness may deliver larger spectral efficiency than OFDM. The use of FBMC for 5G cellular

[TABLE 1] PARAMETER SETTINGS FOR THE DISCUSSED MODULATION FORMATS IN VIEW OF THE SIGNAL (3).

	FBMC-QAM	FBMC-OQAM	SCM	TFS-QAM	TFS-OQAM
N	$N > 1$	> 1	1	> 1	> 1
$\delta_t\delta_f$	≥ 1	≥ 0.5	≥ 1	< 1	< 0.5
$\{d_{k,l}\}$	QAM SYMBOLS	$j^{k+l}a_{k,l}$	QAM SYMBOLS WITH CP	QAM SYMBOLS	$j^{k+l}a_{k,l}$
$\{a_{k,l}\}$	N.A.	REAL-VALUED PAM SYMBOLS	QAM SYMBOLS	N.A.	REAL-VALUED PAM SYMBOLS

networks is mainly endorsed for its ability (due to signal bandlimitedness) to cope with network asynchronicity that naturally arises in the uplink and/or in the downlink with coordinated transmission [11], for its greater robustness to frequency misalignments among users when compared to OFDM [12], and for its more flexible exploitation of frequency white spaces in cognitive radio networks [6], [8]. FBMC is usually either coupled with QAM or with offset-QAM (OQAM) modulation formats. For FBMC-QAM, we have $\delta_t\delta_f \geq 1$ and the transmitted symbols $d_{k,\ell} = a_{k,\ell}$ are drawn from an $\mathcal{M}$-ary QAM constellation. For FBMC-OQAM, symbols are instead half-spaced in time with respect to FBMC-QAM, and consequently we have $\delta_t\delta_f \geq 0.5$. The transmitted symbols are related to the data symbols by the relation $d_{k,\ell} = j^{k+\ell}a_{k,\ell}$, and the data symbols $a_{k,\ell}$ are real-valued $\sqrt{\mathcal{M}}$-ary PAM symbols.

■ *Faster-than-Nyquist (FTN)/Time-frequency-packed (TFS) signaling*: FTN signaling, first discussed by Mazo as early as 1975 in [13], is a technique to increase the spectral efficiency of a communication system by letting $\delta_t < 1$, thus introducing intentional interference among data symbols at the transmitter side. For a long time, FTN was studied only as a single carrier technique [14], and over time it stood clear that FTN can exploit the excess bandwidth of the single carrier signal. The rate gains of FTN in single carrier systems spurred a number of extensions of FTN into multicarrier setups [15]–[19], and the resulting modulation formats have also been named as TFS. Let us first lay down the model for the transmitted signal of TFS. The system model we use is a generalized version of either an FBMC-OQAM model or an FBMC-QAM model. In view of (3), the TFS system has all parameters identical to its FBMC counterpart except for the product $\delta_t\delta_f$. When an FBMC-OQAM system underlies the TFS system, this product should satisfy $\delta_t\delta_f < 0.5$, while for an underlying FBMC-QAM system, it satisfies $\delta_t\delta_f < 1$. More sophisticated arguments, inspired by time-frequency analysis, highlight how these communication systems are based on Weyl–Heisenberg function sets, also known as Gabor sets [7], [20] and as special cases of packing data on a Grassmannian manifold [21]. Arguments on data packing theory [9], [21], indicate that the best packing is obtained by hexagonal lattices, which provide some spectral efficiency gains with respect to rectangular, or staggered, lattices. For simplicity, we will not consider this special case, although the general framework derived herein, as well as the conclusions, can be easily extended.

■ *SCM*: Letting $N = 1$, the outlined signal model boils down to a linear SCM format. During recent times, multicarrier systems have been the dominant modulation format, the main reason being that optimal equalization can be efficiently carried out in the frequency domain, while optimal equalization of a single carrier system is much more involved and essentially requires the use of a Viterbi algorithm. Recently, however, there has been regained interest in single carrier techniques due to the development of high-performance and low-complexity equalizers operating in the frequency domain [22]–[24]. In this article, we consider a single carrier structure adopting a CP to provide an interblock interference free system and to convert the channel into a cyclic convolution, which simplifies the usage of the frequency domain equalizer, especially in time-invariant or slowly-varying channels. If the time duration of the channel impulse response is at most T_{ch} seconds and the symbol time is T_s, then the CP needs to be at least $G_{cp} = \lceil T_{ch}/T_s \rceil$ symbols long. Hence, in the data block in (1), only $2G + 1 - G_{cp}$ symbols $d_{k,l}$ corresponds to data symbols, while the other G_{cp} represent the redundancy of the CP.

Table 1 provides an overview of the values of the parameters characterizing the discussed modulations formats.

SPECTRAL EFFICIENCY

When assessing the performance of a given modulation and coding system, a key figure of merit is the spectral efficiency ρ, defined as

$$\rho = \frac{R_c N \zeta_g \log_2(\mathcal{M})}{T_s W_{tot}} \quad \text{b/s/Hz},$$

where R_c is the rate of the employed channel code, W_{tot} is the total frequency occupancy of the signal according to some measure, and $\zeta_g \leq 1$ is the inefficiency due to possible guard bands in multicarrier systems, or dually, guard time in single carrier systems. We remind the reader that $\mathcal{M}$ denotes the cardinality of the employed modulation, and N is the number of subcarriers. Note that spectral efficiency denotes here the data rate that can be transmitted for each bandwidth unit used for transmission, regardless of the underlying bit error rate (BER). Later on, instead, we will focus on the achievable spectral efficiency (ASE), a much more insightful performance measure, representing the spectral efficiency that a system may attain under the constraint of arbitrarily small BER.

For OFDM, with an $\mathcal{M}$-ary QAM ($\mathcal{M}$-QAM) constellation, the spectral efficiency has the expression

$$\rho_{OFDM} = \frac{R_c N \zeta_g \log_2(\mathcal{M})}{N + G_{cp}} \quad \text{b/s/Hz}.$$

Regarding FBMC, with an $\mathcal{M}$-QAM for FBMC-QAM, an $\sqrt{\mathcal{M}}$-PAM constellation for the FBMC-OQAM system, and strict equality

for $\delta_t\delta_f$ in Table 1, the spectral efficiency, as N grows large, in both cases becomes

$$\text{FBMC} = \log_2(\mathcal{M})R_c \quad \text{b/s/Hz}. \tag{4}$$

Thus, compared with OFDM, the loss due to the CP and spectral guard bands has vanished.

Regarding spectral efficiency for SCM, the ideal choice for $p(t)$ is a sinc pulse with double-sided bandwidth $W = 1/T_s$ Hz. However, in practice this is not possible, so a smoother pulse in frequency is used. Let the bandwidth of $p(t)$ be $W_{tot} = (1+\delta)W = (1+\delta)/T_s$, where δ measures the excess bandwidth in comparison to the sinc pulse. Then, the spectral efficiency becomes

$$\rho_{SC} = \frac{R_c\zeta_g\log_2(\mathcal{M})}{(1+\delta)} \quad \text{b/s/Hz},$$

where $\zeta_g = (2G+1-G_{cp})/(2G+1)$ is the inefficiency due to the CP.

Finally, regarding TFS, in a multicarrier system like FBMC, the two parameters δ_t and δ_f control the amount of compression of time and frequency, respectively. In the special case of $\delta_t = \delta_f = 1$, the subcarrier spacing in the case of a pulse shape with no roll-off is exactly the reciprocal of the symbol time T_s, which means that the time-frequency occupancy per complex input symbol $d_{k,\ell}$ becomes exactly 1 Hz/s, which is the smallest possible occupancy if an orthogonal set of pulses is desired. For TFS-QAM with time and frequency packing activated, i.e., $\delta_t\delta_f < 1$, an $\mathcal{M}$-QAM constellation, and a rate R_c code, the spectral efficiency becomes

$$\rho_{\text{TFS}-\text{QAM}} = \frac{R_c\log_2(\mathcal{M})}{\delta_t\delta_f} \quad \text{b/s/Hz}. \tag{5}$$

For TFS-OQAM with an $\sqrt{\mathcal{M}}$-PAM constellation, the spectral efficiency becomes

$$\rho_{\text{TFS}-\text{OQAM}} = \frac{R_c\log_2(\mathcal{M})}{2\delta_t\delta_f} \quad \text{b/s/Hz}. \tag{6}$$

Thus, a spectral efficiency gain proportional to $1/\delta_t\delta_f$ is achieved, compared with an FBMC system, at the cost of increased interference among the symbols $\{d_{k,\ell}\}$. Note that, by setting the limit values of $\delta_t\delta_f$ in the two equations ($\delta_t\delta_f = 1$ and $\delta_t\delta_f = 0.5$, respectively), (5) and (6) collapse into (4). In small cells, where the signal-to-noise ratio (SNR) can be very high, the current trend is to employ high-order constellations, such as 256-QAM or 1,024-QAM. This is in sharp contrast to TFS, which maintains a small constellation size, such as quaternary PSK (QPSK) or 16-QAM, but increases the degree of time-frequency compression to achieve higher spectral efficiencies.

DISCUSSION AND LITERATURE OVERVIEW

As already stated, OFDM is a multicarrier modulation format wherein the use of a CP and a proper spacing among subcarriers ensure orthogonality of the waveforms modulated by different data symbols. In general, the amount of interference among adjacent (both in time and frequency) data symbols is ruled by the sampling on the time-frequency plane of the ambiguity function associated to the prototype pulse shape $p(t)$, expressed by [7]–[9]

$$A_p(\tau,\nu) = \int_t p(t)p^*(t-\tau)e^{-j2\pi\nu t}\,dt. \tag{7}$$

Thus, to minimize the interference from adjacent symbols in time [intersymbol interference (ISI)], and in frequency [interchannel interference (ICI)], several research efforts have been dedicated to designing pulse shapes with good ambiguity functions, i.e., according to an orthogonal design in both the domains, as expressed by $A_p(\ell\delta_t T, k\delta_f/T) = \delta[k]\delta[\ell]$, where $\delta[i]$ is the Kronecker delta function. An excellent overview in this respect is provided in [9]. However, double orthogonal designs exist only when $\delta_t\delta_f > 1$. Furthermore, multipath channels could destroy orthogonality, and mismatched filtering may be preferable in this case [8]. Anyway, also with mismatched filtering, double-domain orthogonality can be granted [with some SNR penalty in additive white Gaussian noise (AWGN)] only when $\delta_t\delta_f > 1$ [8], [9]. For instance, OFDM preserves orthogonality in frequency-selective (multipath) channels by adding a guard time between successive symbols, by means of a CP or zero padding (ZP) [25], [26] which leads to $\delta_t\delta_f = 1 + T_{cp}/T > 1$. Furthermore, constraining the symbols to be real (or imaginary), i.e., using PAMs rather than QAMs, orthogonality can be granted also when $\delta_t\delta_f = 1/2$, as already noticed in the first studies about multicarrier systems, [27], [28], and successively called offset-QAM-based OFDM (OFDM-OQAM) [29], which realizes a rectangular lattice staggering in the time-frequency plane.

Regarding OFDM, orthogonality is lost in the presence of frequency synchronization errors or phase noise, which cause nonnegligible performance loss to OFDM(A) systems [3], [30]–[32]. Furthermore, orthogonality is also lost (and performance significantly degrades) if any carrier frequency offset (CFO) is present [30] or the multipath channel is significantly time-varying (doubly-selective) within the symbol period T_s. In this case, interference cancellation/mitigation techniques should be considered also for the orthogonally designed OFDM systems [33]–[36]. This fact is one of the main motivations for recent research efforts on FBMC schemes that, exploiting similar approaches to combat ISI and ICI by proper pulse-shape designs, may combat the sensitivity to CFO and doubly selective channels, still preserving spectral efficiency with $\delta_t\delta_f = 1$ [37]. Actually, the same philosophy can be used also when $\delta_t\delta_f < 1$ [17], e.g., with the generalization of FBMC according to an FTN principle [13], [15], [16], [18]. The idea is that relaxing the orthogonality constraints [38], the pulse-shape design has higher degrees of freedom to reduce ISI and ICI sensitiveness under doubly selective channels or CFO effects.

Turning back to the issue of spectral efficiency maximization, we note that, ideally, a sinc pulse should be used in single carrier FTN. In practice a smoother spectrum with roll-off δ is instead employed. In an AWGN channel and without TFS, the adoption of these pulses would result in a loss of a factor

$1/(1+\delta)$ from the Shannon limit in terms of spectral efficiency. However, it can be proved that with FTN, the maximum overall spectral efficiency, even when $\delta > 0$, tends to the Shannon limit. Hence, with FTN the excess bandwidth is not imposing any loss at high SNR [39].

Unfortunately, the impressive rate gains of single carrier TFS do not in general carry over to FBMC systems. The reason is that with FBMC, the excess bandwidth is typically smaller and limited to the last subcarriers at the band edge. Yet, there are reasons that show why TFS may still be attractive in multicarrier systems, which we discuss next.

First, one of the constraints in the design of a classical FBMC system is that the time and frequency translated pulses should form an orthonormal basis so that the data symbols can be demodulated independently in an AWGN channel. However, in all channels, save for the special case of an AWGN channel, orthogonality of the pulses is lost at the receiver. This requires some form of an equalizer structure at the receiver side, and such an equalizer can just as well be designed so that it also equalizes the self-induced interference. Still, the constraint of orthogonality puts heavy restrictions on the FBMC design and, by relaxing it, additional degrees of freedom in the pulse design are made available at the cost of a controlled amount of interference.

In cases where the allocated bandwidth to one user is small, the amount of excess bandwidth at the band edge can still be relatively large. In such cases, TFS can beneficially exploit the sidelobes to increase the spectral efficiency. Take the LTE system as an illustrative example: in LTE's 1.4-MHz downlink mode, only 1.08 MHz of the band is used for transmission. The remaining 0.32 MHz is a guard band and does not contribute to the data rate. With TFS, the guard band starts contributing to the data rate, giving an improved spectral efficiency.

Finally, TFS offers a flexible method to adapt the spectral efficiency through varying the two compression parameters δ_t and δ_f. With FBMC, the data rate can only be adapted in discrete steps by changing the constellation size and the coding rate. With TFS, a much finer granularity is achieved. As an extra bonus, TFS may also reduce the number of error correcting codes as it can maintain the same code rate but adapt the spectral efficiency by controlling the parameters δ_t and δ_f. Moreover, TFS creates a shaping effect of the input constellation, so that an SNR gain is typically achieved over standard QAM-type constellations.

DISCRETE-TIME MODEL

In what follows, we outline a discrete-time model for the considered modulations, which can be obtained by sampling the general waveform in (2); moreover, we give an expression for the discrete-time received signal after it has passed through a (possibly) time-varying channel with impulse response $h_c(t,\tau)$.

Thus, by employing a sampling frequency $F_c = 1/T_c = (N_s/T_s)$, with $N_s \geq N$ an integer representing a possible oversampling factor N_s/N, the discrete-time signal $s_\ell[n] = s_\ell(nT_c)$ associated to the ℓth symbol is expressed by

$$\begin{aligned} s_\ell[n] &= \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \tilde{d}_{k,\ell} \underbrace{p[n] e^{j2\pi k \frac{\delta_f \delta_t T_c}{T_s} n}}_{p_k[n]} \\ &= \underbrace{\frac{p[n]}{\sqrt{N}} \sum_{k=0}^{N-1} \tilde{d}_{k,\ell} e^{j2\pi k \frac{\delta_f \delta_t T_c}{T_s} n}}_{d_\ell[n]}, \end{aligned} \tag{8}$$

where $p[n] = p(nT_c)$ is the discrete-time pulse shape during the time support $[0,(Q-1)T_c]$. Note that, when $\delta_f \delta_t \neq 1$, this corresponds to transmitting phase-rotated symbols (differently for each subcarrier and symbol period), as expressed by $\tilde{d}_{k,\ell} = d_{k,\ell} e^{j2\pi\delta_f\delta_t k\ell}$. The first equality in (8) highlights that the signal is obtained by multiplexing the data by a bank of filters $p_k[n]$, while the second shows that the signal is also a time-domain windowing $p[n]$, independent of ℓ, of a multicarrier (OFDM-like) signal $d_\ell[n]$. Collecting the transmitted samples in a $Q \times 1$ vector $\mathbf{s}_\ell = [s_\ell(0), \ldots, s_\ell(nT_c), \ldots, s_\ell((Q-1)T_c)]^T$ the two equalities suggest equivalent block-matrix representations, as expressed by

$$\begin{aligned} \mathbf{s}_\ell &= \sum_{k=0}^{N-1} \tilde{d}_{k,\ell} \mathbf{p}_{t_k} = \underbrace{[\mathbf{p}_{t_0}, \ldots, \mathbf{p}_{t_k}, \ldots, \mathbf{p}_{t_{N-1}}]}_{P_t} \tilde{\mathbf{d}}_\ell \\ &= \mathbf{F}_Q^H \underbrace{[\mathbf{p}_{f_0}, \ldots, \mathbf{p}_{f_k}, \ldots, \mathbf{p}_{f_{N-1}}]}_{P_f} \tilde{\mathbf{d}}_\ell \\ &= \mathrm{diag}(\mathbf{p}_{t_0}) \underbrace{[\tilde{\mathbf{f}}_0, \ldots, \tilde{\mathbf{f}}_k, \ldots, \tilde{\mathbf{f}}_{N-1}]}_{\tilde{\mathbf{F}}^H} \tilde{\mathbf{d}}_\ell, \end{aligned} \tag{9}$$

where $\tilde{\mathbf{d}}_\ell$ represents the $N \times 1$ transmitted data with $[\tilde{\mathbf{d}}_\ell]_k = \tilde{d}_{k,\ell}$, $\mathbf{p}_{t_k}$ is the $Q \times 1$ kth discrete-time pulse shape with $[\mathbf{p}_{t_k}]_n = p_k[n]$, $\mathbf{F}_Q$ is a $Q \times Q$ unitary DFT matrix with $[\mathbf{F}_Q]_{k+1,n+1} = 1/\sqrt{Q}\, e^{-j\frac{2\pi}{Q}kn}$, $\mathbf{p}_{f_k} = \mathbf{F}_Q \mathbf{p}_{t_k}$ represents the kth pulse shape in the discrete frequency domain, and $\tilde{\mathbf{F}}$ is a $N \times Q$ pseudo-DFT matrix with $[\tilde{\mathbf{F}}]_{k+1,n+1} = (1/\sqrt{N}) e^{-j2\pi\delta_f\delta_t(T_c/T_s)kn}$, whose row-vectors $\tilde{\mathbf{f}}_k^H$ represent the modulation frequencies. Note that the signal vector $\mathbf{s}_\ell$ is obtained by a prototype pulse-shaping filter $p(t)$ that spans $\lceil Q/N_s \rceil$ consecutive blocks, which are transmitted every $T_s = N_s T_c$ seconds. Thus, each symbol would generate ISI to the adjacent ones, unless it is designed according to an orthogonal paradigm, e.g., by a Nyquist principle.

Observing (8), and that a time-domain multiplication induces a circular convolution in the DFT domain, if the signal parameters are chosen such that the DFT frequency bins are aligned with the modulation frequencies, i.e., if

$$Q = \frac{1}{\delta_f \delta_t} M N_s \qquad Q, M, N_s, \in \mathbb{N}$$

the matrix $\tilde{\mathbf{F}}^H$ is obtained collecting the equispaced (by M) rows of the $Q \times Q$ IDFT matrix $\mathbf{F}_Q^H$.

We define $\mathbf{Z}^n (\mathbf{Z}^{-n})$ as the Toeplitz matrix with all zeroes, but ones in the nth subdiagonal (superdiagonal). The transmitted vector during the ℓth symbol period is thus expressed as

$$\mathbf{x}_\ell = \sum_m \mathbf{Z}^{mN_s} \mathbf{s}_{\ell+m} = \sum_m \mathbf{Z}^{mN_s} \mathbf{F}_Q^H \mathbf{P}_f \tilde{\mathbf{d}}_{\ell+m}. \tag{10}$$

Denoted as already specified, by $h_c(t,\tau)$, the (possibly) time-varying channel $h_{i,j}^{(c)} = h_c(iT_c, jT_c)$ is the discrete-time counterpart and $\mathbf{H}_{c,\ell}^{(t)}$ the $Q \times Q$ channel matrix that processes the signal transmitted at the ℓth symbol period, with

COMPUTATION OF ACHIEVABLE RATES

We sketch here a methodology for computing the ASE, i.e., the maximum attainable spectral efficiency with the constraint of arbitrarily small BER. For notational simplicity, we omit the dependence of ASE on the SNR. The ASE takes the particular constellation and signaling parameters into consideration, so it does not qualify as a normalized capacity measure (it is often called *constrained capacity*). We evaluate only ergodic rates so the ASE is computed given the channel realization $\mathbf{H}_c^{(f)}$ and averaged over it—remember that we are assuming perfect channel state information at the receiver. The spectral efficiency of any practical coded modulation system operating at a low PER is upper bounded by the ASE, i.e., $\rho \le \text{ASE}$, where

$$\text{ASE} = \frac{1}{T_s F_{\text{tot}}} E_{\mathbf{H}_c^{(f)}}\left[I\left(\{\mathbf{d}_\ell\}; \{\mathbf{y}_\ell^{(f)}\} \middle| \mathbf{H}_c^{(f)}\right)\right] \text{b/s/Hz}, \tag{S1}$$

$I\left(\{\mathbf{d}_\ell\}; \{\mathbf{y}_\ell^{(f)}\} \middle| \mathbf{H}_c^{(f)}\right)$ being the mutual information given the channel realization, and the expectation is with respect to the channel statistics.

The computation of mutual information requires the knowledge of the channel conditional probability density function (pdf) $p\left(\{\mathbf{y}_\ell^{(f)}\} \middle| \{\mathbf{d}_\ell\}, \mathbf{H}_c^{(f)}\right)$. In addition, only the optimal detector for the actual channel is able to achieve the ASE in (S1). We are instead interested in the achievable performance when using suboptimal low-complexity detectors. For this reason, we resort to the framework described in [70, Sec. VI]. We compute proper lower bounds on the mutual information (and thus on the ASE) obtained by substituting $p\left(\{\mathbf{y}_\ell^{(f)}\} \middle| \{\mathbf{d}_\ell\}, \mathbf{H}_c^{(f)}\right)$ in the mutual information definition with an arbitrary auxiliary channel law $q\left(\{\mathbf{y}_\ell^{(f)}\} \middle| \{\mathbf{d}_\ell\}, \mathbf{H}_c^{(f)}\right)$ with the same input and output alphabets as the original channel (mismatched detection [70]). There is not a strict need for $q\left(\{\mathbf{y}_\ell^{(f)}\} \middle| \{\mathbf{d}_\ell\}, \mathbf{H}_c^{(f)}\right)$ to be a valid conditional pdf, as it suffices that $q\left(\{\mathbf{y}_\ell^{(f)}\} \middle| \{\mathbf{d}_\ell\}, \mathbf{H}_c^{(f)}\right)$ is nonnegative for this result to hold [71]. If the auxiliary channel law can be represented/described as a finite-state channel, $q\left(\{\mathbf{y}_\ell^{(f)}\} \middle| \{\mathbf{d}_\ell\}, \mathbf{H}_c^{(f)}\right)$ and $q_p\left(\{\mathbf{y}_\ell^{(f)}\} \middle| \mathbf{H}_c^{(f)}\right) = \sum_{\{\mathbf{d}_\ell\}} q\left(\{\mathbf{y}_\ell^{(f)}\} \middle| \{\mathbf{d}_\ell\}, \mathbf{H}_c^{(f)}\right) P\left(\{\mathbf{d}_\ell\}\right)$ can be computed, this time, by using the optimal maximum a posteriori symbol detector for that auxiliary channel [70]. This detector, that is clearly suboptimal for the actual channel, has at its input the sequence $\mathbf{y}_\ell^{(f)}$ generated by simulation according to the actual channel model [70]. If we change the adopted receiver (or, equivalently, if we change the auxiliary channel) we obtain different lower bounds on the constrained capacity but, in any case, these bounds are achievable by those receivers, according to mismatched detection theory [70]. We thus say, with a slight abuse of terminology, that the computed lower bounds are the SE values of the considered channel when those receivers are employed.

This technique thus allows us to take reduced complexity receivers into account. In fact, it is sufficient to consider an auxiliary channel, which is a simplified version of the actual channel in the sense that only a portion of the actual channel memory and/or a limited number of impairments are present.

In particular, in this article we only consider auxiliary channel laws of the form

$$q\left(\{\mathbf{y}_\ell^{(f)}\} \middle| \{\mathbf{d}_\ell\}, \mathbf{H}_c^{(f)}\right) = \prod_\ell q\left(\mathbf{y}_\ell^{(f)} \middle| \mathbf{d}_\ell, \mathbf{H}_c^{(f)}\right), \tag{S2}$$

i.e., the processing is made on frequency-domain symbols independently and it is also assumed that the receiver is based on a frequency-domain equalizer $\mathbf{G}$ and a symbol-by-symbol detector and thus

$$q\left(\mathbf{y}_\ell^{(f)} \middle| \mathbf{d}_\ell, \mathbf{H}_c^{(f)}\right) \propto \exp\left\{-\frac{\|\mathbf{G}\mathbf{y}_\ell^{(f)} - \text{diag}(\varepsilon_\ell)\mathbf{d}_\ell\|^2}{N_0}\right\}, \tag{S3}$$

where N_0 is the noise variance at the receiver.

The modulation formats are compared in terms of ASE without taking into account specific coding schemes, being understood that, with a properly designed channel code, the information-theoretic performance can be closely approached.

$[\mathbf{H}_{c,\ell}^{(t)}]_{i+1,j+1} = h_{\ell N_s+i,i-j}^{(c)}$. To recover the data $\tilde{\mathbf{d}}_\ell$ transmitted with the ℓth data-block, it is necessary to observe the channel output for (at least) QT_c seconds, and the associated received vector is expressed by

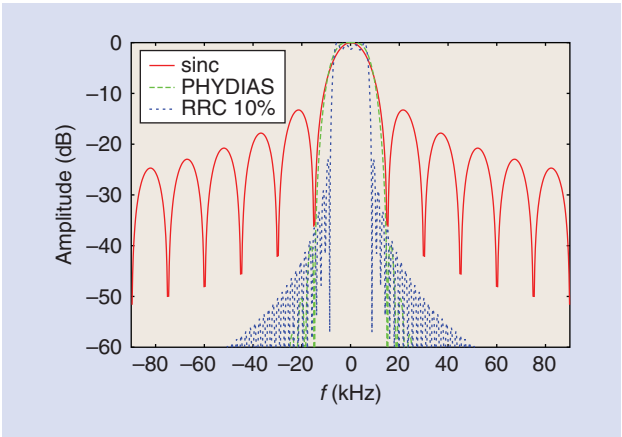
$$\begin{aligned}\mathbf{y}_\ell &= \mathbf{H}_{c,\ell}^{(t)}\mathbf{x}_\ell + \mathbf{w}_\ell = \sum_m \mathbf{H}_{m,\ell}\tilde{\mathbf{d}}_{\ell+m} + \mathbf{w}_\ell \\ &= \mathbf{H}_{\text{tot},\ell}^{(t)}\mathbf{d}_\ell^{(\text{long})} + \mathbf{w}_\ell,\end{aligned} \tag{11}$$

where $\mathbf{d}_\ell^{(\text{long})} = [\ldots, \tilde{\mathbf{d}}_{\ell-1}^T, \tilde{\mathbf{d}}_\ell^T, \tilde{\mathbf{d}}_{\ell+1}^T, \ldots]^T$ is the vector containing both the data of interest and the interference, $\mathbf{H}_{m,\ell} = \mathbf{H}_{c,\ell}^{(t)}\mathbf{Z}^{mN_s}\mathbf{P}_t$, $\mathbf{H}_{\text{tot},\ell}^{(t)} = [\ldots, \mathbf{H}_{-1,\ell}, \mathbf{H}_{0,\ell}, \mathbf{H}_{1,\ell}\ldots]$, and $\mathbf{w}_\ell$ represents the noise at the receiver. In the light of the multicarrier modulation format, the observation model can also be conveniently expressed in the frequency domain by projecting $\mathbf{y}_\ell$ on the same DFT grid of size Q, obtaining

$$\begin{aligned}\mathbf{y}_\ell^{(f)} &= \mathbf{F}_Q\mathbf{y}_\ell = \mathbf{H}_{\text{tot}}^{(f)}\mathbf{d}_\ell^{(\text{long})} + \mathbf{F}_Q\mathbf{w}_\ell \\ &= \mathbf{H}_c^{(f)}\sum_m \mathbf{C}_m\mathbf{P}_f\tilde{\mathbf{d}}_{\ell+m} + \mathbf{w}_\ell^{(f)},\end{aligned} \tag{12}$$

where $\mathbf{H}_{\text{tot}}^{(f)} = \mathbf{F}_Q\mathbf{H}_{\text{tot}}^{(t)}$ is the total observation matrix in the frequency domain, $\mathbf{H}_c^{(f)} = \mathbf{F}_Q\mathbf{H}_c^{(t)}\mathbf{F}_Q^H$ is the frequency-domain channel matrix, $\mathbf{C}_m = \mathbf{F}_Q\mathbf{Z}^{mN}\mathbf{F}_Q^H$ is a full (diagonally dominant) matrix that modifies the pulse-shaping matrix $\mathbf{P}_f$ (note that $\mathbf{C}_0 = \mathbf{I}_Q$), and we omit in the following the dependence on ℓ of the channel matrices for notation compactness.

It is worth noting that the observation models in (11) and (12) share high similarities with the equalization of OFDM signals in doubly selective channels, and several linear and nonlinear data receiver structures may be borrowed, possibly including (data-aided) ISI and ICI cancellation [10], [34]–[36], as



[FIG1] The spectrum of the sinc, PHYDIAS, and RRC 10% pulses.

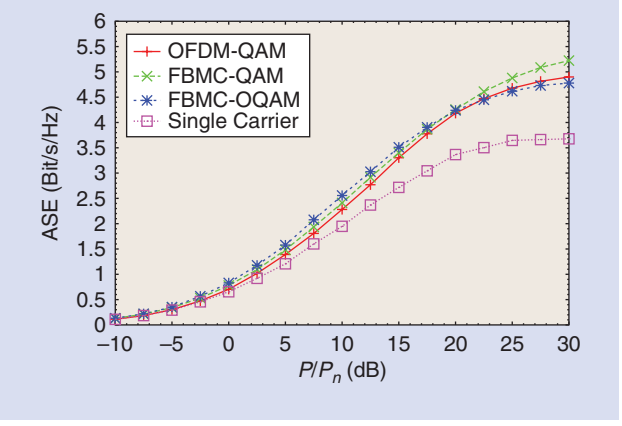
well as joint iterative (turbo) equalization and decoding, [33], [40]. Here, for simplicity, we assume separate data equalization from decoding, and we only consider linear approaches such as matched-filtering (MF), least squares (LS), and linear minimum mean square error (MMSE). Focusing on the frequency-domain observation model, the (soft) estimates of the transmitted data is generally expressed by

$$\hat{\mathbf{d}}_\ell = \mathrm{diag}(\boldsymbol{\epsilon}_\ell)\mathbf{G}\mathbf{y}_\ell^{(\mathrm{f})},$$

where $[\boldsymbol{\epsilon}_\ell]_k = e^{-j2\pi\delta_\mathrm{t}\delta_\mathrm{f}k\ell}$ compensates the phase-rotation when $\delta_\mathrm{t}\delta_\mathrm{f} \neq 1$, $\mathbf{G}$ contains the central rows of the full equalization matrices $\mathbf{G}_{\mathrm{MF}} = \mathbf{H}_{\mathrm{tot}}^{(\mathrm{f})H}$, $\mathbf{G}_{\mathrm{LS}} = \mathbf{H}_{\mathrm{tot}}^{(\mathrm{f})\dagger}$, or $\mathbf{G}_{\mathrm{MMSE}} = \mathbf{H}_{\mathrm{tot}}^{(\mathrm{f})H}(\mathbf{H}_{\mathrm{tot}}^{(\mathrm{f})}\mathbf{H}_{\mathrm{tot}}^{(\mathrm{f})H} + \sigma_n^2\mathbf{I}_Q)^{-1}$, where $\dagger$ is the pseudo inverse operator [41].

EXTENSIONS TO OTHER MODULATION FORMATS AND DISCUSSION

As anticipated in the section "System Model," the data vector may contain some signal processing of the real information vector $\mathbf{a}_\ell$, which in the linear case can be captured by a precoding matrix, e.g., $\mathbf{d}_\ell = \Theta\mathbf{a}_\ell$, and the equalization/detection strategy modified accordingly. A sort of precoding Θ_p on nonfinite alphabets, actually a prefiltering, may be also applied to each vector $\mathbf{s}_\ell$. This way, by proper definition of the prefiltering/precoding matrices Θ_p, Θ, also generalized-FDM (GFDM) [42] as well as universal-FDM (UFDM) [43], can be cast in this framework. Note that, classical FBMC in (8) and (9) performs a prefiltering in the time domain by a diagonal matrix and, consequently, a circular precoding on the data $\mathbf{d}_\ell$. However, a different structure can be imposed to the prefiltering matrix, such as to be full in the time-domain and block-diagonal in the frequency domain, jointly performing spectrum shaping on a block of subcarriers rather than separately on each one, as proposed for UFDM in the ongoing research project 5GNow [43], and somehow reminescent of subblocks precoding in [25]. Adaptation to multiantenna systems is also straightforward by collecting data and observation vectors at each antenna, leading to an observation matrix whose size increases proportionally to the number of antennas. Obviously, the overall complexity, as well as the amount of (interantenna) interference will increase, making the use of MIMO and space-time coded (Alamouti) systems, which heavily rely on the orthogonality (e.g., absence of ISI/ICI) offered by OFDM in frequency-selective channels more challenging. This has probably been one of the strongest objections for employment of FBMC-like systems so far. However, several researchers have already proposed algorithms to deal with these problems, within the great effort of the Physical Layer for Dynamic Access (PHYDIAS) project [44] to establish and promote FBMC-based wireless communications (see [8] and [45] and references therein). Generally, observing that also OFDM faces almost the same problem in doubly selective channels, where it suffers only ICI, channel estimation algorithms, receiver structures, and overall system design can take inspiration by the abundant literature on this subject [46], [47]. For instance, receiver time-domain windowing is effective in this sense to boost the signal-to-noise plus interference ratio (SINR), as proposed in [33] and [35] for pure OFDM. Transmitter and receiver time-domain windowing have been jointly optimized in [48] in multicarrier communications without CP. Recently, maximum SINR approaches for MIMO-FBMC have been investigated in [49] and [50], showing negligible performance degradation with respect to OFDM and significant performance gain with respect to the first attempts to MIMO-FBMC [8], [44].



[FIG2] The ASE for a low-mobility scenario: ETU channel, $f_\mathrm{d} = 0$ Hz, with 64-QAM and bandwidth 1.92 MHz.

PERFORMANCE ASSESSMENT

The modulation formats here discussed will now be compared when used in a typical cellular environment. In particular, we considered the extended typical urban (ETU) channel defined in [51]. This is an example of time- and frequency-selective channel whose continuous-time impulse response can be modeled as

$$h_\mathrm{c}(t,\tau) = \sum_k c_k(t)\,\delta(\tau - \tau_k), \qquad (13)$$

where fading coefficients $c_k(t)$ and continuous-time delays τ_k are typical of each analyzed scenario, and $\delta(\tau)$ is the Dirac delta. From the model (13), the following discrete-time model has been adopted:

$$h_{i,j}^{(c)} = \sum_k c_{k,i}\,\delta[j - j_k],$$

where coefficients $c_{k,i}$ have been generated according to [52], and the discrete-time delays j_k have been chosen as an approximation of continuous-time delays in (13) to their closest integer-multiple of T_c.

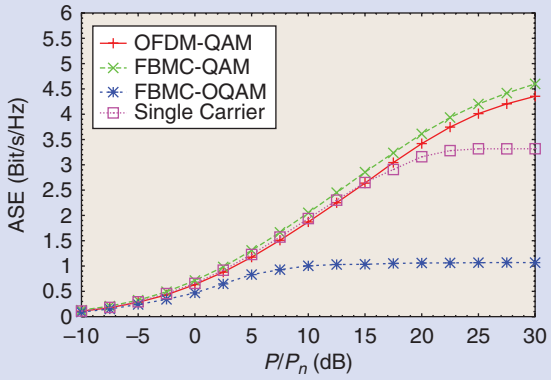
As mentioned, a key figure of merit is represented by the ASE; see "Computation of Achievable Rates" for details on its definition and computation. By properly tailoring the channel code to the considered modulation and the actual channel characteristics, the ASE performance can be closely approached. On the contrary, a comparison based on the bit or packet error rate (PER) performance for a given code does not result in fairness since the code must be specifically tailored to the considered modulation format. Generally speaking and assuming a quasistatic channel, a code designed for the AWGN channel is expected to work well jointly with OFDM or FBMC with offset (or other orthogonal signaling formats). On the contrary, this kind of code will exhibit a significant performance degradation when used with modulation formats that explicitly introduce ISI and/or ICI, as TFS or FBMC without offset. By considering the aforementioned channel, we are also implicitly assessing the robustness of the considered modulation schemes against multipath.

We will assume perfect channel state information at the receiver. Thus, our analysis does not take into account the degradation due to an imperfect channel estimation and the different losses, in terms of spectral efficiency, due to possible different requirements in terms of training or pilot sequences inserted for an accurate channel estimation.

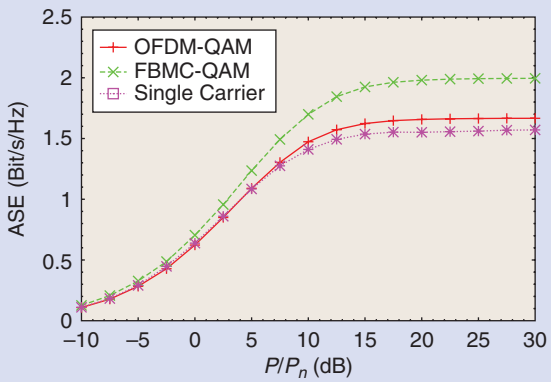
The ASE results will be reported as a function of the ratio between the signal power P and the noise power P_n computed on a reference bandwidth of 1.92 MHz. The spectra of the considered pulses $p(t)$ are reported in Figure 1. In the figure, we see the sinc pulse adopted by OFDM, a pulse with RRC spectrum and excess of bandwidth of 10% (filter length $Q = 1280$, $M = 10$) adopted for FBMC-OQAM and SCMs, and the pulse proposed in the PHYDIAS project [53] for FBMC (filter length $Q = 640$, $M = 5$), where its improved frequency selectivity has been accomplished by using a longer and spectrally well-shaped prototype filter.

In all cases, for OFDM, we will set the losses due to the CP and to the insertion of a number of guard tones compliant with the LTE standard, i.e., 16% in terms of ASE. Moreover, the transmitted symbols, for all the waveforms, will be affected by a random error vector magnitude (EVM) of 4%, with the aim of modeling various imperfections in the implementation (such as carrier leakage, phase noise, etc.). The fractional MMSE equalizer $\mathbf{G}_{\text{MMSE}}$ is adopted at the receiver for all schemes with the exception of FBMC-OQAM, for which we used a matched filter followed by an MMSE equalizer. Indeed, the required length of the filter, for orthoghonality in FBMC-OQAM, makes the receiver complexity too high to use a fractional MMSE.

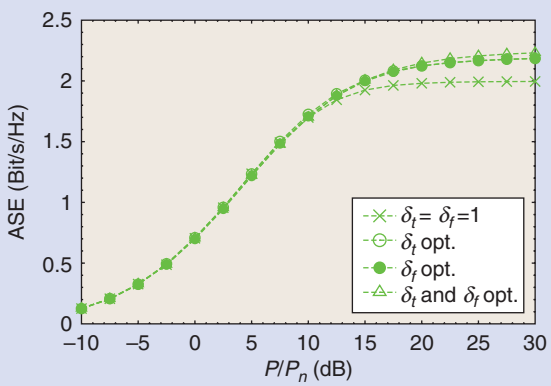
Figures 2 and 3 show the ASE performance for the two extreme scenarios of very low- and high-mobility, characterized by Doppler frequencies $f_d = 0$ and $f_d = 30$ kHz on the ETU channel. We consider QAM with high cardinality $\mathcal{M} = 64$. The figures compare OFDM with FBMC, when $N = 128$ carriers are spaced by 15 kHz. For comparison, we also show the ASE curve of a single carrier with CP system with the same bandwidth 1.92 MHz. We see that, for the low-mobility case, OFDM, FBMC-QAM and



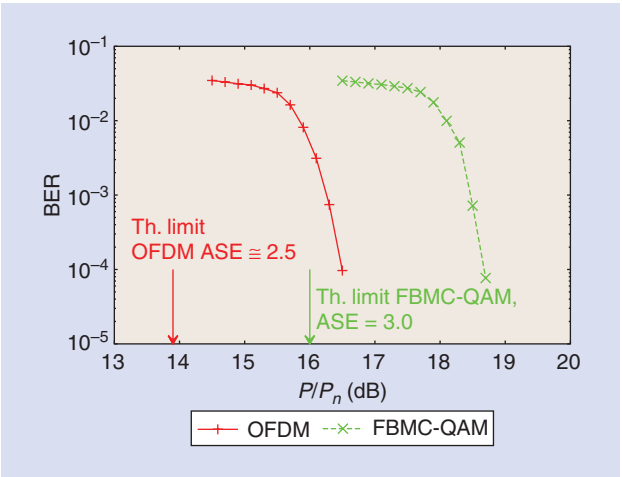
[FIG3] The ASE for a high-mobility scenario: ETU channel, $f_d = 30$ kHz, with 64-QAM and bandwidth 1.92 MHz.



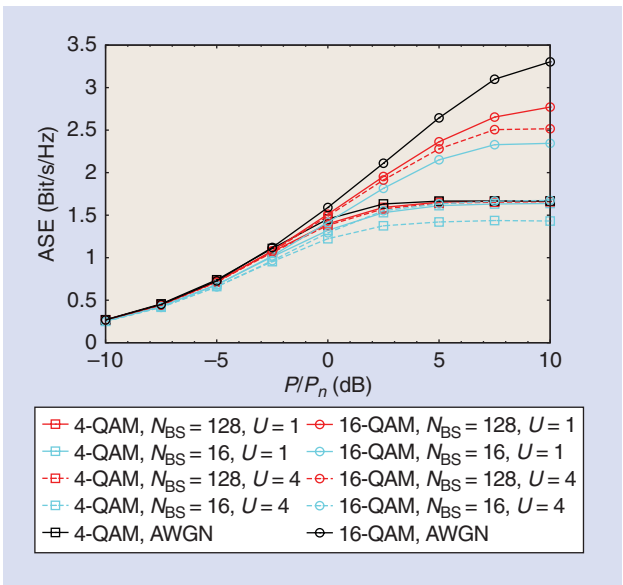
[FIG4] The ASE for ETU channel, $f_d = 30$ kHz, with 4-QAM and bandwidth 1.92 MHz.



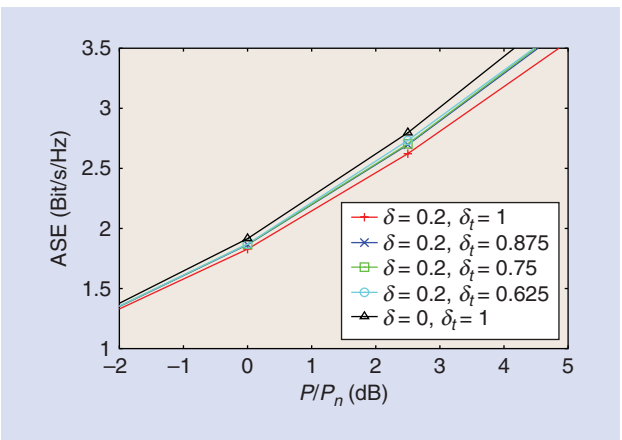
[FIG5] The ASE for TFS: ETU channel, $f_d = 30$ kHz with 4-QAM and bandwidth 1.92 MHz.



[FIG6] The BER for OFDM and FBMC-QAM with 64-QAM on ETU channel, $f_d = 30$ kHz.



[FIG7] The averaged ASE per user for massive MIMO single-carrier FTN systems with different numbers of users and antennas, for 4-QAM and 16-QAM.



[FIG8] The averaged ASE per user for massive MIMO single-carrier FTN systems, with Gaussian inputs, $N_{BS} = 128$, $U = 4$.

FBMC-OQAM have similar performance: FBMC-OQAM achieves a higher spectral efficiency with respect to (w.r.t.) other modulations for low and medium P/P_n, but at high P/P_n values it is outperformed since it has a limited complexity receiver. Instead, performance of SCM is quite limited, since its waveform is strongly affected by the frequency selectivity of the channel, and ASE is limited by the CP loss. In case of high mobility, we see that FBMC-OQAM performance collapses, since its orthogonality is completely destroyed. Instead, FBMC-QAM is more resistant to Doppler and it also gains w.r.t. OFDM.

Since high-order constellations are more sensitive to the impact of the interference present when nonorthogonal signaling is adopted, we also studied the same scenario when the modulation has cardinality $\mathcal{M} = 4$. We can see from Figure 4 that, for this scenario, FBMC outperforms all other modulation formats. We also point out that for all the considered channels, FBMC gains can be even higher by means of a properly designed pulse [9], [17].

We now consider the same scenario when TFS is adopted. Figure 5 shows the performance of TFS when $\mathcal{M} = 4$. Different spacing values δ_t and δ_f have been considered and, to have a wider insight on the possible benefits of this technique, we report the highest ASE achievable when packing in the time domain only ($\delta_t = 0.90$), in the frequency domain only ($\delta_f = 0.95$), and in both domains ($\delta_t = 0.90$ and $\delta_f = 0.95$) is adopted. We can see that TFS gains w.r.t. FBMC are limited, and only at high P/P_n. This is, in some way, expected: further gains could be obtained with more complex receivers (techniques of advanced trellis processing [13], [54]) but, on the other hand, it could be difficult to find substantial gains, since FBMC is already a sort of time-frequency packing. Our own feeling is that these are first results and more research on this topic is required.

As already discussed, the ASE can be approached in practice with proper modulation and coding formats. Figure 6 shows the BER of OFDM and FBMC-QAM for the scenario of Figure 3. The adopted codes are low-density parity-check codes with rate 1/2 and blocklength 64,800 bits. In all cases, a maximum of 50 decoder iterations were performed. We can notice that performance is in accordance with the ASE results. We point out that the loss from the theoretical limit is twofold: first, the adopted code has finite length. Second, it is not designed for the considered channel: the use of codes properly designed for this kind of channels can considerably reduce the loss.

To summarize, as already anticipated in the introduction, we are far from establishing which should be the preferred system, because results highly depend on the specific scenario. Thus, extensive work still has to be done to identify optimal design strategies, which include 1) setting the optimal number of carriers (possibly different for different signal waveforms, especially in the presence or the absence of CP); 2) the optimal pulse-shaper design, which may strongly depend on the information available at the transmitter about the channel maximum delay spread, maximum Doppler spread, and amplitude statistics; 3) the optimal length of the CP, as suggested, e.g., in [55], wherein an OFDM system with tunable length of the CP is proposed.

SINGLE CARRIER FTN MULTIUSER MODULATION WITH MASSIVE MIMO

Fifth-generation macro base stations will certainly be equipped with large-scale antenna arrays, a technology also known as massive MIMO [56], [57]. Using a large number of antennas will help to boost the network throughput, since accurate beamforming will permit serving several users in the same cell and on the same bandwidth, and to stabilize the propagation channel by reducing channel outages by virtue of diversity. The joint design of interference coordination schemes and modulation formats for massive MIMO systems is a topic that will certainly gain momentum in the coming years.

OFDM/OFDMA ARE NOT EXEMPT OF DEFECTS, AND THEIR ADOPTION IN THE FORTHCOMING GENERATION OF WIRELESS NETWORKS IS NOT TAKEN FOR GRANTED.

Let us discuss the uplink between U single antenna users and a base station equipped with N_{BS} antennas. The impulse response, assumed time-invariant for simplicity, between the uth user and the nth base station antenna is denoted by $h_{u,n}(t)$, and we assume these to be perfectly known at the base station side. Each user transmits a, CP-free, single carrier FTN signal according to

$$x_u(t) = \sqrt{\frac{PT_s}{N_{\mathrm{BS}}}} \sum_{\ell} d_{u,\ell} p(t - \ell T_s).$$

The array gain is here harvested as a power saving at the user side and not as increased signal strength at the base station side. The received signal at the nth antenna becomes

$$y_n(t) = \sum_{u=1}^{U} \sqrt{\frac{PT_s}{N_{\mathrm{BS}}}} \sum_{\ell} d_{u,\ell} z_{u,n}(t - \ell T_s) + n(t),$$

where $z_{u,n}(t)$ is the received pulse from the uth user at the nth antenna, i.e., $z_{u,n}(t) = p(t) \star h_{u,n}(t)$, where "$\star$" denotes convolution. To keep complexity low, we consider only single-user detection and construct a discrete-time sequence $\mathbf{y}_u = \{y_{u,\ell}\}$ for the detection of user u according to

$$y_{u,\ell} = \sum_{n=1}^{N_{\mathrm{BS}}} y_n(t) \star z^*_{u,n}(-t)\big|_{t=\ell T_s}.$$

In the case of no FTN, the receiver model is $y_{u,\ell} = \gamma_u d_{u,\ell} + \eta_{u,\ell}$, where γ_u is a measure of signal strength for the uth user and $\eta_{u,\ell}$ collects noise, intersymbol interference, and interuser interference. Under the assumptions that all channel impulse responses are independent and that rich scattering is present, the effect of letting N_{BS} grow is that the impact of intersymbol and interuser interference becomes less and less; asymptotically they both vanish. In such favorable propagation environments, there is no need for any multicarrier system to mitigate multipath as one-tap equalizers can be used, and several users can be spatially multiplexed, which increases spectral efficiency.

While a single carrier system has lower PAPR compared with FBMC systems, there is a reduction of spectral efficiency since pulses with excess bandwidth of an amount δ must be used. To reduce the loss of the excess bandwidth, we make use of FTN. Also in this case it holds that ICI vanishes as N_{BS} grows, but it is no longer true that ISI vanishes. Therefore, we must model the sequence $\mathbf{y}_u$ as

$$\mathbf{y}_u = \mathbf{g}_u \star \mathbf{d}_u + \eta_u,$$

where $\mathbf{g}_u$ is the effective impulse response for user u and η_u collects interuser interference and noise for user u. A sequence detector is now needed to equalize the channel $\mathbf{g}_u$.

PERFORMANCE RESULTS WITH LIVE MASSIVE MIMO CHANNEL MEASUREMENTS

We next report results for SCM in measured massive MIMO channels. Several channel measurement campaigns on massive MIMO has been conducted at Lund University, and more information about the particular one we make use of here can be found in [58]. In brief, four users were placed outdoors around the electrical engineering building at Lund University separated by roughly 30 m, and a linear 128-element antenna array was placed on the roof of the building. The users were placed without any line-of-sight to the base station. The measurement bandwidth is 50 MHz and several snapshots of the propagation channel were taken. In Figure 7, we report results for the no FTN case, i.e., $\delta_t = 1$. The pulse $p(t)$ is RRC shaped with 20% excess bandwidth. We report averaged ASE values over the four users for the system described previously for 4-QAM and 16-QAM, using 16 or 128 antenna elements. The curves marked with "AWGN" show the results obtained when we artificially remove all intersymbol and interuser interference and therefore constitute upper bounds. As can be seen, there is a clear gain in going from $N_{\mathrm{BS}} = 16$ to $N_{\mathrm{BS}} = 128$, and for 4-QAM, the gap to the upper bound is closed. With 16-QAM, the intersymbol and interuser interference has not fully vanished, which shows up as a loss compared with the upper bound. To see how strong the interuser interference is, we also test

[TABLE 2] THE SUITABILITY OF CONSIDERED MODULATION FORMATS TO 5G REQUIREMENTS AND TECHNOLOGIES.

	EASE OF HARDWARE IMPLEMENTATION	LOW LATENCY	IMMUNITY TO PAPR	ROBUSTNESS TO SYNCH. ERRORS	COUPLING WITH MASSIVE MIMO	USE WITH MM-WAVE
OFDM	✓				✓	✓
FBMC				✓	✓	✓
TFS					✓	
SCM		✓	✓		✓	✓

the case of a single user, i.e., $U = 1$. In this case, the gap to the bound reduces, but is not fully closed. This means that the intersymbol interference is stronger than what the single-tap equalizer used can handle. Moreover, the ASE values can be boosted by switching to FTN transmission.

In Figure 8, we repeat the experiments from Figure 7, but we use complex Gaussian modulation symbols and activate FTN; in all cases we use $N_{BS} = 128$ and $U = 4$. In this test, we assume an equalizer that can optimally deal with the intersymbol interference but treats the interuser interference as noise. As a benchmark system, we show the ASE for the impractical but optimal sinc pulse. As we can see, there is a loss in ASE by using the RRC pulse with $\delta = 0.2$. By using FTN, part of this loss is overcome as the ASE curve moves closer to the curve for the sinc pulse. Altogether, we have demonstrated that with massive MIMO, much of the intersymbol and interuser interference can vanish, so that a single-tap equalizer works well for SCM systems. With FTN activated, the loss of the excess bandwidth is reduced. With more advanced transceiver schemes, for example based on interference cancellation, the gap to the upper bounds in Figure 7 can be reduced, and taken together with the favorable PAPR of single carrier, this modulation format seems to be a good choice for uplinks of 5G whenever large antenna arrays can be facilitated.

Although we presented here results for the single carrier only, it is reasonable to foresee that a similar behavior can be observed also for multicarrier modulation formats, such as FBMC and OFDM: in fact, in a massive MIMO system when the number of receiving antennas is sufficiently high, the interuser and intersymbol interference introduced by the channel tend to vanish, no matter the modulation adopted. Such a property has been called *self-equalization* and reported in [59].

REGARDING OFDM, ORTHOGONALITY IS LOST IN THE PRESENCE OF FREQUENCY SYNCHRONIZATION ERRORS OR PHASE NOISE, WHICH CAUSE NONNEGLIGIBLE PERFORMANCE LOSS TO OFDM(A) SYSTEMS.

INTERACTIONS WITH 5G ARCHITECTURE AND REQUIREMENTS

In this section, we finally discuss the interactions between the reviewed modulation formats and some key requirements and features of 5G networks. Although a complete description of how a 5G cellular system will look like is not yet available, some pieces of the puzzles are already known and almost unanimously taken for granted [55]. Some of the concepts discussed here are also summarized in Table 2.

LARGE DATA RATES

Fifth-generation networks will have to support very large data rates; such a goal will be accomplished through a combination of technologies such as the use of multiple antennas (as already discussed), the use of adaptive modulation schemes and, of course, the use of larger bandwidths. This fact tends to promote the use of a multicarrier modulation for two reasons: 1) adaptive modulation is easily implemented with multicarrier schemes, wherein smart bit loading algorithms may permit to tune the modulation cardinality and the coding rate according to the channel status on each subcarrier; and 2) the use of larger bandwidths leads to increased multipath distortion, thus implying that using a multicarrier scheme simplifies the task of equalization with respect to an SCM.

SMALL CELLS AND MM-WAVE COMMUNICATIONS

The use of small cells is a key technique aimed at increasing the overall capacity of wireless networks, intended as offered throughput per square kilometer; recently, there has also been a growing interest for mm-wave communications [60], [61] for supporting short-range cellular communications. Although there is still little knowledge about mm-wave propagation in urban areas, studies are ongoing [62]. It is anticipated that mm-wave will be used on short distances, thus implying that line-of-sight links might be available. In this case, we will have large bandwidths, rather stable propagation environments, and low Doppler offsets. The design of a modulation scheme suited for these conditions is still an open problem, although again multicarrier schemes appear to be much more suited than single-carrier schemes. Due to their anticipated stable propagation environments and low Doppler levels, small cell networks may be especially suitable application areas for nonorthogonal modulation formats. For FBMC, channel estimation gets inherently more challenging due to the interference at the receiver side. However, with increased stability of the propagation environment and low Dopplers, this burden gets significantly simplified. The same arguments also apply to, e.g., advanced FBMC equalizers that equalize the interference among the symbols. Such equalizers need to be updated frequently in the case of nonnegligible Doppler levels, which may impose hefty complexity increases compared with OFDM where only a single tap per detected symbol needs to be updated. On the other hand, there is also a line of thought that foresees, for these high frequencies and large bandwidths, the use of simple modulations formats with low spectral efficiencies, deferring to future generations of cellular systems the task of optimizing the spectrum usage in these bands. The recent paper [63], instead, proposes the use of a single-carrier modulation with CP as a remedy to the PAPR problem of multicarrier schemes.

UNCOORDINATED ACCESS—INTERNET OF THINGS

In the coming years, there will be a tremendous increase in the number of connected devices [64], [65].The current trend is to include a wireless transceiver in almost every electronic gadget/equipment, and researchers have been investigating for some years the so-called Internet of Things—this is also called machine-to-machine communications. A large number of connected devices will require to access the network to transmit short messages. The challenge posed by the Internet of Things lies, rather than in a capacity shortage, in the overwhelming

burden that it produces on the signaling functions of the network. Regarding this aspect, the use of FBMC modulations is preferable with respect to classical OFDM since it allows uncoordinated (i.e., asynchronous) access to the subcarriers. This is one of the main messages conveyed by the ongoing 5GNOW European research project.

THE USE OF FBMC MODULATIONS IS PREFERABLE WITH RESPECT TO CLASSICAL OFDM SINCE IT ALLOWS UNCOORDINATED (I.E., ASYNCHRONOUS) ACCESS TO THE SUBCARRIERS.

LOW LATENCY

Another requirement for 5G wireless cellular systems is the possibility to ensure low-latency communications with a target roundtrip delay of 1 ms. This is seen as a major change of 5G network with respect to existing LTE networks, since it will enable the so-called tactile Internet [66], which will permit the development of brand new real-time applications for monitoring and control. To reduce latency at the physical layer, a single-carrier modulation seems to be preferable, since it avoids block-processing of the data that introduces additional delays. A tunable OFDM system, with an adaptive choice of the length of the data block would also be an option.

ENERGY EFFICIENCY

It is expected that 5G cellular networks will be far more energy efficient than previous cellular systems [67]. Energy saving is mainly a matter that regards a higher layer of the network protocol stack, since it involves adaptive base station switch on/off algorithms, use of renewable energy sources, design of energy-harvesting protocols, base station sharing among network operators during off-peak hours, etc. However, at the physical layer, adaptively switching off unused carriers is a key strategy that may be used to save energy from the radio-frequency (RF) transceiver chain of base stations. This thus once again promotes the use of multicarrier systems with respect to single-carrier modulation.

CLOUD TECHNIQUES AND SOFTWARE RADIO

Another fascinating feature of future wireless networks is the possibility of having a cloud-based radio access network [68], [69]. In practice, base stations will be substituted by light devices, performing baseband-to-RF conversion and signal transmission, and connected through wired optical links to a data center, wherein data coding/decoding and higher-layer functionalities such as resource allocation will take place. The advantages of this structure are represented by the fact that centralized/cooperative strategies (such as the well-known coordinated multipoint) can be readily implemented, as well as by the fact that data modulation can be implemented by a software running in a data center. This adds a lot of flexibility to the choice of the modulation format in the sense that paves the way to adaptive modulation schemes, wherein not only the cardinality and the coding rate may be tuned, but even the waveform itself, including the CP; the recent 5G overview [55] thus proposes the use of "tunable OFDM," a sort of adaptive scheme with parameters chosen based on the instantaneous operating conditions.

Thus, according to the channel conditions, to the requested throughput, and to the available resources in terms, e.g., of number of antennas, adaptive schemes may be designed wherein the modulation format itself is a parameter to be optimized. We believe that, of all the key characteristics of 5G networks, the integration of cloud and software-defined networking strategies within the 5G architecture will be the one to have the greatest impact on the definition of the future modulation format.

CONCLUSIONS

This article provided a review of some linear modulation schemes alternative to OFDM and deemed as suitable candidates for the implementation of the air interface of future 5G cellular communications. A comparison of these modulation schemes in terms of ASE in a cellular environment has been carried out. Our results have shown that there are alternatives to OFDM offering increased values of spectral efficiency, as well as that there is no definite winner, in the sense that the preferable modulation format depends on the considered scenario in terms of channel Doppler spread, channel delay spread, and some other parameters, such as, e.g., the allowed receiver complexity. In this sense, the virtualization of the air interface and the implementation of a cloud radio access network may pave the way towards the adoption of a tunable, adaptive modulation, wherein waveform parameters are chosen based on the specific considered scenario. The article has also reported some discussion on the use of TFS in massive MIMO systems, and has presented a discussion on how the modulation format impacts and is impacted by key technologies and requirements of future 5G networks.

ACKNOWLEDGMENT

We would like to thank Prof. Fredrik Tufvesson and his research group for providing the massive MIMO channel measurements.

AUTHORS

Paolo Banelli (paolo.banelli@unipg.it) received the Laurea degree (cum laude) in electronics engineering and the Ph.D. degree in telecommunications from the University of Perugia, Italy, in 1993 and 1998, respectively. In 2005, he was appointed associate professor in the Department of Electronic and Information Engineering, University of Perugia, where he has been an assistant professor since 1998. In 2001, he joined the SpinComm group, as a visiting researcher, in the Electrical and Computer Engineering Department, University of Minnesota, Minneapolis. His research interests mainly focuses on signal processing for wireless communications, with emphasis on multicarrier transmissions and signal processing for biomedical applications. He is currently an associate editor of *IEEE Transactions on Signal Processing* and was a member (2011–2013) of the IEEE Signal Processing Society's Signal Processing for Communications and Networking Technical Committee. In 2009, he was a general

cochair of the IEEE International Symposium on Signal Processing Advances for Wireless Communications.

Stefano Buzzi (buzzi@unicas.it) is currently an associate professor at the University of Cassino and Lazio Meridionale, Italy. He received his Ph.D. degree in electronic engineering and computer science from the University of Naples "Federico II" in 1999, and he has had short-term visiting appointments in the Department of Electrical Engineering, Princeton University, New Jersey, in 1999–2001 and 2006. His research and study interest lies in the wide area of statistical signal processing and resource allocation for communications, with emphasis on wireless communications. He is the author/coauthor of more than 50 journal and 90 conference papers and is a former associate editor of *IEEE Communications Letters* and *IEEE Signal Processing Letters*. Most recently, he was the lead guest editor for the special issue on "5G Wireless Communications Systems," *IEEE Journal on Selected Areas in Communications* (September 2014).

Giulio Colavolpe (giulio@unipr.it) received the Dr. Ing. degree in telecommunications engineering (cum laude) from the University of Pisa, in 1994 and the Ph.D. degree in information technologies from the University of Parma, Italy, in 1998. Since 1997, he has been at the University of Parma, where he is now an associate professor of telecommunications in the Dipartimento di Ingegneria dell'Informazione. He received the Best Paper Award at the 13th International Conference on Software, Telecommunications, and Computer Networks, Split, Croatia (September 2005), the Best Paper Award for Optical Networks and Systems at the IEEE International Conference on Communcations, Beijing, China (May 2008), and the Best Paper Award at the Fifth Advanced Satellite Mobile Systems Conference and 11th International Workshop on Signal Processing for Space Communications, Cagliari, Italy (September 2010). He is currently an editor of *IEEE Transactions on Communications* and *IEEE Wireless Communications Letters*. He was also an editor of *IEEE Transactions on Wireless Communications* and an executive editor of *Transactions on Emerging Telecommunications Technologies*.

Andrea Modenini (modenini@tlc.unipr.it) received the Dr. Eng. degree in telecommunications engineering (cum laude) in December 2010 from the University of Parma, Italy, where he is currently a Ph.D. student in the Dipartimento di Ingegneria dell'Informazione. His main research interests include information theory and digital transmission theory, with particular emphasis on the optimization of detection algorithm from an information theoretic point of view. He participates in several research projects funded by the European Space Agency (ESA-ESTEC) and important telecommunications companies. In the spring of 2012 he was a visiting Ph.D. student at the University of Lund, Sweden, for research on channel shortening detection for spectrally efficient modulations.

Fredrik Rusek (fredrik.rusek@eit.lth.se) received the M.S. degree in electrical engineering in 2002 and the Ph.D. degree in digital communication theory in 2007, both from Lund Institute of Technology. In 2007, he joined the Department of Electrical and Information Technology at Lund Institute, where he has held an associate professorship since 2012. He has been employed part time as an algorithm expert at Huawei Technologies, Lund, Sweden, since 2012. His research interests include modulation theory, equalization, wireless communications, and applied information theory.

Alessandro Ugolini (alessandro.ugolini@unipr.it) received the Dr. Eng. degree in telecommunications engineering (cum laude) in 2012 from the University of Parma, Italy. Since 2013 he has been a Ph.D. student in the Dipartimento di Ingegneria dell'Informazione at the same university. His main research interests include digital communications, information theory, and spectrally efficient systems. He participates in several research projects funded by the European Space Agency (ESA-ESTEC).

REFERENCES

[1] A. Ghosh, J. Zhang, J. G. Andrews, and R. Muhamed, *Fundamentals of LTE*. Upper Saddle River, NJ: Pearson Education, 2010.

[2] H. Ochiai and H. Imai, "On the distribution of the peak-to-average power ratio in OFDM signals," *IEEE Trans. Commun.*, vol. 49, no. 2, pp. 282–289, Feb. 2001.

[3] M. Morelli, "Timing and frequency synchronization for the uplink of an OFDMA system," *IEEE Trans. Commun.*, vol. 52, no. 2, pp. 296–306, Feb. 2004.

[4] T. Hwang, C. Yang, G. Wu, S. Li, and G. Ye Li, "OFDM and its wireless applications: A survey," *IEEE Trans. Veh. Technol.*, vol. 58, no. 4, pp. 1673–1694, May 2009.

[5] R. Irmer, H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H.-P. Mayer, L. Thiele, and V. Jungnickel, "Coordinated multipoint: Concepts, performance, and field trial results," *IEEE Commun. Mag.*, vol. 49, no. 2, pp. 102–111, Feb. 2011.

[6] S. Haykin, "Cognitive radio: Brain-empowered wireless communications," *IEEE J. Select. Areas Commun.*, vol. 23, no. 2, pp. 201–220, Feb. 2005.

[7] G. Matz, H. Bolcskei, and F. Hlawatsch, "Time-frequency foundations of communications: Concepts and tools," *IEEE Signal Processing Mag.*, vol. 30, no. 6, pp. 87–96, Nov. 2013.

[8] B. Farhang-Boroujeny, "OFDM versus filter bank multicarrier," *IEEE Signal Processing Mag.*, vol. 28, no. 3, pp. 92–112, May 2011.

[9] A. Sahin, I. Güvenç, and H. Arslan, "A survey on multicarrier communications: Prototype filters, lattice structures, and implementation aspects," *IEEE Commun. Surv. Tutorials*, Dec. 2013. doi: 10.1109/SURV.2013.121213.00263

[10] G. Matz and F. Hlawatsch, *Fundamentals of Time-Varying Communication Channels*, F. Hlawatsch and G. Matz, Eds. New York: Academic Press, 2011.

[11] G. Wunder, M. Kasparick, S. ten Brink, F. Schaich, T. Wild, I. Gaspar, E. Ohlmer, S. Krone, N. Michailow, A. Navarro, G. Fettweis, D. Ktenas, V. Berg, M. Dryjanski, S. Pietrzyk, and B. Eged, "5GNOW: Challenging the LTE design paradigms of orthogonality and synchronicity," in *Proc. Vehicular Technology Conf.*, Dresden, Germany, June 2013, pp. 1–5.

[12] T. Fusco, A. Petrella, and M. Tanda, "Sensitivity of multi-user filter-bank multicarrier systems to synchronization errors," in *Proc. IEEE Int. Symp. Communications, Control and Signal Processing*, St. Julian's, Malta, Mar. 2008, pp. 393–398.

[13] J. E. Mazo, "Faster-than-Nyquist signaling," *Bell Syst. Tech. J.*, vol. 54, pp. 1450–1462, Oct. 1975.

[14] A. Liveris and C. N. Georghiades, "Exploiting faster-than-Nyquist signaling," *IEEE Trans. Commun.*, vol. 47, pp. 1502–1511, Sept. 2003.

[15] F. Rusek and J. B. Anderson, "The two dimensional Mazo limit," in *Proc. IEEE Int. Symp. Information Theory*, Adelaide, Australia, Nov. 2005, pp. 970–974.

[16] F. Rusek and J. B. Anderson, "Multistream faster than Nyquist signaling," *IEEE Trans. Commun.*, vol. 57, no. 5, pp. 1329–1340, May 2009.

[17] F.-M. Han and X.-D. Zhang, "Wireless multicarrier digital transmission via Weyl-Heisenberg frames over time-frequency dispersive channels," *IEEE Trans. Commun.*, vol. 57, no. 6, pp. 1721–1733, June 2009.

[18] A. Barbieri, D. Fertonani, and G. Colavolpe, "Time-frequency packing for linear modulations: Spectral efficiency and practical detection schemes," *IEEE Trans. Commun.*, vol. 57, pp. 2951–2959, Oct. 2009.

[19] A. Modenini, F. Rusek, and G. Colavolpe, "Faster-than-Nyquist signaling for next generation communication architectures," in *European Signal Processing Conf. (EUSIPCO)*, Lisbon, Portugal, Sept. 2014.

[20] D. Gabor, "Theory of communication," *IET J. IEE*, vol. 93, no. 26, pp. 429–457, Nov. 1946.

[21] T. Strohmer and R. W. Heath, Jr., "Grassmannian frames with applications to coding and communication," *Elsevier Appl. Computat. Harmonic Anal.*, vol. 14, no. 3, pp. 257–275, May 2003.

[22] E. Ohlmer, M. Jar, and G. P. Fettweis, "Model and comparative analysis of reduced-complexity receiver designs for the LTE-advanced SC-FDMA uplink," *Elsevier Phys. Commun.*, vol. 8, pp. 5–21, Sept. 2012.

[23] W. Gerstacker, F. Adachi, H. Myung, and R. Dinis, "Broadband single-carrier transmission techniques," *Elsevier Phys. Commun.*, vol. 8, pp. 1–4, Sept. 2013.

[24] N. Benvenuto, R. Dinis, D. Falconer, and S. Tomasin, "Single carrier modulation with nonlinear frequency domain equalization: An idea whose time has come again," *Proc. IEEE*, vol. 98, no. 1, pp. 69–96, Jan. 2010.

[25] Z. Wang and G. B. Giannakis, "Wireless multicarrier communications: Where Fourier meets Shannon," *IEEE Signal Processing Mag.*, vol. 17, no. 3, pp. 29–48, May 2000.

[26] B. Muquet, Z. Wang, G. B. Giannakis, M. De Courville, and P. Duhamel, "Cyclic prefixing or zero padding for wireless multicarrier transmissions?," *IEEE Trans. Commun.*, vol. 50, no. 12, pp. 2136–2148, Dec. 2002.

[27] R. W. Chang, "Synthesis of band-limited orthogonal signals for multichannel data transmission," *Bell System Tech. J.*, vol. 45, pp. 1775–1796, July 1966.

[28] B. R. Saltzberg, "Performance of an efficient parallel data transmission system," *IEEE Trans. Commun. Technol.*, vol. 15, no. 6, pp. 805–811, Dec. 1967.

[29] B. Hirosaki, "An orthogonally multiplexed QAM system using the discrete Fourier transform," *IEEE Trans. Commun.*, vol. 29, no. 7, pp. 982–989, July 1981.

[30] L. Rugini and P. Banelli, "BER of OFDM systems impaired by carrier frequency offset in multipath fading channels," *IEEE Trans. Wireless Commun.*, vol. 4, no. 5, pp. 2279–2288, Sept. 2005.

[31] X. Cai and G. B. Giannakis, "Bounding performance and suppressing intercarrier interference in wireless mobile OFDM," *IEEE Trans. Commun.*, vol. 51, no. 12, pp. 2047–2056, Dec. 2003.

[32] L. Tomba, "On the effect of Wiener phase noise in OFDM systems," *IEEE Trans. Commun.*, vol. 46, no. 5, pp. 580–583, May 1998.

[33] P. Schniter, "Low-complexity equalization of OFDM in doubly selective channels," *IEEE Trans. Signal Processing*, vol. 52, no. 4, pp. 1002–1011, Apr. 2004.

[34] Y. Mostofi and D. C. Cox, "ICI mitigation for pilot-aided OFDM mobile systems," *IEEE Trans. Wireless Commun.*, vol. 4, no. 2, pp. 765–774, Mar. 2005.

[35] L. Rugini, P. Banelli, and G. Leus, "Low-complexity banded equalizers for OFDM systems in Doppler spread channels," *EURASIP J. Appl. Signal Processing*, vol. 2006, pp. 1–13, Aug. 2006.

[36] L. Rugini, P. Banelli, and G. Leus, "OFDM communications over time-varying channels," in *Wireless Communications Over Rapidly Time-Varying Channels*, F. Hlawatsch and G. Matz, Eds. New York: Academic Press, 2011, pp. 285–336.

[37] T. Fusco, A. Petrella, and M. Tanda, "Data-aided symbol timing and CFO synchronization for filter bank multicarrier systems," *IEEE Trans. Wireless Commun.*, vol. 8, no. 5, pp. 2705–2715, May 2009.

[38] W. Kozek and A. F. Molisch, "Nonorthogonal pulseshapes for multicarrier communications in doubly dispersive channels," *IEEE J. Select. Areas Commun.*, vol. 16, no. 8, pp. 1579–1589, Oct. 1998.

[39] F. Rusek and J. B. Anderson, "On information rates of faster than Nyquist signaling," in *Proc. IEEE Global Telecommunication Conf.*, San Francisco, CA, Nov. 2006, pp. 1–4.

[40] K. Fang, L. Rugini, and G. Leus, "Low-complexity block turbo equalization for OFDM systems in time-varying channels," *IEEE Trans. Signal Processing*, vol. 56, no. 11, pp. 5555–5566, Nov. 2008.

[41] S. M. Kay, *Fundamentals of Statistical Signal Processing, Vol. 1: Estimation Theory*, Englewood Cliffs, NJ: Prentice Hall, 1998.

[42] G. Fettweis, M. Krondorf, and S. Bittner, "GFDM—Generalized frequency division multiplexing," in *Proc. Vehicular Technology Conf.*, Barcelona, Spain, Apr. 2009, pp. 1–4.

[43] M. Kasparick, G. Wunder, C. F. Schaich, T. Wild, V. Berg, N. Cassiau, J. Dor, D. Ktnas, M. Dryjaski, S. Pietrzyk, I. S. Gaspar, and N. Michailow, "5G waveform candidate selection," Tech. Rep., D3.1 of 5G-Now, FP7 European Research Project, Nov. 2013.

[44] J. Louveaux, L. Baltar, D. Waldhauser, M. Renfors, M. Tanda, C. Bader, and E. Kofidis, "Equalization and demodulation in the receiver (single antenna)," Tech. Rep., D3.1 of PHYsical layer for DYnamic AccesS and cognitive radio (PHYDYAS), FP7-ICT Future Networks, July 2008.

[45] T. Ihalainen, A. Ikhlef, J. Louveaux, and M. Renfors, "Channel equalization for multi-antenna FBMC/OQAM receivers," *IEEE Trans. Veh. Tech.*, vol. 60, no. 5, pp. 2070–2085, June 2011.

[46] A. Stamoulis, S.N. Diggavi, and N. Al-Dhahir, "Intercarrier interference in MIMO OFDM," *IEEE Trans. Signal Processing*, vol. 50, no. 10, pp. 2451–2464, Oct. 2002.

[47] I. Barhumi, G. Leus, and M. Moonen, "Time-domain and frequency-domain per-tone equalization for OFDM over doubly selective channels," *Elsevier Signal Process.*, vol. 84, no. 11, pp. 2055–2066, Nov. 2004.

[48] S. Das and P. Schniter, "Max-SINR ISI/ICI-shaping multicarrier communication over the doubly dispersive channel," *IEEE Trans. Signal Processing*, vol. 55, no. 12, pp. 5782–5795, Dec. 2007.

[49] M. Caus and A. I. Perez-Neira, "Multi-stream transmission for highly frequency selective channels in MIMO-FBMC/OQAM systems," *IEEE Trans. Signal Processing*, vol. 62, no. 4, pp. 786–796, Feb. 2014.

[50] M. Caus and A. I. Perez-Neira, "Transmitter-receiver designs for highly frequency selective channels in MIMO FBMC systems," *IEEE Trans. Signal Processing*, vol. 60, no. 12, pp. 6519–6532, Dec. 2012.

[51] ETSI, "LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Base Station (BS) radio transmission and reception (3GPP TS 36.104 version 11.6.0 Release 11)," Oct. 2013.

[52] Y. R. Zheng and C. Xiao, "Simulation models with correct statistical properties for Rayleigh fading channels," *IEEE Trans. Commun.*, vol. 51, no. 6, pp. 920–928, June 2003.

[53] A. Viholainen, M. Bellanger, and M. Huchard, "Prototype filter and structure optimization," Tech. Rep., D3.1 of PHYsical layer for DYnamic AccesS and cognitive radio (PHYDYAS), FP7-ICT Future Networks, Jan. 2008.

[54] A. Piemontese, A. Modenini, G. Colavolpe, and N. Alagha, "Improving the spectral efficiency of nonlinear satellite systems through time-frequency packing and advanced processing," *IEEE Trans. Commun.*, vol. 61, no. 8, pp. 3404–3412, Aug. 2013.

[55] J. G. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. C.K. Soong, and J. C. Zhang, "What will 5G be?," *IEEE J. Select. Areas Commun.*, vol. 32, no. 6, June 2014, pp. 1065–1082.

[56] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Processing Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[57] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *IEEE J. Select. Areas Commun.*, vol. 31, no. 2, pp. 160–171, Feb. 2013.

[58] X. Gao, O. Edfors, F. Rusek, and F. Tufvesson, "Massive MIMO in real propagation environments," Mar. 2014. [Online]. Available: http://arxiv.org/abs/1403.3376

[59] A. Farhang, N. Marchetti, L. Doyle, and B. Farhang-Boroujeny, "Filter bank multicarrier for massive MIMO," Feb. 2014. [Online]. Available: http://arxiv.org/abs/1402.5881

[60] T. S. Rappaport, Shu Sun, R. Mayzus, Hang Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, May 2013.

[61] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 101–107, June 2011.

[62] T. S. Rappaport, F. Gutierrez, E. Ben-Dor, J. Murdock, Y. Qiao, and J. I. Tamir, "Broadband millimeter-wave propagation measurements and models using adaptive-beam antennas for outdoor urban cellular communications," *IEEE Trans. Antennas Propag.*, vol. 61, no. 4, pp. 1850–1859, Apr. 2013.

[63] A. Ghosh, T. A. Thomas, M. Cudak, R. Ratasuk, P. Moorut, F. W. Vook, T. Rappaport, Jr., G. R MacCartney, S. Sun, and S. Nie, "Millimeter wave enhanced local area systems: A high data rate approach for future wireless networks," *IEEE J. Select. Areas Commun.*, vol. 32, no. 6, June 2014, pp. 1152–1163.

[64] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey," *Elsevier Comput. Netw.*, vol. 54, no. 15, pp. 2787–2805, Oct. 2010.

[65] G. Wu, S. Talwar, K. Johnsson, N. Himayat, and K. D. Johnson, "M2M: From mobile to embedded Internet," *IEEE Commun. Mag.*, vol. 49, no. 4, pp. 36–43, Apr. 2011.

[66] G. Fettweis, "The tactile Internet: Applications and challenges," *IEEE Veh. Technol. Mag.*, vol. 9, no. 1, pp. 64–70, Mar. 2014.

[67] S. Tombaz, A. Vastberg, and J. Zander, "Energy- and cost-efficient ultra-high-capacity wireless access," *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 18–24, Oct. 2011.

[68] Z. Zhu, P. Gupta, Q. Wang, S. Kalyanaraman, Y. Lin, H. Franke, and S. Sarangi, "Virtual base station pool: Towards a wireless network cloud for radio access networks," in *Proc. ACM International Conf. Computing Frontiers*, Ischia, Italy, May 2011, pp. 1–10.

[69] Y. Lin, L. Shao, Z. Zhu, Q. Wang, and R. K. Sabhikhi, "Wireless network cloud: Architecture and system requirements," *IBM J. Res. Develop.*, vol. 54, no. 1, pp. 1–12, Feb. 2010.

[70] D. M. Arnold, H.-A. Loeliger, P. O. Vontobel, A. Kavc˘ic´, and W. Zeng, "Simulation-based computation of information rates for channels with memory," *IEEE Trans. Inform. Theory*, vol. 52, no. 8, pp. 3498–3508, Aug. 2006.

[71] F. Rusek and D. Fertonani, "Bounds on the information rate of intersymbol interference channels based on mismatched receivers," *IEEE Trans. Inform. Theory*, vol. 58, no. 3, pp. 1470–1482, Mar. 2012.

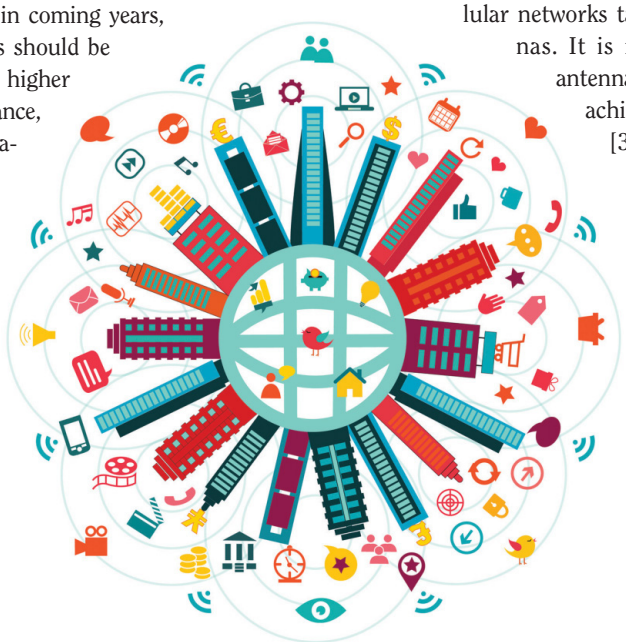
[SP]

S. Mohammad Razavizadeh, Minki Ahn, and Inkyu Lee

Three-Dimensional Beamforming

A new enabling technology for 5G wireless networks

It is anticipated that the mobile data traffic will grow 1,000 times higher from 2010 to 2020 with a rate of roughly a factor of two per year [1]. This increasing demand for data in next-generation mobile broadband networks will lead to many challenges for system engineers and service providers. To address these issues and meet the stringent demands in coming years, innovative and practical solutions should be identified that are able to provide higher spectral efficiency, better performance, and broader coverage. Next generations of wireless cellular networks, which are known as *fifth generation* (*5G*) or *beyond fourth generation* (*B4G*) wireless networks, are expected to produce higher data rates for mobile subscribers in the order of tens of gigabits per second (Gbit/s) and support a wide range of services. Despite the absence of official standards for the 5G, the data rate of 1 Gbit/s per user anywhere for 5G mobile networks is expected to be deployed beyond 2020.



THE 5G REVOLUTION

In this context, several potential technologies have been proposed in recent years that enable 5G systems. Some of these promising methods are small-cell networks, filter bank multicarrier (FBMC), nonorthogonal multiple access (NOMA), massive multiple-input, multiple-output (MIMO), device-to-device (D2D) communication, superwideband frequency spectrums, heterogeneous networks, cognitive radio, millimeter-wave transmission, and three-dimensional beamforming (3DBF). Some basic versions of these techniques have been introduced in the recent releases of mobile system standards such as long-term evolution (LTE) and LTE-Advanced. The current research direction and aim is toward developing these methods for next-generation standards [2].

A number of possible technologies in next-generation cellular networks take advantage of multiple antennas. It is now well known that multiple antennas in wireless systems allow us to achieve higher data rate and reliability [3]–[5]. Beamforming is a signal processing method that generates directional antenna beam patterns using multiple antennas at the transmitter. It is possible to steer the transmitted signal toward a desired direction and, at the same time, avoid receiving the unwanted signal from an undesired direction.

Most beamforming schemes currently employed in wireless cellular networks control the beam pattern radiation in the horizontal plane. In contrast to such two-dimensional beamforming (2DBF), 3DBF adapts the radiation beam pattern in both elevation and azimuth planes to provide more degrees of freedom in supporting users. Higher user capacity, less intercell and intersector interference, higher energy efficiency, improved coverage, and increased spectral efficiency are some of the advantages of 3DBF.

OVERVIEW OF 2DBF

In most of the current cellular networks, the antenna elements at the base station (BS) or the access point are placed along the horizontal axis. Therefore, beamforming and MIMO schemes currently employed in these networks are based on controlling

Digital Object Identifier 10.1109/MSP.2014.2335236
Date of publication: 15 October 2014

1053-5888/14©2014IEEE

the beam pattern radiation in the horizontal plane. This type of beamforming is referred to as 2DBF, which is often combined with cell sectorization to exploit frequency reuse, reduce interference among users, and increase cell capacity. In this method, rather than adopting an omnidirectional antenna at the BS, each cell is divided into sectors (e.g., three sectors) and each sector is served by a directional antenna. The antenna that supports each sector is a one-dimensional array of antenna elements that provides a fan-shaped radiation pattern. These patterns have a wide beamwidth (e.g., 70°) in the horizontal or azimuth plane and a relatively narrow beamwidth (e.g., 10°) in the vertical or elevation plane (see Figure 1). The number of horizontal radiation patterns that can be produced in each sector depends on the number of antennas in that sector. If more than one antenna is placed in each sector, it is possible to form and direct multiple beams to different angles in the horizontal plane. Then, each of these patterns can support one user or a group of users in a specific direction.

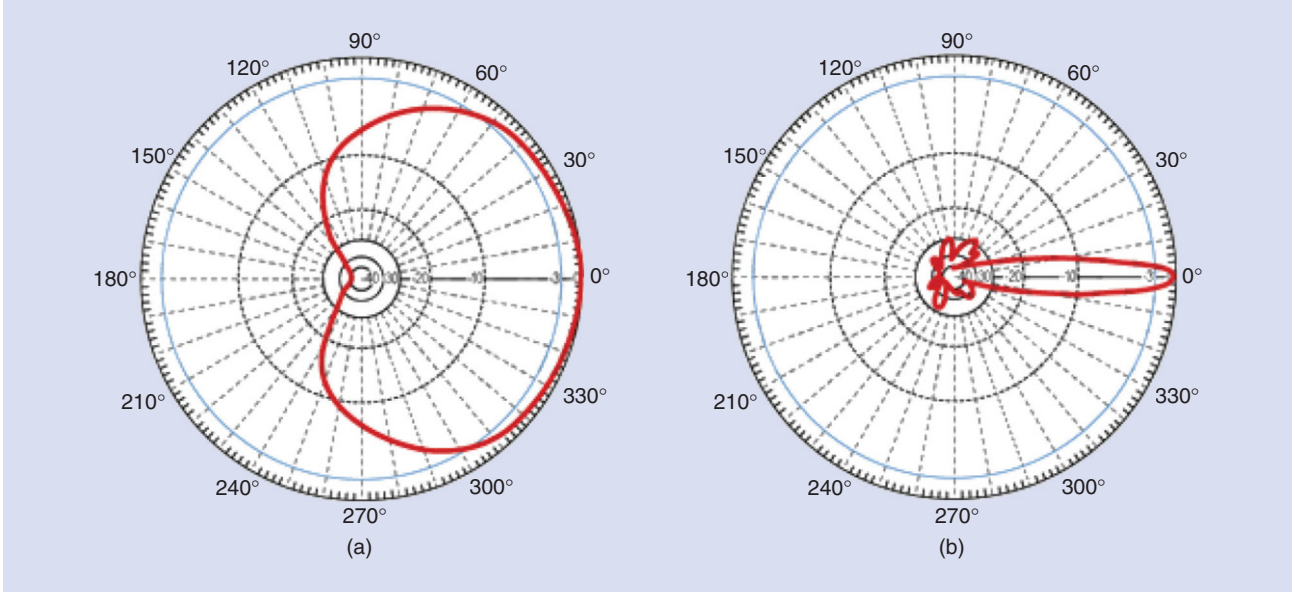
HIGHER USER CAPACITY, LESS INTERCELL AND INTERSECTOR INTERFERENCE, HIGHER ENERGY EFFICIENCY, IMPROVED COVERAGE, AND INCREASED SPECTRAL EFFICIENCY ARE SOME OF THE ADVANTAGES OF 3DBF.

BASICS OF 3DBF

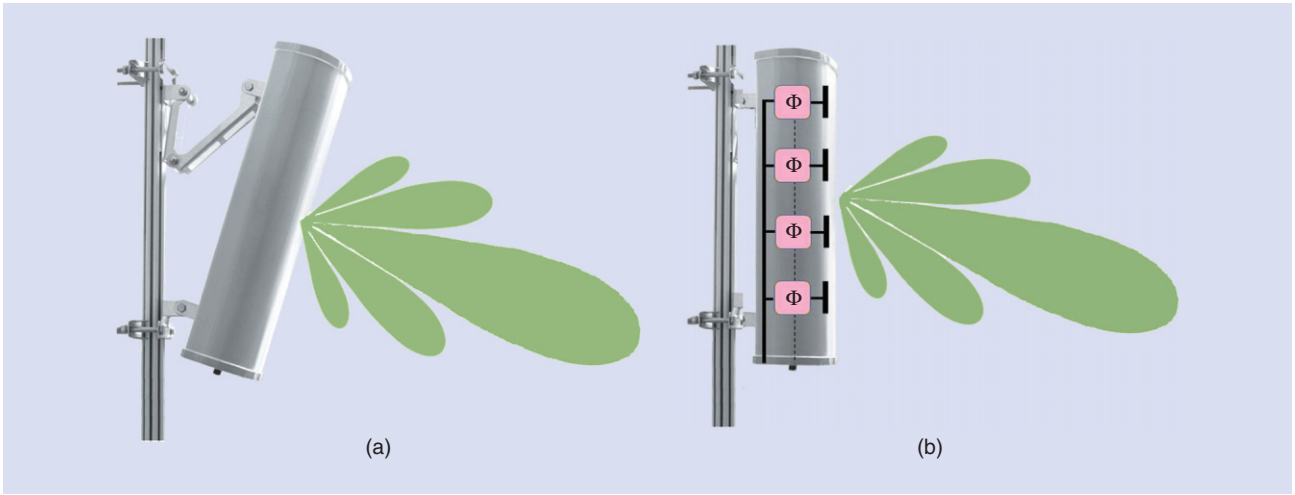
As just mentioned, in 2DBF, the beam pattern is designed only in the horizontal plane. To utilize the vertical domain, antenna tilt can be considered in the vertical axis. The antenna tilting angle is defined as the angle between the horizontal plane and the boresight direction of the antenna pattern. To adjust the tilting angle of the antenna along the vertical axis, mechanical alignment of the antenna can be adopted. In fact, as depicted in Figure 2(a) (which represents mechanical tilt), some adjustable brackets are used to change the tilting angle of the antenna. On the other hand, in some antennas, it is also possible to control the tilting angle electrically (which is called electrical tilt). As shown in Figure 2(b), electrical downtilting is realized by applying an overall phase shift to all antenna elements in the array [6].

An active antenna system (AAS) is a recent technology that helps in getting more control on antenna elements individually. In the AAS, each array element is integrated with a separate radio-frequency (RF) transceiver unit that provides remote control to the elements electronically. By employing AAS at the BS of cellular networks, the vertical radiation pattern can also be adjusted dynamically in each sector, and multiple elevation beams can also be generated to support multiple users or cover multiple regions. A design of 3DBF is achieved by appending this type of vertical beamforming and conventional horizontal beamforming [7].

Depending on the way that the antenna downtilt is changed, 3DBF can be classified into static 3DBF and dynamic 3DBF. The static 3DBF refers to a system where the antenna tilt at the BS is set to a fixed value according to some statistical metrics, e.g., the mean value of the vertical angles of users [7]. This method cannot be adapted to the changing location of the users, i.e., once the tilting angle is selected, it will remain unchanged. In contrast, the dynamic 3DBF is a technique that steers the BS antenna tilting angle instantaneously according to specific user locations. Thus, the mechanical tilt is considered a special case of the static 3DBF, since the tilting angle is determined by the long-term average sense. On the other hand, the dynamic 3DBF



[FIG1] A typical radiation pattern of a BS antenna in (a) horizontal plane and (b) vertical plane.



[FIG2] A comparison of different antenna downtilting methods: (a) mechanical tilting and (b) electrical tilting.

includes the electrical tilt as a special case. We can expect that the dynamic 3DBF offers additional degrees of freedom for performance optimization compared to the static 3DBF.

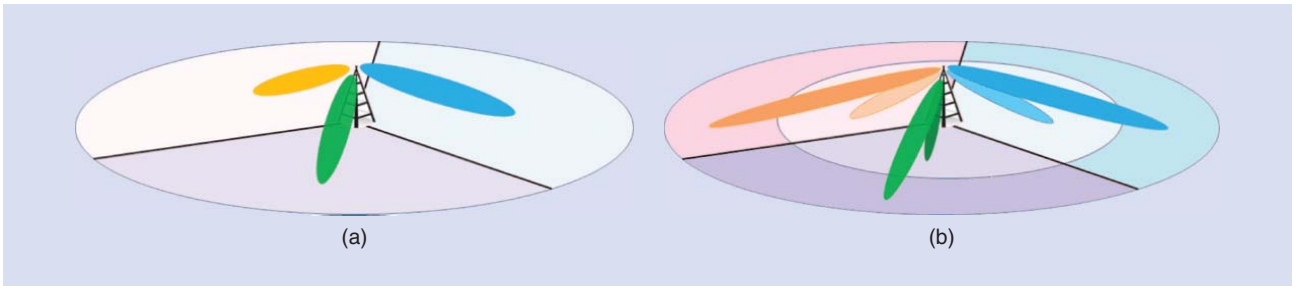
In comparison to the 2DBF, the 3DBF can provide improved capabilities on managing intercell interference in multicell scenarios. When vertical beamforming is applied, different powers can be allocated to the beam patterns that serve cell-edge and cell-center regions separately. This prevents extra power radiation to adjacent cells and decreases the intercell interference in the network. Thus, the 3DBF achieves a higher capacity gain compared to the 2DBF, and its performance gain will be confirmed in the "Simulation Results" section later.

The standardization of the 3DBF has been started in the Third Generation Partnership Project (3GPP) Release 12. In this release, the activities have been limited to a study of feasibility of the 3DBF and its potential gains. The 3DBF implementation is expected to begin at the end of 2015 or even later. It is also foreseen that enhancements to the specifications of the 3DBF will continue at the 3GPP Releases 14 and 15. Works on these two releases will probably start in 2016 and be finalized in 2020. We observe a similar trend in other technologies such as massive MIMO and small cells, which are currently examined in 4G networks and also introduced as promising technologies for 5G networks. In spite of the aforementioned advantages of the 3DBF, there are still some challenges with this technology that need to be addressed. Some of these challenges include three-dimensional (3-D) channel modeling, the overhead related to channel state information (CSI) feedback in frequency division duplexing (FDD) scenarios, power control, antenna designs, and complexity of RF chains.

APPLICATIONS OF DYNAMIC AND STATIC 3DBF

Applying the dynamic 3DBF, a BS can direct the main lobe of the antenna beam to a specific user. As a consequence, the desired signal strength at the intended users is maximized. The 3DBF can also be used to suppress intercell interference in a multicell scenario by adjusting the antenna beams properly. By adopting coordination between neighboring cells, it is possible to achieve a tradeoff between the desired signal power at the intended users and the intercell interference level, and then a further performance improvement is expected.

On the other hand, the static 3DBF in a cellular network can be combined with cell sectorization, which also improves the network performance. As we see in Figure 3(a), the traditional sectorization method provides sectors that are formed along the tangential direction in the horizontal plane. However, as depicted in Figure 3(b), employing two tilting angles at each sector enables additional sectorization along the radial direction, which is called vertical sectorization. Splitting the cell into several sectors can utilize more frequency reuse and



[FIG3] (a) Conventional sectorization. (b) Vertical sectorization.

significantly increase the network capacity. For example, a capacity gain (in terms of the mean throughput) of 70 and 140% can be achieved by moving from three sectors to six and 12 sectors, respectively [8]. Similar gains are also reported in the scenarios that employ vertical sectorization with 3DBF [9]. In addition, vertical sectorization by 3DBF provides more flexibility in traffic load balancing compared to the conventional sectorization by adaptively changing the number of sectors to serve variable traffic loads in the cell [10].

Another important issue in 3DBF is how to acquire CSI to design precoders at the BS. In time-division duplexing (TDD) systems, by exploiting channel reciprocity, CSI of the downlink is obtained from an estimate of the uplink. In FDD systems, the CSI is estimated by a user and then the BS obtains this information through feedback from the user to the BS. In this case, if the number of antenna elements at the BS increases, the feedback overhead is a challenging problem. Hence, in some precoding methods, precoding designs are based on partial or reduced-dimensional CSI [11]. In addition, another problem in TDD operation is the number of training symbols that need to be sent in each coherence block of transmission. This training overhead problem becomes more challenging if the number of users in the network increases [12], and these issues are still open problems in 3DBF.

REVISIT OF 3DBF IN OTHER AREAS

So far, we have discussed the advantages and challenges of 3DBF. It is interesting to note that although 3DBF is a relatively new topic in wireless cellular networks, its history can be traced back to the 1970s when some early methods were proposed for applying multiple elevation radiation beams in radar and underwater sonar systems. For example, in [13], which was published in 1978, the concept of the vertical beamforming was employed for an array of antennas or acoustic elements in synthetic aperture radar systems to increase the resolution of target detection. In this system, multiple contiguous elevation beams with different frequencies are adopted, which can be considered as frequency reuse in cellular networks.

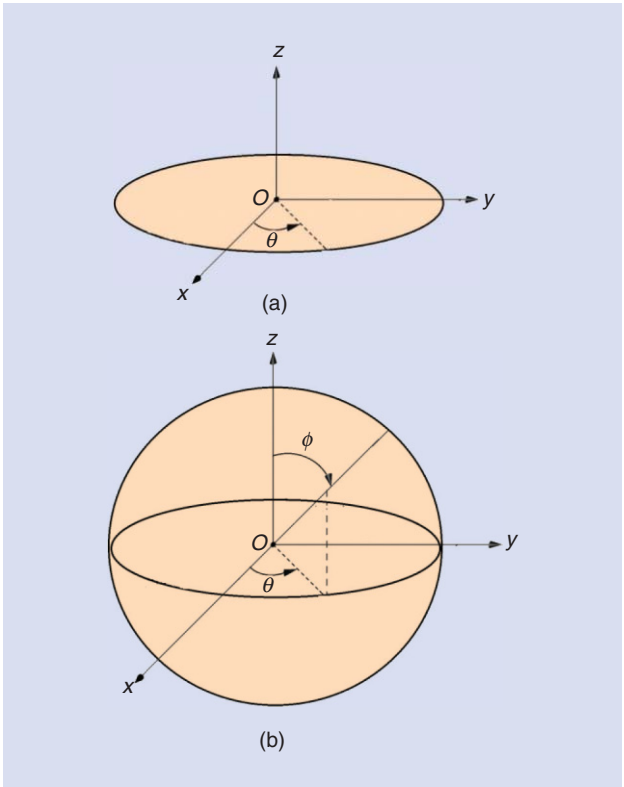
Apart from radio communication systems, techniques similar to the 3DBF have been applied in other areas of signal processing. It is worth studying similarities between those application areas and cellular communication to gain insights on current challenges of 3DBF. One popular application of 3DBF can be found in imaging. Three-dimensional imaging is widely adopted in medical systems using ultrasound, microwave, and X-ray [14], and 3DBF can provide higher-resolution and image quality compared to the conventional methods. For example, in [15], an ultrasound imaging scheme employed 3DBF to synthesize a 3-D volume with a 16 $\times$ 16 planar receiving array and a 6 $\times$ 6 transmit array, and the 3DBF algorithm was decomposed into two simpler 2DBF processes. The idea of these works can be used in designing the antenna arrays and also adaptive beamforming and vertical sectorization methods for cellular systems.

Audio and speech processing is another major application area that has already adopted the 3DBF. In this case, the 3DBF is utilized in a two-dimensional (2-D) or 3-D array of loudspeakers to focus sound radiation to a specific location. Similarly, an array of microphones and the 3DBF can be applied to receive sound from a desired direction in the 3-D space. For example, a moving person can be tracked while she/he speaks or a sound source can be localized in video conference applications [16]. In [17], a 3DBF technique was proposed to use a spherical array consisting of pressure microphones where different frequencies and wave propagation effects of audio and radio waves are taken into account. An example of such an array is illustrated in Figure 4. These configurations can provide inspiration for designing 3-D antenna arrays topologies for wireless communication. The scheme in [17] is based on a spherical harmonic decomposition of the sound field for acoustic applications and is capable of steering the beam pattern to any direction in the 3-D space without changing the shape of the pattern. This can also give ideas for designing beam patterns in the 3DBF with per-user beam steering, considering that modifying the antenna beam shape with the elevation and azimuth angles may increase interuser interference and affect the performance of the 3DBF in the network. These abundant prior works related to the 3DBF demonstrate that this technique has a strong connection to other signal processing areas. By exploiting these results already developed in other applications, a successful deployment of 3DBF in 5G systems can be made possible.

3-D WIRELESS CHANNEL MODELING

To evaluate the performance of the 3DBF in cellular systems, a proper 3-D channel model is required. In particular, the power spectral density of the received signals and the channel capacity are dependent on the adopted channel model. Similar to other multiple antenna technologies, a channel model for the 3DBF must accurately describe the spatial environment along with

[FIG4] A spherical array of loudspeakers (figure used with permission of [30]).



[FIG5] The (a) 2-D and (b) 3-D channel models.

time and frequency characteristics. To capture the spatial properties of 3DBF, we need two types of information: antenna configurations and the propagation field properties [18].

Most of the basic channel models in cellular networks are 2-D models. The 2-D models such as the 3GPP spatial channel model (SCM) [19] or International Telecommunications Union (ITU) double directional channel model [20] consider a distribution of scatterers only in the azimuth plane and do not take the elevation angle into account. For example, the widely used Clark's model [21] is also a 2-D model that assumes all signals received from uniform directions in the azimuth plane as shown Figure 5(a). However, some practical measurements in the urban environments show that up to 65% of energy is incident with elevation angles larger than 10° [22].

To improve the 2-D models, several studies have been conducted on 3-D channel models to include nonzero elevation angles and accurately evaluate the systems. For example, some works have been initiated by 3GPP to define the 3-D channel models for vertical beamforming in LTE [23]. One of the first research works on the statistical modeling of the 3-D fading channels was described in [24], which is an extension of Clark's 2-D model. In this model, the scatterers are assumed to be distributed in a cylinder around the receiver and, in fact, both azimuth and elevation angles of arrival are taken into account. Here, the azimuth angle is distributed uniformly in a circle, while the elevation angle has a nonuniform distribution. This model was later extended to a two-cylinder model in which both transmitter and receiver are assumed to be placed inside two cylinders of scatterers. In a general case, the scatterers can exist inside a sphere around the receiver [as in Figure 5(b)].

One of the most popular 3-D channel models in literature is the WINNER channel model [25], which is an extension of a previous 3GPP 2-D SCM and 2-D ITU model. In addition to different elevation and azimuth angles, this model also includes other parameters such as the number of antennas, the vertical and horizontal radiation patterns, and dual-polarization transmissions. The model can accommodate the bandwidth of up to 100 MHz and carrier frequencies between 2 and 6 GHz. Most of the model's parameters are deduced from empirical measurements. More details about this model can be found in [25].

Although there are some 3-D models that can be used for current wireless networks, a complete and accurate 3-D channel model will be highly desirable for actual evaluation of the 3DBF and other novel technologies in 5G networks. Current channel models such as WINNER+ have some limitations. For example, although the WINNER is an antenna independent model, its implementation is now only applicable to uniform linear arrays [26]. New 3-D models need to support massive and ultradense antenna deployments, greater bandwidth, higher carrier frequencies, high-speed users, new deployment scenarios such as D2D communication, and transitions between different propagation environments (e.g., urban, rural, outdoor, or indoor).

ANTENNA ARRAY DESIGN

As mentioned before, the antenna at a BS of cellular systems is usually implemented as a linear array of a limited number of antennas in the azimuth plane. However, these geometries can shape the radiation pattern only in the horizontal plane, and hence to change the beam in the elevation plane for 3DBF, more general 2-D or 3-D arrays topologies are necessary. Those arrays are active antenna systems that are spaced in both azimuth and vertical planes with different configurations such as planar, circular, spherical, or cylindrical structures. In addition, the array may include copolarized or cross-polarized antenna elements. The active antenna arrays placed in 2-D are also called full-dimension MIMO (FD-MIMO) or 3-D MIMO [27]. Massive MIMO, which employs up to several hundreds of antenna elements, can be a potential extension of 3DBF for 5G systems.

In general, adding more antenna elements to the array provides more flexibility in beam steering designs and increases the number of radiation beams of the array. For vertical sectorization in which the number of vertical sectors is usually small (e.g., two or three), only a small number of antennas are required in the vertical plane. However, in 3DBF with per-user beam pattern adaptation (i.e., user tracking), a large number of antennas are needed. Hence, one of the challenges of the 3DBF is physical constraints and placement of a large number of antennas at a BS. This problem may be alleviated in higher frequencies that are expected for 5G networks. Also FD-MIMO can be utilized to address this issue. For example, in [27] a 2-D array of 32 antennas comprising eight antenna ports in the horizontal plane and four antenna ports in

the vertical dimension is proposed. Each antenna port consists of a subarray of four antennas in the vertical dimension, which results in a total of 128 elements in this antenna. The size of this antenna in the 2.5 GHz band is 1 m × 0.5 m, which is appropriate to fit on a BS tower. A similar study shows that the same number of antennas can be placed on a cylinder with a diameter of 1 m [28]. Another challenge of adding more antenna elements at the BS is cost related to RF chains required for all antennas, which needs to be addressed in future research.

THROUGH NUMERICAL SIMULATIONS, IT IS SHOWN THAT THE 3DBF OUTPERFORMS THE CONVENTIONAL 2DBF, AND THUS IT IS EXPECTED THAT 3DBF WILL PLAY A CRUCIAL ROLE IN 5G SYSTEM DESIGNS.

One more important factor in the BS antenna design is the pattern shape of the array. A desired pattern in the 3-D space can be determined by methods such as the Fourier–Bessel series [29]. Most of these pattern synthesis techniques were proposed for planar arrays and symmetric arrays such as circular arrays. Their aim is to determine the main-lobe shape and side-lobe levels. Besides, most of the beam pattern designs (or array design methods) are based on narrowband systems, which are not appropriate for next-generation wireless systems. Hence, it is important to consider more advanced array configurations and wideband beam shaping designs. Generally, to widen the bandwidth of the arrays, different structures in different frequencies can be employed. It is also possible to apply frequency invariant antenna arrays [29].

In the model introduced by the ITU [20], the antenna radiation pattern in a given user's direction is defined by

$$A_P(\theta,\phi) = G_{\max} - \min\{A_H(\theta) + A_V(\phi), A_m\}, \tag{1}$$

where $G_{\max}$, θ, and ϕ denote the maximum antenna gain at the main beam (boresight) direction, the angles of the user direction from the boresight direction in the horizontal plane, and the tilting angle of the user, respectively, as shown in Figure 6. Here A_H and A_V stand for the relative antenna gains in the horizontal and vertical planes, respectively, expressed as

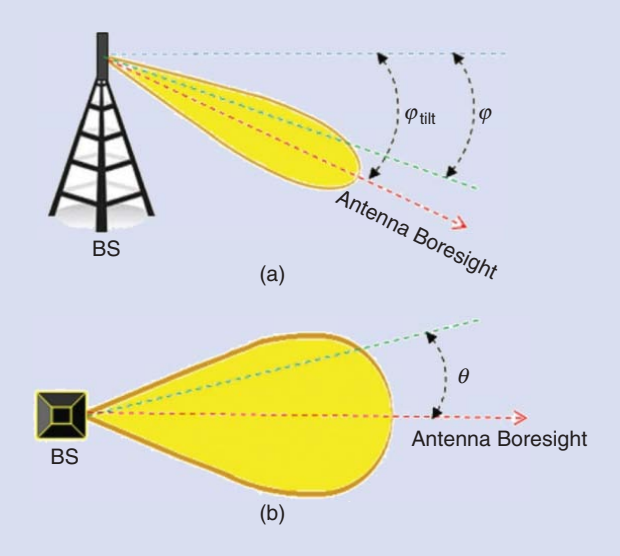
$$A_H(\theta) = \min\left[12\left(\frac{\theta}{\theta_{3\text{dB}}}\right)^2, A_m\right] \tag{2}$$

$$A_V(\phi) = \min\left[12\left(\frac{\phi - \phi_{\text{tilt}}}{\phi_{3\text{dB}}}\right)^2, A_m\right], \tag{3}$$

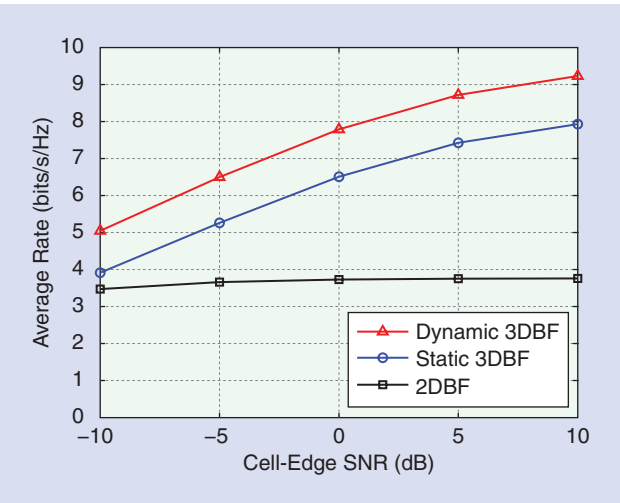
where ϕ_{tilt} represents the main beam tilting angle, $\theta_{3\text{dB}}$ and $\phi_{3\text{dB}}$ indicate the 3-dB beamwidth of the horizontal and vertical patterns, respectively, and A_m equals the side-lobe level attenuation of the antenna pattern. In normal situations, we usually have $A_m = 20\sim30$ dB, $\theta_{3\text{dB}} = 60^\circ\sim70^\circ$, and $\phi_{3\text{dB}} = 8^\circ\sim15^\circ$.

SIMULATION RESULTS

In this section, we present the efficiency of the 3DBF compared to the 2DBF through Monte Carlo simulation. For simulations, we adopt the WINNER+ channel model and the urban microcell environment in [20] and [25] with slight modifications. We assume that a single user is active per sector at each time slot. Also, users are uniformly distributed in all cells, and the BS has perfect CSI to compute transmit precoding vectors. Table 1 illustrates the simulation settings. We assume that the BS antenna structure is the uniform linear array with 10λ antenna spacing where λ denotes the wavelength of the system. The 2DBF is evaluated by the horizontal gain in (2). In these simulations, for the case of the dynamic 3DBF, the tilting angle ϕ_{tilt} is set to the actual vertical angle of the user ϕ. This means that the vertical beam pattern is adapted to the user locations in a dynamic way. We also present the performance of



[FIG6] Three-dimensional antenna pattern modeling (a) the vertical plane and (b) the horizontal plane.



[FIG7] The average rate performance comparison as a function of cell-edge SNR.

[TABLE 1] THE SIMULATION SETTINGS.

CELL TYPE	URBAN MICRO
CELL LAYOUT	SEVEN CELLS WITH THREE SECTORS PER CELL
INTERSITE DISTANCE	200 m
CHANNEL MODEL	WINNER+
MINIMUM DISTANCE BETWEEN A USER AND A BS	10 m
NUMBER OF Tx ANTENNAS AT BS	4
NUMBER OF Rx ANTENNAS OF A USER	1
SCHEDULER	ROUND-ROBIN
TRANSMIT PRECODING	MAXIMAL RATIO TRANSMISSION
BS ANTENNA HEIGHT	10 m
USER ANTENNA HEIGHT	1.5 m
PATH LOSS EXPONENT	3.4
MAXIMUM ANTENNA GAIN	17 dB
MAXIMUM ATTENUATION OF THE BS ANTENNA	20 dB
VERTICAL 3-dB BEAMWIDTH	8°
HORIZONTAL 3-dB BEAMWIDTH	65°

the static 3-D where the tilting angle is fixed. In this case, the optimal tilting angle ϕ^*_{tilt} is obtained from exhaustive search.

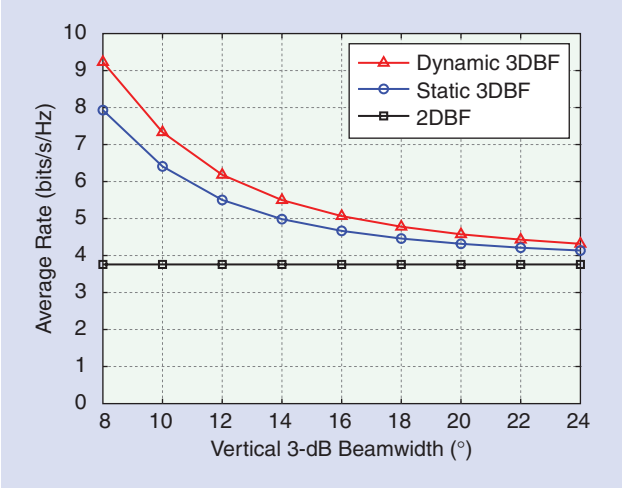
Figure 7 exhibits the average rate performance with respect to the cell-edge signal-to-noise ratio (SNR), which is defined as the received SNR at the cell boundary. Compared to the 2DBF, the average rate is enhanced by 75% in the static 3DBF with the fixed tilting angle $\phi^*_{\text{tilt}} = 11^\circ$ when the cell-edge SNR is equal to 0 dB. These performance improvements are due to a reduction of intercell interference that results from narrow vertical beam adjustment. Also, the dynamic 3DBF achieves a performance gain of 109% over the 2DBF at the cell-edge SNR of 0 dB. We can see that the dynamic 3DBF outperforms the static 3DBF, since the vertical beam in the dynamic beamforming can be adaptively aligned to the user position. This performance gap between the 3DBF and the 2DBF increases as the cell-edge SNR grows.

To illustrate the effect of the vertical 3-dB beamwidth of the BS antennas on the performance, Figure 8 presents the average rate of the 3DBF and the 2DBF for different $\phi_{3\,\text{dB}}$ when the cell-edge SNR is set to 10 dB. Since the 2DBF is not affected by the vertical 3-dB beamwidth, the average rate performance remains constant. However, the performance gain of the 3DBF compared to the 2DBF grows as the antenna vertical beamwidth becomes smaller, since a larger vertical 3-dB beamwidth results in more intercell interference.

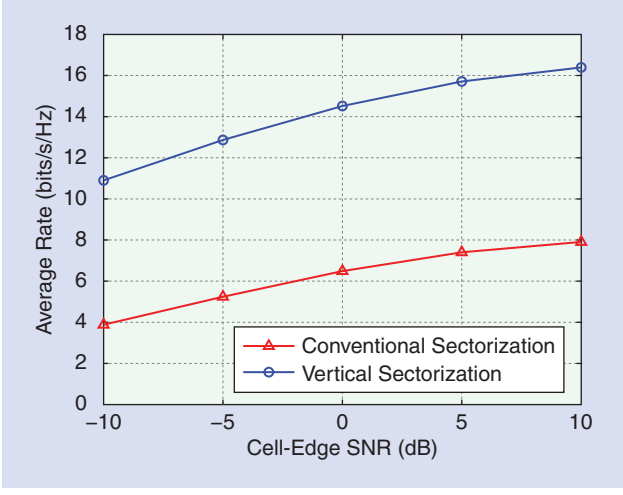
We also compare vertical sectorization with the conventional horizontal sectorization. In this case, the vertical sectorization is evaluated by adopting two fixed tilting angles $\phi_{\text{tilt}\,1}$ and $\phi_{\text{tilt}\,2}$ in the cell. In a similar manner as the static 3DBF case, the optimum tilting angles for the vertical sectorization can be found by the exhaustive search method. For a fair comparison, we assume that each beam power of the vertical sectorization is the half of that of the conventional case. Figure 9 shows the average rates of the conventional sectorization with $\phi^*_{\text{tilt}} = 11^\circ$ and the vertical sectorization with $\phi^*_{\text{tilt}\,1} = 25^\circ, \phi^*_{\text{tilt}\,2} = 11^\circ$, respectively. It is observed that the vertical sectorization leads to an average rate performance gain of 124% compared to the conventional sectorization at the cell-edge SNR of 0 dB. In summary, throughout the simulations, we can see that the 3DBF is able to reduce intercell interference and improve the average sum rate, and thus it brings out a significant performance gain over 2DBF. This makes the 3DBF as a promising technology for 5G systems.

CONCLUSIONS

In this article, we have investigated the 3DBF as a candidate technique that enables 5G wireless systems. In the 3DBF, the radiation beam pattern is adapted in both an elevation and horizontal plane that provides more degrees of freedom in system designs. Compared to the 2DBF, the 3DBF has many advantages including higher user capacity and less intercell and intersector interference. By employing the 3-D antenna pattern modeling, we have evaluated Monte Carlo simulation and then provided the performance comparison between the 3DBF and the 2BDF. Through numerical simulations, it is shown that the 3DBF outperforms the conventional 2DBF, and thus it is expected that 3DBF will play a crucial role in 5G system designs.



[FIG8] The average rate performance comparison as a function of vertical 3-dB beamwidth.



[FIG9] The average rate performance comparison as a function of cell-edge SNR.

AUTHORS

S. Mohammad Razavizadeh (smrazavi@iust.ac.ir) received his B.Sc., M.Sc., and Ph.D. degrees from the Iran University of Science and Technology (IUST), Tehran, Iran, in 1997, 2000, and 2006, respectively, all in electrical engineering. From June 2004 to April 2005, he was a visiting scholar with Coding and Signal Transmission Laboratory, University of Waterloo, Ontario, Canada. From 2005 to 2011, he was with Iran Telecommunication Research Center, Tehran, as a research assistant professor. Since 2011, he has been with the Department of Electrical Engineering at IUST, where he is currently an assistant professor. During the summer of 2013, he was with the Wireless Communications Laboratory at Korea University, Seoul, South Korea, as a visiting professor. His research interests are in the area of signal processing for wireless communication systems and cellular networks. He is a Senior Member of the IEEE.

Minki Ahn (amk200@korea.ac.kr) received the B.S. and M.S. degrees in electrical engineering from Korea University, Seoul, South Korea, in 2010 and 2012, where he is currently working toward the Ph.D. degree in the School of Electrical Engineering. During the winter of 2011, he was a visiting student at the Queen's University, Kingston, Ontario, Canada. His research interests include information theory and signal processing for wireless communication systems, such as the relay network, three-dimensional beamforming, and the multicell network.

Inkyu Lee (inkyu@korea.ac.kr) received the B.S. degree in control and instrumentation engineering from Seoul National University, South Korea, in 1990, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, California, in 1992 and 1995, respectively. He is currently a professor in the School of Electrical Engineering, Korea University, Seoul, South Korea. He has published over 110 journal papers with the IEEE, and has 30 U.S. patents granted or pending. His research interests include digital communications and signal processing techniques applied for next-generation wireless systems. He has served as an associate editor of *IEEE Transactions on Communications* and *IEEE Transactions on Wireless Communications*. He was a chief guest editor of *IEEE Journal on Selected Areas in Communications* (Special Issue on 4G Wireless Systems) in 2006. He received numerous awards including the Best Young Engineer Award of the National Academy of Engineering of Korea in 2013.

REFERENCES

[1] 3GPP, "Report of 3GPP RAN Workshop on Release 12 and onwards," RWS-120052, 3GPP Workshop on Release 12 Onward, June 2012.

[2] 3GPP, "Requirements, Candidate Solutions and Technology Roadmap for LTE Rel-12 Onward," RWS-120010, 3GPP Workshop on Release 12 Onward, June 2012.

[3] D. Gesbert, M. Kountouris, R. W. Heath, Jr., C. B. Chae, and T. Salzer, "From single user to multiuser communications: Shifting the MIMO paradigm," *IEEE Signal Processing Mag.*, vol. 24, no. 5, pp. 36–46, Oct. 2007.

[4] H. Sung, S.-R. Lee, and I. Lee, "Generalized channel inversion methods for multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 57, no. 5, pp. 3489–3499, Nov. 2009.

[5] H. Lee, B. Lee, and I. Lee, "Iterative detection and decoding with an improved V-BLAST for MIMO-OFDM systems," *IEEE J. Select. Areas Commun.*, vol. 24, pp. 504–513, Mar. 2006.

[6] 3GPP, "Text proposal on mechanical and electrical antenna tilting," Tech. Rep. R1-135708, Fraunhofer IIS, 3GPP RAN WG1 Meeting no. 75, Nov. 2011.

[7] W. Lee, S. R. Lee, H. B. Kong, and I. Lee, "3D beamforming designs for single user MISO systems," in *Proc. IEEE Global Communications Conf.*, Dec. 2013, pp. 3914–3919.

[8] H. Huang, O. Alrabadi, J. Daly, D. Samardzija, C. Tran, R. Valenzuela, and S. Walker, "Increasing throughput in cellular networks with higher-order sectorization," in *Proc. Asilomar Conf. Signals, Systems and Computers*, Nov. 2010, pp. 630–635.

[9] C. S. Lee, M. C. Lee, C. J. Huang, and T. S. Lee, "Sectorization with beam pattern design using 3D beamforming techniques," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conf.*, Oct. 2013, pp. 1–5.

[10] P. Kang, Q. Cui, S. Chen, and Y. Liu, "Performance evaluation on coexistence of LTE with active antenna array systems," in *Proc. IEEE Int. Symp. Personal Indoor and Mobile Radio Communications*, Sept. 2012, pp. 1066–1070.

[11] A. Adhikary, J. Nam, J. Ahn, and G. Caire, "Joint spatial division and multiplexing—The large-scale array regime," *IEEE Trans. Inform. Theory*, vol. 59, no. 10, pp. 6441–6463, Oct. 2013.

[12] T. L. Marzetta, "How much training is required for multiuser MIMO?," in *Proc. Asilomar Conf. Signals, Systems and Computers*, Oct./Nov. 2006, pp. 359–363.

[13] H. E. Lee, "Synthetic array processing for underwater mapping applications," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Aug. 1978, pp. 148–151.

[14] M. Karaman, I. O. Wygant, O. Oralkan, and B. T. Khuri-Yakub, "Minimally redundant 2-D array designs for 3-D medical ultrasound imaging," *IEEE Trans. Med. Imaging*, vol. 28, no. 7, pp. 1051–1061, July 2009.

[15] A. C. Dhanantwari, S. Stergiopoulos, L. Song, C. Parodi, F. Bertora, P. Pellegretti, and A. Questa, "An efficient 3D beamformer implementation for real-time 4D ultrasound systems deploying planar array probes," in *Proc. IEEE Ultrasonics Symp.*, Aug. 2004, vol. 2, pp. 1421–1424.

[16] J. M. Valin, F. Michaud, and J. Rouat, "Robust 3D localization and tracking of sound sources using beamforming and particle filtering," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, May 2006, pp. 841–844.

[17] J. Meyer and G. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, May 2002, vol. 2, pp. 1781–1784.

[18] H. Kanj, P. Lusina, S. M. Ali, and F. Kohandani, "A 3D-to-2D transform algorithm for incorporating 3D antenna radiation patterns in 2D channel simulators," *IEEE Antenna Propagat. Wireless Lett.*, vol. 8, pp. 815–818, June 2009.

[19] 3GPP, "Spatial channel model for multiple input multiple output (MIMO) simulations (Release 9)," 3GPP, Tech. Rep. TR 25.996 V9.0.0, Dec. 2009.

[20] ITU-R Report M.2135-1, "Guidelines for evaluation of radio interface technologies for IMT-advanced," 2009.

[21] R. H. Clarke, "A statistical theory of mobile radio reception," *Bell Syst. Tech. J.*, vol. 47, pp. 957–1000, July/Aug. 1968.

[22] A. Kuchar, J.-P. Rossi, and E. Bonek, "Directional macro-cell channel characterization from urban measurements," *IEEE Trans. Antennas Propagat.*, vol. 48, no. 2, pp. 137–146, Feb. 2000.

[23] 3GPP, "Study on 3D channel model for LTE (Release 12)," Tech. Rep. R1-134980, Sept. 2013.

[24] T. Aulin, "A modified model for the fading signal at a mobile radio channel," *IEEE Trans. Veh. Technol.*, vol. 28, no. 3, pp. 182–203, Aug. 1979.

[25] J. Meinila, P. Kyosti, L. Hentila, T. Jamsa, E. Suikkanen, E. Kunnari, and M. Narandzié, "D5.3: WINNER+ final channel models," in *Wireless World Initiative New Radio—WINNER+*, P. Heino, Ed., CELTIC/CP5-026, 2010.

[26] M. Narandzic, C. Schneider, R. S. Thoma, T. Jamsa, P. Kyosti, and X. Zhao, "Comparison of SCM, SCME, and WINNER channel models," in *Proc. IEEE Vehicular Technology Conf.*, Apr. 2007, pp. 413–417.

[27] Y. H. Nam, B. L. Ng, K. Sayana, Y. Li, J. Zhang, Y. Kim, and J. Lee, "Full-dimension MIMO (FD-MIMO) for next-generation cellular technology," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 172–179, June 2013.

[28] E. Larsson, O. Edfors, F. Tufvesson, and T. Marzetta, "Massive MIMO for next-generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[29] Y. Li, K. C. Ho, and C, Kwan, "3-D array pattern synthesis with frequency invariant property for concentric ring array," *IEEE Trans. Signal Processing*, vol. 54, no. 2, pp. 780–784, Feb. 2006.

[30] F. Zotter and B. Bank, "Geometric error estimation and compensation in compact spherical loudspeaker array calibration," in *Proc. IEEE Int. Instrumentation and Measurement Technology Conf.*, pp. 2710–2715, May 2012.

[SP]

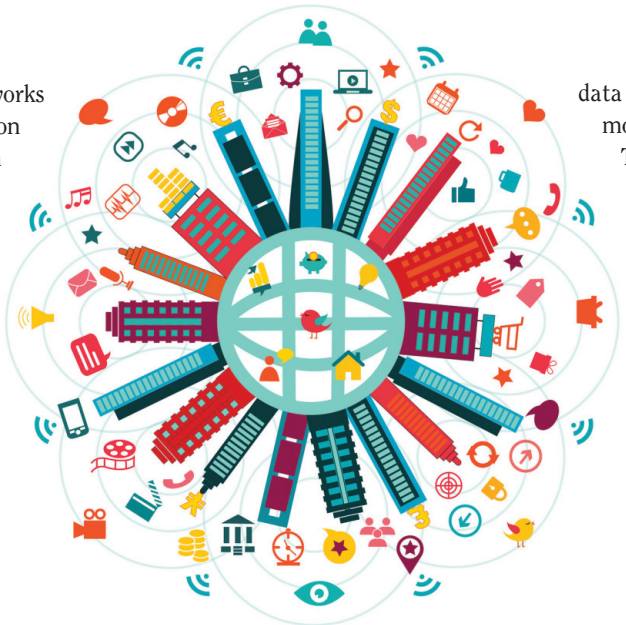
IEEE SignalProcessing Previous Page | Contents | Zoom in | Zoom out | Front Cover | Search Issue | Next Page Qmags THE WORLD'S NEWSSTAND®

Rocco Di Taranto, Srikar Muppirisetty, Ronald Raulefs, Dirk T.M. Slock, Tommy Svensson, and Henk Wymeersch

Location-Aware Communications for 5G Networks

How location information can improve scalability, latency, and robustness of 5G

Fifth-generation (5G) networks will be the first generation to benefit from location information that is sufficiently precise to be leveraged in wireless network design and optimization. We argue that location information can aid in addressing several of the key challenges in 5G, complementary to existing and planned technological developments. These challenges include an increase in traffic and number of devices, robustness for mission-critical services, and a reduction in total energy consumption and latency. This article gives a broad overview of the growing research area of location-aware communications across different layers of the protocol stack. We highlight several promising trends, tradeoffs, and pitfalls.



THE 5G REVOLUTION

INTRODUCTION AND CHALLENGES

Fifth-generation will be characterized by a wide variety of use cases, as well as orders-of-magnitude increases in mobile data volume per area, number of connected devices, and typical user data rate, all compared to current mobile communication systems [1]. To cope with these demands, a number of challenges must be addressed before 5G can be successfully deployed. These include the demand for extremely high data rates and much lower latencies, potentially down to 1 ms end-to-end for certain applications [2]. Moreover, scalability and reduction of signaling overhead must be accounted for, as well as minimization of (total) energy consumption to enable affordable cost for network operation. To fulfill these requirements in 5G, network densification is key, calling for a variety of coordination and cooperation techniques between various kinds of network elements in an ultradense heterogeneous network. Moreover, by implementing sharing and coexistence approaches, along with new multi-GHz frequency bands, spectrum efficiency can be improved. An overview of a number of disruptive technologies for 5G is provided in [1].

It is our vision that context information in general and location information in particular can complement both traditional and disruptive technologies in addressing several of the challenges in 5G networks. While location information was

Digital Object Identifier 10.1109/MSP.2014.2332611
Date of publication: 15 October 2014

1053-5888/14©2014IEEE

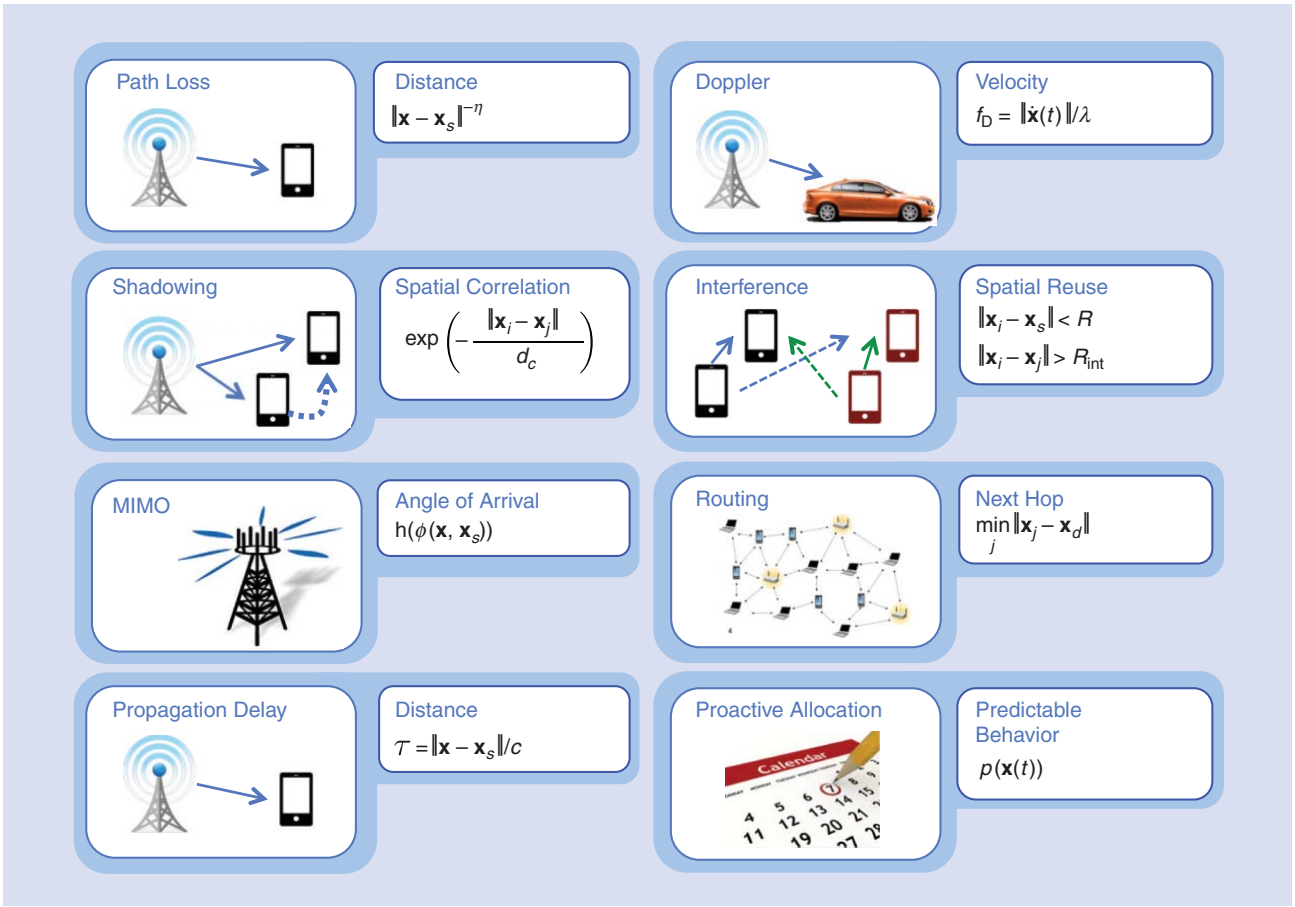
available in previous generations of cellular mobile radio systems, e.g., cell-identifier (ID) positioning in second generation (2G), timing-based positioning using communication-relevant synchronization signals in third generation (3G), and additionally dedicated positioning reference signals in fourth generation (4G), accuracy ranged from hundreds to tens of meters, rendering position information insufficiently precise for some communications operations. In 5G, for the first time, a majority of devices can benefit from positioning technologies that achieve a location accuracy on the order of 1 m.

FIFTH-GENERATION NETWORKS WILL BE THE FIRST GENERATION TO BENEFIT FROM LOCATION INFORMATION THAT IS SUFFICIENTLY PRECISE TO BE LEVERAGED IN WIRELESS NETWORK DESIGN AND OPTIMIZATION.

In this article, we argue why and how such precise location awareness can be harnessed in 5G networks. We first present technologies providing seamless and ubiquitous location awareness for 5G devices, identify associated signal processing challenges, and describe at a high level how location information can be utilized across the protocol stack. We then zoom in on each layer of the protocol stack and provide an overview of recent and relevant research on location-aware communications. We conclude the article by identifying a number of issues and research challenges that must be addressed before 5G technologies can successfully utilize location information and achieve the predicted performance gains.

LOCATION AWARENESS IN 5G NETWORKS

A majority of 5G devices will be able to rely on ubiquitous location awareness, supported through several technological developments: a multitude of global navigation satellite systems (GNSS) are being rolled out, complementing the current global positioning system (GPS). Combined with ground support systems and multiband operation, these systems aim to offer



[FIG1] Communication systems are tied to location information in many ways, including through distances, delays, velocities, angles, and predictable user behavior. The notations are as follows (starting from the top left downward): $\mathbf{x}$ is the user location, $\mathbf{x}_s$ is the base station or sender location, and η is the path loss exponent; $\mathbf{x}_i$ and $\mathbf{x}_j$ are the two-user location and d_c is a correlation distance; $\phi(.)$ is an angle of arrival between a user and a base station and $\mathbf{h}$ is a multiple-input, multiple-output (MIMO) channel; c is the speed of light and τ a propagation delay; f_D is a Doppler shift, $\dot{\mathbf{x}}(t)$ is the user velocity, and λ is the carrier wavelength; R is a communicate range and R_{int} is an interference range; $\mathbf{x}_d$ is a destination; and $p(\mathbf{x}(t))$ is a distribution of a user position at a future time t.

location accuracies around 1 m in open sky [3]. In scenarios where GNSS is weak or unavailable (in urban canyons or indoors), other local radio-based technologies such as ultrawideband (UWB), Bluetooth, ZigBee, and radio frequency identification (RFID), will complement current Wi-Fi-based positioning. Together, they will also result in submeter accuracy.

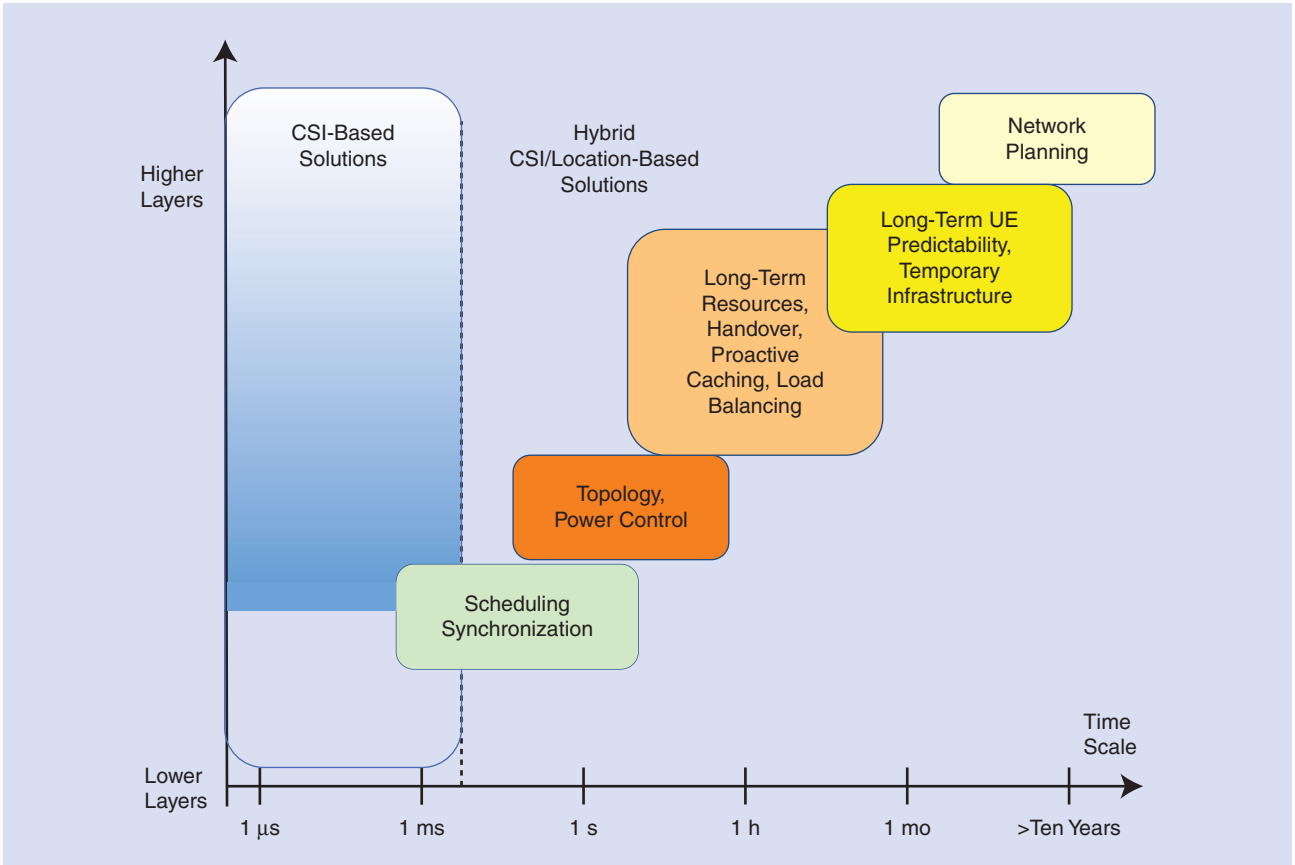
RECENT STUDIES HAVE REVEALED THAT LOCATION INFORMATION CAN BE HARNESSED NOT ONLY BY COGNITIVE NETWORKS, BUT ALSO CELLULAR AND AD HOC CONFIGURATIONS.

Accurate location information can be utilized by 5G networks across all layers of the communication protocol stack [4]. This is due to a number of reasons (see Figure 1), which will be detailed in later sections. First of all, signal-to-noise ratio (SNR) reduces with distance due to path loss, so that location knowledge and thus distance knowledge can serve as an indication of received power and interference level. Thus, if shadowing is neglected, the optimal multihop path between a source-destination pair in a dense network is the one that is shortest in terms of distance. Second, while path loss is the dominant effect in wireless communications, shadowing creates significant localized power differences due to signal propagation through objects. Since shadowing often exhibits decorrelation distances larger than the positioning uncertainty, local channel information can be extrapolated to nearby terminals. Third, most 5G user terminals will largely be predictable in their mobility patterns, since they will be either associated with people or fixed or controllable entities. Finally, at the highest layers, location information is often crucial, not only for location-based services, but also for a variety of tasks in cyberphysical systems, such as robotics and intelligent transportation systems.

Location awareness can be harnessed in a variety of ways to address several of the challenges in 5G networks. In particular, location-aware resource allocation techniques can reduce overheads and delays due to their ability to predict channel quality beyond traditional time scales. In Figure 2, we provide a top-level view of how location information may be utilized, inspired by research activities in 3G and 4G communication networks, while details will be provided in the subsequent sections.

LOCATION AWARENESS ACROSS THE PROTOCOL STACK

Location awareness has received intense interest from the research community, in particular with respect to cognitive radio [5], where location databases are being used to exploit TV



[FIG2] At very short time scales, resource allocation (especially in the lower layers) must rely on instantaneous channel-state information (CSI). At longer time scales, position information can be harnessed to complement CSI.

white spaces. However, recent studies have revealed that location information can be harnessed not only by cognitive networks, but also cellular and ad hoc configurations [6].

In this section, we aim to group a number of representative works in this developing area, based on the layer of the protocol stack to which they pertain. Since many of the works below are inherently cross-layer, sometimes we had to make choices among two layers. We start by a description of how channel quality metrics can be predicted through a suitable database and inference engine.

THE CHANNEL DATABASE

To predict the channel quality in locations where no previous channel quality measurement was available, a flexible predictive engine is needed. As different radio propagation environments have different statistical model parameters, this engine should be able to learn and adapt. Regression techniques from machine learning can be used for this purpose. Among these techniques, we focus on Gaussian processes (GPs) [7]. GPs have been used to predict location-dependent channel qualities in [8] and [9] in the following manner: users send a channel quality metric (CQM) to the database, along with the time and location at which it was acquired. After a training stage, the GP can provide an estimate of the CQM along with the uncertainty for any other receiver location. Hence, the output of the GP can be considered as a prior distribution on the channel quality. The construction and utilization of such a GP database is shown in Figure 3. The CQM can take on a variety of forms (see also Figure 1), including received power, root mean square delay spread, interference levels, or angular spread and rank profile for multiantenna systems [6], [4]. For the sake of simplicity, we will consider received power and disregard any temporal correlation of the CQM.

To model the received power CQM, we recall that a radio signal is affected mainly by three major components of the wireless propagation channel: distance dependent path-loss, shadowing due to obstacles in the propagation medium, and small-scale fading due to multipath effects. Small scale-fading decorrelates over very short distances for target operational frequencies. Hence, even with highly accurate position information, predictions of small-scale fading in new locations are not possible. This implies that we can only provide coarse channel information, which in many cases must be complemented with instantaneous small-scale information (see Figure 2). We let $P_{\mathrm{RX}}(\mathbf{x}_s, \mathbf{x}_i)$ be the power at a receiver node (located at $\mathbf{x}_i \in \mathbb{R}^2$), averaged over the small-scale fading in either time or frequency, from a source node (located at $\mathbf{x}_s \in \mathbb{R}^2$), which can be expressed in a dB scale as

$$P_{\mathrm{RX}}(\mathbf{x}_s, \mathbf{x}_i) = L_0 - 10\eta \log_{10}(\|\mathbf{x}_s - \mathbf{x}_i\|) + \Psi(\mathbf{x}_s, \mathbf{x}_i), \quad (1)$$

where η is the path-loss exponent, $\Psi(\mathbf{x}_s, \mathbf{x}_i)$ is the location-dependent shadow fading between the source and the receiver (expressed in dB), and L_0 is a constant that captures antenna and other propagation gains. Although L_0 is assumed to be common to all users, additional user-specific biases, such as different antenna types or transmit powers can be calculated by the user and sent back to the base station. A common choice for shadow fading is to assume a log-normal distribution, i.e., $\Psi(\mathbf{x}_s, \mathbf{x}_i) \sim \mathcal{N}(0, \sigma_\Psi^2)$, where σ_Ψ^2 is the shadowing variance. While the location dependence on path loss is clear from (1), the shadowing also has well-established spatial correlation models, such as [10] for cellular networks, wherein the spatial autocovariance function of shadowing is given by

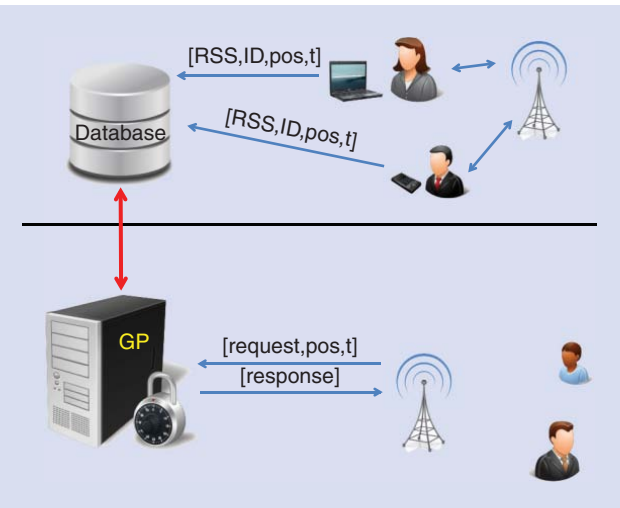
$$1C(\mathbf{x}_i, \mathbf{x}_j) = \mathbb{E}\{\Psi(\mathbf{x}_s, \mathbf{x}_i)\Psi(\mathbf{x}_s, \mathbf{x}_j)\} = \sigma_\psi^2 \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|}{d_c}\right), \quad (2)$$

where d_c denotes the correlation distance.

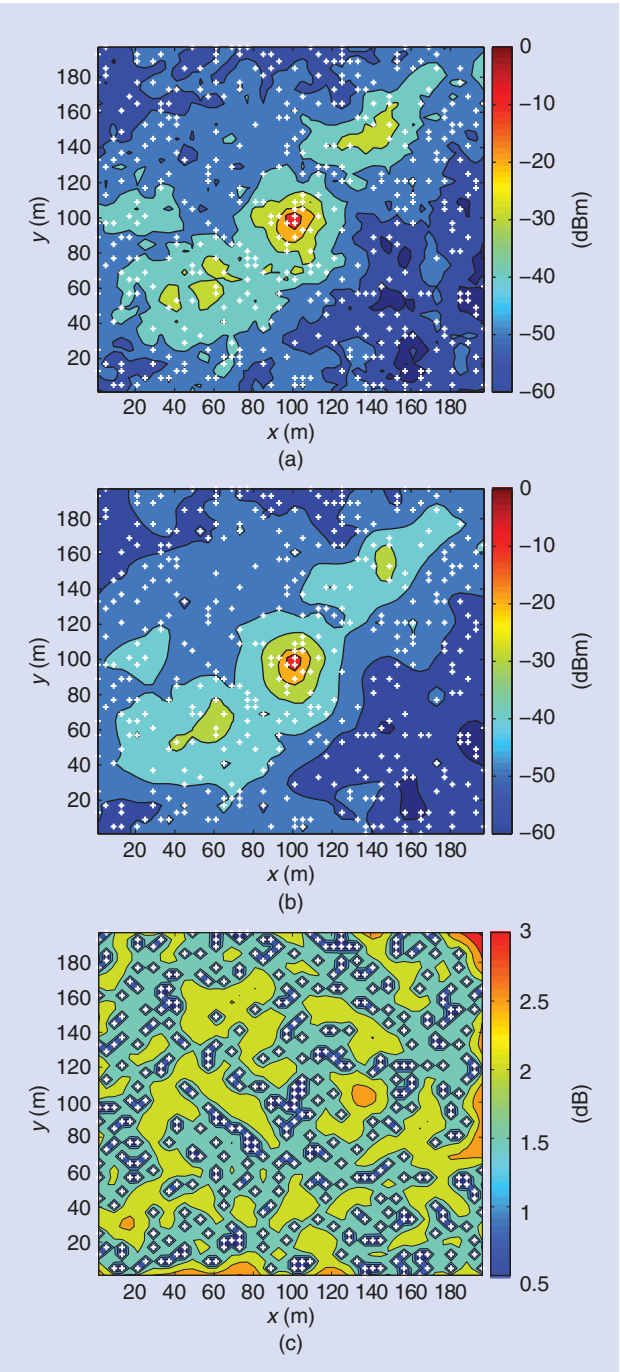
In the case of a common transmitter (e.g., a base station with location $\mathbf{x}_s$), the GP framework operates as follows. The power $P_{\mathrm{RX}}(\mathbf{x}_s, \mathbf{x}_i)$ is considered to be a GP as a function of $\mathbf{x}_i$, with mean function $\mu(\mathbf{x}_i)$ and covariance function $C(\mathbf{x}_i, \mathbf{x}_j)$. If we choose the mean function to be $\mu(\mathbf{x}_i) = L_0 - \eta 10\log_{10}(||\mathbf{x}_s - \mathbf{x}_i||)$, then the covariance function is exactly as defined in (2). To train the GP, let $y_i = P_{\mathrm{RX}}(\mathbf{x}_s, \mathbf{x}_i) + n_i$ be the noisy (scalar) observation of the received power at node i, where n_i is a zero mean additive white Gaussian noise random variable with variance σ_n^2. We introduce $\mathbf{X} = [\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_N^T]^T$, $\mathbf{y} = [y_1, y_2, \ldots, y_N]^T$, and $\mathcal{D} = \{\mathbf{X}, \mathbf{y}\}$. The joint distribution of the N training observations now exhibits a Gaussian distribution [7]

$$\mathbf{y} \,|\, \mathbf{X}; \theta \sim \mathcal{N}(\mu(\mathbf{X}), \mathbf{K}), \quad (3)$$

where $\mu(\mathbf{X}) = [\mu(\mathbf{x}_1), \mu(\mathbf{x}_2), \ldots, \mu(\mathbf{x}_N)]^T$ is the mean vector and $\mathbf{K}$ is the covariance matrix with entries $[\mathbf{K}]_{ij} = C(\mathbf{x}_i, \mathbf{x}_j) + \sigma_n^2 \delta_{ij}$, where $\delta_{ij} = 1$ for $i = j$ and zero otherwise. The Gaussian distribution (3) depends on a number of parameters $\theta = [\sigma_n^2, d_c, L_0, \eta, \sigma_\Psi^2]$, which can be learned using the training



[FIG3] Users upload their location (pos) and time-tagged (t) channel quality metrics [e.g., the received signal strength, (RSS)], possibly along with their user ID, to a channel database. The information can be extrapolated to future users, requesting a channel quality metric in other locations for the same base station, using techniques such as GPs.



[FIG4] **Radio channel prediction in decibel scale, with hyperparameters $\theta = [\sigma_n^2 = 0.01, d_c = 70\ m, L_0 = 10$ dB, $\eta = 3, \sigma\psi = 9$ dB], $N = 400$ measurements (+ signs). The channel prediction is performed at a resolution of 4 m. (a) shows the true channel field, (b) the mean [obtained from (4)] of the predicted channel field; and (c) the standard deviation [obtained from the square root of (5)] of the predicted channel field.**

database $\mathcal{D}$ by minimizing negative log-likelihood $-\log(p(\mathbf{y}\,|\,\mathbf{X};\theta))$ with respect to θ. This completes the training process. The predictive distribution of the noise-free signal power $P_{RX}(\mathbf{x}_s, \mathbf{x}_*)$ at a new node location $\mathbf{x}_*$, given the training database $\mathcal{D}$, is a Gaussian distribution with mean $\overline{P}_{RX}(\mathbf{x}_s, \mathbf{x}_*)$ and variance $\Sigma_{RX}(\mathbf{x}_s, \mathbf{x}_*)$, given by [7]

$$\overline{P}_{RX}(\mathbf{x}_s, \mathbf{x}_*) = \mu(\mathbf{x}_*) + \mathbf{k}_*^T\mathbf{K}^{-1}(\mathbf{y} - \mu(\mathbf{X})) \quad (4)$$

$$\Sigma_{RX}(\mathbf{x}_s, \mathbf{x}_*) = C(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^T\mathbf{K}^{-1}\mathbf{k}_*, \quad (5)$$

in which $\mathbf{k}_*$ is the $N \times 1$ vector of cross-covariances $C(\mathbf{x}_*, \mathbf{x}_i)$ between $\mathbf{x}_*$. and the training inputs $\mathbf{x}_i$.

Figure 4 demonstrates an example of radio channel prediction using a GP. A base station is placed in the center and a two-dimensional radio propagation field is simulated through a computer model according to (3) with sampling points on a square grid of 200 m × 200 m and a resolution of 4 m. Based on measurements at marked locations, the mean and standard deviation of the prediction are obtained for any location. Observe the increased uncertainty in Figure 4(c) in regions where few measurements are available.

In the case where links rely on different transmitters, the model above can still be applied [8], though more advanced models exist. For instance, in the case of ad hoc networks, [9] proposes a model where shadow fading is due to an underlying spatial loss field.

GPs can thus provide a statistical description of the CQM in any location and any time. This description can be used in resource allocation at different layers, e.g., to reduce delays and/or overheads. In the following, we present specific examples that are useful mainly in one layer. We will start with the physical layer.

THE PHYSICAL LAYER

In the lowest layer of the protocol stack, location information can be harnessed to reduce interference and signaling overhead, to avoid penalties due to feedback delays, or to synchronize coordinated communication schemes.

The best known application is spatial spectrum sensing for cognitive radio [11], where a GP allows the estimation of power emitted from primary users at any location through collaboration among secondary users. The resulting power density maps enable the secondary users to choose the frequency bands that are not crowded and to adapt their transmit power to minimize the interference to the primary users. These techniques can be adapted in 5G to perform interference coordination. For instance, significant potential for the exploitation of location information in multiantenna techniques arises in spatial cognitive radio paradigms (underlay, overlay, interweave) [6]. Such location-aided techniques could be compatible with some very recent developments in massive MIMO, where the exploitation of slow fading subspaces in the multiantenna propagation has been advocated.

The GP database also provides useful information in any application that relies on a priori channel information, such as slow adaptive modulation and coding or channel estimation. This is investigated in [12], where location-aware adaptive mobile communication uses both channel and spatial movement coherence in combination with location prediction and a fingerprint database. When at time t a user reports future predicted locations $\mathbf{x}(t), \mathbf{x}(t+1), \ldots, \mathbf{x}(t+T)$ to the database, the

corresponding received powers can be determined $\overline{P}_{\mathrm{RX}}(t)$, $\overline{P}_{\mathrm{RX}}(t+1), \ldots, \overline{P}_{\mathrm{RX}}(t+T)$. For each time, the predicted capacity is then

$$C(t) = W \log_2\left(1 + \frac{\overline{P}_{\mathrm{RX}}(t)}{N_0 W}\right), \tag{6}$$

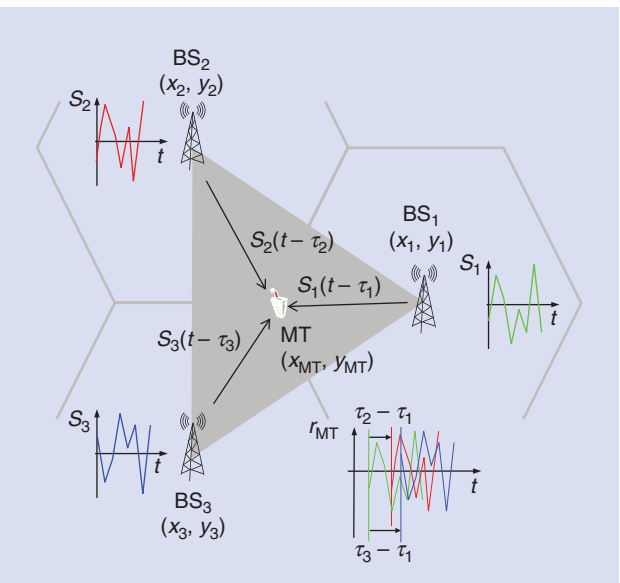
where W is the signaling bandwidth and N_0 is the noise power spectral density. The communication rate is then adapted to not exceed the predicted capacity. It is demonstrated that location-aware adaptive systems achieve large capacity gains compared to state-of-the-art adaptive modulation schemes for medium to large feedback delays. Such delays are especially important in 5G application with fast-moving devices, such as transportation systems, which are also the topic of [13], where the short channel coherence time precludes adaptation based on the fast fading channel. Instead, link adaptation based on path loss is considered, which in turn depends on the locations of the vehicles. Expressions for large-scale coherence time and velocity are derived, and it is found that feeding back location information can substantially reduce feedback overhead without compromising data rate. Location-based feedback latency reduction is also discussed in [6], e.g., for fast relay selection.

Significant opportunities for location-aware communications concern resource allocation aspects, especially for multiuser (MU) and MIMO systems. In such systems, recent information theory progress shows that an optimized handling may lead to significant system capacity increases, though only in the presence of very precise channel-state information at the transmitter (CSIT). In the single-cell case (or at the cell center), one can consider location-aware downlink MU-MIMO. Multiple antennas at the user side do not allow a base station with M antennas serving M users to send more streams in a cell, but a user can use its N antennas to suppress the effect of $N-1$ multipath components. Hence, if the overall propagation scenario involves a line-of-sight (LoS) path and up to $N-1$ multipath components, the user can use receive beamforming (BF) to transform its channel into a pure LoS channel, allowing the base station to perform zero-forcing (ZF) transmission with only location information [14]. In the multicell case, which in information-theoretic terms corresponds to the interference channel (IC) and in practice to the macrocellular environment or to HetNets (coexistence of macro and femto/small cells), there are opportunities for location-aided MIMO interference channels [14]. In particular the feasibility of joint transmitter/receiver (Tx/Rx) ZF BF is of interest in the case of reduced rank MIMO channels (with LoS being the extreme case of rank one). Whereas in the full rank MIMO case, the joint Tx/Rx design is complicated by overall coupling between all Tx and all Rx, i.e., a requirement of overall system CSI at all base stations, some simplifications may occur in the reduced rank case. In particular, for the LoS case (the easiest location-aided scenario, higher rank cases requiring databases), the Tx/Rx design gets decoupled, leading to only local (e.g., location-based) CSI requirements [14].

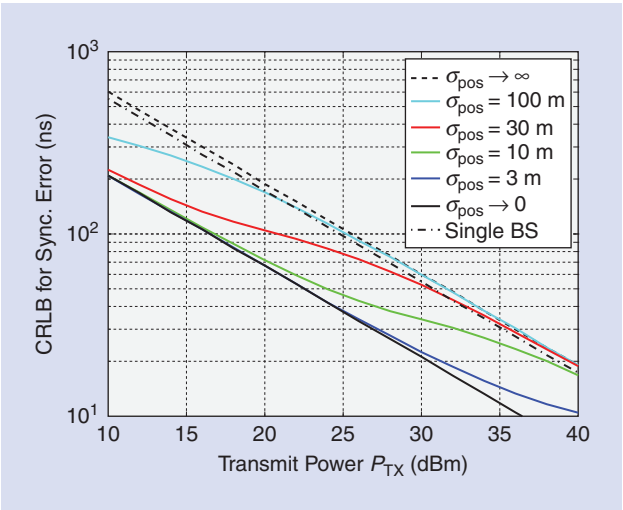
Locations can also be utilized in a different manner, by converting them not to a CQM, but to other physical quantities, such as Doppler shifts (proportional to the user's relative velocity), arrival angles (used in [4] for location-based spatial division multiple access), or timing delays (which are related to the distance between transmitter and receiver). This latter idea is taken up in [6] and applied for coordinated multipoint (CoMP) transmission, illustrated in Figure 5, showing a mobile node receiving synchronization signals from three base stations, deployed with a frequency reuse of 1. CoMP transforms interference experienced by the mobile users to signal power, especially at the cell edge, by coordinating the signals of all involved base stations.

CoMP relies on accurately synchronized signals, a process that can be aided through a priori location information, which determines the potential window to exploit the synchronization signals from different base stations. Figure 6 shows the potential gains in terms of required transmit power at the base station to achieve a certain synchronization performance for different values of the location uncertainty of the mobile node. The communication system benefits if the synchronization requirements are at least in the range of the location accuracy (1-ns timing uncertainty corresponds to 30-cm position uncertainty). For example, comparing a system with multiple base stations for a desired synchronization accuracy of 20 ns, 40 dBm is required when no position information is available, while less than 32 dBm is required when positioning accuracy is around 3 m.

As the aforementioned works indicate, location information provides valuable side-information about the physical layer. It can be harnessed to reduce delays and feedback overhead, and even to improve performance. Determining when to utilize location-based CQM and when to rely on instantaneous CSI is an important topic in the optimization of 5G communications. Next, we move up to the medium access control (MAC) layer, where even richer opportunities for the use of location information arise than



[FIG5] By knowing the location of the mobile node (x_{MT}, y_{MT}), the propagation delays τ_i of the signal components $s_i(t)$ coming from different base stations [with known locations (x_i, y_i)] can be related to each other to improve the CoMP transmission.



[FIG6] The Cramér–Rao lower bound (CRLB) for the synchronization error (standard deviation) versus the base station transmit power P_{TX} for different positioning uncertainties (σ_{pos}) of the mobile node. The Third-Generation Partnership Project (3GPP) long-term evolution secondary synchronization signals were used with IDs 142, 411, and 472. (Figure based on [6].)

at the physical layer, especially without the need to estimate channel gains based on position and/or distance information.

THE MAC LAYER

With more devices communicating with each other, scalability, efficiency, and latency are important challenges in designing efficient protocols for MAC. In this section, we provide an overview of some of the existing works on the use of location information at the MAC layer to address these design challenges. In particular, multicasting, scheduling, and selection protocols are considered. Again, we can make a distinction between approaches that tie locations to channel and approaches where locations are exploited in a different way.

In the first group, we find works such as [6] and [15]–[17]. The basic premise is that a link between transmitter with position $\mathbf{x}_s$ and receiver with position $\mathbf{x}_i$ can be scheduled with the same resource as an interfering transmitter with position $\mathbf{x}_j$, provided that

$$\frac{P_{RX}(\mathbf{x}_s, \mathbf{x}_i)}{N_0 W + P_{RX}(\mathbf{x}_j, \mathbf{x}_i)} > \gamma, \tag{7}$$

where γ is a signal-to-interference-plus-noise ratio (SINR) threshold. SINR expressions such as (7) can easily be combined with a CQM database. In [6], a location-aided round-robin scheduling algorithm for fractional frequency reuse is proposed, where allowing temporary sharing of resources between cell-center and cell-edge users is shown to achieve higher total throughput with less and less frequent feedback than the conventional method. In the same paper, location-based long-term power setting in heterogeneous cochannel deployments of macro and femto base stations is investigated. In [15], location-based multicasting is considered, assuming a disk model, and is shown to both reduce the number of contention phases and increase the reliability of packet delivery, especially in dense networks. Time division with spatial reuse is considered in [17], which investigates location-aware joint scheduling and power control for IEEE 802.15.3, leading to lower latencies and higher throughput compared to a traditional round-robin type scheduling mechanism. Location information is also beneficial in reducing the overhead associated with node selection mechanisms (e.g., users, relays), by allowing base stations to make decisions based solely on the users' positions [6]. Finally, location information is a crucial ingredient in predicting interference levels in small/macrocell coexistence, in multicell scenarios, and in all cognitive radio primary/secondary systems. For example, [6] and [14] demonstrate the use of location information to allow to significantly improve intercell interference coordination techniques. Location-based modeling of attenuation and slow fading components will bring about progress in the design of multicellular systems, complementing the recent significant progress that has focused almost exclusively on the fast fading component (e.g., interference alignment). For underlay cognitive radio systems, location-based prediction of interference caused to primary users may be a real enabling approach. These works indicate that significant gains in terms of throughput and latency can be reaped from location-aware MAC in 5G networks, provided appropriate channel models are used.

In the second group, we find approaches that utilize location information in a different way [16], [18], [19]. All turn out to relate to vehicular networks. In [16], a family of highly efficient location-based MAC protocols is proposed, whereby vehicles broadcast information to other vehicles only when they pass through predetermined transmission areas. When the traffic flow rate increases, the proposed location-based protocols have a smaller message delivery time compared with conventional random access schemes. A similar idea is proposed in [18], where a decentralized location-based channel access protocol for intervehicle communication is studied. Channels are allocated based on vehicles' instantaneous geographic location, and unique channels are associated to geographic cells. Using a prestored cell-to-channel mapping, vehicles know when to transmit on which channel, alleviating the need for a centralized coordinator for channel allocation. This leads to efficient bandwidth use and avoids hidden node problems, since neighboring cells do not use the same channel. In addition, communication delay is bounded and fairness among the vehicles is maintained as each vehicle gets a channel regularly to transmit. Finally, [19] introduces the concept of geocasting, whereby multicast regions are formed based on the geographical location of the nodes and packets are sent to all the nodes in the group. Specialized location-based multicasting schemes are proposed to decrease the delivery overhead of packets when compared to multicast flooding mechanisms.

We observe that in the MAC layer there is a more varied use of location information than in the physical layer, especially without direct need of the channel database. In all cases, improvements in terms of latency, overhead, or throughput were reported. The

emphasis on dense mobile networks, and the limited need for centralized infrastructure make these techniques promising for 5G networks. We now move up to the network and transport layers, where geographic routing plays an important role.

WITH MORE DEVICES COMMUNICATING WITH EACH OTHER, SCALABILITY, EFFICIENCY, AND LATENCY ARE IMPORTANT CHALLENGES IN DESIGNING EFFICIENT PROTOCOLS FOR MAC.

NETWORK AND TRANSPORT LAYERS

At the network and transport layers, location information has been shown to improve scalability and to reduce overhead and latency. A full-fledged location-based network architecture is proposed in [5] for cognitive wireless networks, dealing with dynamic spectrum management, network planning and expansion, and in handover. In particular, a location-aided handover mechanism significantly reduces the number of handovers compared with signal strength-based methods [20], which are subject to delay and hysteresis effects.

Location-aided techniques, especially using mobility information to forecast future channel capacities for the mobile, become particularly powerful when vertical temporary handovers are considered to systems with larger channel capacity to offload data. Such large capacity systems may exhibit short windows of opportunity due to their limited coverage.

Most other works at the network layer have focused on the routing problem. A well-known technique in this area is geographic routing (georouting), which takes advantage of geographic information of nodes (actual geographic coordinates or virtual relative coordinates) to move data packets to gradually approach and eventually reach their intended destination. In its most basic form, given a destination d, a node i with neighbors $\mathcal{N}_i$ will choose to forward data to a neighbor closest to the destination:

$$j^* = \arg\min_{j\in\mathcal{N}_i} \|\mathbf{x}_j - \mathbf{x}_d\|. \tag{8}$$

Recently, georouting has gained considerable attention, as it promises a scalable, efficient, and low-latency solution for information delivery in wireless ad hoc networks. For a comprehensive survey of the existing literature on georouting, investigating how location information can benefit routing, we refer to [21].

Georouting is mainly limited due to two factors: it is sensitive to localization errors and it does not exploit CQM, favoring latency (measured in this context in terms of progress toward the destination) over throughput. The first issue is investigated in [22], where it is shown that georouting quickly degrades as location information becomes imprecise. More robust routing mechanisms are proposed, combining progress toward the destination with an error measure in the locations. The second issue is treated in [23] and [24]. In [23], where positions are mapped to a CQM, a centralized routing algorithm aims to maximize end-to-end flow. The mismatch between the estimated and true channels is mitigated using a distributed algorithm, whereby nodes locally adjust their rate, but not the routes. While [23] no longer directly optimizes progress toward the destination, [24] considers both throughput and latency in a fully distributed manner. In [24], the network consists of power-constrained nodes that transmit over wireless links with adaptive transmission rates. Packets randomly enter the system at each node and wait in output queues to be transmitted through the network to their destinations. The data flows from source to destination according to the enhanced dynamic routing and power control (EDRPC) algorithm, which is proven to stabilize the network with a bounded average delay. In EDRPC, each of the N nodes in the network maintains N queues, $Q_i^{(d)}$ denoting the queue at node i with stored information destined to node d (note that $Q_d^{(d)} = 0$ for all destinations). Each link, say (i, j), locally decides the destination to serve, such as

$$d_{ij}^* = \arg\max_{d\in\{1,\ldots,N\}} (\tilde{Q}_i^{(d)} - \tilde{Q}_j^{(d)}), \tag{9}$$

where $\tilde{Q}_i^{(d)} = Q_i^{(d)} + V_i^{(d)}$, in which $V_i^{(d)} \geq 0$ is a design parameter. When $V_i^{(d)} = 0$, $\forall i$, the destination with the largest backlog will be served over link (i, j). Setting the values $V_i^{(d)} = f(\|\mathbf{x}_i - \mathbf{x}_d\|)$, where $f(\cdot)$ is a monotonically increasing function will incentivize data to flow toward the geographic position of the destination (i.e., given equal backlogs, the destination will be chosen that maximizes $f(\|\mathbf{x}_i - \mathbf{x}_d\|) - f(\|\mathbf{x}_j - \mathbf{x}_d\|)$, favoring small $\|\mathbf{x}_j - \mathbf{x}_d\|$). Following the choice of d_{ij}^*, EDRPC performs a (centralized) power allocation for each link, leading to an allowable rate per link. Finally, each node i will serve destination d_{ij}^* over link (i, j) with an amount of data at the allowable rate and thus reduces its queue length $Q_i^{(d_i^*)}$.

The focus in [22]–[24] is on relatively static networks, where there are no drastic topology changes. In certain applications, such as vehicular networks, this assumption is no longer valid, as is treated in [25] and [26]. In [25], the use of mobility prediction to anticipate topology changes and perform rerouting prior to route breaks is considered. The mobility prediction mechanism is applied to some of the most popular representatives of the wireless ad hoc routing family, mainly an on-demand unicast routing protocol, a distance vector routing protocol, and a multicast routing protocol. Routes that are the most stable (i.e., routes that do not become invalid due to node movements) and stay connected longest are chosen by utilizing the mobility prediction. The mobility characteristics of the mobile nodes are taken into account in [26], and a velocity-aided routing algorithm is proposed, which determines its packet forwarding scheme based on the relative velocity between the intended forwarding node and the destination node. The routing performance can further be improved by the proposed predictive mobility and location-aware routing algorithm, which incorporates the predictive moving behaviors of nodes in protocol design. The region for packet forwarding is determined by predicting the future trajectory of the destination node.

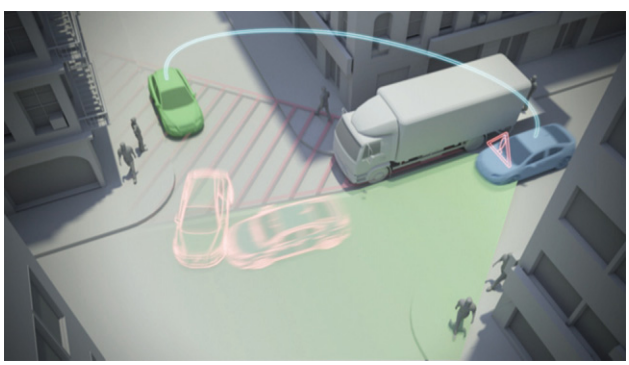
We clearly see that at the network and transport layer, harnessing location information appropriately can aid in reducing overhead and latency, while offering scalable solutions, even for highly mobile networks. In such networks, location information also plays an important role in the higher layers, as we will detail next.

HIGHER LAYERS

At the higher layers, location information will naturally be critical to provide navigation and location-based services. While we do not aim to provide a complete overview of such services, we briefly detail several applications of importance in the context of 5G networks.

FIFTH-GENERATION MOBILE AND WIRELESS COMMUNICATION SYSTEMS WILL REQUIRE A MIX OF NEW SYSTEM CONCEPTS TO BOOST SPECTRAL EFFICIENCY, ENERGY EFFICIENCY, AND THE NETWORK DESIGN.

First, we have classical context awareness, which finds natural applications in location-aware information delivery [27] (e.g., location-aware advertising) and multimedia streaming [28]. For the latter application, [28] tackles the problem of guaranteeing continuous streaming of multimedia services while minimizing the overhead involved, by capturing correlated mobility patterns, predicting future network planning events. A second class of applications is in the context of intelligent transportation systems. Several car manufacturers and research centers are investigating the development of intervehicle communication protocols. In this context, [29] focuses on the problem of providing location-aware services (e.g., traffic-related, time-sensitive information) to moving vehicles by taking advantage of short-range, intervehicle wireless communication and vehicular ad hoc networks. Location information is also critical for autonomous vehicles to coordinate and plan the vehicle's actions with respect to the environment and current traffic conditions (see Figure 7). Highly related are the tactile Internet [2] and other mobile cyber-physical systems, such as groups of unmanned aerial vehicles or robots [8], where localization and communication are closely intertwined.



[FIG7] The use of location information in intelligent transportation systems. After self-positioning, the vehicles become aware of each other through wireless communication and are able to avoid an accident.

Finally, location information also has implications in the context of security and privacy. For example, [30] studies the management of encryption keys in large-scale clustered sensor networks. In particular, a novel distributed key management scheme is proposed that reduces the potential of collusion among compromised sensor nodes by factoring the geographic location of nodes in key assignment. In [6], location information is utilized to detect wormhole attacks, which disrupt the network topology, as perceived by the benign nodes.

While we have focused on existing applications, we can expect novel, unforeseen location-based services in 5G networks, following us at all times, anticipating our needs, and providing us with information when and where we need it. With this comes a number of risks related to security and privacy, which should be addressed explicitly.

RESEARCH CHALLENGES AND CONCLUSIONS

Fifth-generation mobile and wireless communication systems will require a mix of new system concepts to boost spectral efficiency, energy efficiency, and the network design. There are many open issues to be addressed before these systems will be able to enter the market. In the following, we focus our attention on challenges related to the use of location information in 5G networks.

- *Achieving location awareness*: Throughout this article, we have assumed accurate location information is available. However, to realize the predicted position accuracies, significant signal processing challenges must be addressed so that seamless and ubiquitous localization can be made possible. The challenges include 1) handover, fusion, and integration of different positioning technologies; 2) coping with errors due to harsh propagation environments and interference; and 3) decentralization and reduction of complexity. In addition, 5G technologies themselves may have tight interactions with positioning. For example, millimeter wave systems may require accurate user tracking through BF; novel waveforms such as those used in filter bank multicarrier have relaxed synchronization demands, and may therefore reduce time-based positioning accuracy.
- *Ad hoc networking*: In ad hoc and certain machine-to-machine (M2M) networks, availability of a CQM database is questionable. In addition, accessing the database would require a preexisting communication infrastructure. Hence, distributed databases (or database-free methods) may be required in such networks, to capitalize on location awareness. The construction, maintenance, and exploitation of these databases will rely on distributed signal processing and deserves further study. Location knowledge can also be leveraged to find low-latency control and data paths in ad hoc networks, enabling wireless control systems. The appropriate storage, utilization, and combination of location-based with pilot-based CQM is an open issue.

■ *Signaling overhead*: While the ratio of signaling overhead with respect to data payload is generally increasing, this is particularly apparent in M2M and Internet of Things signaling, as the typically used protocols are inefficient for such traffic. Even in combination with location awareness, overhead will be a major bottleneck, and dedicated representation and compression mechanisms as well as localized protocols need to be designed. The choice of CQM also plays an important role as more precise information can yield better gains, but at costs in terms of complexity, robustness, and overhead.

LOCATION AWARENESS BEARS GREAT PROMISE TO THE 5G REVOLUTION, PROVIDED WE CAN UNDERSTAND THE RIGHT TRADEOFFS FOR EACH OF THE POSSIBLE USE CASES.

■ *Spatial channel modeling*: The wide variety of use cases requires a flexible and robust inference engine. GPs, as presented earlier, are a promising candidate, but they are faced with challenges in terms of storage and computational complexity. Sparsifying techniques to build and maintain the database, decentralized processing, as well as structured approaches in the prediction are among the main signal processing challenges. In addition, various sources of uncertainty must be accounted for explicitly in the GP framework (e.g., in terms of the position), as well as inherent nonstationarities in the channel statistics. Yet another challenge is to keep the database of the different CQMs updated and synchronized. The updates may be delivered by different 5G radio devices and could drive the synergy between the different radio types, such as M2M or mobile radio devices. Compared to today's drive tests, the autonomous refinement of network resources would allow to increase the coherence time of the database content.

■ *PHY/MAC/NET layers*: Location information can be exploited in a number of ways, both through databases and channel modeling, as well as more directly at the PHY/MAC/NET layers. An important challenge is to identify the right tradeoff between relying on location-based information and on pilot-based CQM information. A second challenge involves the amount of centralized versus decentralized processing. An open question on the network level is how to best utilize location information for identifying when network-assisted device-to-device communication is beneficial and aiding neighbor discovery. Finally, the issue of energy-efficiency deserves further study. For example, location information could be used to decide when to power down certain small cell base stations.

■ *Higher layers*: In 5G networks using location information, there are great possibilities for resource allocation (power, bandwidth, rate) based on prediction of user behaviors/trajectories, predicted load levels at various network nodes, channel statistics, and interference levels (previously stored in databases, for example). Some initial research has been done, but there is a large space for designing completely new algorithms and solutions. In addition, sharing location information raises important privacy and security issues. Secure and private computing in a location-aware context are promising, but they pose technical challenges in which signal processing can aid in masking and hiding information and in developing attack-resistant algorithms and protocols.

In summary, location awareness bears great promise to the 5G revolution, provided we can understand the right tradeoffs for each of the possible use cases. In this article, we have given an overview of how location awareness can be leveraged across the different layers of the (traditional) protocol stack, and we highlighted a number of important technical challenges.

ACKNOWLEDGMENTS

This work is supported, in part, by the European Research Council under grant number 258418 (COOPNET), the European Union under ICT-248894 FP7 project WHERE2, and by the Swedish Research Council under grant numbers 2011-6864 and 2009-4555.

AUTHORS

Rocco Di Taranto (taranto@chalmers.se) received the Laurea in Ingegneria delle Telecomunicazioni from Politecnico di Torino in 2005 and the Ph.D. degree in wireless communications from Aalborg University in 2010. From March to June 2009 he was a visiting researcher at the New Jersey Institute of Technology, United States. From February 2011 to May 2012 he was a postdoctoral fellow at the University of Waterloo, Ontario, Canada. From June 2012 to May 2014, he was a postdoctoral fellow at Chalmers University of Technology, Sweden. His research focuses on resource allocation in wireless and fiber-optic communications.

Srikar Muppirisetty (srikar.muppirisetty@chalmers.se) received his B.Tech degree in electronics and communication engineering in 2005 from R.V.R. and J.C. College of Engineering, India. He received his M.Sc. degree in communication engineering in 2009 and is currently pursuing a Ph.D. degree at Chalmers University of Technology. He has over four years of industrial experience in physical layer algorithm development. From 2005 to 2007, he worked as a member of technical staff at Digibee Microsystems, Bangalore. During 2009–2012, he worked as a senior system software engineer and as a technical leader on a simulation team at ST-Ericsson R&D Center, Bangalore. His current research focuses on developing algorithms for resource allocation using position information in wireless communications.

Ronald Raulefs (r.raulefs@dlr.de) is senior researcher and project manager at the Institute of Communications and Navigation of the German Aerospace Center (DLR) in Wessling, Germany. He received the Dr.-Ing. degree from the University of Erlangen-Nuremberg, Germany, in 2008. He was a visiting researcher at the City University of Hong Kong in September 2004. He initiated and coordinated the European Union Seventh Framework Programme for Research projects WHERE and WHERE2. He holds numerous patents in the area of mobile communications

and radio-based localization. His current research interests include various aspects of mobile radio communications and localization, such as dynamic resource allocation for cooperative localization.

Dirk T.M. Slock (slock@eurecom.fr) received an engineering degree from the University of Ghent, Belgium, in 1982, and the M.S. degree in electrical engineering (1986), M.S. degree in statistics (1989), and Ph.D. degree in electrical engineering (1989), all from Stanford University (Fulbright grant). He has been a professor at EURECOM since 1991. He invented semiblind channel estimation, the chip equalizer-correlator receiver (3G), MIMO-CDD now part of LTE, and the Single Antenna Interference Cancellation (SAIC) (GSM standard). He has (co) authored 420 papers. He received a Best Paper Award from the IEEE Signal Processing Society and from EURASIP in 1992, and two IEEE Globecom'98, one IEEE SIU'04 and one IEEE SPAWC'05 Best Student Paper Award.

Tommy Svensson (tommy.svensson@chalmers.se) received his Ph.D. degree in information theory from Chalmers University of Technology in 2003. He is an associate professor of communication systems at the same university, leading the research on air interface and wireless backhaul networking technologies for wireless systems. He actively contributes to European projects, currently within the EU FP7 METIS project targeting solutions for fifth generation. His main research interests are in design and analysis of physical layer algorithms, multiple access, resource allocation, cooperative systems, and moving relays/cells/networks. He is chair of the IEEE Sweden joint Vehicular Technology/Communications/Information Theory Societies Chapter and the coordinator of the Communication Engineering Master's Program at Chalmers. He is a Senior Member of the IEEE.

Henk Wymeersch (henkw@chalmers.se) is an associate professor of communication systems in the Department of Signals and Systems, Chalmers University of Technology, Sweden. He received the Ph.D. degree in 2005 from Ghent University. For his Ph.D. dissertation, he was awarded the 2006 Alcatel Bell Scientific Award. Between 2005 and 2009, he was a postdoctoral researcher at the Massachusetts Institute of Technology. His research interests include Bayesian inference, cooperative networks, and optical communications. He is the author of *Iterative Receiver Design* (Cambridge University Press, 2007), and is an associate editor of *IEEE Transactions on Wireless Communications* and *IEEE Transactions on Emerging Telecommunications Technologies*.

REFERENCES

[1] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[2] G. Fettweis, "The tactile internet: Applications and challenges," *IEEE Veh. Technol. Mag.*, vol. 9, no. 1, pp. 64–70, Mar. 2014.

[3] M. de Reuver, D. Skournetou, and E.-S. Lohan, "Impact of Galileo commercial service on location-based service providers: Business model analysis and policy implications," *J. Location Based Serv.*, vol. 7, no. 2, pp. 67–78, 2013.

[4] D. Slock, "Location aided wireless communications," in *Proc. Int. Symp. Communications Control and Signal Processing*, 2012, pp. 1–6.

[5] H. Celebi and H. Arslan, "Utilization of location information in cognitive wireless networks," *IEEE Wireless Commun.*, vol. 14, no. 4, pp. 6–13, 2007.

[6] A. Dammann, G. Agapiou, J. Bastos, L. Brunel, M. García, J. Guillet, Y. Ma, J. Ma, J. J. Nielsen, L. Ping, R. Raulefs, J. Rodriguez, D. Slock, D. Yang, and N. Yi, "WHERE2 location aided communications," in *Proc. European Wireless Conf.*, Apr. 2013, pp. 1–8.

[7] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press, 2006.

[8] J. Fink, "Communication for teams of networked robots," Ph.D. thesis, Elect. Syst. Eng., Univ. Pennsylvania, Philadelphia, PA, Aug. 2011.

[9] P. Agrawal and N. Patwari, "Correlated link shadow fading in multi-hop wireless networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 8, pp. 4024–4036, 2009.

[10] M. Gudmundson, "Correlation model for shadow fading in mobile radio systems," *Electron. Lett.*, vol. 27, no. 23, pp. 2145–2146, 1991.

[11] I. Nevat, G. W. Peters, and I. B. Collings, "Location-aware cooperative spectrum sensing via Gaussian processes," in *Proc. Australian Communications Theory Workshop*, 2012, pp. 19–24.

[12] S. Sand, R. Tanbourgi, C. Mensing, and R. Raulefs, "Position aware adaptive communication systems," in *Proc. Asilomar Conf. Signals, Systems and Computers*, 2009, pp. 73–77.

[13] R. C. Daniels and R. W. Heath, "Link adaptation with position/motion information in vehicle-to-vehicle networks," *IEEE Trans. Wireless Commun.*, vol. 11, no. 2, pp. 505–509, 2012.

[14] A. Dammann et al., A. Dammann, J. Bastos, L. Brunel, M. García, J. Guillet, J. Ma, J. J. Nielsen, L. Ping, R. Raulefs, J. Rodriguez, D. Slock, D. Yang, and S. Zazo, "Location aided wireless communications," *IEEE Commun. Mag.,* to be published.

[15] M. T. Sun, L. Huang, S. Wang, A. Arora, and T. H. Lai, "Reliable MAC layer multicast in IEEE 802.11 wireless networks," *Wireless Commun. Mobile Comput.*, vol. 3, no. 4, pp. 439–453, 2003.

[16] N. Wen and R. Berry, "Information propagation for location-based MAC protocols in vehicular networks," in *Proc. Annu. Conf. Information Sciences and Systems*, 2006, pp. 1242–1247.

[17] S. B. Kodeswaran and A. Joshi, "Using location information for scheduling in 802.15.3 MAC," in *Proc. Int. Conf. Broadband Networks*, 2005, pp. 718–725.

[18] S. Katragadda, C. N. S. G. Murthy, R. Rao, S. M. Kumar, and R. Sachin, "A decentralized location-based channel access protocol for inter-vehicle communication," in *IEEE Vehicular Technology Conf.*, 2003, vol. 3, pp. 1831–1835.

[19] Y. B. Ko and N. H. Vaidya, "Geocasting in mobile ad hoc networks: Location-based multicast algorithms," in *Proc. IEEE Workshop on Mobile Computing Systems and Applications*, 1999, pp. 101–110.

[20] J. J. Nielsen, "Location based network optimizations for mobile wireless networks—A study of the impact of mobility and inaccurate information," Ph.D. dissertation, Dept. Electronic Systems, Aalborg Univ., Mar 2011.

[21] F. Cadger, K. Curran, J. Santos, and S. Moffett, "A survey of geographical routing in wireless ad-hoc networks," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 2, pp. 621–653, 2013.

[22] A. M. Popescu, N. Salman, and A. H. Kemp, "Geographic routing resilient to location errors," *IEEE Wireless Commun. Lett.*, vol. 2, no. 2, pp . 203–206, 2013.

[23] R. Di Taranto and H. Wymeersch, "Simultaneous routing and power allocation using location information," in *Proc. Asilomar Conf. Signals, Systems and Computers*, 2013, pp. 1700–1704.

[24] M. J. Neely, E. Modiano, and C. E. Rohrs, "Dynamic power allocation and routing for time-varying wireless networks," *IEEE J. Select. Areas Commun.*, vol. 23, no. 1, pp. 89–103, 2005.

[25] W. Su, S. J. Lee, and M. Gerla, "Mobility prediction and routing in ad hoc wireless networks," *Int. J. Network Manage.*, vol. 11, no. 1, pp. 3–30, 2001.

[26] K. T. Feng, C. H. Hsu, and T. E. Lu, "Velocity-assisted predictive mobility and location-aware routing protocols for mobile ad-hoc networks," *IEEE Trans. Veh. Technol.*, vol. 57, no. 1, pp. 448–464, 2008.

[27] N. Marmasse and C. Schmandt, "Location-aware information delivery with ComMotion," in *Proc. Handheld and Ubiquitous Computing*, 2000, pp. 157–171.

[28] B. Li and K. H. Wang, "NonStop: Continuous multimedia streaming in wireless ad hoc networks with node mobility," *IEEE J. Select. Areas Commun.*, vol. 21, no. 10, pp. 1627–1641, 2003.

[29] M. D. Dikaiakos, A. Florides, T. Nadeem, and L. Iftode, "Location-aware services over vehicular ad-hoc networks using car-to-car communication," *IEEE J. Select. Areas Commun.*, vol. 25, no. 8, pp. 1590–1602, 2007.

[30] M. F. Younis, K. Ghumman, and M. Ektiweissy, "Location-aware combinatorial key management scheme for clustered sensor networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 17, no. 8, pp. 865–882, 2006.

[SP]

sp HISTORY

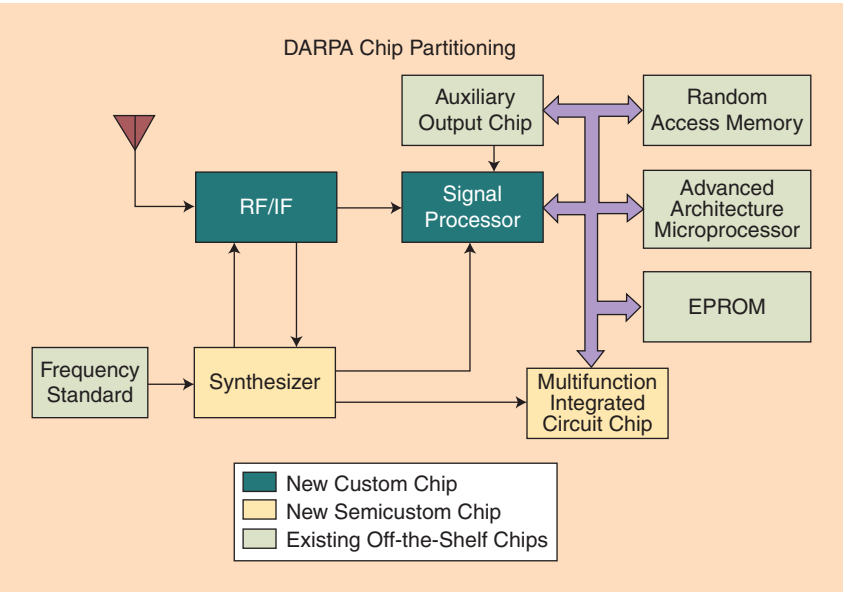
Larry B. Stotts, Sherman Karp, and Joseph M. Aein

The Origins of Miniature Global Positioning System-Based Navigation Systems

For the past decade or so, the Navigation Signal Timing and Ranging Global Positioning System (NAVSTAR GPS) has been synonymous with personal navigation. We find GPS in our smartphones, tablets, cars, aircraft, and boats. But how did it get that way? The U.S. Department of Defense (DoD) established the NAVSTAR GPS Joint Program Office (JPO) (GPS JPO) in 1973 to make the GPS a military and civilian reality [1]–[3]. However, its receivers were large and heavy and not very attractive to military and civilian personnel without vehicles. It wasn't until the mid-1980s that miniature GPS receiver (GPSR) and inertial navigation system (INS) technologies came into being, whose architectures became the standards for both the military and commercial precise navigation markets after the 1991 Desert Storm military campaign [4].

The reason for this remarkable change was the Miniature GPS Receiver (MGR) program, circa 1985, managed by the Defense Advanced Research Projects Agency (DARPA) [4]–[7]. The fear of GPS jamming led DARPA to initiate a companion program in 1988, called the GPS Guidance Package (GGP). Using a Kalman filter (KF), GGP tightly coupled its MGR with a navigation-grade miniature inertial measurement unit (MIMU) for precise navigation in electromagnetic interference (EMI) conditions [5]–[7]. These two programs are the topic of this article.

Early in 1983, the primary military GPSR developed by the GPS JPO was Receiver IIIa. Receiver IIIa was to serve as the core capability for deployment to a variety of DoD (powered) platforms: ships, planes, trucks, etc. Also at that time, the U.S. Army was sponsoring development of a (large) battery-operated "Man Pack" GPSR that would be of a size similar to a WWII Company/Platoon level, man portable tactical network radio. Unfortunately, the resulting Man Pack receiver was cumbersome and heavy (~50 lb), further burdening a soldier or Marine with his backpack, rifle, and helmet.



[FIG1] **The MGR architecture.**

THE U.S. DEPARTMENT OF DEFENSE ESTABLISHED THE NAVSTAR GPS JOINT PROGRAM OFFICE IN 1973 TO MAKE GPS A MILITARY AND CIVILIAN REALITY.

Dr. Sherman Karp, a program manager in DARPA's Strategic Technology Office, approached Dr. Anthony Tether, the office director in early 1983, to present his idea to develop a handheld "all-digital" GPSR to include military, cryptographic precision capability (P-Code). Specifically, the MGR would exploit both very high-speed integrated circuit (VHSIC) signal processing and hybrid analog/digital radio-frequency (RF)/intermediate frequency (IF) monolithic microwave integrated circuit (MMIC) components. When Dr. Tether asked just how small this new receiver would be, Dr. Karp grabbed a pack of Virginia Slims cigarettes off a nearby desk and said it would fit inside the box (100 cc). So, DARPA's MGR program was born (unofficially, MGR was called *Virginia Slims*). At the same time, the U.S. Marine Corps (USMC) had decided that the Man Pack GPSR would be too large to meet its needs. After learning of the

Digital Object Identifier 10.1109/MSP.2014.2343985
Date of publication: 15 October 2014

1053-5888/14©2014IEEE

DARPA MGR effort, the USMC came to DARPA in late 1983 with a formal requirement for a miniature GPSR.

The MGR program began with five defense contractors competing for a hardware development contract, and of those five, only Rockwell Collins went on to the fabrication and testing the MGR. After Dr. Karp retired in 1986, Neil Dougherty stepped into the MGR leadership position at DARPA. Dougherty left in late 1987, and Dr. Larry Stotts became the MGR program manager [5]–[7] and led the effort to its very successful completion in 1991 [4]. He also persuaded DARPA to start the GGP program taking the program through its initial development [5]–[7]. At DARPA, when Dr. Stotts was assigned new responsibilities, Lt. Col. Beth Kaspar replaced him and continued GGP development to successful flight testing on a Navy combat F-18C fighter jet in late 1996. Finally, Steven Welby at DARPA completed the GGP program circa 2000.

Rockwell Collins achieved several major technical breakthroughs in the MGR: 1) the first hybrid analog/digital, GaAs MMIC chip, 2) the first "all-digital" GPSR (except for the MMIC RF/IF front end) having secure, precision military P-Code security encryption, and 3) the first all-ADA GPSR software. Figure 1 shows the MGR functional layout and chip partitioning.

After successfully passing all the required tests, the MGR was transitioned to the GPS JPO, and its productization resulted in the PLGR in 1990—20,000 PLGRs were procured by the U.S. Army at US$1,700 per unit, a factor of three less than the MGR price goal. Figure 2 is a photo of the PLGR; it shows the five-channel PLGR version in (a) and its one-channel predecessor in (b). Selected MGR chip parts also were used in the Tomahawk Land Attack Missile, Conventional guidance system, in the Miniature Airborne GPS Receiver (MAGR), two space-borne MGR receivers launched into space, and in a program for material tracking. Today, the more modern Defense Advanced GPS Receiver (DAGR) has replaced the PLGR in the military.

The key characteristics of the MGR chip set are summarized in Table 1. Subsequent Silicon (Si) technology improvements increased DSP chip density and clock rate, lowered power drain, and increased, to over four the number of GPS signal channels processed simultaneously. Initially, Si speed had not progressed to where amplification of the GPS L-band signals was possible. Consequently a Gallium arsenide (GaAs) MMIC was needed. With the continuing increase in Si clock rate to over 2 GHz, the GaAs MMIC was replaced by a lower-cost Si MMIC. This early progress in miniature GPSR compared to the JPO's IIIa receiver is summarized in Table 2.

AFTER SUCCESSFULLY PASSING ALL THE REQUIRED TESTS, THE MGR WAS TRANSITIONED TO THE GPS JPO, AND ITS PRODUCTIZATION RESULTED IN THE PLGR IN 1990.

The MGR GaAs MMIC was high risk. Both analog (RF/IF) and digital (A/D) circuitry were placed on one large, commonly foundered, GaAs chip. The high frequency A/D sampling output, at relatively large voltage, needed to be electrically isolated from the very sensitive RF input amplification stages.

[FIG2] (a) The next-generation version DAGR of the (b) Precision Lightweight GPS Receiver (PLGR). (Image courtesy of Rockwell Collins and used with permission.)

The other feature of the MGR was its all-digital signal processing of the GPS 10 MHz security encryption (P-Code). Additionally, the MGR was the first GPSR to employ a very low power, microprocessor on a chip to do the navigation computations and manage the MGR functionality modes (including display and user interface) and resources. Needing military GPS code security, the MGR had to employ an existing GPS security certified part, i.e., two auxiliary onboard chips (AOCs) that put considerable pressure on the rest of the MGR parts to meet the 100 cc volume goal.

Once synchronized to a set of four or more received GPS radio navigation signals, the GPSR maintains track of these signals over time. After processing the received GPS signals in combination with the broadcasted GPS system data, the GPSR provides output measurements **P**, **V**, and t of user position and velocity (**P** and **V** are three-dimensional vectors) to an

[TABLE 1] THE MGR CHIP SET SUMMARY.

CHIP TYPE	DEVICE COUNT	DIMENSIONS (in)	IMPLEMENTATION TECHNOLOGY	POWER (mW)
RF/IF TRANSLATOR L-BAND	300–400	0.200 × 0.240	GaAs MMIC	1,700
GPS DSP 10 Msamples/s	20,000	0.185 × 0.220	1.25-μm VHISC COMPLEMENTARY METAL–OXIDE–SEMICONDUCTOR (CMOS)	90
MULTIFUNCTION INTERFACE CHIP	29,000	0.370 × 0.370	1.6-μm BULK CMOS	20
RECEIVER MANAGEMENT MICROPROCESSOR	60,000	0.214 × 0.261	2-μm BULK CMOS	80
FREQUENCY SYNTHESIZER	600–700	0.250 × 0.250	BIPOLAR SILICON	500
TOTAL POWER				2,390

[TABLE 2] THE EVOLUTION OF MILITARY GPSRS USING P(Y) CODES.

GPSR	IIIA	DARPA MGR	PLGR	GPS GUIDED
DEVELOPMENT STARTS	1982	1991	1993	BOMB GPSR 1995
RF/IF/AD	DISCRETE	GaAs MMIC	GaAs MMIC	Si MMIC
CORRELATOR	ANALOG	DIGITAL VHISC	DIGITAL ASIC	DIGITAL ASIC
SECURITY		PPS/SM	PPS/SM	PPS/SM
	SINGLE CHANNEL	MULTICHANNEL	MULTICHANNEL	MULTICHANNEL
	AOC	AOC	AOC	AOC
CAPABILITY				
FREQUENCIES	L1 AND L2	L1 OR L2 USER SELECTABLE	L1 ONLY	L1 AND L2
NUMBER OF CHANNELS	FIVE	TWO	FIVE	12
NAVIGATION PROCESSING	PARALLEL	SEQUENTIAL	PARALLEL	PARALLEL
SIZE: VOLUME (in^3)	9,600	6.1 (100 cm^3)	90	20
WEIGHT (lb)	36		<4	<1
POWER (W)	110	2.5	5	<3
PRODUCT YEAR	1982	1988	1993	1998
QUANTITY	3,000+	TWO PROTOTYPES	200,000+	40,000+

accuracy well under 10 m and 0.5 m/s every 20 ms (with a time of validity accurate to 1 μs or better). Finally, the navigation frame of reference that the GPS employs is the Earth centered, Earth fixed frame (ECEF) (ECEF is fixed to and rotates with the Earth) known as WGS-84.

As the MGR development showed promise, DARPA initiated the GGP development. Specifically, DARPA developed the first miniaturized, navigation-grade, interferometric, fiber-optic gyroscopes and silicon accelerometers, using three of each to form a precision INS. This unit then was integrated with the MGR to create an embedded GPS/INS (EGI). The INS calculated the resulting position, velocity, and acceleration from the various components data streams using KF, and used the MGR to correct any bias and other errors. This form of processing is known as *tight coupling* [8], [9]. GGP was the first to demonstrate a navigation-grade, tight-coupled GPS/INS. The MGR carrier-loops used precisely time-tagged velocity aiding provided by the MIMU, and the MIMU sensor error bias-drift was reduced by the MGR position solutions. Size (<300 in^3) and weight (25 lb) goals were met, as was the goal for inertial navigator performance of less than 0.5 nautical mile per hour. Litton Industries was the INS and integration contractor on the effort, and Rockwell Collins provided the MGR. Figure 3 shows the GGP unit [10].

[FIG3] The GPS Guidance Package (Image courtesy of [9], used with permission.)

DARPA HAD A STRONG ROLE IN THE DEVELOPMENT OF PRECISE MGR AND MGR/INS EQUIPMENT, AND ITS LEGACY IS STILL SEEN TODAY.

The GGP was successfully flight tested at the Patuxent Naval Air Station 26 November–17 December 1996 on a high dynamic (7g) F/A-18C fighter jet as part of the F-18 Special Project Office's EGI competition. However, its productization did not occur until the prime contractor, Litton Industries [now Northrop Grumman Corporation (NGC)], created its family of GPS/INSs to sell to the military and commercial world [10], [11]. The first system out of the box was the LN-251, which NGC stated "resulted from a DARPA-funded project (i.e., GGP) to produce the next generation of navigation-grade inertial system that would provide the smallest volume, lowest weight, lowest power consumption, and highest system reliability compared to any other approach using alternate technologies such as mechanical or ring laser gyros" [10]. The first military transition occurred in October 2002, when the Turkish Army became the first large purchaser of LN-270s, another GGP-based product. (Northrop-Grumman produced the LN-270 with its partner company ASELSAN of Ankara, Turkey.) On 9 January 2006, the U.S. Air Force selected NGC's newest fiber optic gyro INS at the time, the LN-260, also GGP derived, as the avionics upgrade for the F-16 Multinational Fighter Program aircraft fleet.

In summary, DARPA had a strong role in the development of precise MGR and MGR/INS equipment, and its legacy is still seen today.

AUTHORS

Larry B. Stotts (lbstotts@gmail.com) is currently a private consultant. He was the deputy director for the Strategic Technology Office at the Defense Advanced Research Agency. He coauthored *Fundamentals of Electro-Optic Systems Design: Communications, Lidar, and Imaging* (Cambridge Press, New York). He is a fellow of the IEEE, a life fellow of the SPIE and a senior member of the Optical Society of America.

Sherman Karp (shermankarp@msn.com) is currently a private consultant. He was the principal scientist of the Defense Advanced Research Agency in the early 1980s. He also coauthored *Fundamentals of Electro-Optic Systems Design: Communications, Lidar, and Imaging*, (Cambridge Press, New York). He is a Life Fellow of the IEEE.

Joseph M. Aein (joe.aein0097@verizon.net) is now retired. He was employed at the Institute for Defense Analyses, Arlington, Virginia, followed by the RAND Corp., Santa Monica, California, where he participated in technology and systems evaluations in support of the U.S. Department of Defense. Among these were his participation in the Defense Advanced Research Agency Miniature GPS Receiver and GPS Guidance Package efforts. He is a Life Fellow of the IEEE.

REFERENCES

[1] B. Parkinson and S. T. Powers, "The origins of GPS: Part 1" *GPS World*, vol. 21, no. 5, pp. 30–41, May 2010.

[2] B. Parkinson and S. T. Powers, "The origins of GPS: Part 2" *GPS World*, vol. 21, no. 6, pp. 8–18, June 2010.

[3] S. Pace, G. P. Frost, I. Lachow, D. R. Frelinger, D. Fossum, D. Wassem, and M. M. Pinto. (1995). The global positioning system: Assessing national policies. Santa Monica, CA: RAND Corporation. [Online]. Available: http://www.rand.org/pubs/monograph_reports/MR614

[4] R. B. Langley, "The evolution of the GPS receiver," *GPS World*, vol. 11, no. 4, pp. 54–58, Apr. 2000.

[5] L. B. Stotts and J. Aein, "Status of DARPA guidance and control program," invited paper in *Proc. Cruise Missile Association Annu. Meeting and Symp.*, Apr. 1989, Washington, DC, pp. 205–228.

[6] L. B. Stotts, J. Aein, and N. Doherty, "Miniature GPS-based guidance technology," in *Proc. Guidance and Control Panel 48th Symp. Advances in Techniques and Technologies for Air Vehicle Navigation and Guidance*, May 1989, Lisbon, Portugal.

[7] L. B. Stotts and J. Aein, "Guidance technology useful for military communications," presented at Military Communications Conf. (MILCOM 89), Classified Session, 15–18 Oct. 1989.

[8] N. Dahlen, T. Caylor, and E. Goldner, "High performance GGP for multiple dual-use applications," in *Proc. 1996 National Technical Meeting of the Institute of Navigation*, Santa Monica, CA, Jan. 1996, pp. 63–74.

[9] N. J. Dahlen, T. L. Caylor, and E. L. Goldner, "Tightly coupled IFOG-based GPS guidance package," *J. Inst. Navigation*, vol. 43, no. 3, Fall 1996, pp. 257–272.

[10] C. Volk, J. Lincoln, and D. Tazartes, "Northrop Grumman's family of fiber-optic based inertial navigation systems," in *Proc. IEEE/ION PLANS 2006*, Apr. 25–27, 2006, San Diego, CA, pp. 382–389.

[11] G. Pavlath, "Fiber optic gyros past, present, and future," in *Proc. 22nd Int. Conf. Optical Fiber Sensors (OFS)*, Y. Liao, H. Ho, W. Jin, D. D. Sampson, R. Yamauchi, Y. Chung, K. Nakamura, and Y. Rao, Eds., *Proc. SPIE*, vol. 8421. Bellingham, WA: SPIE, 2012, p. 842102.

[SP]

[special REPORTS] (continued from page 11)

Center module. The outputs of this module are two images, the initial image interpolated and the linearly adjusted image after interpolation.

The Find Center module attempts to more accurately locate the cell's center. It finds the center of the cell by converting input images into binary image and counting the number of nonzero pixels in each row and column. The module processes the two output images from the Interpolation module and averages both to identify the center point. This is done to improve accuracy as specular noise can affect the results of either in-put. The Find Center module transforms these images into binary images by adaptively thresholding at different intensity values to separate the inner cell area and cell wall.

At the last stage, the system determines morphological properties of the cell using the interpolated image and its corresponding center point. It converts the resized image from Cartesian coordinates into polar coordinates. The darkest pixels found on a line from the cell center at each angle are considered the cell wall.

The researchers found that they obtained significantly faster performance with the FPGA than with GPU. The result didn't come as a total surprise, since FPGAs, unlike GPUs, can be custom-tailored to match the algorithm.

While designing the FPGA the researchers carefully studied each step and made changes designed to enhance efficiency and performance. Kastner notes that when mapping to custom hardware, it's important to balance algorithm complexity against result accuracy. "Algorithms incorporating a large number of decisions points, or that have to make multiple passes over the data, can lead to a slow and inefficient FPGA," he says.

Still, when correctly implemented, an FPGA can be used to perform operations at stunning speeds (the UCSD algorithm needs fewer than 500 µs to detect a cell and calculate its radius).

The researchers' ultimate goal is to analyze cell properties in real time and then use the information to accurately sort the cells. To achieve this capability, the sorting decision must be made in fewer than 10 ms. With the new approach promising sorting rates as low as 11.94 ms that target is now tantalizingly close.

Kastner is optimistic that the new technology will eventually be used in wide range of clinical applications. "This has to potential to lead to numerous breakthroughs," he says. "We are collaborating with UCLA and their industrial partners to commercialize the technology."

AUTHOR

John Edwards (jedwards@johnedwardsmedia.com) is a technology writer based in the Phoenix, Arizona, area.

[SP]

applications CORNER

Viet Anh Nguyen, Jiangbo Lu, Shengkui Zhao, Douglas L. Jones, and Minh N. Do

Teleimmersive Audio-Visual Communication Using Commodity Hardware

Natural human communication involves complex visual and audio behavior, and often context and joint interaction with the surrounding environment, to create a rich and satisfying experience. However, widely used virtual meeting systems such as WebEx and Skype still provide rather limited functionalities and hardly maintain the experience of an in-person meeting. In particular, traditional systems lack a sense of colocation and interaction as in a face-to-face meeting due to the separate displays of remote participants and poor integration with the shared collaborative contents. As a result, teleimmersive (TI) systems that aim to provide natural user experiences and interaction have attracted increasing research interest [1]. High-end telepresence products such as Cisco TelePresence or HP's Halo were expressly designed to create the perception of meeting in the same physical space. But to achieve such an experience, these systems require a proprietary installation and high setup costs. Recently, some three-dimensional (3-D) TI systems have been developed to enhance remote collaboration by merging remote participants into the same 3-D virtual space [2]–[4]. However, these systems still fall short of simulating a face-to-face collaboration with the presence of shared contents. Also, the required bulky and expensive hardware with nontrivial calibration and setup hinders their wide adoption. With the wide availability of low-cost, commodity computing devices with embedded video cameras, microphones, and ubiquitous Internet access adequate for real-time media, the dream of high-quality TI communication should finally be within our reach.

Achieving the dream of natural interactive communication at a distance with low-cost personal computing devices presents several new challenges that high-cost, dedicated systems can avoid. Since (unlike the Halo system with its dedicated, carefully purpose-designed physical environment, acquisition and display hardware, and dedicated network communication link) such a system may be used in any distracting environment and background, so a commodity telepresence system must detect and segment the user(s) from any type of background scene. It must also successfully and reliably do so with any common camera device in any variety of illumination conditions without prior calibration. Since the communication must utilize the Internet, it must be robust to network bandwidth variations, latency, and temporary dropouts, and the data must be efficiently compressed into a low bandwidth. We also desire to the ability to enable multiparty participation and interaction with virtual electronic objects such as presentations or imagery. A natural and high-quality audio experience is critical to achieve a sense of truly "being there," as well as to enable users to parse a complex scene to know who is speaking and when, so the audio should maintain a correct 3-D spatial rendering and remove background noise and interference. Finally, all of these elements must be seamlessly integrated in a system of low enough computational complexity to operate in real time on commodity hardware.

> THE DREAM OF HIGH-QUALITY TI COMMUNICATION SHOULD FINALLY BE WITHIN OUR REACH.

In this column, we present a real-time immersive telepresence system for entertainments and meetings (ITEM) based on a low-cost, flexible setup (e.g., a Webcam and/or a depth camera and a desktop/laptop connected to the Internet). The system puts remote participants into the same virtual space and seamlessly integrates them with any shared contents for a more natural person-to-person interaction. With the addition of a depth camera and a low-cost microphone array, ITEM supports spatialized 3-D audio, active speaker detection, and gesture-based controls to reproduce nonverbal signals and interactions with the shared contents in an intuitive and effective manner. The key points of ITEM are highlighted in Table 1 compared to existing TI solutions.

The remainder of this column describes a complete design and implementation of such a system by addressing the challenges and key components in the whole pipeline of media processing, communication, and display.

SYSTEM OVERVIEW

Our focus is on the system aspects in building such a lightweight practical TI system that maximizes the end-user experience, optimizes the system and network resources, and enables a variety of TI application scenarios. We consider major practical requirements in our design to build a system that supports

Digital Object Identifier 10.1109/MSP.2014.2340232
Date of publication: 15 October 2014

1053-5888/14©2014IEEE

[TABLE 1] A COMPARISON BETWEEN AN ITEM AND THE EXISTING VIDEO TELECONFERENCING SOLUTIONS.

SOLUTIONS	SETUP COST	HARDWARE	NETWORK	AUDIO/VIDEO QUALITY	QUALITY OF EXPERIENCE
HIGH-END TELEPRESENCE (CISCO, HP's HALO)	EXTREMELY HIGH	DEDICATED SETUP, PROPRIETY HARDWARE	DEDICATED BANDWIDTH	LIFE-SIZE VIDEO QUALITY, STUDIO ROOM QUALITY	IMMERSIVE ILLUSION, PERIPHERAL AWARENESS
NTII [2], TEEVE [3], BEING THERE [4]	HIGH	BULKY, EXPENSIVE 3-D CAMERA SETUP, BLUE SCREEN	INTERNET2 NETWORK	RELIABLE VIDEO CUTOUT, LOW 3-D VIDEO QUALITY, LOW FRAME RATE, STANDARD AUDIO	3-D IMMERSIVE RENDERING, BODY INTERACTION AND COLLABORATION
2-D TI SYSTEMS (VIRTUAL MEETING [5], COLISEUM [6], CUTE CHAT [7])	LOW	STANDARD PCs—AUDIO, VIDEO PERIPHERALS, MULTIPLE CAMERAS (IN COLISEUM)	INTERNET	UNRELIABLE VIDEO CUTOUT, LOW VIDEO RESOLUTION, LOW FRAME RATE, STANDARD AUDIO	IMMERSIVE DISCUSSION, WITHOUT SUPPORTING NONVERBAL SIGNALS/CUES AND COLLABORATION
STANDARD VIDEO CONFERENCING (SKYPE)	LOW	STANDARD PCs—AUDIO, VIDEO PERIPHERALS	INTERNET	MEDIUM-TO-HIGH VIDEO QUALITY, STANDARD AUDIO	NONIMMERSIVE, WITHOUT NONVERBAL COLLABORATION
ITEM	LOW	STANDARD PCs—AUDIO, VIDEO PERIPHERALS, DEPTH CAM (OPTIONAL), MICROPHONE ARRAY	INTERNET	ROBUST, RELIABLE CUTOUT, SUPPORT HD RESOLUTION HIGH FRAME RATE, SPATIAL AUDIO, SPEAKER DETECTION	IMMERSIVE, NATURAL CONVERSATION WITH NONVERBAL COLLABORATION

multimodality (audio/video, shared contents), scalability for a large number of participants and concurrent meetings, flexibility in a system setup [two-dimensional (2-D) color Webcam and/or 3-D depth camera], and desirable functionality to best suit different application contexts. As an end result, we present several interesting applications and user experiences created by ITEM. Figure 1(a) gives an overview of the ITEM system, where only the pair of a sender and a receiver is shown for simplicity. At the sender site, a commodity Webcam (or a depth camera) captures a live video stream, which is processed with our video object cutout technique to segment out object A in real time. The foreground object stream is then encoded using chroma key-based object coding prior to transmission over the Internet, reaching the receiver site under the management of an enhanced video delivery scheme. After decoding the video object, a clean segmentation map recovery method is applied to reconstruct a clean foreground segmentation map, which would otherwise contain boundary artifacts caused by compressing object video with background chroma keying. Finally, the user has a range of options on how to compose the final video to be displayed. He or she can choose to merge his/her own object video into the final frames, while the background (e.g., slides, photos) can be selected either from the local store, or streamed dynamically from the Internet as shared resources. In a group teleconferencing scenario, a low-cost compact microphone array is used to provide 3-D audio capture and reproduction as well as active speaker detection and tracking. The new communication experiences and compelling functionalities created by ITEM can be seen in a video on YouTube (http://youtu.be/cuRvXcLUIR4). For all technical details and quantitative evaluation of our technologies and the comparison with other existing approaches, we refer readers to a technical report [8].

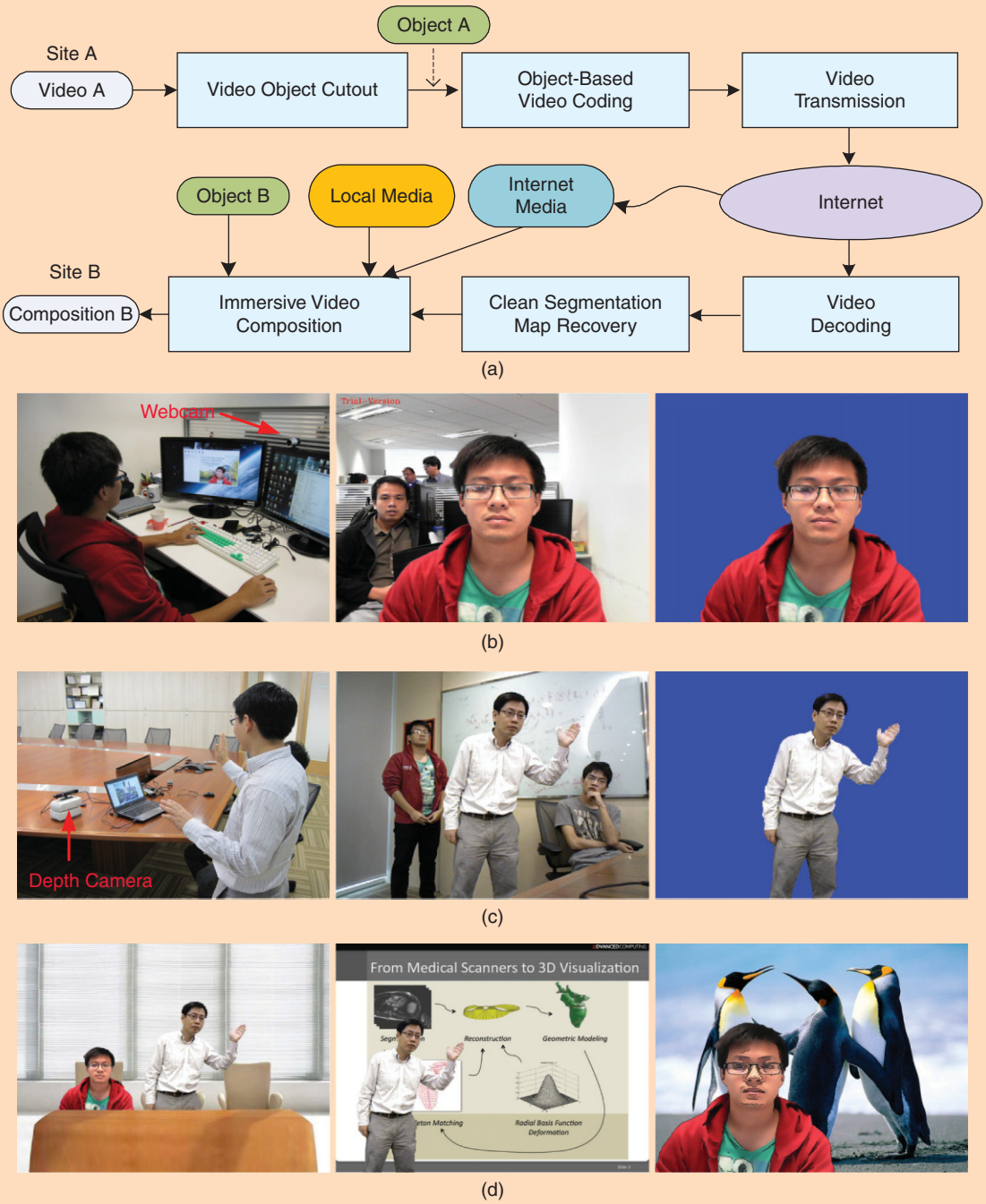
SEPARATE CODING AND DELIVERY OF MEANINGFUL FOREGROUND OBJECTS DECOMPOSED FROM CAPTURED SIGNALS IS ALSO CRUCIAL IN TI SYSTEMS.

VIDEO OBJECT CUTOUT

Video object cutout is essential to enable the immersive experience in the ITEM system. Assuming that the background is known and the Webcam is static, we have developed a practical solution for segmenting a foreground layer in real time from a live video captured by a single Webcam. Though this assumption appears somewhat constrained, a good solution can be widely deployed and it enables the aforementioned exciting systems with no additional cost. In fact, segmenting the foreground layer accurately from a complex scene where various changes can happen in the background is still rather challenging. Compared with state-of-the-art segmentation techniques, our technology has the following advantages: 1) reliable segmentation with high accuracy under challenging conditions, 2) real-time speed [18–25 frames per second (FPS) for video graphics array (VGA) resolution, 14–18 FPS for high-definition (HD)-resolution] on a commodity hardware such as a laptop/desktop, and 3) ease of use with little or no user intervention in the initialization phase. Based on a unified optimization framework, our technology probabilistically fuses different cues together with spatial and temporal priors for accurate foreground segmentation. In particular, the proposed technology consists of two major steps, i.e., layer estimation and labeling refinement. When a depth camera is available, the current framework can also be automatically configured to utilize the important depth information for more reliable inference, while leveraging several other key components also shared by the Webcam-based object cutout flow. Figure 1(b) and (c) shows foreground segmentation results using different setups (e.g., with/without using a depth camera) under challenging test conditions.

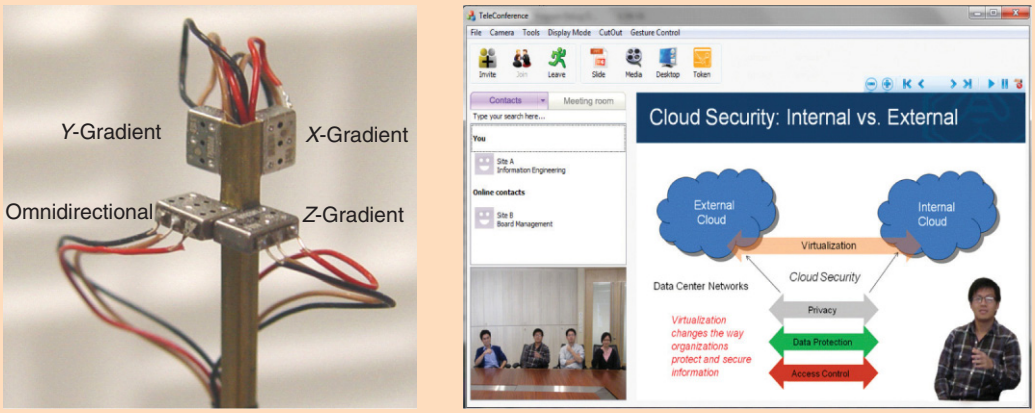
OBJECT-BASED CODING

Separate coding and delivery of meaningful foreground objects decomposed from captured signals is also crucial in TI systems.



[FIG1] An overview of the ITEM system (a) with the key real-time video object cutout technology using (b) a normal Webcam or (c) a depth (plus RGB) camera. From left to right: system setup, input video frame, and cutout result. Some video composition effects are shown in (d).

This not only facilitates object-based processing such as immersive rendering in a shared environment but also eases the network traffic by discarding irrelevant background information. In ITEM, we use an efficient object-based video coding using a chroma-key-based scheme with a high coding efficiency H.264 codec, where a chroma-key color is used as the background [9]. A new, fast-mode decision method speeds up the encoding process by considering the characteristics of real-life conferencing videos (e.g., containing sudden, complex motion such as hand gestures, facial expressions) to eliminate unnecessary coding modes, which reduces both the bandwidth and the computational requirements by factors of three to four [7],



[FIG2] Speaker detection based on visual cue and localization using our (a) low-cost compact microphone array to enhance the communication experience by immersively rendering the active speaker with (b) the shared background.

[9]. This is important to reduce the complexity of highly computational H.264 encoder, leaving more central processing unit (CPU) resources for other tasks. For the incoming object videos, a nonlinear neighborhood filter is used in the binary mask recovery to attain a clean segmentation map by removing speckle labelling noise due to video coding quantization artifacts.

3-D SOUND CAPTURE, RENDER, LOCALIZATION

Multiparty TI systems are greatly enhanced by spatial sound and the detection of speakers, especially with multiple participants at a single site. Unlike the existing systems requiring a large spatially separated microphone array, we built a very compact microphone array [10], where four collocated miniature microphones approximate an acoustic vector sensor (AVS) [Figure 2(a)]. This AVS array consists of three orthogonally mounted acoustic particle velocity gradient microphones X, Y, and Z and one omnidirectional acoustic pressure microphone O, which is referred to as the XYZO array. Gradient microphones provide additional spatial acoustic information in the amplitude as well as the time difference compared to standard microphone arrays. As a result, the XYZO array offers better performance in a much smaller size. In this system, we, for the first time, deploy and evaluate the XYZO array for the 3-D capture and sound source localization (SSL) Our 3-D audio capture and reproduction utilizes beamforming techniques to reconstruct each beam through the corresponding head-related transfer function (HRTF) to emulate human sound localization based on the filtering effects of the human ear. Meanwhile, 3-D SSL is based on a frequency-domain 3-D spatial search to estimate direction of arrival (DOA). Interested readers can refer to [10] and [11] for more details.

In group teleconferencing, our system not only supports 3-D audio perception but also active speaker detection. With the addition of a depth sensor, the visual content of the active speaker can be accurately tracked and segmented by fusing both audio and visual cues. This will enable a compelling and effective communication experience by immersively rendering the active speaker with the shared contents [see Figure 2 (b)] [12].

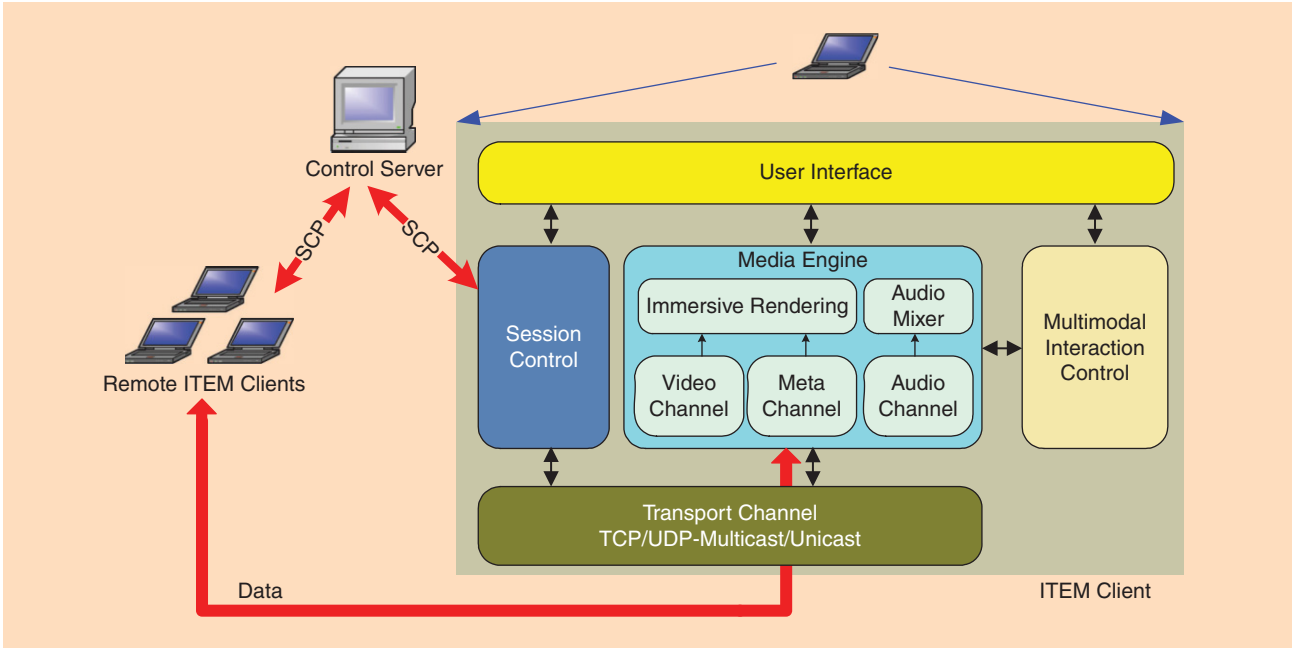
MULTIPARTY NETWORKING STRUCTURE

Media data between two clients in ITEM are exchanged in a peer-to-peer (P2P) manner. To provide the scalability, our system design supports a mechanism to flexibly specify the transmission structure for media data during a session initialization. Currently, we support two architectures for data transmission among multiple users: 1) decentralized ad hoc structure for a small group meeting, and 2) multicast-based structure for one-to-many connections. In the decentralized ad hoc structure, we use a node as a designated translator, which establishes P2P connections to other nodes. Each node in a session will only transmit data to the designated translator, which in turn relays the data back to all nodes. The design is simple and inexpensive compared with a centralized solution with a dedicated multipoint control unit (MCU), while it avoids the computing and bandwidth bottleneck with an increased number of concurrent sessions. Compared with a full-mesh connection, the uplink bandwidth at each node is significantly reduced and independent of the number of users, except for the translator node. Meanwhile, the multicast-based structure is used to support a large number of passive users (e.g., in e-learning), where overlay multicast techniques are employed, if necessary. The current design makes it easy to enhance and extend the networking capabilities in the future.

SYSTEM DESIGN AND IMPLEMENTATION

A modular design approach is used to improve reusability, extensibility, and reconfigurability in various application contexts. Figure 3 depicts the simplified data flows and major components of an ITEM client.

The session control manages the initialization and control of a communication



[FIG3] The simplified diagram of the key components of an ITEM client and the main data flow in the ITEM system architecture.

session including both resource information (e.g., media capabilities, transmission paths) and process management (e.g., initiation, termination) by communicating with a control server through session control protocol (SCP).

The role of the media engine is to process both the local media prior to transmission and the incoming media from remote users for immersive composition. The engine provides seamless audio/video communication among users through a video/audio channel, while shared contents (e.g., documents, media-rich information) are processed through a meta channel. Object-based video processing (e.g., video cutout and object coding) are processed within video channel while audio channel accommodates 3-D sound processing and SSL for spatialized audio and speaker detection. To meet the low-latency requirement and to improve the system performance, multithreading techniques are employed to independently handle different channels and incoming data from each remote user. The segmented user objects from different sources are merged with shared contents from the metachannel in an immersive and interactive manner [Figure 1(d)] through the immersive rendering module. While refreshing the composed frame upon receiving new data from any source is desired for low-latency rendering, such a rendering strategy will overload the CPU usage due to a high rendering frame rate incurred. Thus, a master clock is used to update and render the composed frame at some frame rate (e.g., 30 FPS) without introducing any noticeable delay.

For a more natural interaction with the shared contents, the use of a keyboard and a mouse to interact with the system should be avoided whenever appropriate. With the addition of a depth camera, we employ hand gestures to interact with the system and provide users a comfortable experience through the multimodal interaction control module. Currently, we support several hand gestures to control the shared contents (e.g., paging through the slides or changing the virtual room background).

The transport channel communicates with the session control and media engine modules to create an appropriate connection for data transmission in various channels based on the transmission architectures and data types. The module assigns and manages the list of destinations (e.g., a remote user address or a multicast address, if available). Real-time audio/video data is transmitted using real-time transport protocol (RTP) in conjunction with real-time transport control protocol (RTCP) built on top of user datagram protocol (UDP) packetization.

[TABLE 2] AN ANALYSIS OF LATENCY (MS).

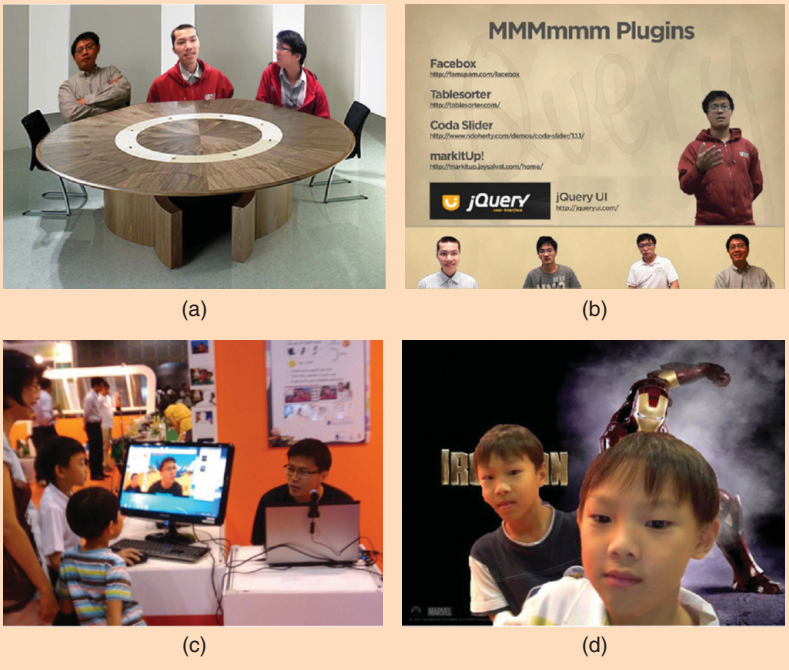
VIDEO OBJECT CUTOUT	38–54
VIDEO OBJECT ENCODING	24–38
NETWORK (JITTER, RELAY, ETC.)	28–43
RENDERING AND DISPLAY	12–30
END-TO-END LATENCY	102–165

SYSTEM PERFORMANCE

We deployed ITEM to conduct multiparty conferencing over the Internet using the decentralized ad hoc structure for the overall system performance evaluation. With the compressed video bit rate of 500 kbits/s, ITEM can easily support up to six participants within a session and run comfortably at a real-time speed. The typical end-to-end latency is reported in Table 2. We also measured the total CPU usage of about 35–40% (about 15% for video object cutout, 10% for video coding/decoding and rendering, 10% for other tasks). With an increased number of participants in a session, we observed about a 10–15% increase in CPU workload [for ten connections over the local area network (LAN) that is consumed by

the decoding and rendering processes]. The results show that our system has better scalability compared with [4], where almost 100% of the CPU is utilized for only about three participants, leading to a significantly lower frame rate with more participants.

In the case of group conferencing with multiple participants at a site, we observed that the 3-D SSL module could detect the angle of a speaker with an accuracy of 6° in real time. Together with the user-tracking information obtained through a depth sensor, this always led to accurate active speaker identification and his/her correct video content extraction. It is worth mentioning that the speaker's video switched too often whenever there was a noise sound source within a short period, which was distracting. To eliminate this undesired effect, we used a threshold to verify the 3-D SSL new detection output. The system only switches to the new speaker video when a consistent detection output is presented within a certain period.



[FIG4] **(a)–(d) The multiparty immersive video communication for an effective business online meeting and fun video chat is shown.**

USER EXPERIENCE

To validate the system from a user perspective, we customized and deployed ITEM in a variety of application scenarios such as business meetings, group teleconferencing, and video chats to conduct extensive user studies.

For business meeting scenarios, we designed several rendering modes to naturally put participants in the same designed virtual meeting space or shared contents, allowing participants the freedom to navigate around the virtual background (Figure 4). The system supports a variety of collaborative contents from slides, documents, media-rich contents, and even desktop windows through the metachannel. We deployed our system for internal trials and collected some useful initial feedback. Users liked the virtual meeting room design, which gave them a strong sense of presence in the same physical space without any distracting, private backgrounds. Although users were not fully satisfied with the current layout in the content-sharing mode when there were many remote participants, this mode was often preferred due to the need of sharing collaborative content during a meeting and its effectiveness for conveying the gesture signals to the shared contents. It was observed that when there was a single participant at a location, the user preferred a simple setup of a Webcam without using a depth camera for gesture-based control.

In the case of group teleconferencing, we employed spatial audio and active speaker detection. Users liked the immersive visual composition of the active speaker into the shared presentation slides. They felt the quality of the meeting was more effective due to how simple it was to keep track of both the slide contents and the speaker video at the same time. Meanwhile, in discussions without any shared contents, users felt that immersive audio could assist them in quickly identifying who was talking and perceiving the direct conversation flow.

In addition to teleconferencing solutions for the business domain, ITEM was also deployed in the consumer space as a lightweight TI video chat application to bring fun, exciting additions to the video chat experience. The system allowed friends and distant family members to experience a sense of togetherness by creating a virtual space to let them see, interact, and do something fun together. Being able to separate users from the background, the application let users apply video effects such as blurring the background or stylizing the user video (see right-most image in Figure 1(d)]. We have demonstrated our application at various technical festivals and conducted user experience surveys (Figure 4). Users liked this new feature of instantly sharing something as fun and exciting as the background while conducting video chats at the same time. They were impressed by the high quality of the real-time segmentation of the foreground from live videos, and felt that his/her Webcam had been transformed into one that was intelligent. Users also enjoyed the immersive video chat features, where they felt more tightly connected with remote friends.

ACKNOWLEDGMENTS

This study is supported by the research grant for the Human Sixth Sense Programme at

(continued on page 136)

[lecture NOTES]

Konstantinos Slavakis,
Seung-Jun Kim,
Gonzalo Mateos,
and Georgios B. Giannakis

Stochastic Approximation vis-à-vis Online Learning for Big Data Analytics

We live in an era of data deluge, where data translate to knowledge and can thus contribute in various directions if harnessed and processed intelligently. There is no doubt that signal processing (SP) is of uttermost relevance to timely big data applications such as real-time medical imaging, smart cities, network state visualization and anomaly detection (e.g., in the power grid and the Internet), health informatics for personalized treatment, sentiment analysis from online social media, Web-based advertising, recommendation systems, sensor-empowered structural health monitoring, and e-commerce fraud detection, just to name a few. Accordingly, abundant chances unfold to SP researchers and practitioners for fundamental contributions in big data theory and practice.

With such big blessings, however, come big challenges. The sheer volume and dimensionality of data often make it impossible to run analytics and traditional batch inferential methods on standalone processing units. With regards to scalability, online data processing is well motivated as the computational complexity of jointly processing the entire data set as a batch is prohibitive. Furthermore, there are many applications in which data themselves are made available in a streaming fashion, meaning that smaller chunks of data are acquired sequentially in time, e.g., nodes of a large network transmitting small blocks of data to a central unit continuously and incoherently in time. As information sources unceasingly produce data in real time, analytics must often be performed on the fly, typically without a chance to revisit previous data. In addition, big data tasks are often subject to stringent time constraints so that a high-quality answer obtained slowly via batch techniques can be less useful than a medium-quality answer that is obtained fast in an online fashion.

Digital Object Identifier 10.1109/MSP.2014.2345536
Date of publication: 15 October 2014

RELEVANCE

In this context, this lecture note presents recent advances in online learning for big data analytics. It is demonstrated that many of these approaches, mostly developed within the machine-learning discipline, have strong ties with workhorse statistical SP tools such as stochastic approximation (SA) and stochastic gradient (SG) algorithms. Important differences and novel aspects are highlighted as well. A key message conveyed is that seminal works on SA, such as by Robbins–Monro and Widrow, which go back half a century, can play instrumental roles in modern online learning tasks for big data analytics. Consequently, ample opportunities arise for the SP community to contribute in this growing and inherently cross-disciplinary field, spanning multiple areas across science and engineering.

THIS LECTURE NOTE PRESENTS RECENT ADVANCES IN ONLINE LEARNING FOR BIG DATA ANALYTICS.

PREREQUISITES

The required background includes basics of linear algebra, probability theory, convex analysis, and stochastic optimization.

STOCHASTIC APPROXIMATION BASICS

Consider the prototypical statistical learning problem in the realm of stochastic optimization (SO) [2], [3] where given a loss function f, one aims at minimizing the expected loss $\mathbb{E}_y\{f(\boldsymbol{w};\boldsymbol{y})\}$, possibly augmented with a complexity-controlling convex regularizer $r(\boldsymbol{w})$, with respect to (w.r.t.) a deterministic parameter (weight) vector $\boldsymbol{w}\in\mathscr{W}$. An example of $r(\boldsymbol{w})$ is the recently popular sparsity-promoting l_1-norm of the $p\times 1$ vector $\boldsymbol{w}$ where $r(\boldsymbol{w})=\|\boldsymbol{w}\|_1:=\sum_{i=1}^{p}|w_i|$. Expectation $\mathbb{E}_y\{\cdot\}$ is taken w.r.t. the typically unknown probability distribution of data $\boldsymbol{y}$ describing, e.g., input-response pairs in a supervised learning setting, and $\mathscr{W}$ denotes a subset of some Euclidean space, introduced here to cover general cases where constraints are imposed on $\boldsymbol{w}$. In lieu of the aforementioned distributional information, given training data $\{\boldsymbol{y}_t\}_{t=1}^{T}$ one can instead opt for solving the empirical risk minimization problem

$$\min_{\boldsymbol{w}\in\mathscr{W}}\frac{1}{T}\sum_{t=1}^{T}f(\boldsymbol{w};\boldsymbol{y}_t)+r(\boldsymbol{w}),\qquad(1)$$

which is an approximation of its ensemble counterpart, specifically $\min_{\boldsymbol{w}\in\mathscr{W}}[\mathbb{E}_y\{f(\boldsymbol{w};\boldsymbol{y})\}+r(\boldsymbol{w})]$. Beyond a purely learning paradigm, one should appreciate the generality offered by (1), since it can subsume, e.g., (constrained) maximum-likelihood problems with f identified as the log-likelihood function and data assumed statistically independent.

In big data settings, T can be huge, potentially infinite in a real-time paradigm where t identifies time instances of data acquisition. Moreover, the search space $\mathscr{W}$ can be excessively high-dimensional with complex structure. This observation justifies the inclusion of a regularizer in (1) to effectively reduce the dimensionality

1053-5888/14©2014IEEE

and/or size of $\mathscr{W}$ and yield parsimonious models that are interpretable and have satisfactory predictive performance. Unsurprisingly, there has been growing interest over the last decade in devising scalable and fast online algorithms for big data learning tasks such as (1).

The main premise of SO is centered around solving the minimization task [cf. (1)]

$$\min_{w\in\mathbb{R}^p}[\varphi(w):=\mathbb{E}_y\{f(w;y)\}] \qquad (2)$$

without having $\mathbb{E}_y\{\cdot\}$ available; see, e.g., [3]. (Compared to (1) and its ensemble version, both $\mathscr{W}$ and the regularizer r have been dropped here for brevity.) Key features present in SO algorithms are: 1) The data comprise a sequence of either dependent vectors with (asymptotically) vanishing covariance, or, independent identically distributed (i.i.d.) realizations $\{y_t\}_{t=1}^T$ of y; and, 2) given (w, y_t), there is a means of obtaining an unbiased "stochastic" gradient estimate $\nabla f(w; y_t)$, so that $\mathbb{E}_y\{\nabla f(w; y_t)\} = \nabla\varphi(w)$.

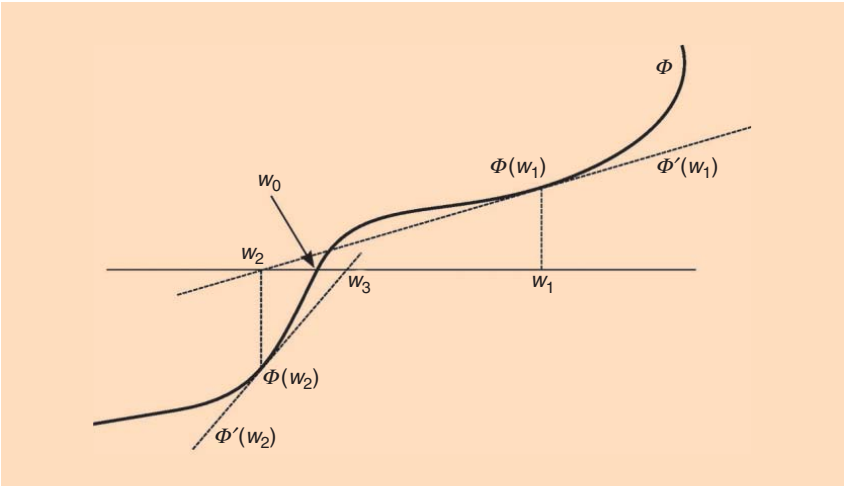
For φ smooth, minimizing φ in (2) amounts to searching for a zero of $\Phi(w):=\nabla\varphi(w)$, i.e., a w_0 such that (s.t.) $\Phi(w_0) = 0$ [3]. The classical Newton–Raphson (N-R) algorithm provides the means to achieve this goal. For w scalar and with $'$ denoting differentiation, the sequence generated by the recursion $w_{k+1}:=w_k-\Phi(w_k)/\Phi'(w_k)=w_k-\varphi'(w_k)/\varphi''(w_k)$ converges under mild conditions to a root of $\Phi(w)$, and thus to a minimizer of $\varphi(w)$. An illustration of the N-R iteration can be seen in Figure 1. Starting from w_1 and using the derivatives $\{\Phi'(w_k)\}_{k=1}^{+\infty}$ in the N-R iteration, the resultant updates $\{w_k\}_{k=2}^{+\infty}$ gradually approach w_0, where $\Phi(w_0) = 0$. Such a simple recursion can be readily extended to the $p\times 1$ vector case as $w_{k+1}:=w_k-H_\varphi^{-1}(w_k)\nabla\varphi(w_k)$, where now $H_\varphi(w_k)$ stands for the $p\times p$ Hessian matrix of φ at w_k with (i, j)th entry $\partial^2\varphi(w_k)/(\partial w_i\partial w_j)$.

Clearly, the N-R algorithm cannot be applied if $\mathbb{E}_y\{\cdot\}$ is not available; e.g., if the probability density function (pdf) of y is unknown, or, when computing $\mathbb{E}_y\{\cdot\}$ entails cumbersome integration over high-dimensional domains. To alleviate this burden, SA through the celebrated Robbins–Monro algorithm relies on the sequence of realizations $\{y_t\}$ and ingeniously uses the instantaneous $\nabla f(w_t; y_t)$ instead of the ensemble $\nabla\varphi(w_k)$ (indexes have been changed from k to t, for time-adaptive operation). With μ_t denoting the step-size, SA generates the online (or stochastic) gradient descent (OGD) iteration

$$w_{t+1} = w_t - \mu_t\nabla f(w_t; y_t), \qquad (3)$$

which "learns" expectations on the fly. This point is better illustrated in "Online Averaging as SA."

Several well-known adaptive SP and online learning algorithms stem from OGD.



[FIG1] The N-R method for finding a w_0 s.t. $\Phi(w_0) = 0$.

LMS AS SA

Consider, for instance, scalar d_t and vector x_t processes that comprise the training data collected in $y_t:=[d_t, x_t^\top]^\top$, and let $f(w; y_t):=(d_t-w^\top x_t)^2/2$, where $\top$ stands for transposition. It can be readily verified that $\nabla f(w; y_t) = (w^\top x_t - d_t)x_t$, and application of OGD yields $w_{t+1} = w_t - \mu_t(w^\top x_t - d_t)x_t$, which is nothing but the celebrated least mean-squares (LMS) algorithm [3].

RLS AS SA

The OGD class can be further broadened by allowing matrix step-sizes $\{M_t\}$ instead of scalar ones $\{\mu_t\}$ to obtain $w_{t+1} = w_t - M_t\nabla f(w_t; y_t)$. To highlight the potential of this extension, consider (jointly) wide sense stationary $\{d_t, x_t\}_{t=1}^\infty$, with $C_{xx}:=\mathbb{E}_x\{x_t x_t^\top\}$, as well as $r_{dx}:=\mathbb{E}_{d,x}\{d_t x_t\}$. It turns out that the solution of $\min_w\mathbb{E}_{d,x}\{(d_t-w^\top x_t)^2\}$ is the linear minimum mean-square error estimator $w_0 = C_{xx}^{-1}r_{dx}$. However, without knowing C_{xx} one relies on the sample average estimate $\hat{C}_t:=(1/t)\sum_{\tau=1}^t x_\tau x_\tau^\top$, and on OGD with $M_t:=(1/t)\hat{C}_t^{-1}$ to obtain

$$w_{t+1} = w_t - \frac{1}{t}\hat{C}_t^{-1}x_t(w_t^\top x_t - d_t) \qquad (4a)$$

$$\hat{C}_{t+1}^{-1} = \frac{t+1}{t}\Big[\hat{C}_t^{-1} - \hat{C}_t^{-1}x_{t+1}x_{t+1}^\top\hat{C}_t^{-1}/(t + x_{t+1}^\top\hat{C}_t^{-1}x_{t+1})\Big], \qquad (4b)$$

where the matrix inversion lemma is applied to carry out efficiently the inversion in (4b). Recursions (4) comprise the well-known recursive least-squares (RLS) algorithm [3].

ONLINE AVERAGING AS SA

The solution of $\min_w\mathbb{E}_y\{\|w-y\|_2^2/2\}$ is clearly $w_0 = \mathbb{E}_y\{y\}$. Following the SA rationale, consider $f(w; y_t):=\|w-y_t\|_2^2/2$. The OGD iteration is $w_{t+1} = w_t - \mu_t(w_t - y_t)$, and if $w_1:=0$ as well as $\mu_t:=1/t$, simple mathematical induction yields $w_{t+1} = (1/t)\sum_{\tau=1}^t y_\tau$, which in accordance with the law of large numbers converges to $w_0 = \mathbb{E}_y\{y\}$ as $t\to+\infty$ [3].

PERFORMANCE OF SA ALGORITHMS

Based on the samples $\{y_t\}$, SA algorithms produce estimates $\{w_t\}$ that allow for estimation, tracking, and out-of-sample inference tasks, such as prediction. Performance analysis of SA schemes has leveraged advances in martingale and ordinary differential equation theories to establish, e.g., in the stationary case, convergence of $\{w_t\}$ to a time-invariant w_0 in probability, or with probability one, or in the mean-square sense [3]. In this stationary setting, convergence of OGD requires step-sizes selected to diminish with a certain rate. Specifically, $\{\mu_t\}$ must satisfy 1) $\mu_t \geq 0$, 2) $\lim_{t\to\infty}\mu_t = 0$, and 3) $\sum_{t=1}^{\infty}\mu_t = +\infty$. Clearly, 1)–3) are satisfied for $\mu_t := 1/t$, which vanishes as $t \to +\infty$ but not too fast so that 3) enables $\{w_t\}$ to reach asymptotically the desired w_0.

Departing from the standard route of SA convergence analysis [3], recent results take advantage of convexity if it is present in the objective function. Specifically for convex costs, the OGD recursion (3) generalizes to: $w_{t+1} = \mathcal{P}_{\mathcal{W}}[w_t - \mu_t \nabla f(w_t; y_t)]$, where $\mathcal{P}_{\mathcal{W}}(w) := \operatorname{argmin}_{w' \in \mathcal{W}} \|w - w'\|_2$ stands for the projection mapping onto a closed and convex constraint set $\mathcal{W}$. For φ differentiable and strongly convex with index $c > 0$, it holds that $\varphi(w') \geq \varphi(w) + (w' - w)^{\top}\nabla\varphi(w) + (c/2)\|w' - w\|_2^2$, for all (w, w). With step-sizes selected as $\mu_t := \mu/t$ with $\mu > 1/(2c)$, and for bounded stochastic gradients as in $\sup_w \mathbb{E}_y\left\{\|\nabla f(w; y)\|_2^2\right\} \leq \Delta$, it can be verified that the error $\mathbb{E}_y\left\{\|w_t - w_0\|_2^2\right\}$, where $w_0 = \operatorname{argmin}_{w\in\mathcal{W}}\mathbb{E}_y\{f(w; y)\}$, satisfies the following finite-sample bound [2]:

$$\mathbb{E}_y\{\|w_t - w_0\|_2^2\} \leq \frac{Q(\mu)}{t},$$

with

$$Q(\mu) := \max\{\mu^2\Delta^2/(2\mu c - 1), \|w_1 - w_0\|^2\}.$$

If, in addition, $\nabla\varphi$ is L-Lipschitz continuous, i.e., $\|\nabla\varphi(w) - \nabla\varphi(w')\|_2 \leq L\|w - w'\|_2$, $\forall w, w'$, then a similar finite-sample bound holds also for the sequence of function values $\{\varphi(w_t)\}$ [2]

$$\mathbb{E}_y\{\varphi(w_t) - \varphi(w_0)\} \leq \frac{LQ(\mu)}{2t},$$

where expectation is taken over $\{w_t\}$, which involves stochastic gradients.

Performance analysis of SA algorithms deals with convergence of $\{w_t\}$, whereas the online convex optimization framework outlined in a subsequent section starts from (1), invokes fewer or no assumptions on the underlying pdfs, and asserts convergence of the costs $\{f(w_t; y_t)\}$, rather than $\{w_t\}$.

THE OCO FRAMEWORK CAN BE VIEWED AS A MULTIROUND GAME BETWEEN A PLAYER (LEARNER) AND AN ADVERSARY.

Recently, SA was combined with the alternating direction method of multipliers (ADMM) which is attractive for offline optimization of composite costs [4]. The resultant SA-ADMM solver [5] is suitable for online optimization of composite costs such as $\min_{w\in\mathcal{W}}[\mathbb{E}_y\{f(w; y)\} + r(w)]$, in a fully distributed fashion—an operational mode that is highly desirable for big data applications.

SEQUENTIAL OPTIMIZATION AND DATA SKETCHING

The importance of sequential optimization along with the attractive operation of random sampling (also known as *sketching*) of big data will be illustrated in this subsection in the context of the familiar LS task:

$$\min_{w\in\mathbb{R}^p}\left[\frac{1}{2T}\|X^{\top}w - d\|_2^2 = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{2}(x_t^{\top}w - d_t)^2\right], \quad (5)$$

where $X := [x_1, \ldots, x_T]$ denotes the $p \times T$ matrix that gathers all available regressor or input vectors, and $d := [d_1, \ldots, d_T]^{\top}$ the $T \times 1$ vector of desired outputs (responses). Although irrelevant to the minimization in (5), the normalization with T is included to draw connections with (1). In this sense, the loss function becomes $f(w; y_t) = (x_t^{\top}w - d_t)^2/2$, with $y_t := [d_t, x_t^{\top}]^{\top}$, and its gradient $\nabla f(\cdot; y_t)$ is Lipschitz continuous with constant $L_t = \|x_t\|_2^2$. Different from the previous discussion, here T is fixed, and "online" means processing $\{d_t, x_t\}_{t=1}^{T}$ sequentially.

Searching for a solution w_0 of (5) requires eigen-decomposition of $XX^{\top}$, which incurs complexity $\mathcal{O}(Tp^2)$. Alternatively, the standard gradient descent recursion $w_{k+1} = w_k - \mu_k(XX^{\top}w_k - Xd)$ entails $\mathcal{O}(p^2)$ computations per iteration k. Both cases are prohibitive in big data settings where the number of samples, T, is massive and/or the data dimensionality, p, can be huge. To surmount these obstacles, solving for w_0 can rely on subsampling (also known as *sketching* to obtain a subset of) the rows of $X^{\top}$, along with the corresponding entries of d, to reduce complexity w.r.t. T, while visiting them in a sequential fashion that scales linearly with p.

Kaczmarz's algorithm, a special case of the projections onto convex sets (POCS) method [6], produces a sequence of estimates $\{w_k\}$ to solve (5). For an arbitrary initial estimate w_1, the kth iteration of Kaczmarz's algorithm selects a row $t(k)$ of $X^{\top}$, together with the corresponding entry $d_{t(k)}$, and projects the current estimate w_k onto the set of all minimizers $\mathcal{H}_{t(k)} := \{w | x_{t(k)}^{\top}w = d_{t(k)}\}$ of $f(w; y_{t(k)})$, which is nothing but a hyperplane (a closed and convex set). Hence, the $(k+1)$st estimate is

$$w_{k+1} := \mathcal{P}_{\mathcal{H}_{t(k)}}(w_k) = w_k - \frac{x_{t(k)}^{\top}w_k - d_{t(k)}}{\|x_{t(k)}\|_2^2}x_{t(k)}, \quad (6)$$

where $\mathcal{P}_{\mathcal{H}_{t(k)}}$ stands for the projection mapping onto $\mathcal{H}_{t(k)}$. Notice here that the complexity of computing $\mathcal{P}_{\mathcal{H}_{t(k)}}(w_k)$ scales linearly with p. If every (d_t, x_t) is visited infinitely often, then under several conditions (6) converges to a solution of (5) [6]. Visiting each (d_t, x_t) a large number of times is prohibitive with big data since T can be excessively large. In contrast, poor selection of rows can slow down convergence; see Figure 2. Nevertheless, randomly drawing rows with equal probabilities has been shown empirically to accelerate convergence relative to cyclic revisits of rows [7]. Judicious sampling schemes can yield further speedups, as highlighted in "Accelerating SG via Nonuniform Sampling."

LEARNING VIA ONLINE CONVEX OPTIMIZATION

Recently, online learning approaches based on a online convex optimization (OCO) framework have attracted significant attention, as they do not require elaborate statistical models for data and yet can provide robust performance guarantees. This is true even under an adversarial setup, where the data sequence $\{\boldsymbol{y}_t\}$ may be generated strategically in reaction to the learner's iterates $\{\boldsymbol{w}_t\}$, as in the humans-in-the-loop applications such as the Web advertising optimization.

The OCO framework can be viewed as a multiround game between a player (learner) and an adversary [10]. In the context of the learning formulation in (1), the learner plays an action $\boldsymbol{w}_t \in \mathcal{W}$ in round t, where $\mathcal{W}$ is assumed to be closed and convex. Based on the action $\boldsymbol{w}_t$ that the player took, the adversary provides some feedback information $\mathcal{F}_t$, manifested in the data (feature) vector $\boldsymbol{y}_t$, based on which a convex loss function $\mathcal{L}_t: \mathcal{W} \to \mathbb{R} \cup \{+\infty\}$ is constructed, such as $\mathcal{L}_t(\boldsymbol{w}) := f(\boldsymbol{w}; \boldsymbol{y}_t) + r(\boldsymbol{w})$. The learner then suffers the loss at $\boldsymbol{w}_t$, specifically, $\mathcal{L}_t(\boldsymbol{w}_t)$. The overall process is depicted in Figure 3.

The learner's goal is to minimize the so-termed *regret* $R(T)$ over T rounds, defined as

$$R(T) := \sum_{t=1}^{T} \mathcal{L}_t(\boldsymbol{w}_t) - \min_{\boldsymbol{w} \in \mathcal{W}} \sum_{t=1}^{T} \mathcal{L}_t(\boldsymbol{w}), \tag{7}$$

which captures how much worse the learner performed cumulatively, compared to the case where a single best action is chosen with the knowledge of the entire sequence of cost functions $\{\mathcal{L}_t\}_{t=1}^{T}$ in hindsight. In particular, OCO aims at producing a sequence $\{\boldsymbol{w}_t\}$, which gives rise to sublinear regret, that is, the one with $R(T)/T \to 0$ as T grows. The key question now for the learner is how to pick $\boldsymbol{w}_t$ in each round t.

OCO ALGORITHMS AND PERFORMANCE

An important class of algorithms that can achieve the desired sublinear regret bound is based on the online mirror descent (OMD) iteration [11]. In a nutshell, the method minimizes a first-order approximation of $\mathcal{L}_t$ at the current iterate $\boldsymbol{w}_t$, while encouraging the search in the vicinity of $\boldsymbol{w}_t$. Specifically, OMD computes the next round iterate $\boldsymbol{w}_{t+1}$ as

$$\boldsymbol{w}_{t+1} = \arg\min_{\boldsymbol{w} \in \mathcal{W}} (\boldsymbol{w} - \boldsymbol{w}_t)^{\top} \mathcal{L}_t'(\boldsymbol{w}_t) + \frac{1}{\mu} D_{\psi}(\boldsymbol{w}, \boldsymbol{w}_t), \tag{8}$$

where $'$ denotes a (sub)gradient of a function, $\mu > 0$ is a learning rate parameter, and $D_{\psi}(\boldsymbol{w}, \boldsymbol{v})$ is the Bregman divergence associated with a continuously differentiable and strongly convex ψ, defined as

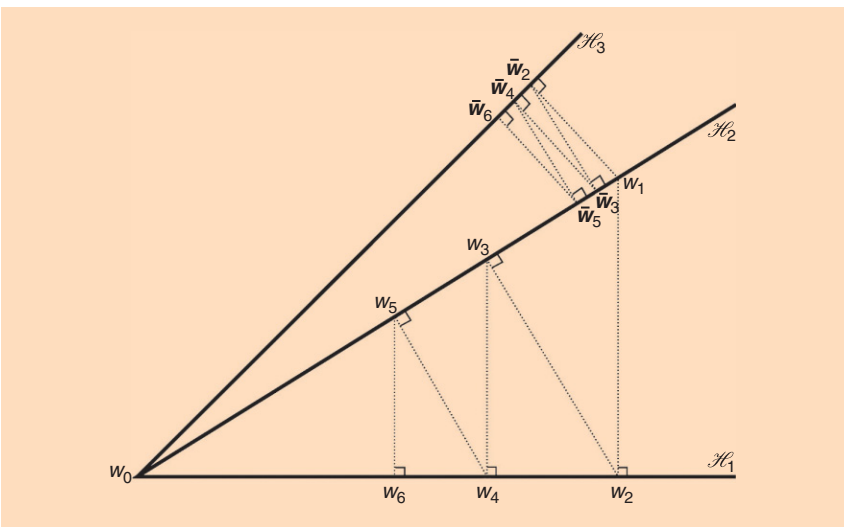
$$D_{\psi}(\boldsymbol{w}, \boldsymbol{v}) := \psi(\boldsymbol{w}) - \psi(\boldsymbol{v}) - (\boldsymbol{w} - \boldsymbol{v})^{\top} \nabla \psi(\boldsymbol{v}). \tag{9}$$

ACCELERATING SG VIA NONUNIFORM SAMPLING

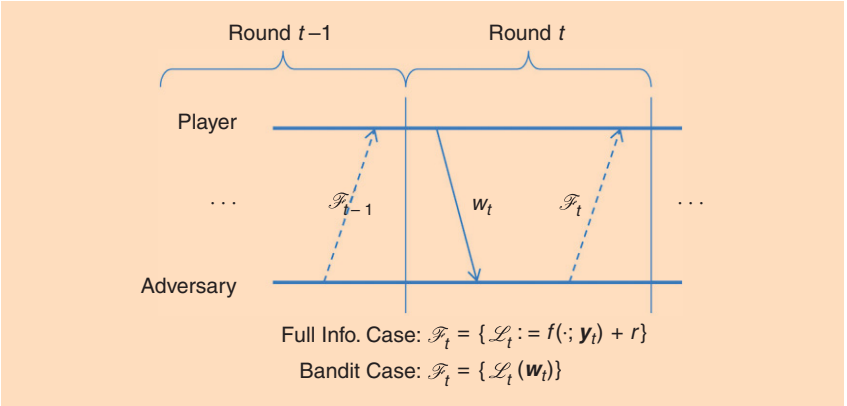
In the noiseless case ($\boldsymbol{X}^{\top}\boldsymbol{w} = \boldsymbol{d}$), randomly drawing rows in proportion to their Lipschitz constants L_t is known to provide finite-sample bounds of the form [7]

$$\mathbb{E}_{\mathcal{R}}\{\|\boldsymbol{w}_k - \boldsymbol{w}_0\|_2^2\} \le [1 - \kappa(\boldsymbol{X})^{-2}]^k \|\boldsymbol{w}_1 - \boldsymbol{w}_0\|_2^2,$$

where $\kappa(\boldsymbol{X})$ stands for the condition number of $\boldsymbol{X}$, and $\mathbb{E}_{\mathcal{R}}\{\cdot\}$ denotes expectation w.r.t. the distribution over which $\{d_t, \boldsymbol{x}_t\}$ are selected. The previous nonuniform sampling scheme yields better convergence rates than those resulting from uniform sketching [7]. More information on (non)uniform sketching and its application to SG descent methods can be found in [8] and [9].



[FIG2] Kaczmarz's algorithm for three hyperplanes $\{\mathcal{H}_t\}_{t=1}^{3}$ with the nonempty intersection $\{\boldsymbol{w}_0\} = \cap_{t=1}^{3} \mathcal{H}_t$. Row (hyperplane) selection affects convergence rate; $\{\boldsymbol{w}_k\}$, which alternates between $\mathcal{H}_1$ and $\mathcal{H}_2$ approaches $\boldsymbol{w}_0$ faster than $\{\bar{\boldsymbol{w}}_k\}$, which is generated via $\mathcal{H}_2, \mathcal{H}_3$.



[FIG3] OCO as a multiround game.

In the special case of using $\psi(\boldsymbol{w}) := \|\boldsymbol{w}\|_2^2/2$, the corresponding $D_\psi(\boldsymbol{w},\boldsymbol{v}) = \|\boldsymbol{w}-\boldsymbol{v}\|_2^2/2$, and the OMD update in (8) boils down to OGD [10], establishing an immediate link between OCO and SA. In general, a judicious choice of ψ can capture the structure of the search space $\mathscr{W}$, leading to an efficient update formula for $\boldsymbol{w}_t$. For example, when $\mathscr{W}$ is the probability simplex, i.e., $\mathscr{W} = \{\boldsymbol{w} \mid w_i \geq 0, \sum_i w_i = 1\}$, setting $\psi(\boldsymbol{w}) := \sum_i w_i \log w_i$ in (8) and (9) yields the exponentiated gradient algorithm, which obviates the need to explicitly impose the probability simplex constraints [10]. Aiming at an efficient use of prior information on $\boldsymbol{w}$, a notable generalization of OMD is offered by the "COMID Algorithm."

Both COMID and OMD (which is a special case of COMID) can attain sublinear regret bounds. Specifically, $R(T) = \mathcal{O}(\sqrt{T})$ in general, and the bound becomes $\mathcal{O}(\log T)$ when $\mathscr{L}_t$ is strongly convex [10], [12]. Noteworthy differences between SA and OCO are outlined in "SA vis-à-vis OCO."

ONLINE LEARNING WITH BANDIT FEEDBACK

The bandit setup of OCO refers to the case where the feedback $\mathscr{F}_t$ from the adversary does not explicitly reveal the cost function $\mathscr{L}_t(\cdot)$ but only the sample cost $\mathscr{L}_t(\boldsymbol{w}_t)$ due to action $\boldsymbol{w}_t$; refer also to Figure 3. For example, $\boldsymbol{w}_t$ may represent the advertising budget allocated to different media channels, and $\mathscr{L}_t(\boldsymbol{w}_t)$ the corresponding overall cost (e.g., the total advertising expense minus the resulting income). In this case, it may be difficult to know the explicit form of $\mathscr{L}_t$, but $\mathscr{L}_t(\boldsymbol{w}_t)$ can be easily observed.

SEQUENTIAL OR ONLINE LEARNING SCHEMES TOGETHER WITH RANDOM SAMPLING OR DATA SKETCHING METHODS ARE EXPECTED TO PLAY A PRINCIPAL ROLE IN SOLVING LARGE-SCALE OPTIMIZATION TASKS.

The idea of bandit OCO is to estimate the necessary gradient using SA in the context of OGD. Specifically, a key observation is that if one can evaluate a function $f: \mathbb{R}^p \to \mathbb{R}$ at $\boldsymbol{w}$ perturbed by a small $\delta\boldsymbol{v}$, where $\delta > 0$ and $\boldsymbol{v}$ is uniformly distributed on the surface of a unit sphere, then $(p/\delta) f(\boldsymbol{w}+\delta\boldsymbol{v})\boldsymbol{v}$ offers an unbiased estimate of the gradient at $\boldsymbol{w}$ of a locally smoothed version of f [14]. Thus, plugging this noisy gradient directly into the OGD update in the spirit of SA, one can still establish a sublinear regret bound. However, the best bound found in [14] is $\mathcal{O}(T^{3/4})$, slower than the $\mathcal{O}(\sqrt{T})$-bound for the full information case, illustrating the price to pay for the lack of information.

LESSONS LEARNED AND FUTURE AVENUES

This lecture note offered a short exposition of recent advances in online learning for big data analytics, highlighting their differences and many similarities with prominent statistical SP tools such as SA and SO methods. It was demonstrated that the seminal Robbins–Monro algorithm, the workhorse behind several classical SP tools such as the LMS and RLS algorithms, carries rich potential for solving large-scale learning tasks under low computational budget. It was also explained that sequential or online learning schemes together with random sampling or data sketching methods are expected to play a principal role in solving large-scale optimization tasks. A short description of the OCO framework revealed its flexibility on the variety of optimization tasks that can be accommodated, including scenarios where data are provided in an adversarial fashion or with limited feedback. Yet, such a flexibility comes at a price; OCO-based statistical analysis refers mostly to bounds of the regret cost. Based on the common ground between OCO and SA, OCO can only benefit from the rich theoretical armory of SA, e.g., the martingale theory, where results pertain also to convergence of the primal (random) variables of the optimization task at hand. Vice versa, SA can also profit from the powerful toolbox of convex analysis, the engine behind OCO, for establishing strong analytical claims in the big data context. In closing, Figure 4 depicts the unique and complementary strengths SA, SO, and OCO offer to online learning, as well as adaptive SP theory and big data applications.

COMID ALGORITHM

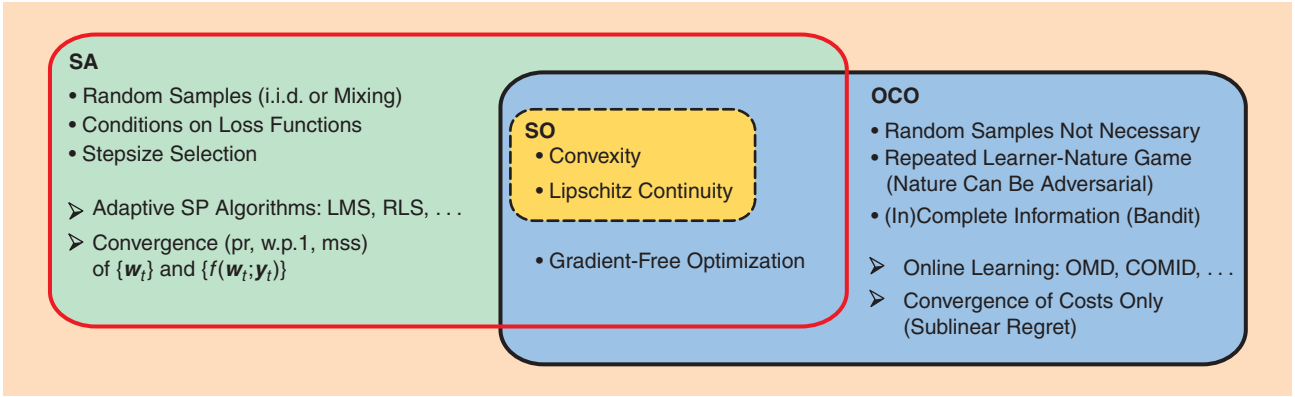
While the OMD update provides a computationally attractive solution to (1), the linearization involved often defeats one of the purposes of the regularizer r, which is to promote a priori known structure in the solution. For example, setting $r(\boldsymbol{w})$ proportional to the ℓ_1-norm of $\boldsymbol{w}$ encourages sparsity in $\boldsymbol{w}$. To properly capture such a benefit, one has to respect the composite structure of $\mathscr{L}_t$, which decomposes into the data-dependent part $f_t(\boldsymbol{w}) := f(\boldsymbol{w};\boldsymbol{y}_t)$ and the invariant part $r(\boldsymbol{w})$ [12], [13]. In particular, the composite objective mirror descent (COMID) algorithm relies on [12]

$$\boldsymbol{w}_{n+1} = \arg\min_{\boldsymbol{w}\in\mathscr{W}} (\boldsymbol{w}-\boldsymbol{w}_t)^\top f_t'(\boldsymbol{w}_t) + r(\boldsymbol{w}) + \frac{1}{\mu} D_\psi(\boldsymbol{w},\boldsymbol{w}_t), \tag{S1}$$

where it is seen that the regularizer is not linearized.

SA VIS-A-VIS OCO

Compared to the SA approaches, the OCO framework does not require stochastic models. This is a salient departure from typical SA setups, since the regret bounds are guaranteed even for $\{\boldsymbol{y}_t\}$ that may have been generated adversarially, i.e., with $\boldsymbol{y}_t$ arbitrary correlated to past actions $\{\boldsymbol{w}_\tau\}_{\tau\leq t}$ and past data $\{\boldsymbol{y}_\tau\}_{\tau<t}$. On the other hand, the bounds pertain to convergence of the sequence of costs rather than the iterates $\{\boldsymbol{w}_t\}$ themselves. Nonetheless, building upon the flexibility offered by OCO, certain limited feedback learning tasks are feasible as elaborated in the "Online Learning with Bandit Feedback" section, where, interestingly, the SA ideas prove instrumental once again.



[FIG4] SA/SO vis-à-vis OCO: features and implications.

ACKNOWLEDGMENTS

The work in this lecture note was supported by the National Science Foundation grants EARS-1343248 and EAGER-1343860, and the MURI grant AFOSR FA9550-10-1-0567.

AUTHORS

Konstantinos Slavakis (kslavaki@umn.edu) is a research associate professor in the Department of Electrical and Computer Engineering and Digital Technology Center, University of Minnesota, United States.

Seung-Jun Kim (sjkim@umbc.edu) is an assistant professor in the Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County, United States.

Gonzalo Mateos (gmateosb@ece.rochester.edu) is an assistant professor in the Department of Electrical and Computer Engineering, University of Rochester, New York, United States.

Georgios B. Giannakis (georgios@umn.edu) is a professor in the Department of Electrical and Computer Engineering and director of the Digital Technology Center, University of Minnesota, Minneapolis, United States.

REFERENCES

[1] K. Slavakis, G. B. Giannakis, and G. Mateos, "Modeling and optimization for big data analytics," *IEEE Signal Processing Mag.*, vol. 31, no. 5, pp. 18–31, Sept. 2014.

[2] A. Nemirovski, A. Juditski, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM J. Optim.*, vol. 19, no. 4, pp. 1574–1609, 2009.

[3] H. J. Kushner and G. G. Yin, *Stochastic Approximation Algorithms and Applications*. New York: Springer, 1997.

[4] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, 2nd ed. Belmond, MA: Athena Scientific, 1997.

[5] I. D. Schizas, G. Mateos, and G. B. Giannakis, "Distributed LMS for consensus-based in-network adaptive processing," *IEEE Trans. Signal Processing*, vol. 57, no. 6, pp. 2365–2381, 2009.

[6] H. H. Bauschke and J. M. Borwein, "On projection algorithms for solving convex feasibility problems," *SIAM Review*, vol. 38, no. 3, pp. 367–426, Sept. 1996.

[7] T. Strohmer and R. Vershynin, "A randomized Kaczmarz algorithm with exponential convergence," *J. Fourier Anal. Appl.*, vol. 15, no. 2, pp. 262–278, 2009.

[8] D. Needell, N. Srebro, and R. Ward. (2013, Feb.). Stochastic gradient descent and the randomized Kaczmarz algorithm. ArXiv e-prints. [Online]. Available: arXiv:1310.5715v2

[9] A. Nedić and D. P. Bertsekas, "Incremental subgradient methods for nondifferentiable optimization," *SIAM J. Optim.*, vol. 12, no. 1, pp. 109–138, 2001.

[10] S. Shalev-Shwartz, "Online learning and online convex optimization," *Found. Trends Mach. Learn.*, vol. 4, no. 2, pp. 107–194, Mar. 2012.

[11] A. Beck and M. Teboulle, "Mirror descent and nonlinear projected subgradient methods for convex optimization," *Oper. Res. Lett.*, vol. 31, no. 3, pp. 167–175, 2003.

[12] J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari, "Composite objective mirror descent," in *Proc. Conf. Learning Theory*, Haifa, Israel, June 2010, pp. 14–26.

[13] L. Xiao, "Dual averaging methods for regularized stochastic learning and online optimization," *J. Mach. Learn. Res.*, vol. 11, pp. 2543–2596, Oct. 2010.

[14] A. D. Flaxman, A. T. Kalai, and H. B. McMahan, "Online convex optimization in the bandit setting: Gradient descent without a gradient," in *Proc. ACM-SIAM Symp. Discrete Algorithms*, Vancouver, Jan. 2005, pp. 385–394.

[SP]

life SCIENCES

Raúl Chávez-Santiago
and Ilangko Balasingham

Ultrawideband Signals in Medicine

In the last decade, the utilization of radio technology for a variety of clinical and medical applications has increased dramatically. Advances in microelectronics have enabled the miniaturization and integration of biomedical sensors and radio transceivers into single units. Such wireless sensors can be worn or implanted, and they facilitate the continuous collection and transmission of physiological signals. The centrally coordinated or distributed interconnection of wireless biomedical sensors, referred to as a *body area network* (*BAN*), has already been standardized in IEEE Standard 802.15.6-2012. Besides narrowband (NB), IEEE Standard 802.15.6-2012 supports the use of ultrawideband (UWB) radio interfaces for wearable sensors. The U.S. Federal Communications Commission (FCC) defines UWB signals to be those with fractional bandwidth exceeding 20% of the center frequency, or alternatively, a bandwidth greater than 500 MHz. A number of techniques can be used to generate UWB signals; the most widely used is the one referred to as *impulse radio* (*IR*) and consists of transmitting very narrow pulses in the time domain, commonly in the order of a few nanoseconds with fast rise times reaching 50 ps [1]. Another common approach to generate UWB signals is subcarrier aggregation in orthogonal frequency-domain multiplexing (OFDM), which is generally used for short-range nonmedical indoor communications. The very large bandwidth enables high-data-rate communications. In the presence of noise and power constraints, UWB allows trading a section of the bandwidth for power according to the Shannon–Hartley theorem. This means UWB communication systems can operate with ultralow power and low signal-to-noise ratio (SNR) using different modulation and coding strategies. UWB signals have other inherent characteristics that make them suitable for the wireless interface of wearable biomedical sensors. Noise-like behavior due to the extremely low maximal effective isotropically radiated power (EIRP) spectral density of –41.3 dBm/MHz makes UWB signals difficult to detect by NB systems and robust against jamming, potentially rescinding the need for complex encryption algorithms. Additionally, UWB signals do not represent a threat to patients' safety [2]. Much research has been done toward the use of BAN technology for ubiquitous monitoring of patients suffering from chronic diseases. Nevertheless, the commercial use of UWB for BAN has been limited in part by the FCC regulations, which restrict the communication applications to 3.1–10.6 GHz where propagation conditions in the body environment are rather unfavorable.

ONE OF THE MEDICAL DEVICES THAT WOULD GREATLY BENEFIT FROM THE HIGH DATA RATES PROVIDED BY UWB INTERFACES IS THE WIRELESS CAPSULE ENDOSCOPE.

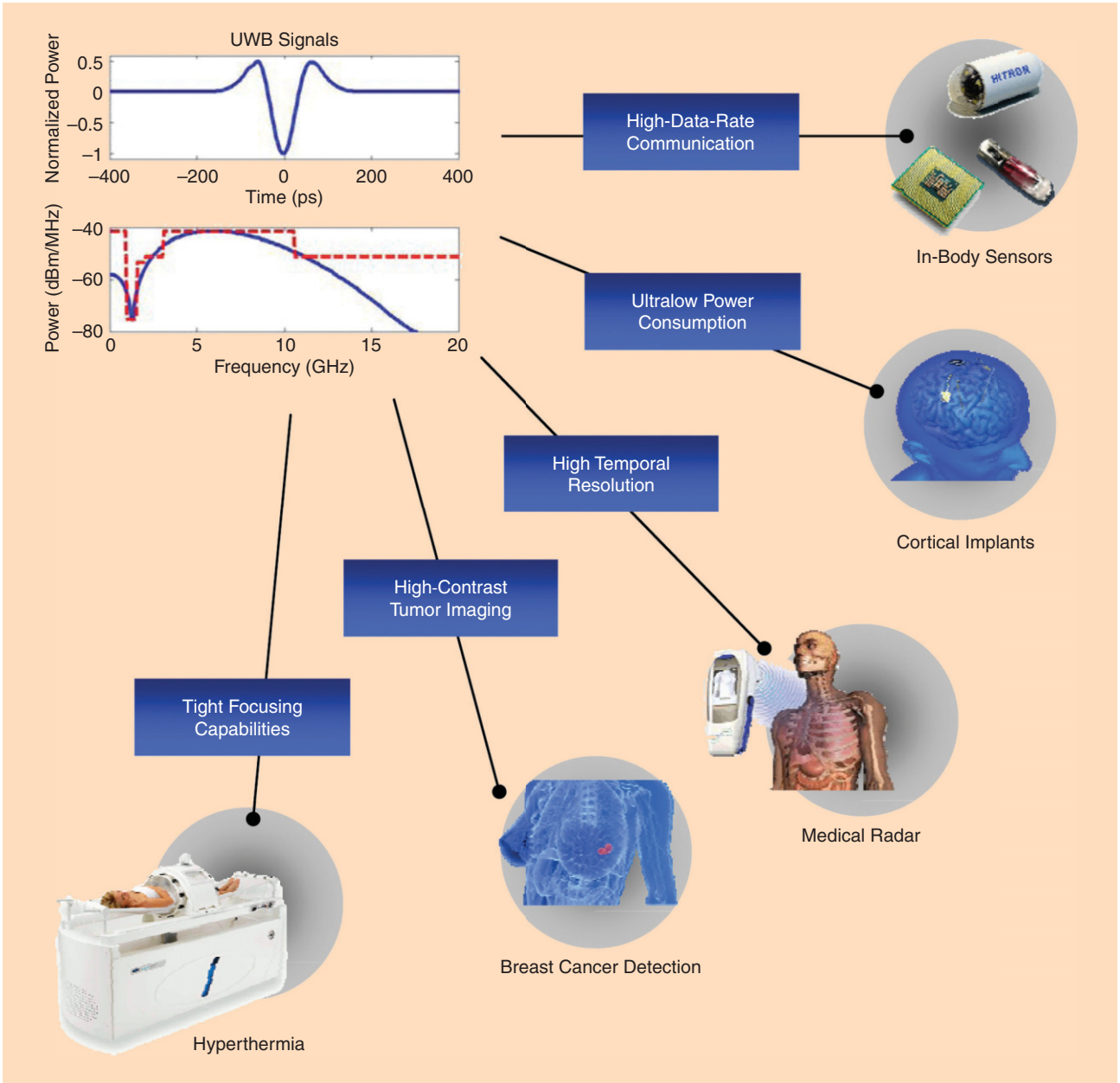
In spite of the scarcity of commercial products, UWB has vast potential to play an important role in BAN technology as a communication interface. However, other areas of interest like medical sensing and medical imaging can benefit from the use of UWB signals too as shown in Figure 1. For instance, the UWB medical radar [1] can enable noninvasive monitoring of respiratory and cardiac motion, blood pressure estimation, and medical imaging for early cancer detection. These are but a few of the applications that UWB technology can have in medical practice. In this column, we survey the different uses of UWB in medicine and health care from a signal processing perspective.

IMPLANTABLE SENSORS

The use of implantable sensors transmitting physiological signals to an external unit for processing and visualization can enable the personalized management of chronic diseases. Although IEEE Standard 802.15.6-2012 does not consider the use of UWB for in-body biomedical devices, the feasibility of high-data-rate communication for implants using this technology has been demonstrated [3]. In an in vivo experiment on a living porcine subject, a 1-Mbit/s wireless UWB link was established for a maximal implantation depth of 12 cm and a bit error rate (BER) of 10^{-2}. The data rate obtained in this experiment can be improved with a variety of practical solutions like the use of a dielectric matching layer and by increasing the transmit power, as long as the regulations for nonionizing radiation exposure inside the body, i.e., specific absorption rate (SAR), and electromagnetic (EM) emissions outside the body are respected [4]. Computer simulations and simple in vitro experiments have predicted the feasibility of transmitting up to 100 Mbit/s for UWB implant communications. However, further research including more elaborated in vivo experiments is needed to verify whether a sensor implanted into the body at a depth of several centimeters can communicate at such high data rates. The

Digital Object Identifier 10.1109/MSP.2014.2340234
Date of publication: 15 October 2014

1053-5888/14©2014IEEE



[FIG1] Some UWB signals' characteristics and possible applications in medicine. (Thumbnail images courtesy of the BSD Medical Corporation, Micrima Ltd., Lawrence Livermore National Laboratory, and Jinan Nefisa Medical Trade Co. Ltd.)

large attenuation suffered by UWB signals propagating through biological tissues is the main hindering factor.

WIRELESS CAPSULE ENDOSCOPES

One of the medical devices that would greatly benefit from the high data rates provided by UWB interfaces is the wireless capsule endoscope (WCE). This device transmits still images and, in some cases, real-time video from inside the gastrointestinal (GI) tract to an external receiver for subsequent processing and clinical analysis. The WCE facilitates the diagnosis of gastroenterological disorders in parts of the GI tract that other endoscopic techniques cannot visualize. The large bandwidth of UWB can enable the transmission of real-time video with higher resolution than currently available. Hence, techniques to counter the large propagation losses should be devised. Moreover, the design of an UWB-WCE communication system has to take into consideration the particularities of the radio channel. For instance, in [4] an UWB on-body single-branch correlation receiver using a suboptimal signal template created from simulations of the WCE radio channel is described. This simple receiver structure does not require channel estimation thereby simplifying its implementation. Spatial diversity reception techniques can further counter the effects of attenuation. With an appropriate arrangement of diversity receivers around the body,

improvement of 10 dB on the E_b/N_0, where E_b is the bit energy and N_0 is the noise power spectral density, at a BER of 10^{-3} can be obtained with the use of maximum ratio combining (MRC).

CORTICAL IMPLANTS

Cortical implants are biocompatible multielectrode arrays in direct connection with the cerebral cortex of the brain. Contrasting other medical implants, the implantation depth of these devices rarely exceeds 3 cm. They are used to receive or transmit neural signals and can provide different benefits depending on their design and placement. For example, a brain–machine interface (BMI) can utilize a multichannel neural recording system to extract a signal from a paraplegic patient's brain to allow the control of some form of external hardware like a prosthetic arm. Other cortical implants are used to return sensation through an implanted signal gathered by an external sensor, e.g., partial restoration of vision by directly stimulating the visual cortex. With the complexity of the brain, a myriad possible applications for these implants are envisaged and the amount of research in this field is growing rapidly.

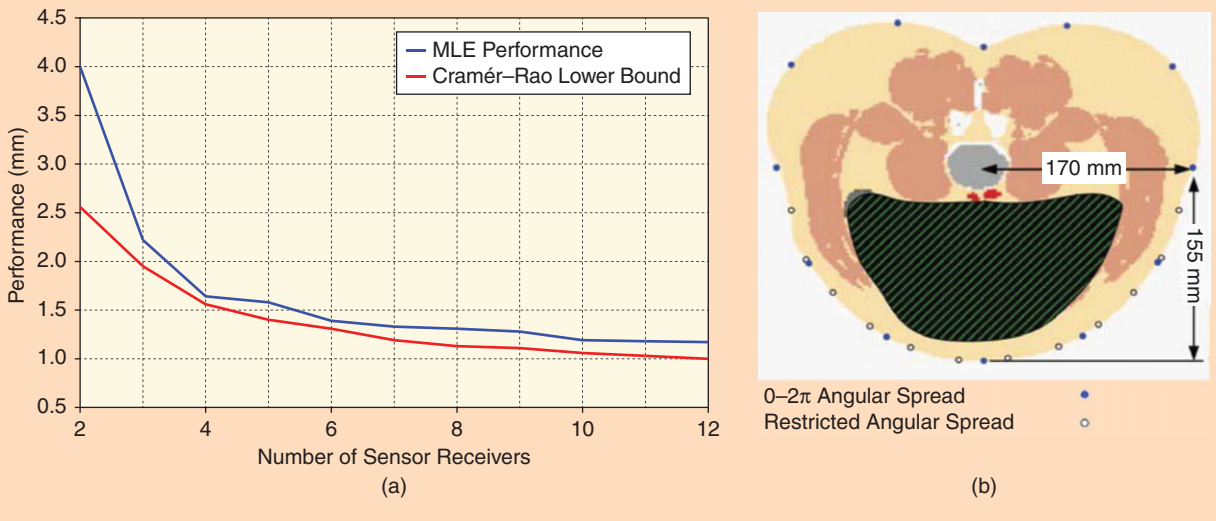
Depending on the application, a cortical implant may require a high-data-rate communication link. Thus, a single-chip multichannel neural recording system has been developed by integrating eight 16-channel front-end blocks, a data serializing circuit, a digital signal processor (DSP) unit for spike detection and feature extraction, a digital multiplexer (MUX), an encoder, and a UWB transmitter providing up to 90 Mbit/s [5]. A more recent (2012) simulation-based study suggested that up to 125 Mbit/s can be achieved with the use of inductive coils in the near-field domain. Besides the high data rate, the use of UWB interfaces can significantly reduce the overall size and power consumption, a key requirement for future and more sophisticated cortical implants.

LOCALIZATION AND TRACKING

The high temporal resolution of UWB signals can enable the localization and real-time tracking of objects with high accuracy. UWB has been demonstrated effective for the continuous tracking of persons in indoor environments, which may be useful for assisted living situations for patients suffering from dementia. The benefits of UWB could be also exploited for the accurate localization and tracking of medical in-body devices like the WCE.

Localization is typically done with a two-step method that uses a combination of physical measurement and estimation. These measurements can include received signal strength (RSS), time of arrival (TOA), time difference of arrival (TDOA), or angle of arrival (AoA). However, accurate localization using one or more of these parameters in a lossy propagation medium like the human body is significantly more difficult than in free space because of the frequency-dependent attenuation and changes in EM wave velocities. These problems can be reduced with a combination of uncorrelated receivers, diversity principles, and a priori channel state information. As an example, Figure 2 shows the simulation-based three-dimensional (3-D) localization performance of a maximum likelihood estimator (MLE) as a function of the number of sensor receivers for a WCE scenario. The MLE used RSS measurements of UWB signals and distance calculations performed using an appropriate path loss model, which can reportedly provide millimeter accuracy [6].

The tracking of a WCE traveling inside the GI tract poses additional challenges because of the device's arbitrary maneuvers like sudden stops and starts. Thus, for this problem the classical tracking algorithms using Kalman filter should be extended, e.g., by the use of multimodels



[FIG2] (a) The performance of an MLE (blue line) for 3-D localization of a WCE using RSS measurements of UWB signals superimposed to the Cramér–Rao lower bound (red line) [6]. (b) Cross-section view of the sensor receiver placements around the waist of a digital human model. The localization performance graphic was obtained with the $0-2\pi$ angular spread.

with advanced detection and estimation techniques and particle filters. To detect the maneuver changes, it will be useful to study impulsive responses of the change signature and the convergence properties of the innovation process. Innovation process is defined as a white Gaussian noise process derived from the original process by a causal and causally invertible transformation. For impulsive changes where a single realization of a Gaussian random vector is added to the plant equation with known variance, the Nyman–Pearson detector is a good choice [6]. In the asymptotic case, i.e., when the Gaussian random vector has infinite variance, this detector has marginal benefit from the knowledge of the change being impulsive. Therefore, it has been shown that the generalized likelihood ratio neglecting a priori information of the variance provides detection probabilities closer to the matched filter.

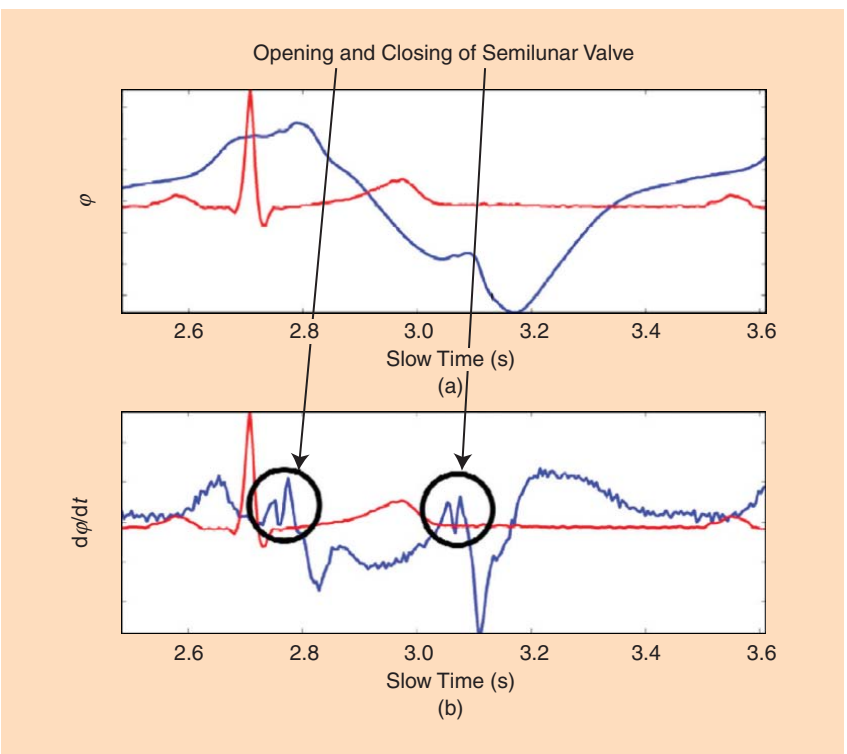
MEDICAL RADAR

The UWB radar has established itself as a useful tool for nonmedical applications like through-the-wall imaging and ground penetrating radar. During the last decade, multiple possible medical applications for the UWB radar have been demonstrated too, which include the detection of internal injuries such as intracranial hematoma, monitoring of respiratory and cardiac functions, and imaging of the human body [1].

One big contribution to the advancement of medical radar technology was the development of a family of UWB sensors known as micropower impulse radar (MIR) at the Lawrence Livermore National Laboratory in the United States. An MIR is composed of a pulse generator, a transmitter, a receiver, a range gate delay circuit, and a signal processing unit. The main problem in processing the received pulse echoes reflected by the body is the separation of the different signals that correspond to the heart activity, respiration, body movement, interference, and noise. Ordinary frequency filtration cannot be easily applied in this case because of very close arrangement of divided signal frequencies usually below 1 Hz. Hence, special algorithms for the extraction of specific physiological signals, e.g., heartbeat, have been proposed [7]. Today, UWB radar enables continuous noncontact monitoring of heart and respiratory motion rate. Although the primary tool used in screening for cardiovascular diseases is still the electrocardiogram (ECG), ongoing research aims to extract similar amount of clinical information from the received UWB pulse echoes as illustrated in Figure 3.

AORTIC BLOOD PRESSURE ESTIMATION

Noninvasive blood pressure measurement methods include the use of the sphygmomanometer, photoplethysmography, tonography, or pulse transit time. Intra-arterial measurement through a cannula, although invasive, is still considered the most accurate measurement method. Noninvasive methods are preferred as they rely on peripheral measurement points, but their accuracy can be affected by flow redistribution due to injuries, body temperature variation, and other factors. Therefore, central measurements, e.g., obtained from the aorta, can avoid the aforementioned problems. Some studies have shown the linear relationship between percentage changes in the instantaneous blood pressure and the diameter variation of the carotid artery. Moreover, there exists a nonlinear relationship between the mean arterial blood pressure and the arterial compliance, i.e., a ratio of change in volume versus change in pressure. Using these principles, UWB radar-based aortic blood pressure estimation has been investigated using digital human models, phantoms, and human subjects [8]. Results have shown that millimeter accuracy in the estimation of the aorta's diameter is highly dependent on the frequency of the transmitted EM signals and the elasticity of the aorta wall. Ongoing experiments with human subjects show that the measurements are highly affected by the mechanical movement of the heart



[FIG3] (a) Phase and (b) instantaneous frequency of the radar recordings (blue line) taken from a person's heartbeat. The corresponding normalized ECG signal (red line) is superimposed for comparison. Notice that advanced processing of the received UWB pulse echoes can make easily identifiable the opening and closing of the heart valves, which is not possible with the sole visual inspection of the phase graphic. [Images courtesy of the Norwegian Defence Research Establishment (FFI) and the MELODY Project (http://www.melody-project.info).]

toward the aorta measurement points propagating along the artery wall. Therefore, advanced signal processing techniques are required to separate the heartbeat and the wall movement without losing the estimation granularity on frequency and phase information. Estimating the absolute blood pressure from the differential pressure requires additional patient's information such as age, gender, weight, aorta elasticity, and compliance.

MEDICAL IMAGING

The use of an array of three or more radar sensors can enable the imaging of the human body. Medical imaging with UWB radar involves transmitting an extremely short pulse into the body and then recording the backscattered signal from different locations. UWB medical imaging has a number of advantages in comparison to current techniques like magnetic resonance imaging (MRI), computerized tomography (CT), X-rays, and ultrasound. UWB signals are not ionizing radiation. Therefore, safe continuous noncontact imaging without the need for bulky and expensive scanners becomes possible.

One medical application that has been extensively researched is the use of UWB radar imaging for breast cancer detection [2], [9]. X-ray mammography is currently the most widely used detection method, but despite its ability to provide high-resolution images, it suffers from a high false-alarm rate and the incapability to distinguish between malignant and benign tumors. On the other hand, the basis for detecting and locating a cancerous tumor with UWB imaging is the different dielectric properties of healthy and malignant breast tissues. Healthy tissues are largely transparent to microwaves, whereas tumors, which contain more water and blood, scatter the pulses back to the probing antenna array. Promising results for the accurate detection and localization of cancerous tumors have been reported using near-field tomographic image reconstruction (TIR), confocal microwave imaging, space-time beamforming, generalized likelihood ratio test based detection, and the TOA data fusion method. The capability to detect tumors with a size as small as 2 mm in diameter with 100% certainty has been demonstrated via computer simulations and experiments on breast phantoms [9]. UWB imaging of the heart has also been demonstrated in a recent proof-of-concept experiment [10]. Sufficient resolution for the observation of different moving parts of the heart was obtained as shown in Figure 4, although it was not fine enough to be used for diagnostic purposes. The imaging was undertaken using a time-domain multistatic backprojection algorithm. Improvement of the radar system and signal processing along with a comparison with established medical heart imaging techniques like echocardiography are still needed.

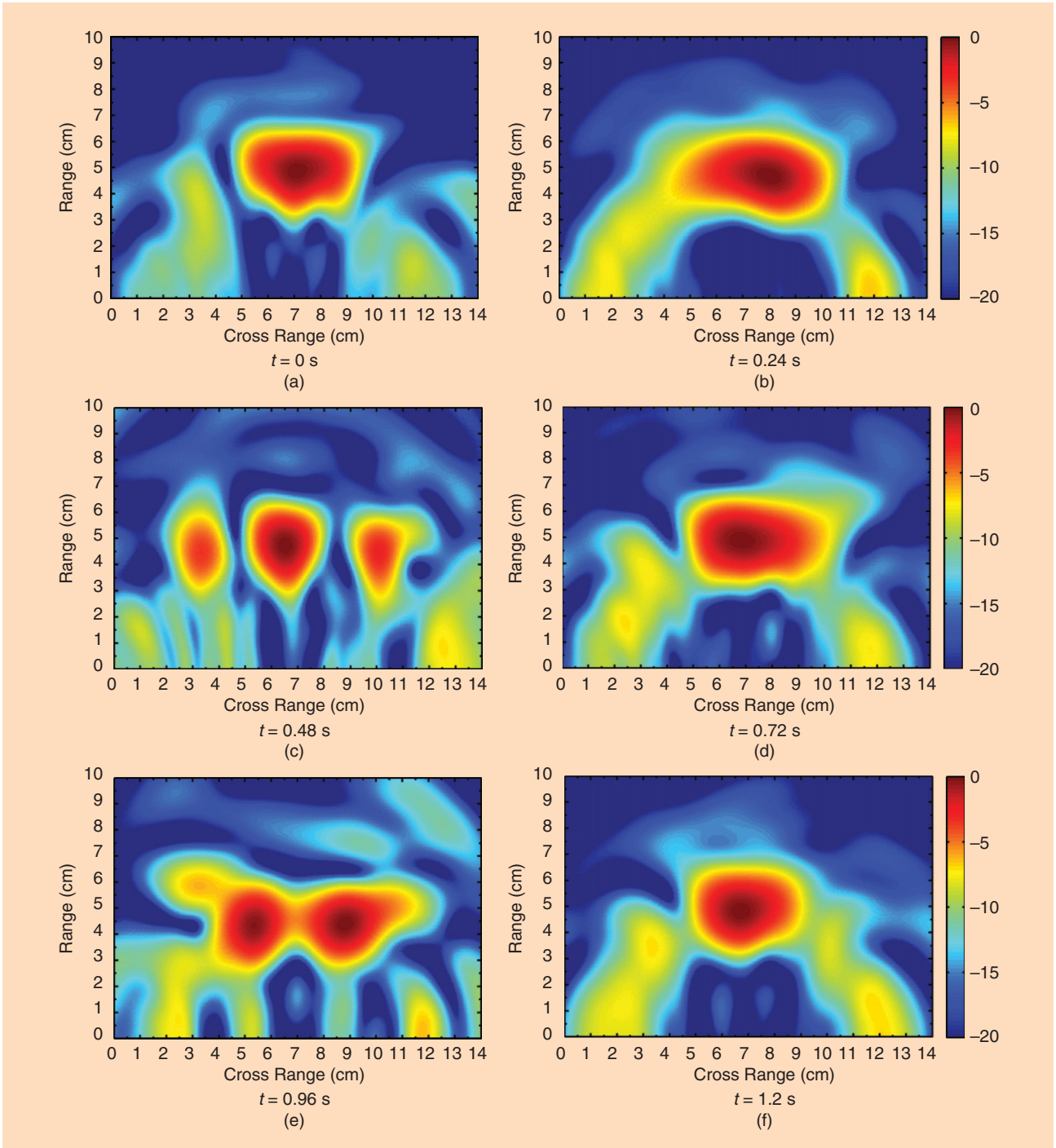
HYPERTHERMIA

UWB is also very promising for the improvement of hyperthermia, a therapeutic procedure in which EM waves are used to elevate temporarily the temperature of malignant tumors. Hyperthermia has been found useful in the treatment of various cancers by directly inducing cell death or by increasing the effectiveness of chemotherapy and radiotherapy. In a hyperthermia procedure, the temperature of the tumor is elevated above 42 °C for a predetermined period of time, while the temperature of surrounding healthy tissues is preserved at a normal physiological value. Confining accurately the EM energy to deeply seated tumors is, however, a standing and challenging problem. Thus, the use of an antenna array can facilitate the transmission of EM signals so that constructive wave interference occurs within the tumor and destructive interference elsewhere. At present, the most widely used method is the annular phased array, which is a circumferential array of transmitters placed around the patient. Constructive wave interference is achieved by changing the amplitude and phase at the feed points of the antennas, whereas focusing is obtained by maximizing the constructive interference produced by individual E-field components at a single point. Focusing can also be achieved by optimizing the SAR or the temperature values.

Because of the limited focusing obtained with the use of NB signals, UWB has gained increasing attention in the last decade. Hyperthermia techniques for the treatment of breast cancer have been introduced, one of which (2004) uses space-time beamforming based on the time delay and impulse response obtained from reflection measurements. In this technique, the signals are time delayed by FIR filters to compensate for the propagation delay of the different signals traveling from their corresponding antenna to the focal spot. Then, the signals from all the antennas are simultaneously transmitted into the breast. A simulation-based study with realistic numerical breast models derived from MRI data has demonstrated that UWB outperforms NB by offering the potential for tighter focusing [11]. A similar method (2005) consists of transmitting a pulse from a single antenna to generate the phase and amplitude information. The backscattered signals received by all the antennas are then time-gated, time-reversed, weighted, amplified, and retransmitted into the breast simultaneously. A more recent time-reversal approach (2010) estimates the phase and amplitudes from simulations of wave transmission instead of measured data [12]. Besides being faster, this approach allows instantaneous sweeping of the focusing point, which can be useful in the case of complex tumor shapes or movements of internal organs during the hyperthermia procedure.

CONCLUSIONS

UWB signals can have a vast application in medical practice as wireless communication interfaces with biomedical sensors, remote monitoring of vital signs, imaging of different organs, and hyperthermia for cancer treatment. Because of the low power consumption and compact size of UWB devices, portable medical sensing and imaging equipment can be fabricated for emergency crews and rural health-care provision. The current advancements in all these areas can open new possibilities for medical diagnosis and treatment procedures. For example, the tight focusing capabilities of UWB signals evidenced in hyperthermia can be exploited for the treatment of neurodegenerative diseases (e.g., Alzheimer's disease),



[FIG4] The time-lapsed imaging of the heart spanning approximately the duration of one cardiac cycle, i.e., T=1.2 s. The absolute value of each image is plotted in dB scale. [Images courtesy of the Norwegian Defence Research Establishment (FFI) and the MELODY Project (http://www.melody-project.info).]

stroke, and head trauma, conditions that deteriorate the patient's cognition capabilities. Currently, researchers try to mimic the natural coding of the brain with electrical stimulation thereby replacing compromised regions of the hippocampus, which may help to restore long-term memory; some cortical implants have been designed for this purpose. Nevertheless, stimulation by induced electric fields through UWB EM radiation can make this therapy significantly less invasive. Epilepsy could potentially be treated in a similar manner.

In this column, we aimed to spark interest in the medical uses of UWB signals. As illustrated throughout the article, the role of the signal processing community is

crucial for the improvement of existing applications and the search for novel procedures for diagnosing and treating effectively life-threatening diseases. Therefore, further research in this field is encouraged.

ACKNOWLEDGMENT

We acknowledge financial support from the Research Council of Norway given to the MELODY Project (Phase II) under contract number 225885.

AUTHORS

Raúl Chávez-Santiago (raul.chavez-santiago@rr-research.no) is a researcher at the Intervention Centre, Oslo University Hospital, Norway.

Ilangko Balasingham (ilangko.balasingham@medisin.uio.no) is the head of the Biomedical Sensor Network Research Group at the Intervention Centre, Oslo University Hospital, Norway, and professor in the Department of Electronics and Telecommunications, Norwegian University of Science and Technology, Trondheim.

REFERENCES

[1] C. N. Paulson, J. T. Chang, C. E. Romero, J. Watson, J. F. Pearce, and N. Levin, "Ultra-wideband radar methods and techniques of medical sensing and imaging," in *Proc. SPIE Int. Symp. Optics East*, Boston, MA, 2005, pp. 96–107.

[2] E. Zastrow, S. K. Davis, and S. C. Hagness, "Safety assessment of breast cancer detection via ultrawideband microwave radar operating in pulsed-radiation mode," *Microw. Opt. Technol. Lett.*, vol. 49, no. 1, pp. 221–225, 2007.

[3] A. Daisuke, K. Katsu, R. Chávez-Santiago, Q. Wang, D. Plettemeier, J. Wang, and I. Balasingham, "Experimental evaluation of implant UWB-IR transmission with living animal for body area networks," *IEEE Trans. Microw. Theory Techn.*, vol. 62, no. 1, pp. 183–192, 2014.

[4] R. Chávez-Santiago, and I. Balasingham, "The ultra wideband capsule endoscope," in *Proc. IEEE Int. Conf. Ultra-Wideband (ICUWB)*, Sydney, Australia, 2013, pp. 72–78.

[5] M. S. Chae, Z. Yang, M. R. Yuce, L. Hoang, and W. Liu, "A 128-channel 6 mW wireless neural recording IC with spike feature extraction and UWB transmitter," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 17, no. 4, pp. 312–321, 2009.

[6] B. Moussakhani, "On localization and tracking for wireless capsule endoscopy," Ph.D. dissertation, Dept. Electronics and Telecommunications, Norwegian Univ. Sci. and Technol., Trondheim, Norway, 2013.

[7] S. N. Pavlov, and S. V. Samkov, "Algorithm of signal processing in ultra-wideband radar designed for remote measuring parameters of patient's cardiac activity," in *Proc. 2nd Int. Workshop Ultra Wideband Ultra Short Impulse Signals*, Sevastopol, Ukraine, 2004, pp. 205–207.

[8] L. E. Solberg, S.-E. Hamran, T. Berger, and I. Balasingham, "Minimum variance signal selection for aorta radius estimation using radar," *EURASIP J. Adv. Signal Process.*, vol. 2010, no. 49, pp. 1–13, 2010.

[9] S. A. Alshehri, S. Khatun, A. B. Jantan, R. S. A. R. Abdullah, R. Mahmood, and Z. Awang, "Experimental breast tumor detection using NN-based UWB imaging," *Prog. Electromagn. Res. (PIER)*, vol. 111, pp. 447–465, 2011.

[10] S. Brovoll, T. Berger, Y. Paichard, Ø. Aardal, T. S. Lande, and S.-E. Hamran, "Time-lapse imaging of human heartbeats using UWB radar," in *Proc. IEEE Biomedical Circuits Systems Conf. (BioCAS)*, Rotterdam, The Netherlands, 2013, pp. 142–145.

[11] M. Converse, E. J. Bond, B. D. Van Veen, and S. C. Hagness, "A computational study of ultra-wideband versus narrowband microwave hyperthermia for breast cancer treatment," *IEEE Trans. Microw. Theory Techn.*, vol. 54, no. 5, pp. 2169–2180, 2006.

[12] H. D. Trefná, J. Vrba, and M. Persson, "Time-reversal focusing in microwave hyperthermia for deep-seated tumors," *Phys. Med. Biol.*, vol. 55, no. 8, pp. 2167–2185, 2010.

[SP]

applications CORNER (continued from page 123)

the Advanced Digital Sciences Center (ADSC) from Singapore's Agency for Science, Technology and Research (A*STAR). We thank Tien Dung Vu, a software engineer at ADSC, for his assistance in implementing our ITEM prototype system.

AUTHORS

Viet Anh Nguyen (vanguyen@adsc.com.sg) is a researcher at the Advanced Digital Sciences Center, Illinois at Singapore.

Jiangbo Lu (jiangbo.lu@adsc.com.sg) is a researcher at the Advanced Digital Sciences Center, Illinois at Singapore.

Shengkui Zhao (shengkui.zhao@adsc.com.sg) is a researcher at the Advanced Digital Sciences Center, Illinois at Singapore.

Douglas L. Jones (dl-jones@illinois.edu) is a professor at the University of Illinois at Urbana-Champaign.

Minh N. Do (minhdo@illinois.edu) is a professor at the University of Illinois at Urbana-Champaign.

REFERENCES

[1] J. G. Apostolopoulos, P. A. Chou, B. Culbertson, T. Kalker, M. D. Trott, and S. Wee, "The road to immersive communication," *Proc. IEEE*, vol. 100, no. 4, pp. 974–990, 2012.

[2] D. E. Ott and K. Mayer-Patel, "Coordinated multistreaming for 3D tele immersion," in *Proc. ACM Multimedia*, 2004, pp. 596–603.

[3] Z. Yang, K. Nahrstedt, C. Yi, B. Yu, J. Liang, S.-H. Jung, and R. Bajscy, "TEEVE: The next generation architecture for tele-immersive environments," in *Proc. 7th IEEE Int. Symp. Multimedia*, 2005, pp. 112–119.

[4] C. Kuster, N. Ranieri, Agustina, H. Zimmer, J. C. Bazin, C. Sun, T. Popa, and M. Gross, "Towards next generation 3D teleconferencing systems," in *Proc. 3DTV-Conf.: True Vision-Capture, Transmission, Display 3D Video (3DTV-CON)*, 2012.

[5] C. W. Lin, Y. J. Chang, C. M. Wang, Y. C. Chen, and M.-T. Sun, "A standard-compliant virtual meeting system with active video object tracking," *EURASIP J. Adv. Signal Process.*, vol. 2002, no. 6, pp. 622–634, 2002.

[6] H. Baker, N. Bhatti, D. Tanguay, I. Sobel, D. Gelb, M. Goss, B. Culbertson, and T. Malzbender, "Understanding performance in coliseum, an immersive videoconferencing system," *ACM Trans. Multimedia Computing Commun. Appl. (TOMCCAP)*, vol. 1, no. 2, pp. 190–210, 2005.

[7] J. Lu, V. A. Nguyen, Z. Niu, B. Singh, Z. Luo, and M. N. Do, "CuteChat: A lightweight tele-immersive video chat system," in *Proc. ACM Multimedia*, 2011, pp. 1309–1312.

[8] V. A. Nguyen, J. Lu, S. Zhao, T. D. Vu, H. Yang, D. L. Jones, and M. N. Do, "ITEM: Immersive telepresence for entertainment and meetings—A practical approach," Tech Rep., Advanced Digital Sciences Center, Aug. 2014, arXiv:1408.0605.

[9] V. A. Nguyen, J. Lu, and M. N. Do, "Efficient video compression methods for a lightweight tele-immersive video chat system," in *Proc. IEEE Int. Symp. Circuits Systems (ISCAS)*, 2012, pp. 149–152.

[10] S. Zhao, A. Ahmed, Y. Liang, K. Rupnow, D. Chen, and D. L. Jones, "A real-time 3D sound localization system with miniature microphone array for virtual reality," in *Proc. IEEE Industrial Electronics Applications (ICIEA)*, 2012, pp. 1853–1857.

[11] S. Zhao, R. Rogowski, R. Johnson, and D. L. Jones, "3D binaural audio capture and reproduction using a miniature microphone array," in *Proc. 15th Int. Conf. Digital Audio Effects (DAFx)*, 2012, pp. 151–154.

[12] V. A. Nguyen, S. Zhao, T. D. Vu, D. L. Jones, and M. N. Do, "Spatialized audio multiparty teleconferencing with commodity miniature microphone array," in *Proc. ACM Multimedia*, 2013, pp. 553–556.

[SP]

REFLECTIONS

H. Vincent Poor

Reflections on Excellence in Research and Education in Signal Processing

NOTE FROM THE EDITORS

In July of this year, Prof. H. Vincent Poor, from Princeton University, New Jersey, was elected Foreign Member of the Royal Society. This addition to the numerous awards he has received is highly prestigious, as the Royal Society is the oldest scientific academy in continuous existence worldwide. Undoubtedly, this recognition is very well-deserved. Prof. Poor has tackled many technical challenges with great success in the field of signal processing in areas such as wireless networks, social networks, smart grids, and statistical signal processing, to name a few. In addition to the numerous technical contributions in signal processing, Prof. Poor has influenced many members of the signal processing community through his educational activities and services to the IEEE. He has been instrumental to the success of *IEEE Signal Processing Magazine* (*SPM*), not only as an active member of the senior editorial board, but also as a guest editor for special issues and the author of numerous articles. His exemplary contributions to *SPM* and the IEEE Signal Processing Society at large warrant a celebration of his election to the Royal Society with the signal processing community through an interview in this "Reflections" column. We hope that you will enjoy reading it and that his insights will be an inspiration to many signal processing practitioners.

Abdelhak Zoubir (zoubir@spg.tu-darmstadt.de) is the editor-in-chief of *SPM* and a professor at Technische Universität Darmstadt.

Andres Kwasinski (axkeec@rit.edu) is an area editor for columns and forum of *SPM* and an associate professor at the Rochester Institute of Technology.

***IEEE SPM*: Your contributions have had a significant impact across multiple IEEE Societies and other associations, such as the Institute of Mathematical Statistics. How do you manage to maintain these outstanding multidisciplinary achievements in both depth and breadth?**

H. Vincent Poor (VP): I've been quite fortunate throughout my career to collaborate with colleagues from many different backgrounds, and also to work with talented students and postdocs who have brought their own creativity and skills to our work. Also, having a background in basic disciplines like probability, statistical inference, information theory, game theory, and so forth has been quite helpful when it comes to finding inroads into new problem areas.

***IEEE SPM*: How did you manage to maintain nonintermittent scholarly research work over decades?**

VP: Again, I would point to my collaborators and students, and also to the institutions where I have spent most of my career—Illinois and Princeton—both of which set very high standards for research and provide environments where it can take place. A big surprise to me when I took on the role as dean of engineering at Princeton was that I was able to maintain a very active research program, and that is also principally because of these factors.

***IEEE SPM*: Your election to the Royal Society is a recognition of your contributions to signal processing. What is your vision on the future of our discipline?**

VP: I am constantly amazed at the breadth of interests within the signal processing community, although it is perhaps not so surprising when one considers how widely applicable many of the basic principles of signal processing are. Essentially, a signal is a quantity that changes with respect to some other quantity, and that definition covers quite a lot of the phenomena arising in science and engineering disciplines. So, I think in the future the discipline will continue to contribute to new fields and applications as they arise. It's hard to predict what those will be, but I think it's safe to say that signal processing will play a major role in many of them.

***IEEE SPM*: How did signal processing applications widen and broaden, and how did they merge with other disciplines?**

VP: I am sure there are many reasons why this happened, but I would say that it arose at least in part when signal processing ideas became more abstract and thus more universally applicable. If you look at the transactions today, you will see some fairly abstract ideas being exposed in the context of signal processing. So, it has come to resemble a fundamental science in that respect. With a core of general tools, moving into other disciplines has been fairly effortless. Another very important factor here, of course, is the rise of digital technologies that has enabled these abstractions to be realized in practical systems. A great example of these factors is in

Digital Object Identifier 10.1109/MSP.2014.2343984
Date of publication: 15 October 2014

1053-5888/14©2014IEEE

my own field of wireless communications where signal processing has had a tremendous impact.

IEEE SPM: There has been some discussion about the dissemination of our discipline among the public at large, referred to as *inside signal processing*. In your view, what should we do to increase the outreach of our activities?

VP: There has been some very good work in promoting signal processing to precollege students, which I think is an important part of outreach. In my experience, it's very hard to get the attention of the general public in any large numbers, but social media provide a natural framework for approaching this today. Short online videos about interesting applications and ideas can have an impact if they're entertaining. Also, Twitter feeds can have a similar effect. These are both things that we do here in the engineering school at Princeton to good effect. But these kinds of things require a lot of time and attention to keep them fresh. Television shows like *Nova* can also reach broader audiences, although certainly not as broad as those that can be reached via social media.

IEEE SPM: In your opinion, what are the most challenging problems signal processing should tackle, and what are the emerging trends?

VP: There are big societal problems in which signal processing can have a significant impact: energy, environment, health, security. Of course, signal processing is already tackling these areas, but they all present very serious challenges and will continue to be sources of important research problems and applications for the foreseeable future.

THERE ARE BIG SOCIETAL PROBLEMS IN WHICH SIGNAL PROCESSING CAN HAVE A SIGNIFICANT IMPACT: ENERGY, ENVIRONMENT, HEALTH, SECURITY.

H. Vincent Poor signs the roll book of the Royal Society, which has been used since the Society's founding in 1660. This book contains the signatures of many scientific luminaries, including Isaac Newton and Charles Darwin. (Photo courtesy and copyright of The Royal Society.)

IEEE SPM: You have been involved in signal processing research for several decades. What have been the biggest changes you noticed on how signal processing research is being conducted?

VP: The abstraction of the field, which I mentioned earlier, is certainly a significant change in how research is being conducted. Also, the open-mindedness of the community in looking at far-ranging problems is another. Needless to say, there has been an explosion of interest in the field over those decades, which is very encouraging. The number of people working in the field has increased greatly and the number of places where good research is being done has grown significantly. This has created some challenges, as the amount of published work and the pace of progress have made it harder to keep up with what's going on across the discipline. But, overall these are very healthy developments for a research field.

IEEE SPM: In 2005, the IEEE recognized your achievements as an educator by awarding you the IEEE James H. Mulligan, Jr. Education Medal. How has your work as an educator influenced your research? Conversely, how did your research contributions help your educational activities?

VP: For someone in academia, research and education are very closely intertwined. Essentially, my entire research career has been conducted in collaboration with students and postdocs, and so these two aspects are almost inseparable. I think anyone who works with young researchers will tell you that learning goes both ways in these relationships, and being around students helps keep research fresh and topical. At the same time, my research has informed my teaching in major ways. What's taught in the classroom is largely the product of research and, in teaching, questions often arise that feed back into new research problems. The academic enterprise is essentially a process of inquiry and discussion that propels research forward. Of course, research takes place in other environments as well, but academia provides a natural setting for creativity and exploration through this process.

IEEE SPM: All colleagues from *IEEE SPM* who worked with you testify that you almost instantly reply to e-mails and requests. Given your large load and many responsibilities, how do you manage to do all this?

VP: I find it hard not to answer e-mail quickly! For me it's much easier to stay on top of things than otherwise. I enjoy almost all aspects of my work, and for that reason I tend to get involved in a lot of different things. I think I would lose track of some of these if I didn't handle them in a fairly timely manner. Also, I'm addicted to having an empty inbox—it's a bit of an obsession, really.

[SP]

dates AHEAD

Please send calendar submissions to:
Dates Ahead, c/o Jessica Barragué
IEEE Signal Processing Magazine
445 Hoes Lane
Piscataway, NJ 08855 USA
e-mail: j.barrague@ieee.org
(Colored conference title indicates SP-sponsored conference.)

2014

OCTOBER

IEEE Workshop on Signal Processing Systems (SIPS)
20–23 October, Belfast, Ireland.

NOVEMBER

48th Asilomar Conference on Signals, Systems, and Computers
2–5 November, Pacific Grove, California, United States.
General Chair: Roger Woods
Technical Program Chair: Geert Leus
URL: http://www.asilomarsscconf.org/

DECEMBER

IEEE Global Conference on Signal and Information Processing (GlobalSIP)
3–5 December, Atlanta, Georgia, United States.
General Chairs: Geoffrey Li and Fred Juang
URL: http://renyi.ece.iastate.edu/globalsip2014/

IEEE International Workshop on Information Forensics and Security (WIFS)
3–5 December, Atlanta, Georgia, United States.
General Chairs: Yan (Lindsay) Sun and Vicky H. Zhao
URL: http://ieeewifs.org/

2014 Workshop on Genomic Signal Processing and Statistics (GENSIPS)
3–5 December, Atlanta, Georgia, United States.
General Chairs: Geoffrey Li and Fred Juang
URL: http://www.ieeeglobalsip.org/gensips/gensips-cfp.html

IEEE Spoken Language Technology Workshop (SLT)
6–9 December, South Lake Tahoe, California, United States.
General Chairs: Murat Akbacak and John Hansen

2014 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)
9–12 December, Chiang Mai, Thailand.
Honorary Cochairs: Sadaoki Furui, K.J. Ray Liu, and Prayoot Akkaraekthalin
General Cochairs: Kosin Chamnongthai, C.-C. Jay Kuo, and Hitoshi Kiya
URL: http://www.apsipa2014.org/home/

2015

APRIL

Data Compression Conference (DCC)
7–9 April, Snowbird, Utah, United States.
URL: http://www.cs.brandeis.edu/~dcc/index.html

14th IEEE International Conference on Information Processing in Sensor Networks (IPSN)
13–17 April, Seattle, Washington, United States.
General Chair: Suman Nath
URL: http://ipsn.acm.org/2015

12th IEEE International Symposium on Biomedical Imaging (ISBI)
16–19 April, Brookyln, New York, United States.
General Chairs: Elsa Angelini and Jelena Kovačević
URL: http://biomedicalimaging.org/2015/

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)
19–24 April, Brisbane, Australia.
General Cochairs: Vaughan Clarkson and Jonathan Manton
URL: http://icassp2015.org/

JUNE

IEEE International Conference on Multimedia and Expo (ICME)
29 June–3 July, Turin, Italy.
General Chairs: Enrico Magli, Stefano Tubaro, and Anthony Vetro
URL: http://www.icme2015.ieee-icme.org/index.php

SEPTEMBER

IEEE International Conference on Image Processing (ICIP)
28 September–1 October, Quebec City, Quebec, Canada.

Digital Object Identifier 10.1109/MSP.2014.2348541
Date of publication: 15 October 2014

[SP]

advertisers INDEX

The Advertisers Index contained in this issue is compiled as a service to our readers and advertisers: the publisher is not liable for errors or omissions although every effort is made to ensure its accuracy. Be sure to let our advertisers know you found them through *IEEE Signal Processing Magazine*.

ADVERTISER	PAGE	URL	PHONE
IEEE Marketing Department	3	www.ieee.org/digitalsubscriptions	
IEEE MDL/Marketing	7	www.ieee.org/go/freemonth	
Mathworks	CVR 4	www.mathworks.com/accelerate	+1 508 647 7040
Mini-Circuits	CVR 2, 5, CVR 3	www.minicircuits.com	+1 718 934 4500
University of Illinois at Urbana-Champaign	13	bioinstrumentation.illinois.edu	
University of Minnesota	129	www.ece.umn.edu	

advertising SALES OFFICES

James A. Vick
Sr. Director, Advertising
Phone: +1 212 419 7767;
Fax: +1 212 419 7589
jv.ieeemedia@ieee.org

Marion Delaney
Advertising Sales Director
Phone: +1 415 863 4717;
Fax: +1 415 863 4717
md.ieeemedia@ieee.org

Susan E. Schneiderman
Business Development Manager
Phone: +1 732 562 3946;
Fax: +1 732 981 1855
ss.ieeemedia@ieee.org

Product Advertising

MIDATLANTIC
Lisa Rinaldo
Phone: +1 732 772 0160;
Fax: +1 732 772 0164
lr.ieeemedia@ieee.org
NY, NJ, PA, DE, MD, DC, KY, WV

NEW ENGLAND/SOUTH CENTRAL/ EASTERN CANADA
Jody Estabrook
Phone: +1 774 283 4528;
Fax: +1 774 283 4527
je.ieeemedia@ieee.org
ME, VT, NH, MA, RI, CT, AR, LA, OK, TX
Canada: Quebec, Nova Scotia, Newfoundland, Prince Edward Island, New Brunswick

SOUTHEAST
Thomas Flynn
Phone: +1 770 645 2944;
Fax: +1 770 993 4423
tf.ieeemedia@ieee.org
VA, NC, SC, GA, FL, AL, MS, TN

Digital Object Identifier 10.1109/MSP.2013.2290965

MIDWEST/CENTRAL CANADA
Dave Jones
Phone: +1 708 442 5633;
Fax: +1 708 442 7620
dj.ieeemedia@ieee.org
IL, IA, KS, MN, MO, NE, ND, SD, WI, OH
Canada: Manitoba, Saskatchewan, Alberta

MIDWEST/ ONTARIO, CANADA
Will Hamilton
Phone: +1 269 381 2156;
Fax: +1 269 381 2556
wh.ieeemedia@ieee.org
IN, MI. Canada: Ontario

WEST COAST/MOUNTAIN STATES/ WESTERN CANADA
Marshall Rubin
Phone: +1 818 888 2407;
Fax: +1 818 888 4907
mr.ieeemedia@ieee.org
AZ, CO, HI, NM, NV, UT, AK, ID, MT, WY, OR, WA, CA. Canada: British Columbia

EUROPE/AFRICA/MIDDLE EAST ASIA/FAR EAST/PACIFIC RIM
Louise Smith
Phone: +44 1875 825 700;
Fax: +44 1875 825 701
les.ieeemedia@ieee.org
Europe, Africa, Middle East
Asia, Far East, Pacific Rim, Australia, New Zealand

Recruitment Advertising

MIDATLANTIC
Lisa Rinaldo
Phone: +1 732 772 0160;
Fax: +1 732 772 0164
lr.ieeemedia@ieee.org
NY, NJ, CT, PA, DE, MD, DC, KY, WV

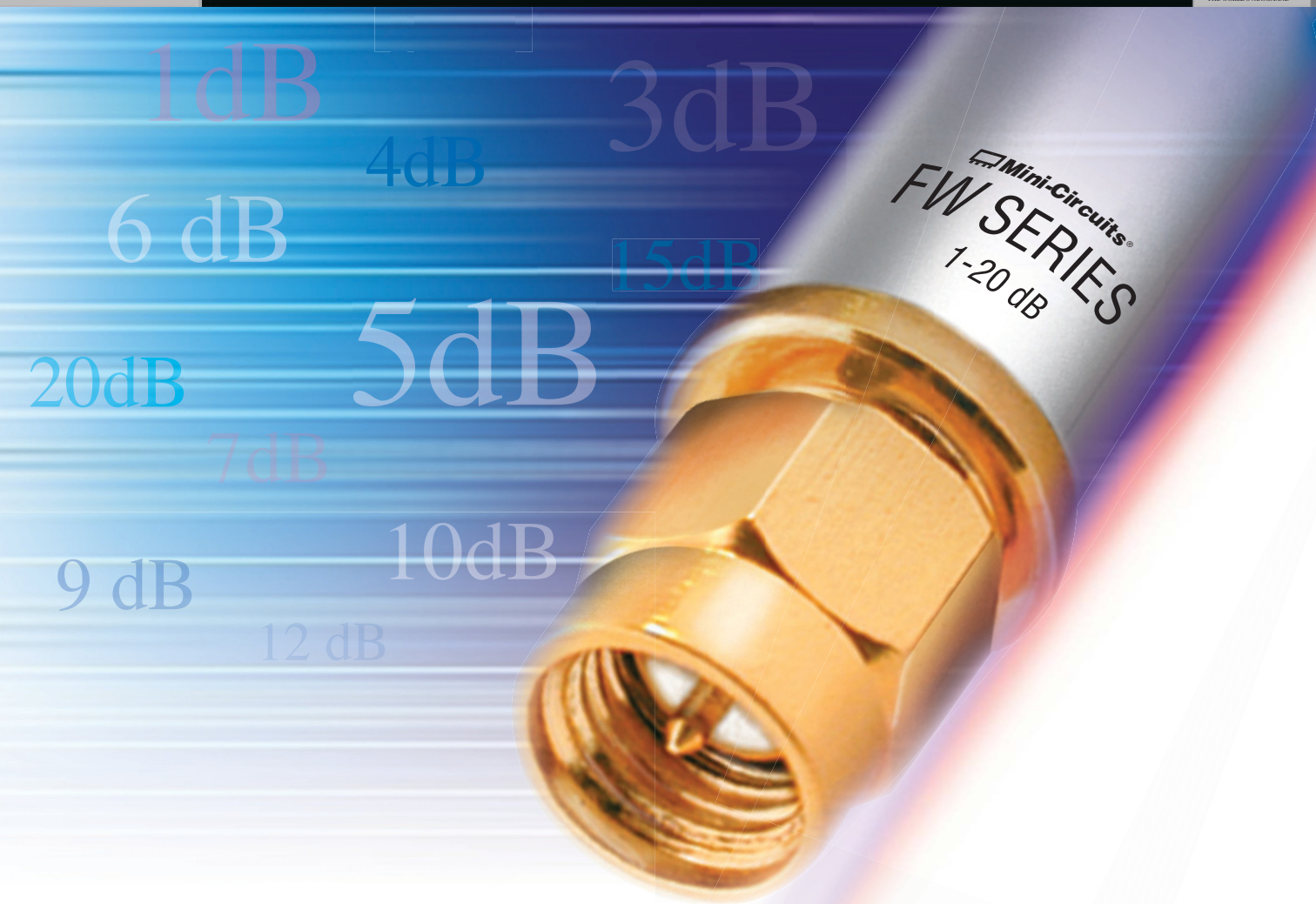
NEW ENGLAND/EASTERN CANADA
Liza Reich
Phone: +1 212 419 7578;
Fax: +1 212 419 7589
e.reich@ieee.org
ME, VT, NH, MA, RI. Canada: Quebec, Nova Scotia, Prince Edward Island, Newfoundland, New Brunswick

SOUTHEAST
Cathy Flynn
Phone: +1 770 645 2944;
Fax: +1 770 993 4423
cf.ieeemedia@ieee.org
VA, NC, SC, GA, FL, AL, MS, TN

MIDWEST/SOUTH CENTRAL/ CENTRAL CANADA
Darcy Giovingo
Phone: +224 616 3034;
Fax: +1 847 729 4269
dg.ieeemedia@ieee.org;
AR, IL, IN, IA, KS, LA, MI, MN, MO, NE, ND, SD, OH, OK, TX, WI. Canada: Ontario, Manitoba, Saskatchewan, Alberta

WEST COAST/SOUTHWEST/ MOUNTAIN STATES/ASIA
Tim Matteson
Phone: +1 310 836 4064;
Fax: +1 310 836 4067
tm.ieeemedia@ieee.org
AZ, CO, HI, NV, NM, UT, CA, AK, ID, MT, WY, OR, WA. Canada: British Columbia

EUROPE/AFRICA/MIDDLE EAST
Louise Smith
Phone: +44 1875 825 700;
Fax: +44 1875 825 701
les.ieeemedia@ieee.org
Europe, Africa, Middle East



DC-12 GHz 1W Attenuators

$18.95 ea. qty. 1-9

Accurate attenuation over ultra-wide frequency range at low cost! Mini-Circuits' new FW-series fixed SMA attenuators cover your applications from DC – 12 GHz with a wide selection of attenuation values to meet your needs. All models offer excellent attenuation flatness versus frequency, low VSWR, and 1W input power handling, making them ideal, cost-saving solutions for impedance matching and signal level adjustment in a broad range of systems.

Built using Mini-Circuits rugged unibody construction with SMA(M) to SMA (F) connectors, FW-Series attenuators provide outstanding reliability in tough operating conditions. Just go to minicircuits.com for full specs and see how these attenuators can add performance and value to your design! They're available off the shelf for immediate shipment. Order today, and have them in hand as soon as tomorrow!

Model	ATTENUATION Nominal (dB)	Flatness (dB)	VSWR (:1) Midband Typ.	Power (W)
FW-1+	1	±0.15	1.15	1.0
FW-2+	2	±0.15	1.15	1.0
FW-3+	3	±0.20	1.15	1.0
FW-4+	4	±0.20	1.15	1.0
FW-5+	5	±0.20	1.15	1.0
FW-6+	6	±0.25	1.15	1.0
FW-7+	7	±0.25	1.15	1.0
FW-8+	8	±0.30	1.15	1.0
FW-9+	9	±0.30	1.15	1.0
FW-10+	10	±0.30	1.15	1.0
FW-12+	12	±0.30	1.20	1.0
FW-15+	15	±0.35	1.20	1.0
FW-20+	20	±0.50	1.20	1.0

RoHS compliant.

Mini-Circuits®

www.minicircuits.com P.O. Box 350166, Brooklyn, NY 11235-0003 (718) 934-4500 sales@minicircuits.com

531 rev. orig.

Find it at
mathworks.com/accelerate
datasheet
video example
trial request

MODEL PHYSICAL SYSTEMS

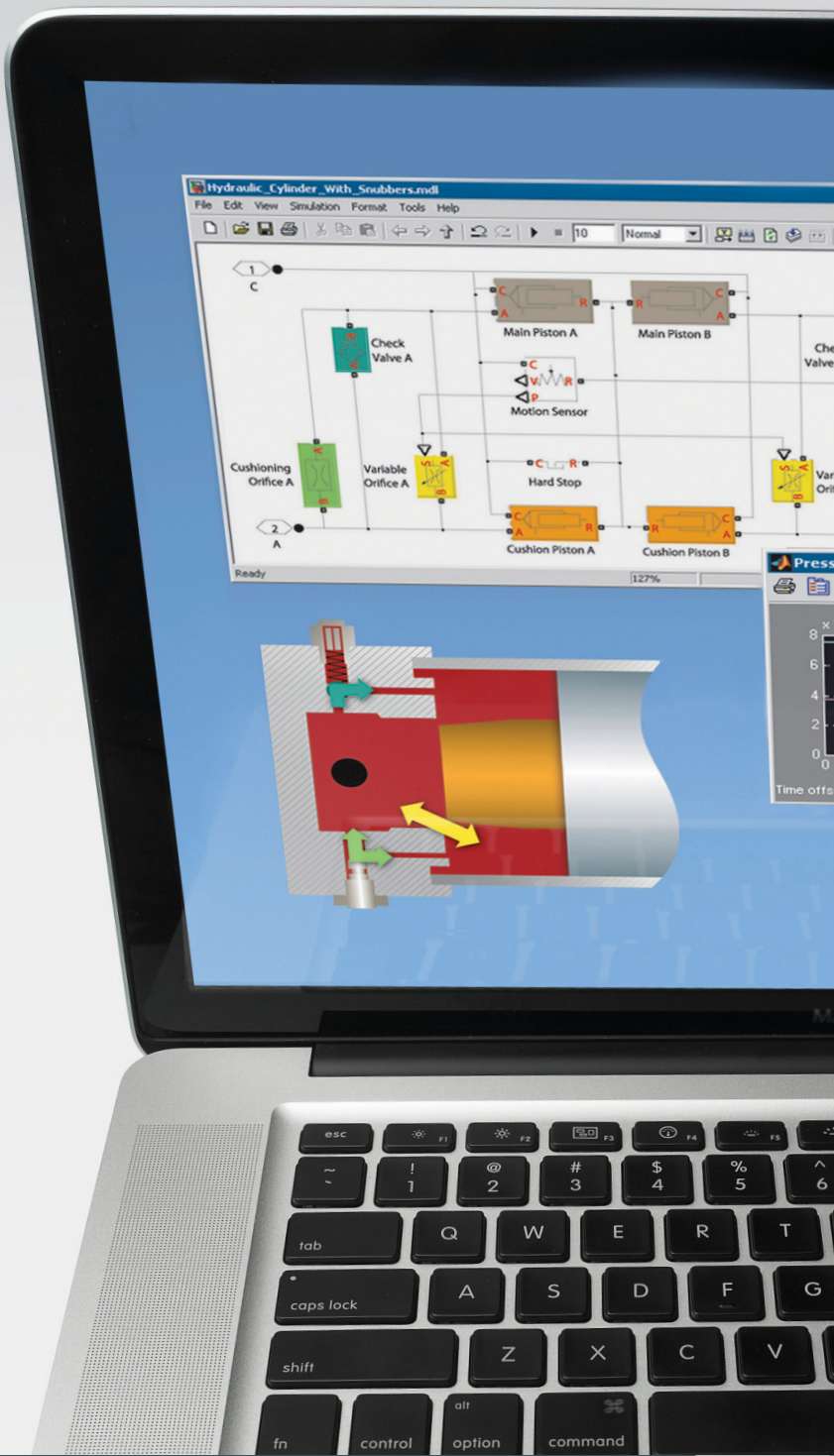
in

Simulink

with **Simscape™**

- Electrical
- Mechanical
- Hydraulic

and more



Use SIMSCAPE with SIMULINK to model and simulate the plant and controller of an embedded system. Assemble your model with a graphical interface, or import physical models from CAD systems. Use built-in components or create your own with the Simscape language.

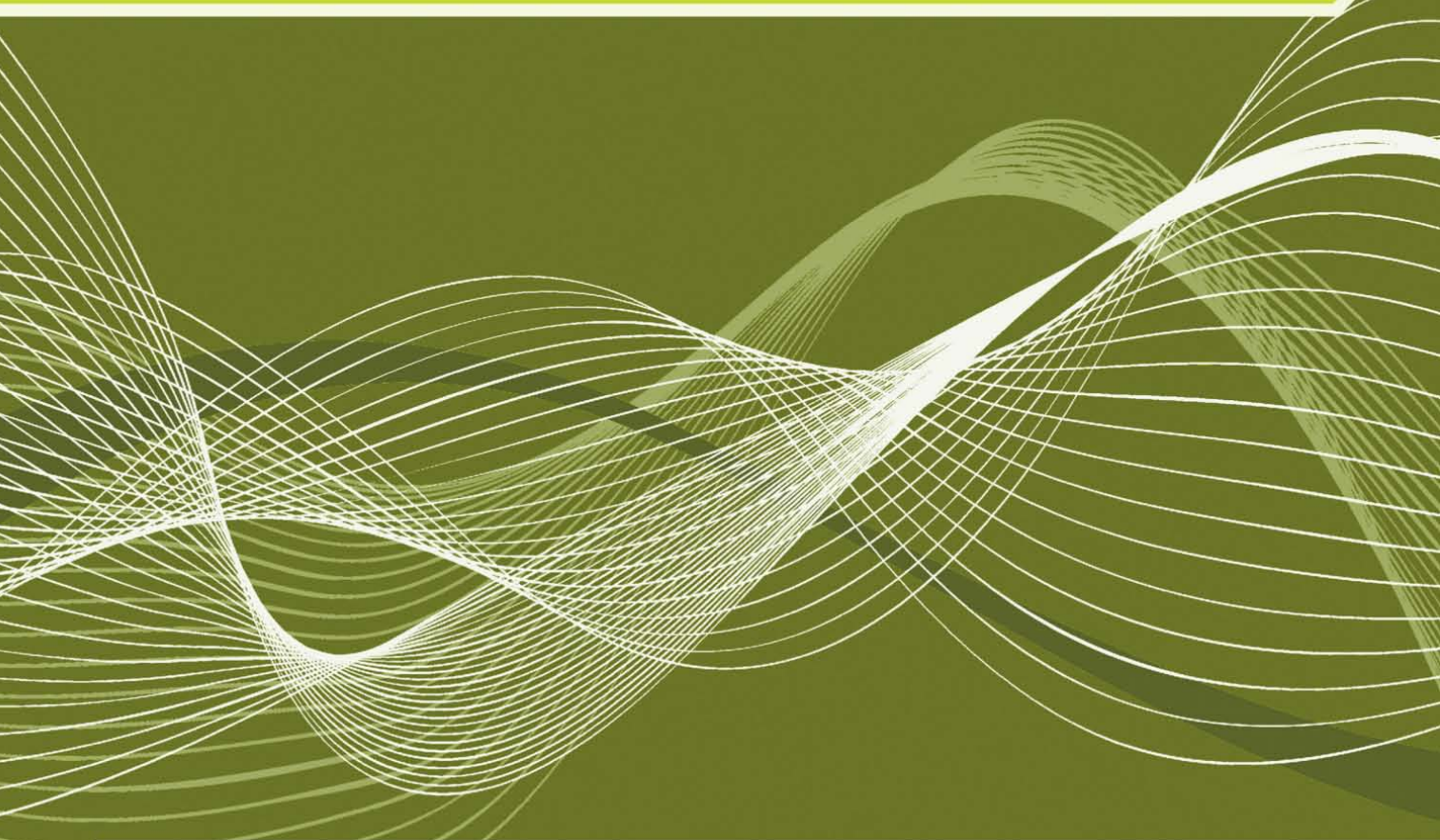
MATLAB® & SIMULINK®

MathWorks®
Accelerating the pace of engineering and science

IEEE SIGNAL PROCESSING SOCIETY

CONTENT GAZETTE

[ISSN 2167-5023]



NOVEMBER 2014

IEEE International Symposium on Biomedical Imaging

April 16th — 19th 2015, Brooklyn, NY USA

Conference Chairs

Elsa Angelini
Telecom ParisTech, France
Columbia University, USA

Jelena Kovačević
Carnegie Mellon University, USA

Program Chairs

Sebastien Ourselin
University College London, UK

Jens Rittcher
Oxford University, UK

Organizing Committee

Stephen Aylward, Kitware
Dana Brooks, Northeastern U.
Qi Duan, NIH
Elisa Konofagou, Columbia U.
Jan Kybic, Czech Tech. University
Erik Meijering, Erasmus MC
Wiro Niessen, Erasmus MC
Ricardo Otazo, NYU
Dirk Padfield, GE Healthcare
Gustavo Rohde, Carnegie Mellon
Badri Roysam, U. of Houston
Ivan Selesnick. Polytech NYU
Dimitri Van De Ville, EPFL
Simon Warfield, Harvard
Ge Yang, Carnegie Mellon

Contact

d.bernstein@ieee.org

The IEEE International Symposium on Biomedical Imaging (ISBI) is a premier interdisciplinary conference encompassing all scales of imaging in medicine and the life sciences. The 2015 meeting will continue its tradition of fostering knowledge transfer among different imaging communities and contributing to an integrative approach to biomedical imaging across all scales of observation.

ISBI is a joint initiative from the IEEE Signal Processing Society (SPS) and the IEEE Engineering in Medicine and Biology Society (EMBS). The 2015 meeting will open with a morning of tutorials, followed by a scientific program of plenary talks, invited special sessions, challenges, as well as oral and poster presentations of peer-reviewed papers.

High-quality papers are requested containing original contributions to mathematical, algorithmic, and computational aspects of biomedical imaging, from nano- to macro-scale. Topics of interest include image formation and reconstruction, computational and statistical image processing and analysis, dynamic imaging, visualization, image quality assessment, and physical, biological, and statistical modeling. We also encourage papers that elucidate biological processes (including molecular mechanisms) or translational ramification through integration of image-based data. Accepted 4-page regular papers will be published in the symposium proceedings and included in IEEE Xplore.

To encourage attendance by a broader audience of imaging scientists (in particular from the biology, radiology, and physics community) and offer additional opportunities for cross-fertilization, ISBI will again propose a second track featuring posters selected from abstract submissions without subsequent archival publication.

Important Dates

Tutorials, Special, Sessions & Challenges
Proposal Submission
June — Sept. 2014

4-Page Paper Submission

*Aug. 1st — **Nov. 10th**, 2014*
Notification
Dec. 20th, 2014
Upload & Registration
Jan. 10th, 2015

1-Page Paper Submission

*Nov. 20th , 2014 — **Dec. 20th** , 2014*
Notification
Feb. 1st , 2015
Upload & Registration
Feb. 15th , 2015

Venue: ISBI 2015 will be held at the ***Marriott hotel at the Brooklyn bridge***, located on Adams street, next to the historical Court House building, with premier shopping, dining, and attractions in the heart of the Dumbo district. A short walk will take you to eight subway lines, a city bike station or a yellow cab to explore Brooklyn or to reach Manhattan just 1.5 miles (2 subway stations) across the East river for memorable nights in the Big Apple.

http://biomedicalimaging.org/2015

IEEE TRANSACTIONS ON SIGNAL PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society

www.signalprocessingsociety.org

Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine

PubMed MEDLINE U.S. National Library of Medicine

SEPTEMBER 15, 2014 VOLUME 62 NUMBER 18 ITPRED (ISSN 1053-587X)

REGULAR PAPER

IEEE TRANSACTIONS ON SIGNAL PROCESSING (ISSN 1053-587X) is published semimonthly by the Institute of Electrical and Electronics Engineers, Inc. Responsibility for the contents rests upon the authors and not upon the IEEE, the Society/Council, or its members. **IEEE Corporate Office:** 3 Park Avenue, 17th Floor, New York, NY 10016-5997. **IEEE Operations Center:** 445 Hoes Lane, Piscataway, NJ 08854-4141. **NJ Telephone:** +1 732 981 0060. **Price/Publication Information**: Individual copies: IEEE Members $20.00 (first copy only), nonmembers $569.00 per copy. (Note: Postage and handling charge not included.) Member and nonmember subscription prices available upon request. **Copyright and Reprint Permissions:** Abstracting is permitted with credit to the source. Libraries are permitted to photocopy for private use of patrons, provided the per-copy fee of $31.00 is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For all other copying, reprint, or republication permission, write to Copyrights and Permissions Department, IEEE Publications Administration, 445 Hoes Lane, Piscataway, NJ 08854-4141.

Periodicals Postage Paid at New York, NY and at additional mailing offices. **Postmaster:** Send address changes to IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE, 445 Hoes Lane, Piscataway, NJ 08854-4141. GST Registration No. 125634188. CPC Sales Agreement #40013087. Return undeliverable Canada addresses to: Pitney Bowes IMEX, P.O. Box 4332, Stanton Rd., Toronto, ON M5W 3J4, Canada. IEEE prohibits discrimination, harassment and bullying. For more information visit http://www.ieee.org/nondiscrimination. Printed in U.S.A.

IEEE

IEEE TRANSACTIONS ON SIGNAL PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society

www.signalprocessingsociety.org

Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine

PubMed | MEDLINE U.S. National Library of Medicine

OCTOBER 1, 2014 VOLUME 62 NUMBER 19 ITPRED (ISSN 1053-587X)

IEEE TRANSACTIONS ON SIGNAL PROCESSING (ISSN 1053-587X) is published semimonthly by the Institute of Electrical and Electronics Engineers, Inc. Responsibility for the contents rests upon the authors and not upon the IEEE, the Society/Council, or its members. **IEEE Corporate Office:** 3 Park Avenue, 17th Floor, New York, NY 10016-5997. **IEEE Operations Center:** 445 Hoes Lane, Piscataway, NJ 08854-4141. **NJ Telephone:** +1 732 981 0060. **Price/Publication Information**: Individual copies: IEEE Members $20.00 (first copy only), nonmembers $569.00 per copy. (Note: Postage and handling charge not included.) Member and nonmember subscription prices available upon request. **Copyright and Reprint Permissions:** Abstracting is permitted with credit to the source. Libraries are permitted to photocopy for private use of patrons, provided the per-copy fee of $31.00 is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For all other copying, reprint, or republication permission, write to Copyrights and Permissions Department, IEEE Publications Administration, 445 Hoes Lane, Piscataway, NJ 08854-4141.

Periodicals Postage Paid at New York, NY and at additional mailing offices. **Postmaster:** Send address changes to IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE, 445 Hoes Lane, Piscataway, NJ 08854-4141. GST Registration No. 125634188. CPC Sales Agreement #40013087. Return undeliverable Canada addresses to: Pitney Bowes IMEX, P.O. Box 4332, Stanton Rd., Toronto, ON M5W 3J4, Canada. IEEE prohibits discrimination, harassment and bullying. For more information visit http://www.ieee.org/nondiscrimination. Printed in U.S.A.

IEEE

IEEE TRANSACTIONS ON
SIGNAL PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society

www.signalprocessingsociety.org

Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine

PubMed MEDLINE U.S. National Library of Medicine

OCTOBER 15, 2014 VOLUME 62 NUMBER 20 ITPRED (ISSN 1053-587X)

REGULAR PAPERS

IEEE

The Ninth IEEE Sensor Array and Multichannel Signal Processing Workshop

10th-13th July 2016,Rio de Janeiro, Brazil

IEEE Signal Processing Society

Call for Papers

General Chairs
Rodrigo C. de Lamare,
PUC-Rio, Brazil and University of York, United Kingdom

Martin Haardt,
TU Ilmenau, Germany

Technical Chairs
Aleksandar Dogandzic,
Iowa State University, USA

Vítor Nascimento,
University of São Paulo, Brazil

Special Sessions Chair
Cédric Richard,
University of Nice, France

Publicity Chair
Maria Sabrina Greco,
Universityof Pisa, Italy

Important Dates

Special Session Proposals
29th January , 2016

Submission of Papers
26thFebruary, 2016

Notification of Acceptance
29th April , 2016

Final Manuscript Submission
16th May, 2016

Advance Registration
16th May, 2016

Technical Program
The SAM Workshop is an important IEEE Signal Processing Society event dedicated to sensor array and multichannel signal processing. The organizing committee invites the international community to contribute with state-of-the-art developments in the field. SAM 2016 will feature plenary talks by leading researchers in the field as well as poster and oral sessions with presentations by the participants.

Welcome to Rio de Janeiro! – The workshop will be held at the Pontifical Catholic University of Rio de Janeiro, located in Gávea, in a superb area surrounded by beaches, mountains and the Tijuca National Forest, the world's largest urban forest. Rio de Janeiro is a world renowned city for its culture, beautiful landscapes, numerous tourist attractions and international cuisine. The workshop will take place during the first half of July about a month before the 2016 Summer Olympic Games when Rio will offer plenty of cultural activities and festivities, which will make SAM 2016 a memorable experience.

Research Areas
Authors are invited to submit contributions in the following areas:
- Adaptive beamforming
- Array processing for biomedical applications
- Array processing for communications
- Blind source separation and channel identification
- Computational and optimization techniques
- Compressive sensing and sparsity-based signal processing
- Detection and estimation
- Direction-of-arrival estimation
- Distributed and adaptive signal processing
- Intelligentsystems and knowledge-based signal processing
- Microphone and loudspeaker array applications
- MIMO radar
- Multi-antenna systems: multiuser MMO, massive MIMO and space-time coding
- Multi-channel imaging and hyperspectral processing
- Multi-sensor processing for smart grid and energy
- Non-Gaussian, nonlinear, and non-stationary models
- Performance evaluations with experimental data
- Radar and sonar array processing
- Sensor networks
- Source Localization, Classification and Tracking
- Synthetic aperture techniques
- Space-time adaptive processing
- Statistical modelling for sensor arrays
- Waveform diverse sensors and systems

Submission of papers – Full-length four-page papers will be accepted only electronically.

Special session proposals – They should be submitted by e-mail to the Technical Program Chairs and the Special Sessions Chair and include a topical title, rationale, session outline, contact information, and list of invited speakers.

IEEE/ACM TRANSACTIONS ON

AUDIO, SPEECH, AND LANGUAGE PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society

acm Association for Computing Machinery

www.signalprocessingsociety.org

Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine

PubMed MEDLINE U.S. National Library of Medicine

SEPTEMBER 2014 VOLUME 22 NUMBER 9 ITASFA (ISSN 2329-9290)

IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING (ISSN 2329-9290) is published bimonthly in print and monthly online by the Institute of Electrical and Electronics Engineers, Inc. Responsibility for the contents rests upon the authors and not upon the IEEE, the Society/Council, or its members. **IEEE Corporate Office:** 3 Park Avenue, 17th Floor, New York, NY 10016-5997. **IEEE Operations Center:** 445 Hoes Lane, Piscataway, NJ 08854-4141. **NJ Telephone:** +1 732 981 0060. **Price/Publication Information:** Individual copies: IEEE Members $20.00 (first copy only), nonmembers $307.00 per copy. (Note: Postage and handling charge not included.) Member and nonmember subscription prices available upon request. Periodicals Postage Paid at New York, NY and at additional mailing offices. **Postmaster:** Send address changes to IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, IEEE, 445 Hoes Lane, Piscataway, NJ 08854-4141. GST Registration No. 125634188. CPC Sales Agreement #40013087. Return undeliverable Canada addresses to: Pitney Bowes IMEX, P.O. Box 4332, Stanton Rd., Toronto, ON M5W 3J4, Canada. IEEE prohibits discrimination, harassment and bullying. For more information visit http://www.ieee.org/nondiscrimination. Printed in U.S.A.

IEEE

IEEE/ACM TRANSACTIONS ON

AUDIO, SPEECH, AND LANGUAGE PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society

acm Association for Computing Machinery

www.signalprocessingsociety.org

Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine

PubMed MEDLINE U.S. National Library of Medicine

OCTOBER 2014 VOLUME 22 NUMBER 10 ITASFA (ISSN 2329-9290)

REGULAR PAPERS

Acoustic Sensor Array Processing

IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING (ISSN 2329-9290) is published bimonthly in print and monthly online by the Institute of Electrical and Electronics Engineers, Inc. Responsibility for the contents rests upon the authors and not upon the IEEE, the Society/Council, or its members. **IEEE Corporate Office:** 3 Park Avenue, 17th Floor, New York, NY 10016-5997. **IEEE Operations Center:** 445 Hoes Lane, Piscataway, NJ 08854-4141. **NJ Telephone:** +1 732 981 0060. **Price/Publication Information:** Individual copies: IEEE Members $20.00 (first copy only), nonmembers $307.00 per copy. (Note: Postage and handling charge not included.) Member and nonmember subscription prices available upon request. Periodicals Postage Paid at New York, NY and at additional mailing offices. **Postmaster:** Send address changes to IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, IEEE, 445 Hoes Lane, Piscataway, NJ 08854-4141. GST Registration No. 125634188. CPC Sales Agreement #40013087. Return undeliverable Canada addresses to: Pitney Bowes IMEX, P.O. Box 4332, Stanton Rd., Toronto, ON M5W 3J4, Canada. IEEE prohibits discrimination, harassment and bullying. For more information visit http://www.ieee.org/nondiscrimination. Printed in U.S.A.

IEEE

IEEE TRANSACTIONS ON

IMAGE PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society

www.signalprocessingsociety.org

Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine

PubMed MEDLINE U.S. National Library of Medicine

SEPTEMBER 2014 VOLUME 23 NUMBER 9 IIPRE4 (ISSN 1057-7149)

PAPERS

IEEE

IEEE TRANSACTIONS ON

IMAGE PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society

www.signalprocessingsociety.org

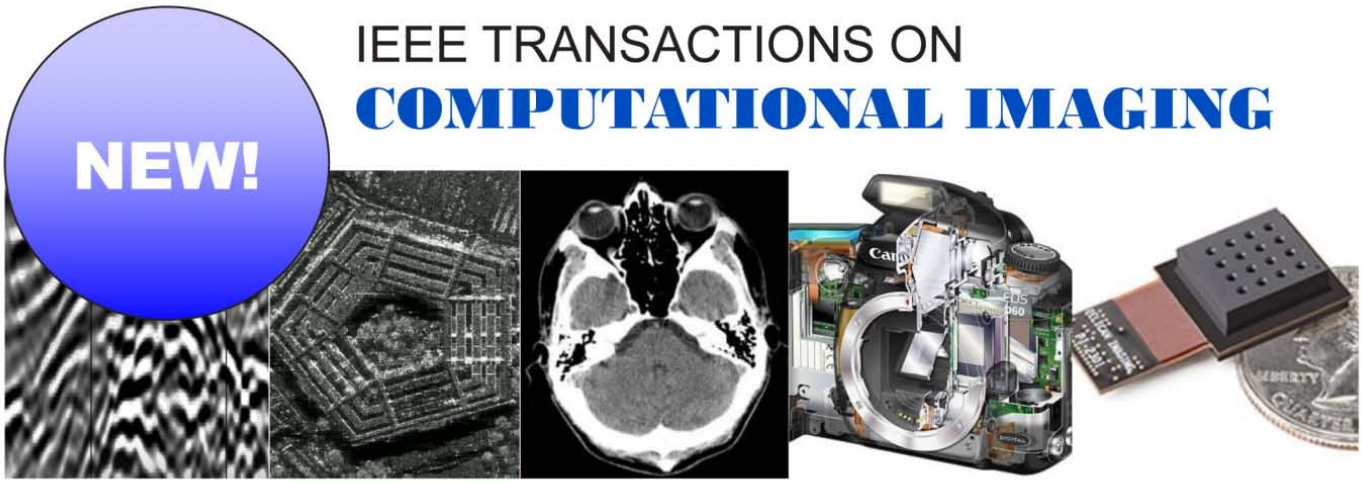
Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine

PubMed MEDLINE U.S. National Library of Medicine

OCTOBER 2014 VOLUME 23 NUMBER 10 IIPRE4 (ISSN 1057-7149)

PAPERS

IEEE



IEEE TRANSACTIONS ON COMPUTATIONAL IMAGING

The new IEEE Transactions on Computational Imaging seeks original manuscripts for publication. This new journal will publish research results where computation plays an integral role in the image formation process. All areas of computational imaging are appropriate, ranging from the principles and theory of computational imaging, to modeling paradigms for computational imaging, to image formation methods, to the latest innovative computational imaging system designs. Topics of interest include, but are not limited to the following:

Imaging Models and Representation

- Statistical-model based methods
- System and image prior models
- Noise models
- Graphical and tree-based models
- Perceptual models

Computational Sensing

- Coded source methods
- Structured light
- Coded aperture methods
- Compressed sensing
- Light-field sensing
- Plenoptic imaging
- Hardware and software systems

Computational Image Creation

- Sparsity-based methods
- Statistically-based inversion methods, Bayesian regularization
- Super-resolution, multi-image fusion
- Learning-based methods, Dictionary-based methods
- Optimization-based methods; proximal iterative methods, ADMM

Computational Photography

- Non-classical image capture, Generalized illumination
- Time-of-flight imaging
- High dynamic range imaging
- Focal stacks

Computational Consumer Imaging

- Cell phone imaging
- Camera-array systems
- Depth cameras

Computational Acoustic Imaging

- Multi-static ultrasound imaging
- Photo-acoustic imaging
- Acoustic tomography

Computational Microscopic Imaging

- Holographic microscopy
- Quantitative phase imaging
- Multi-illumination microscopy
- Lensless microscopy

Tomographic Imaging

- X-ray CT
- PET
- SPECT

Magnetic Resonance Imaging

- Diffusion tensor imaging
- Fast acquisition

Radar Imaging

- Synthetic aperture imaging
- Inverse synthetic imaging
- Terahertz imaging

Geophysical Imaging

- Multi-spectral imaging
- Ground penetrating radar
- Seismic tomography

Multi-spectral Imaging

- Multi-spectral imaging
- Hyper-spectral imaging
- Spectroscopic imaging

Editor-in-Chief: W. Clem Karl, Boston University.
To submit a paper go to: https://mc.manuscriptcentral.com/tci-ieee

IEEE Signal Processing Society | EMB | GRSS

IEEE TRANSACTIONS ON

INFORMATION FORENSICS AND SECURITY

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society

www.signalprocessingsociety.org

SEPTEMBER 2014 VOLUME 9 NUMBER 9 ITIFA6 (ISSN 1556-6013)

PAPERS

IEEE TRANSACTIONS ON

INFORMATION FORENSICS AND SECURITY

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society

www.signalprocessingsociety.org

OCTOBER 2014 VOLUME 9 NUMBER 10 ITIFA6 (ISSN 1556-6013)

PAPERS

ANNOUNCEMENTS

ICME 2015

IEEE International Conference on Multimedia & Expo

Torino, Italy, June 29 – July 3, 2015

http://www.icme2015.ieee-icme.org

General Chairs
Enrico Magli – *Politecnico di Torino*
Stefano Tubaro – *Polit. di Milano*
Anthony Vetro – *MERL*

Technical Program Chairs
Marco Tagliasacchi – *Polit. di Milano*
Yap-Peng Tan – *NTU*
Max Mühlhäuser – *TU Darmstadt*
Sanjeev Mehrotra – *Microsoft Res.*
Tao Mei – *Microsoft Research*

Finance and Local Arrang. Chairs
Marco Grangetto – *Univ. Torino*
Giulio Coluccia – *Politecnico di Torino*
Roberto Rinaldo – *Univ. Udine*

Plenary Chairs
Andrea Cavallaro – *QMUL*
Shipeng Li – *Microsoft Research Asia*

Workshop Chairs
Pascal Frossard - *EPFL*
Gyorgy Dan – *KTH*

Tutorial Chairs
Joao Ascenso – *IT*
Gene Cheung – *NII*

Special Session and Events Chair
Antonio Ortega – *USC*
Ivan Bajic – *SFU*

Panel Chairs
Fernando Pereira – *IT*
Alberto del Bimbo – *Univ. Firenze*

Electronic Media and Pub. Chair
Matteo Cesana – *Polit. Milano*

Publicity Chair
Chia-Wen Lin – *NTHU*
A. Rocha – *Unicamp*

Innovation and Demo Chairs
Carlo Regazzoni – *Univ. Genova*
Bernhard Rinner – *Univ. Klagenfurt*

Exhib. & Industry connection Chairs
Giovanni Cordara - Huawei
Fabrizio Rovati - STMicroelectronics
Thomas Stockhammer - Nomor

Awards Chairs
Riccardo Leonardi – *Univ. Brescia*
Christine Guillemot – *INRIA*

Liaisons Chairs
Chia-Wen Lin – *NTHU*
C.C.-Jay Kuo – *USC*
D. Taubman – *UNSW*
A. Rocha – *Unicamp*

CALL FOR PAPERS

The IEEE International Conference on Multimedia & Expo (ICME) has been the flagship multimedia conference sponsored by four IEEE societies since 2000. It serves as a forum to promote the exchange of the latest advances in multimedia technologies, systems, and applications from both the research and development perspectives of the circuits and systems, communications, computer, and signal processing communities. ICME also features an Exposition of multimedia products and prototypes.

Authors are invited to submit a full paper (two-column format, maximum 6 pages) according to the guidelines available on the conference website at http://www.icme2015.ieee-icme.org/. Topics of interest include, but are not limited to:

- Multimedia content analysis
- Multimedia activity and event understanding
- Multimedia search and retrieval
- Mobile, location-based and other context-based multimedia
- Social, user-generated, and cloud-based multimedia
- 3D immersion and virtual reality
- Multimedia security and forensics
- Human computer interaction based on multimedia
- Multimedia networking and communication
- Multimedia coding and compression
- Multimedia signal processing and enhancement
- Multimedia systems, applications, services and implementations

ICME 2015 aims to have high quality oral and poster presentations. Several awards sponsored by industry and institutions will be granted. The conference will provide student author travel grants.

A number of workshops will be organized by the sponsoring societies. To further foster new emerging topics, ICME 2015 also welcomes researchers, developers and practitioners to organize regular workshops. Industrial exhibitions are held in parallel with the main conference. Proposals for tutorials, panels, and demos are also invited. Please visit the ICME 2015 website for submission details.

Important Dates

Proposals for Workshops:	Nov. 21, 2014
Full Paper Submissions:	**Nov. 28, 2014**
Notification to Authors:	Mar. 15, 2015
Camera-ready Paper Submission:	Apr. 3, 2015

IEEE TRANSACTIONS ON

MULTIMEDIA

A PUBLICATION OF
THE IEEE CIRCUITS AND SYSTEMS SOCIETY
THE IEEE SIGNAL PROCESSING SOCIETY
THE IEEE COMMUNICATIONS SOCIETY
THE IEEE COMPUTER SOCIETY

http://www.signalprocessingsociety.org/tmm/

OCTOBER 2014 VOLUME 16 NUMBER 6 ITMUF8 (ISSN 1520-9210)

IEEE JOURNAL OF
SELECTED TOPICS IN SIGNAL PROCESSING

IEEE Signal Processing Society

www.ieee.org/sp/index.html

OCTOBER 2014 VOLUME 8 NUMBER 5 IJSTGY (ISSN 1932-4553)

EDITORIAL

PAPERS

IEEE

IEEE

SIGNAL PROCESSING LETTERS

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society

www.ieee.org/sp/index.html

NOVEMBER 2014 VOLUME 21 NUMBER 11 ISPLEM (ISSN 1070-9908)

LETTERS

SignalProcessing | Previous Page | Contents | Zoom in | Zoom out | Front Cover | Search Issue | Next Page | Qmags


IEEE
SignalProcessing
MAGAZINE
[VOLUME 31 NUMBER 6 NOVEMBER 2014]
THE 5G REVOLUTION
ADVANCES AND POTENTIAL CHALLENGES
5G
TELEIMMERSIVE AUDIO-VISUAL COMMUNICATION
ULTRAWIDEBAND SIGNALS IN MEDICINE
REFLECTIONS ON EXCELLENCE IN RESEARCH AND EDUCATION
IEEE Signal Processing Society
IEEE

CONTENTS

[VOLUME 31 NUMBER 6]

SPECIAL SECTION—THE 5G REVOLUTION

COLUMNS

DEPARTMENT

Digital Object Identifier 10.1109/MSP.2014.2350051

Call for Papers

www.icip2015.org

The International Conference on Image Processing (ICIP), sponsored by the IEEE Signal Processing Society, is the premier forum for the presentation of technological advances and research results in the fields of theoretical, experimental, and applied image and video processing. ICIP 2015, the twenty second in the series that has been held annually since 1994, brings together leading engineers and scientists in image and video processing from around the world. Research frontiers in fields ranging from traditional image processing applications to evolving multimedia and video technologies are regularly advanced by results first reported in ICIP technical sessions.

Topics include, but are not limited to:

- **Image/video coding and transmission:** Still-image and video coding, stereoscopic and 3-D coding, distributed source coding, source/channel coding, image/video transmission over wireless networks;
- **Image/video processing:** Image and video filtering, restoration and enhancement, image segmentation, video segmentation and tracking, morphological processing, stereoscopic and 3-D processing, feature extraction and analysis, interpolation and super-resolution, motion detection and estimation, color and multispectral processing, biometrics;
- **Image formation:** Biomedical imaging, remote sensing, geophysical and seismic imaging, optimal imaging, synthetic-natural hybrid image systems;
- **Image scanning, display, and printing**: Scanning and sampling, quantization and half toning, color reproduction, image representation and rendering, display and printing systems, image-quality assessment;
- **Image/video storage, retrieval, and authentication:** Image and video databases, image and video search and retrieval, multimodality image/ video indexing and retrieval, authentication and watermarking;
- **Applications:** Biomedical sciences, mobile imaging, geosciences & remote sensing, astronomy & space exploration, document image processing and analysis, other applications.

Paper Submission: Authors are invited to submit papers of not more than four pages for technical content including figures and references, with one optional page containing only references.

Call for Tutorials: Tutorials will be held on Sunday, September 27, 2015. Proposals should be submitted by January 15, 2015 to tutorials@icip2015.org and must include title, outline of the tutorial and its motivation, short description, contact information and credentials for each presenter including name, affiliation, email, mailing address, and a two-page resume.

Call for Special Sessions: Proposals should be submitted by November 27, 2014 in a single PDF document sent to specialsessions@icip2015.org. Please include title, motivation for the special session topic, potential authors and titles of papers, as well as contact information and credentials for each organizer including name, affiliation, email, mailing address, and a short resume.

Important Dates

- Special Sessions Proposals: **27 November 2014**
- Regular Papers Submission: **15 January 2015**
- Tutorials Proposal: **15 January 2015**

Visit icip2015.org for details on paper submission, social events, no-show policy, and more.

Organizing Committee

General Co-Chairs
Jean-Luc DUGELAY
André MORIN

Technical Program Chairs
Fabrice LABEAU
Jean-Philippe THIRAN

Finances
Jean FORTIN

Plenary Sessions
Stéphane COULOMBE
Kenneth ROSE

Special Sessions
Oscar C. AU
Éric DUBOIS

Tutorials
Janusz KONRAD
André ZACCARIN

Local Arrangements
Paul FORTIER

Registration
Xavier MALDAGUE

Exhibit/Industry
Khaled EL-MALEH
Branislav KISACANIN

Publicity
Aishy AMER

Publications
Mireille BOUTIN

Electronic Media
Abdulmotaleb EL SADDIK
Benoît HUET

International Liaison
Carlo S. REGAZZONI
Wan-Chi SIU

Student Activities
Sylvie DANIEL
Guoliang FAN

Awards
Phil CHOU

Québec
Office du tourisme de Québec

Subject: IEEE Signal Processing Cup 2015 at ICASSP2015

Call for Participation: IEEE Signal Processing Cup 2015
http://icassp2015.org/signal-processing-cup-2015/

Challenge: **Heart Rate Monitoring During Physical Exercise Using Wrist-Type Photoplethysmographic (PPG) Signals**
For details of the competition project, please visit: www.zhilinzhang.com/spcup2015/

The IEEE Signal Processing Society organizes the SP Cup competition for undergraduates at ICASSP2015. This competition aims to provide undergraduate students with the opportunity to form teams and work together to solve a challenging and interesting real-world problem using signal-processing techniques. Three teams will be selected to present their work, and the prizes will be awarded at ICASSP 2015.

You are very welcome to participating in the competition. Please also help us to circulate this email to other colleagues or students you know who may be interested in this competition.

Participation in the Competition:

Each team participating in the competition is to be composed of one faculty member (whose role is the supervisor of the team members), at most one graduate student (who will assist the supervisor in supervising the undergraduate team members), and at least 3 but no more than 10 undergraduates. At least three of the undergraduate team members must be either IEEE SP members or student members.

Participating teams must submit their project by February 6, 2015. Each submission should include a report, in the form of an IEEE conference paper, on the technical details of the methods used and the results, as well as the programs developed (MATLAB is preferred). Participating teams must register to join the competition by January 16, 2015. The online registration system will be open in September 2014. By February 27, 2015, the best 3 teams will be identified to participate in the final competition at ICASSP2015.

Important Dates:

January 16, 2015 (Friday): Team registration to join the SP Cup competition
February 6, 2015 (Friday): Submission deadline for participating teams
February 27, 2015 (Friday): Announcement of the best 3 teams
April 20, 2015: Final competition at ICASSP 2015

Team Prizes:

The champion:	$5,000
The first runner-up:	$2,500
The second runner-up:	$1,500

Each team invited to ICASSP2015 will have their travel expenses supported by the SP Society. Each team member is offered up to $1200 for continental travel, or $1,700 for intercontinental travel, and at most 3 people from each team will be supported.

Enquiries:

Technical problems: zhilinzhang@ieee.org
General enquiry: sp-enq-spcup@ieee.org

Organizers:

Bio Imaging and Signal Processing Technical Committee (BISP TC)
IEEE SPS Student Services Committee

12th IEEE International Conference on Advanced Video and Signal-based Surveillance.

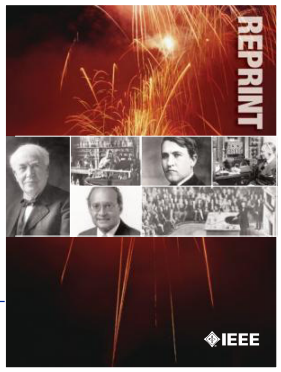
August 25-28 2015

Karlsruhe Germany.

General Chair	Jürgen Beyerer, Fraunhofer IOSB / KIT, Germany.
	Rainer Stiefelhagen, KIT, Germany.
Organization Chair	Eduardo Monari, Fraunhofer IOSB, Germany.
Program Chairs	Roman Pflugfelder, TU Graz/AIT, Austria.
	Peter Tu, GE Global research, USA.
Workshops Chair	Rita Cucchiara, Univ. of Modena/Reggio Emilia, Italy.
	Jordi Gonzales, UA de Barcelona, Spain.
Challenge Chair	James Ferryman, Univ. of Reading, UK.
Publicity Chairs	Marion Staub, Fraunhofer IOSB, Germany.
Publications Chair	Francois Bremond, INRIA, France.
	Jean-Marc Odobez, IDIAP/EPFL, Switzerland.
Local Arrangements	Uwe Böhm, KIT, Germany.
& financial chair	Marion Staub, Fraunhofer IOSB, Germany.
Area Chairs	Fatih Porikli, NICTA, Australia.
	Michael Arens, Fraunhofer IOSB, Germany.
	Bir Bhanu, UC Riverside, USA.
	Rainer Stiefelhagen, KIT, Germany.
	Jörg Hähner, Univ. of Augsburg, Germany.
	Jean-Luc Dugelay, Eurecom, France.
Steering Committee	Fatih Porikli, NICTA, Australia.
	Carlo Regazzoni, Univ. of Genua, Italy.
	Andrea Cavallaro, Queen Mary Univ. of London, UK.
	Massimo Piccardi, Univ. of Tech. Sydney, Australia.
	Kostas Plataniotis, Univ. of Toronto, Canada.
	Bernhard Rinner, Univ. of Klagenfurt, Austria.
	Peter Tu, GE Global research, USA.
	Stefano Tubaro, Politecnico die Milano, Italy.

info@avss2015.org

www.avss2015.org



IEEE ORDER FORM FOR REPRINTS

Purchasing IEEE Papers in Print is easy, cost-effective and quick.

Complete this form, send via our secure fax (24 hours a day) to 732-981-8062 or mail it back to us.

PLEASE FILL OUT THE FOLLOWING

Author: _____

Publication Title: _____

Paper Title: _____

_____

RETURN THIS FORM TO:
IEEE Publishing Services
445 Hoes Lane
Piscataway, NJ 08855-1331

Email the Reprint Department at reprints@ieee.org for questions regarding this form

PLEASE SEND ME

■ 50 ■ 100 ■ 200 ■ 300 ■ 400 ■ 500 or _____(in multiples of 50) reprints.

■ YES ■ NO Self-covering/title page required. COVER PRICE: $74 per 100, $39 per 50.

■ $58.00 Air Freight must be added for all orders being shipped outside the U.S.

■ $21.50 must be added for all USA shipments to cover the cost of UPS shipping and handling.

PAYMENT

■ Check enclosed. Payable on a bank in the USA.

■ Charge my: ■ Visa ■ Mastercard ■ Amex ■ Diners Club

Account # _____ Exp. date _____

Cardholder's Name (please print): _____

_____

■ Bill me (you must attach a purchase order) Purchase Order Number _____

Send Reprints to:

_____

_____

_____

_____

Bill to address, if different:

_____

_____

_____

_____

Because information and papers are gathered from various sources, there may be a delay in receiving your reprint request. This is especially true with postconference publications. Please provide us with contact information if you would like notification of a delay of more than 12 weeks.

Telephone: _____ Fax: _____ Email Address: _____

2012 REPRINT PRICES (without covers)

Number of Text Pages

	1-4	5-8	9-12	13-16	17-20	21-24	25-28	29-32	33-36	37-40	41-44	45-48
50	$129	$213	$245	$248	$288	$340	$371	$408	$440	$477	$510	$543
100	$245	$425	$479	$495	$573	$680	$742	$817	$885	$953	$1021	$1088

Larger quantities can be ordered. Email reprints@ieee.org with specific details.

Tax Applies on shipments of regular reprints to CA, DC, FL, MI, NJ, NY, OH and Canada (GST Registration no. 12534188).
Prices are based on black & white printing. Please call us for full color price quote, if applicable.

2015 IEEE MEMBERSHIP APPLICATION

(students and graduate students must apply online)

IEEE
Advancing Technology for Humanity

Start your membership immediately: Join online www.ieee.org/join

Please complete both sides of this form, typing or **printing in capital letters**. Use only English characters and abbreviate only if more than 40 characters and spaces per line. We regret that incomplete applications cannot be processed.

1 Name & Contact Information

Please PRINT your name as you want it to appear on your membership card and IEEE correspondence. As a key identifier for the IEEE database, circle your last/surname.

☐ Male ☐ Female Date of birth (Day/Month/Year) ____/____/____

Title First/Given Name Middle Last/Family Surname

▼ Primary Address ☐ Home ☐ Business (All IEEE mail sent here)

Street Address

City State/Province

Postal Code Country

Primary Phone

Primary E-mail

▼ Secondary Address ☐ Home ☐ Business

Company Name Department/Division

Street Address City State/Province

Postal Code Country

Secondary Phone

Secondary E-mail

To better serve our members and supplement member dues, your postal mailing address is made available to carefully selected organizations to provide you with information on technical services, continuing education, and conferences. Your e-mail address is not rented by IEEE. Please check box only if you do not want to receive these postal mailings to the selected address. ☐

2 Attestation

I have graduated from a three- to five-year academic program with a university-level degree.
☐ Yes ☐ No

This program is in one of the following fields of study:
☐ Engineering
☐ Computer Sciences and Information Technologies
☐ Physical Sciences
☐ Biological and Medical Sciences
☐ Mathematics
☐ Technical Communications, Education, Management, Law and Policy
☐ Other (please specify): _____

This academic institution or program is accredited in the country where the institution is located. ☐ Yes ☐ No ☐ Do not know

I have _____ years of professional experience in teaching, creating, developing, practicing, or managing within the following field:

☐ Engineering
☐ Computer Sciences and Information Technologies
☐ Physical Sciences
☐ Biological and Medical Sciences
☐ Mathematics
☐ Technical Communications, Education, Management, Law and Policy
☐ Other (please specify): _____

3 Please Tell Us About Yourself

Select the numbered option that best describes yourself. This information is used by IEEE magazines to verify their annual circulation. Please enter numbered selections in the boxes provided.

A. Primary line of business → ☐

1. Computers
2. Computer peripheral equipment
3. Software
4. Office and business machines
5. Test, measurement and instrumentation equipment
6. Communications systems and equipment
7. Navigation and guidance systems and equipment
8. Consumer electronics/appliances
9. Industrial equipment, controls and systems
10. ICs and microprocessors
11. Semiconductors, components, sub-assemblies, materials and supplies
12. Aircraft, missiles, space and ground support equipment
13. Oceanography and support equipment
14. Medical electronic equipment
15. OEM incorporating electronics in their end product (not elsewhere classified)
16. Independent and university research, test and design laboratories and consultants (not connected with a mfg. co.)
17. Government agencies and armed forces
18. Companies using and/or incorporating any electronic products in their manufacturing, processing, research or development activities
19. Telecommunications services, telephone (including cellular)
20. Broadcast services (TV, cable, radio)
21. Transportation services (airline, railroad, etc.)
22. Computer and communications and data processing services
23. Power production, generation, transmission and distribution
24. Other commercial users of electrical, electronic equipment and services (not elsewhere classified)
25. Distributor (reseller, wholesaler, retailer)
26. University, college/other educational institutions, libraries
27. Retired
28. Other_____

B. Principal job function → ☐

1. General and corporate management
2. Engineering management
3. Project engineering management
4. Research and development management
5. Design engineering management—analog
6. Design engineering management—digital
7. Research and development engineering
8. Design/development engineering—analog
9. Design/development engineering—digital
10. Hardware engineering
11. Software design/development
12. Computer science
13. Science/physics/mathematics
14. Engineering (not elsewhere specified)
15. Marketing/sales/purchasing
16. Consulting
17. Education/teaching
18. Retired
19. Other _____

C. Principal responsibility → ☐

1. Engineering and scientific management
2. Management other than engineering
3. Engineering design
4. Engineering
5. Software: science/mngmnt/engineering
6. Education/teaching
7. Consulting
8. Retired
9. Other _____

D. Title → ☐

1. Chairman of the Board/President/CEO
2. Owner/Partner
3. General Manager
4. VP Operations
5. VP Engineering/Dir. Engineering
6. Chief Engineer/Chief Scientist
7. Engineering Management
8. Scientific Management
9. Member of Technical Staff
10. Design Engineering Manager
11. Design Engineer
12. Hardware Engineer
13. Software Engineer
14. Computer Scientist
15. Dean/Professor/Instructor
16. Consultant
17. Retired
18. Other _____

Are you now or were you ever a member of IEEE?
☐ Yes ☐ No If yes, provide, if known:

Membership Number Grade Year Expired

4 Please Sign Your Application

I hereby apply for IEEE membership and agree to be governed by the IEEE Constitution, Bylaws, and Code of Ethics. I understand that IEEE will communicate with me regarding my individual membership and all related benefits. **Application must be signed.**

Signature Date

Over Please

Information for Authors

(Updated/Effective September 17, 2014)

The IEEE TRANSACTIONS ON SIGNAL PROCESSING is published online twice per month (semimonthly) covering advances in the theory and application of signal processing. The scope is reflected in the EDICS: the Editor's Information and Classification Scheme. Please consider the journal with the most appropriate scope for your submission.

Authors are encouraged to submit manuscripts of Regular papers (papers which provide a complete disclosure of a technical premise), or Comment Correspondences (brief items that provide comment on a paper previously published in the TRANSACTIONS). Submissions/resubmissions must be previously unpublished and may not be under consideration elsewhere.

Every manuscript must (a) provide a clearly defined statement of the problem being addressed, (b) state why it is important to solve the problem, and (c) give an indication as to how the current solution fits into the history of the problem, including bibliographic references to related work rather than restating established algorithms and scientific principles.

In order to be considered for review, a paper must be within the scope of the journal and represent a novel contribution. A paper is a candidate for an Immediate Rejection if it is of limited novelty, e.g. a straightforward combination of theories and algorithms that are well established and are repeated on a known scenario, no new experimental data or new application. Experimental contributions will be rejected without review if there is insufficient experimental data. The TRANSACTIONS are published in English. Papers that have a large number of typographical and/or grammatical errors will also be rejected without review.

By submission/resubmission of your manuscript to this TRANSACTIONS, you are acknowledging that you accept the rules established for publication of manuscripts, including agreement to pay all overlength page charges, color charges, and any other charges and fees associated with publication of the manuscript. Such charges are not negotiable and cannot be suspended.

New and revised manuscripts should be prepared following the "Manuscript Submission" guidelines below, and submitted to the online manuscript system ScholarOne Manuscripts. After acceptance, finalized manuscripts should be prepared following the "Final Manuscript Submission Guidelines" below. Do not send original submissions or revisions directly to the Editor-in-Chief or Associate Editors; they will only access your manuscript electronically via the ScholarOne Manuscripts system.

Manuscript Submission. Please follow the next steps.

1. *Account in ScholarOne Manuscripts*. If necessary, create an account in the on-line submission system ScholarOne Manuscripts. Please check first if you already have an existing account which is based on your e-mail address and may have been created for you when you reviewed or authored a previous paper.
2. *Electronic Manuscript*. Prepare a PDF file containing your manuscript in double-column, single-spaced format using a font size of 10 points or larger, having a margin of at least 1 inch on all sides. For a regular paper, the manuscript may not exceed 13 double-column pages, including title; names of authors and their complete contact information; abstract; text; all images, figures and tables, appendices and proofs; and all references.

   Upload this version of the manuscript as a PDF file "double.pdf" to the ScholarOneManuscripts site. You are encouraged to also submit a single-column, double-spaced version (11 point font or larger), but page length restrictions will be determined by the double-column version.

   For regular papers, the *revised* manuscript may not exceed 16 double-column pages (10 point font), including title; names of authors and their complete contact information; abstract; text; all images, figures and tables, appendices and proofs; and all references.

   Proofread your submission, confirming that all figures and equations are visible in your document before you "SUBMIT" your manuscript. Proofreading is critical; once you submit your manuscript, the manuscript cannot be changed in any way. You may also submit your manuscript as a PostScript or MS Word file. The system has the capability of converting your files to PDF, however it is your responsibility to confirm that the conversion is correct and there are no font or graphics issues prior to completing the submission process.
3. *Additional Documents for Review*. Please upload pdf versions of all items in the reference list which are not publicly available, such as unpublished (submitted) papers. Other materials for review such as supplementary tables and figures may be uploaded as well. Reviewers will be able to view these files only if they have the appropriate software on their computers. Use short filenames without spaces or special characters. When the upload of each file is completed, you will be asked to provide a description of that file.
4. *Multimedia Materials*. IEEE Xplore can publish multimedia files (audio, images, video) and Matlab code along with your paper. Alternatively, you can provide the links to such files in a README file that appears on Xplore along with your paper. For details, please see http://www.ieee.org/publications_standards/publications/authors/authors_journals.html#sect6 under "Multimedia." To make your work reproducible by others, the TRANSACTIONS encourages you to submit all files that can recreate the figures in your paper. Files that are to be included with the final paper must be uploaded for consideration in the review process.
5. *Submission.* After uploading all files and proofreading them, submit your manuscript by clicking "Submit." A confirmation of the successful submission will open on screen containing the manuscript tracking number and will be followed with an e-mail confirmation to the corresponding and all contributing authors. Once you click "Submit," your manuscript cannot be changed in any way.
6. *Copyright Form and Consent Form.* By policy, IEEE owns the copyright to the technical contributions it publishes on behalf of the interests of the IEEE, its authors, and their employers; and to facilitate the appropriate reuse of this material by others. To comply with the IEEE copyright policies, authors are required to sign and submit a completed "IEEE Copyright and Consent Form" prior to publication by the IEEE.

   The IEEE recommends authors to use an effective electronic copyright form (eCF) tool within the ScholarOne Manuscripts system. You will be redirected to the "IEEE Electronic Copyright Form" wizard at the end of your original submission; please simply sign the eCF by typing your name at the proper location and click on the "Submit" button.

Comment Correspondence. Comment Correspondences provide brief comments on material previously published in the TRANSACTIONS. A comment correspondence may not exceed 2 pages in double-column, single double-spaced format, using 9 point type, with margins of 1 inch minimum on all sides, and including: title, names and contact information for authors, abstract, text, references, and an appropriate number of illustrations and/or tables. Comment Correspondences are submitted in the same way as regular manuscripts (see "Manuscript Submission" above for instructions).

Manuscript Length. Papers published on or after 1 January 2007 can now be up to 10 pages, and any paper in excess of 10 pages will be subject to over length page charges. The IEEE Signal Processing Society has determined that the standard manuscript length shall be no more than 10 published pages (double-column format, 10 point type) for a regular submission. Manuscripts that exceed these limits will incur mandatory over length page charges, as discussed below. Since changes recommended as a result of peer review may require additions to the manuscript, it is strongly recommended that you practice economy in preparing original submissions.

Exceptions to manuscript length requirements may, under extraordinary circumstances, be granted by the Editor-in-Chief. However, such exception does not obviate your requirement to pay any and all over length or additional charges that attach to the manuscript.

Resubmission of Previously Rejected Manuscripts. Authors of manuscripts rejected from any journal are allowed to resubmit their manuscripts only once. At the time of submission, you will be asked whether your manuscript is a new submission or a resubmission of an earlier rejected manuscript. If it is related to a manuscript previously rejected by any journal, you are expected to submit supporting documents identifying the previous submission and detailing how issues raised in the previous reviews have been addressed. Papers that do not disclose connection to a previously rejected paper or that do not provide documentation as to changes made may be immediately rejected.

Full details of the resubmission process can be found in the Signal Processing Society "Policy and Procedures Manual" at http://www.signalprocessingsociety.org/about/governance/policy-procedure/.

Author Misconduct.

Author Misconduct Policy: Plagiarism includes copying someone else's work without appropriate credit, using someone else's work without clear delineation of citation, and the uncited reuse of an authors previously published work that also involves other authors. Plagiarism is unacceptable.

Self-plagiarism involves the verbatim copying or reuse of an authors own prior work without appropriate citation; it is also unacceptable. Self-plagiarism includes duplicate submission of a single journal manuscript to two different journals, and submission of two different journal manuscripts which overlap substantially in language or technical contribution.

Authors may only submit original work that has not appeared elsewhere in a journal publication, nor is under review for another journal publication. Limited overlap with prior journal publications with a common author is allowed only if it is necessary for the readability of the paper. If authors have used their own previously published work as a basis for a new submission, they are required to cite the previous work and very briefly indicate how the new submission offers substantively novel contributions beyond those of the previously published work.

It is acceptable for conference papers to be used as the basis for a more fully developed journal submission. Still, authors are required to cite related prior work; the papers cannot be identical; and the journal publication must include novel aspects.

Digital Object Identifier 10.1109/TSP.2014.2358875

Author Misconduct Procedures: The procedures that will be used by the Signal Processing Society in the investigation of author misconduct allegations are described in the IEEE SPS Policies and Procedures Manual.

Author Misconduct Sanctions: The IEEE Signal Processing Society will apply the following sanctions in any case of plagiarism, or in cases of self-plagiarism that involve an overlap of more than 25% with another journal manuscript:

1) immediate rejection of the manuscript in question;
2) immediate withdrawal of all other submitted manuscripts by any of the authors, submitted to any of the Society's publications (journals, conferences, workshops), except for manuscripts that also involve innocent co-authors; immediate withdrawal of all other submitted manuscripts by any of the authors, submitted to any of the Society's publications (journals, conferences, workshops), except for manuscripts that also involve innocent co-authors;
3) prohibition against each of the authors for any new submissions, either individually, in combination with the authors of the plagiarizing manuscript, or in combination with new co-authors, to all of the Society's publications (journals, conferences, workshops). The prohibition shall continue for one year from notice of suspension.

Further, plagiarism and self-plagiarism may also be actionable by the IEEE under the rules of Member Conduct.

Submission Format.

Authors are encouraged to prepare manuscripts employing the on-line style files developed by IEEE. All manuscripts accepted for publication will require the authors to make final submission employing these style files. The style files are available on the web at http://www.ieee.org/publications_standards/publications/authors/authors_journals.html#sect2 under "Template for all TRANSACTIONS." (LaTeX and MS Word).

Authors using LaTeX: the two PDF versions of the manuscript needed for submission can both be produced by the IEEEtran.cls style file. A double-spaced document is generated by including \documentclass[11pt,draftcls,onecolumn]{IEEEtran} as the first line of the manuscript source file, and a single-spaced double-column document for estimating the publication page charges via \documentclass[10pt,twocolumn,twoside]{IEEEtran} for a regular submission, or \documentclass[9pt,twocolumn,twoside]{IEEEtran} for a Correspondence item.

- *Title page and abstract:* The first page of the manuscript shall contain the title, names and contact information for all authors (full mailing address, institutional affiliations, phone, fax, and e-mail), the abstract, and the EDICS. An asterisk * should be placed next to the name of the Corresponding Author who will serve as the main point of contact for the manuscript during the review and publication processes.

  An abstract must be a well-written stand-alone paragraph 150-250 words long, with no displayed equations, footnotes, references or tabular material. The abstract should indicate the scope of the paper and summarize the author's conclusions, making it a useful tool for information retrieval. Visit http://www.signalprocessingsociety.org/publications/periodicals/tsp/tsp-author-info/ for specifications and description.
- *EDICS:* All submissions must be classified by the author with an EDICS (Editors' Information Classification Scheme) selected from the list of EDICS published online at http://www.signalprocessingsociety.org/publications/periodicals/tsp/TSP-EDICS/
- NOTE: EDICS are necessary to begin the peer review process. Upon submission of a new manuscript, please choose the EDICS categories that best suit your manuscript. Failure to do so will likely result in a delay of the peer review process.
- The EDICS category should appear on the first page—i.e., the title and abstract page—of the manuscript.
- *Illustrations and tables:* Each figure and table should have a caption that is intelligible without requiring reference to the text. Illustrations/tables may be worked into the text of a newly-submitted manuscript, or placed at the end of the manuscript. (However, for the final submission, illustrations/tables must be submitted separately and not interwoven with the text.)

  Illustrations in color may be used but, unless the final publishing will be in color, the author is responsible that the corresponding grayscale figure is understandable.

  In preparing your illustrations, note that in the printing process, most illustrations are reduced to single-column width to conserve space. This may result in as much as a 4:1 reduction from the original. Therefore, make sure that all words are in a type size that will reduce to a minimum of 9 points or 3/16 inch high in the printed version. Only the major grid lines on graphs should be indicated.
- *Abbreviations:* This TRANSACTIONS follows the practices of the IEEE on units and abbreviations, as outlined in the Institute's published standards. See http://www.ieee.org/portal/cms_docs_iportals/iportals/publications/authors/transjnl/auinfo07.pdf for details.
- *Mathematics:* All mathematical expressions must be legible. Do not give derivations that are easily found in the literature; merely cite the reference.

Final Manuscript Submission Guidelines.

Upon formal acceptance of a manuscript for publication, instructions for providing the final materials required for publication will be sent to the Corresponding Author. Finalized manuscripts should be prepared in LaTeX or MS Word, and are required to use the style files established by IEEE, available at http://www.ieee.org/publications_standards/publications/authors/authors_journals.html#sect2.

Instructions for preparing files for electronic submission are as follows:

- For regular papers, the final manuscript may not exceed 16 double-column pages (10 point font), including title; names of authors and their complete contact information; abstract; text; all images, figures and tables, appendices and proofs; and all references. Without expressed approval from the Editor-in-Chief, papers that exceed 16 pages in length will not publish.
- Files must be self-contained; that is, there can be no pointers to your system setup.
- Include a header to identify the name of the TRANSACTIONS, the name of the author, and the software used to format the manuscript.
- Do not import graphics files into the text file of your finalized manuscript (although this is acceptable for your initial submission). If submitting on disk, use a separate disk for graphics files.
- Do not create special macros.
- Do not send PostScript files of the text.
- File names should be lower case.
- Graphics files should be separate from the text, and not contain the caption text, but include callouts like "(a)," "(b)."
- Graphics file names should be lower case and named fig1.eps, fig2.tif, etc.
- Supported graphics types are EPS, PS, TIFF, or graphics created using Word, Powerpoint, Excel or PDF. Not acceptable is GIF, JPEG, WMF, PNG, BMP or any other format (JPEG is accepted for author photographs only). The provided resolution needs to be at least 600 dpi (400 dpi for color).
- Please indicate explicitly if certain illustrations should be printed in color; note that this will be at the expense of the author. Without other indications, color graphics will appear in color in the online version, but will be converted to grayscale in the print version.

IEEE supports the publication of author names in the native language alongside the English versions of the names in the author list of an article. For more information, please visit the IEEE Author Digital Tool Box at the following URL: http://www.ieee.org/publications_standards/publications/authors/auth_names_native_lang.pdf

Additional instructions for preparing, verifying the quality, and submitting graphics and multimedia files are available via http://www.ieee.org/publications_standards/publications/authors/authors_journals.html.

Open Access.

This publication is a hybrid journal, allowing either Traditional manuscript submission or Open Access (author-pays OA) manuscript submission. Upon submission, if you choose to have your manuscript be an Open Access article, you commit to pay the discounted $1,750 OA fee if your manuscript is accepted for publication in order to enable unrestricted public access. Any other application charges (such as over-length page charge and/or charge for the use of color in the print format) will be billed separately once the manuscript formatting is complete but prior to the publication. If you would like your manuscript to be a Traditional submission, your article will be available to qualified subscribers and purchasers via IEEE Xplore. No OA payment is required for Traditional submission.

Page Charges.

Voluntary Page Charges. Upon acceptance of a manuscript for publication, the author(s) or his/her/their company or institution will be asked to pay a charge of $110 per page to cover part of the cost of publication of the first ten pages that comprise the standard length (six pages, in the case of Technical Correspondences until their publication will be discontinued).

Mandatory Page Charges. The author(s) or his/her/their company or institution will be billed $220 per each page in excess of the first ten published pages for regular papers and six published pages for technical correspondence until their publication will be discontinued. These are mandatory page charges and the author(s) will be held responsible for them. They are not negotiable or voluntary. The author(s) signifies his willingness to pay these charges simply by submitting his/her/their manuscript to the TRANSACTIONS. The Publisher holds the right to withhold publication under any circumstance, as well as publication of the current or future submissions of authors who have outstanding mandatory page charge debt.

Color Charges. Color figures which appear in color only in the electronic (Xplore) version can be used free of charge. In this case, the figure will be printed in the hardcopy version in grayscale, and the author is responsible that the corresponding grayscale figure is intelligible. Color reproduction in print is expensive, and all charges for color are the responsibility of the author. The estimated costs are as follows. There will be a charge of $62.50 for each figure; this charge may be subject to change without notification. In addition, there are printing preparation charges which may be estimated as follows: color reproductions on four or fewer pages of the manuscript: a total of approximately $1045; color reproductions on five pages through eight pages: a total of approximately $2090; color reproductions on nine through 12 pages: a total of approximately $3135, and so on. Payment of fees on color reproduction is not negotiable or voluntary, and the author's agreement to publish the manuscript in the TRANSACTIONS is considered acceptance of this requirement.

2015 IEEE SIGNAL PROCESSING SOCIETY MEMBERSHIP APPLICATION

Mail to: IEEE OPERATIONS CENTER, ATTN: Louis Curcio, Member and Geographic Activities, 445 Hoes Lane, Piscataway, New Jersey 08854 USA
or Fax to (732) 981-0225 (credit card payments only.)
For info call (732) 981-0060 or 1 (800) 678-IEEE or E-mail: new.membership@ieee.org

IEEE

1. PERSONAL INFORMATION

NAME AS IT SHOULD APPEAR ON IEEE MAILINGS: SEND MAIL TO: ☐ **Home Address** OR ☐ **Business/School Address**
If not indicated, mail will be sent to home address. Note: Enter your name as you wish it to appear on membership card and all correspondence.
PLEASE PRINT Do not exceed 40 characters or spaces per line. Abbreviate as needed. Please circle your last/surname as a key identifier for the IEEE database.

TITLE | FIRST OR GIVEN NAME | MIDDLE NAME | SURNAME/LAST NAME

HOME ADDRESS

CITY | STATE/PROVINCE | POSTAL CODE | COUNTRY

2

Are you now or were you ever a member of IEEE? ☐ **Yes** ☐ **No**
If yes, please provide, if known:

MEMBERSHIP NUMBER_____

Grade_____ **Year Membership Expired:**_____

3. BUSINESS/PROFESSIONAL INFORMATION

Company Name

Department/Division

Title/Position | Years in Current Position

Years in the Profession Since Graduation | ☐ PE State/Province

Street Address

City | State/Province | Postal Code | Country

4. EDUCATION

A baccalaureate degree from an IEEE recognized educational program assures assignment of "Member" grade. For others, additional information and references may be necessary for grade assignment.

A. Baccalaureate Degree Received | Program/Course of Study

College/University | Campus

State/Province | Country | Mo./Yr. Degree Received

B. Highest Technical Degree Received | Program/Course of Study

College/University | Campus

State/Province | Country | Mo./Yr. Degree Received

5.

Full signature of applicant

6. DEMOGRAPHIC INFORMATION – ALL APPLICANTS -

Date Of Birth _____ ☐ Male ☐ Female
Day Month Year

7. CONTACT INFORMATION

Office Phone/Office Fax | Home Phone/Home Fax

Office E-Mail | Home E-Mail

8. 2015 IEEE MEMBER RATES

IEEE DUES Residence	16 Aug 14-28 Feb 15 Pay Full Year	1 Mar -15 Aug 15 Pay Half Year**
United States	$193.00 ☐	$96.50 ☐
Canada (incl. GST)	$171.25 ☐	$85.63 ☐
Canada (incl. HST for PEI)	$184.30 ☐	$92.15 ☐
Canada (incl. HST for Nova Scotia)	$185.75 ☐	$92.88 ☐
Canada (incl. HST for NB, NF and ON)	$182.85 ☐	$91.43 ☐
Canada (incl. GST and QST Quebec)	$185.71 ☐	$92.86 ☐
Africa, Europe, Middle East	$158.00 ☐	$79.00 ☐
Latin America	$149.00 ☐	$74.50 ☐
Asia, Pacific	$150.00 ☐	$75.00 ☐

Canadian Taxes (GST/HST): All supplies, which include dues, Society membership fees, online products and publications (except CD-ROM and DVD media), shipped to locations within Canada are subject to the GST of 5% the HST of 13%, 14% or 15%, depending on the Province to which the materials are shipped. GST and HST do not apply to Regional Assessments. (IEEE Canadian Business Number 12563 4188 RT0001)

Value Added Tax (VAT) in the European Union: In accordance with the European Union Council Directives 02/38/EC and 77/388/EEC amended by Council Regulation (EC)792/2002, IEEE is required to charge and collect VAT on electronic/digitized products sold to private consumers that reside in the European Union. The VAT rate applied is the EU member country standard rate where the consumer is resident. (IEEE's VAT registration number is EU826000081)

U.S. Sales Taxes: Please add applicable state and local sales and use tax on orders shipped to **Alabama, Arizona, California, Colorado, District of Columbia, Florida, Georgia, Illinois, Indiana, Kentucky, Massachusetts, Maryland, Michigan, Minnesota, Missouri, New Jersey, New Mexico, New York, North Carolina, Ohio, Oklahoma, West Virginia, Wisconsin.** Customers claiming a tax exemption must include an appropriate and properly completed tax-exemption certificate with their first order.

IEEE Signal Processing Society

2015 SPS MEMBER RATES

		16 Aug-28 Feb Pay Full Year	1 Mar-15 Aug Pay Half Year
Signal Processing Society Membership Fee*		$ 20.00 ☐	$ 10.00 ☐
Fee includes: **IEEE Signal Processing Magazine** (electronic and digital), **Inside Signal Proc. eNewsletter** (electronic) and **IEEE Signal Processing Society Content Gazette** (electronic).			
Add $15 to enhance SPS Membership and also receive:		$15.00 ☐	$ 7.50 ☐
IEEE Signal Processing Magazine (print) and **SPS Digital Library**: online access to Signal Processing Magazine, Signal Processing Letters, Journal of Selected Topics in Signal Processing, Trans. on Audio, Speech, and Language Processing, Trans. on Image Processing, Trans. on Information Forensics and Security and Trans. on Signal Processing.			
Publications available only with SPS membership:			
Signal Processing, IEEE Transactions on:	Print	$190.00 ☐	$ 95.00 ☐
Audio, Speech, and Lang. Proc., IEEE/ACM Trans. on:	Print	$145.00 ☐	$ 72.50 ☐
Image Processing, IEEE Transactions on:	Print	$188.00 ☐	$ 94.00 ☐
Information Forensics and Security, IEEE Trans. on:	Print	$163.00 ☐	$ 81.50 ☐
IEEE Journal of Selected Topics in Signal Processing:	Print	$160.00 ☐	$ 80.00 ☐
Affective Computing, IEEE Transactions on:	Electronic	$ 35.00 ☐	$ 17.50 ☐
Biomedical and Health Informatics, IEEE Journal of:	Print	$ 55.00 ☐	$ 27.50 ☐
	Electronic	$ 40.00 ☐	$ 20.00 ☐
	Print & Electronic	$ 65.00 ☐	$ 32.50 ☐
IEEE Cloud Computing	Electronic and Digital	$ 39.00 ☐	$ 19.50 ☐
***New!* IEEE Trans. on Cognitive Comm. &Networking**	Electronic	$ 26.00 ☐	$ 13.00 ☐
***New!* IEEE Trans. on Computational Imaging**	Electronic	$ 28.00 ☐	$ 14.00 ☐
***New!* IEEE Trans. on Big Data**	Electronic	$ 25.00 ☐	$ 12.50 ☐
***New!* IEEE Trans. on Molecular, Biological, & Multi-scale Communications**	Electronic	$ 24.00 ☐	$ 12.00 ☐
IEEE Internet of Things Journal	Electronic	$ 26.00 ☐	$ 13.00 ☐
IEEE Trans. on Cloud Computing	Electronic	$ 42.00 ☐	$ 21.00 ☐
IEEE Trans. on Computational Social Systems	Electronic	$ 30.00 ☐	$ 15.00 ☐
***New!* IEEE Trans. on Signal & Info Proc. Over Networks**	Electronic	$ 28.00 ☐	$ 14.00 ☐
IEEE Biometrics Compendium:	Online	$ 30.00 ☐	$ 15.00 ☐
Computing in Science & Engrg. Mag.:	Electronic and Digital	$ 39.00 ☐	$ 19.50 ☐
	Print	$ 149.00 ☐	$ 74.50 ☐
Medical Imaging, IEEE Transactions on:	Print	$ 74.00 ☐	$ 37.00 ☐
	Electronic	$ 53.00 ☐	$ 26.50 ☐
	Print & Electronic	$ 89.00 ☐	$ 44.50 ☐
Mobile Computing, IEEE Transactions on:	ELE/Print Abstract/CD-ROM	$ 40.00 ☐	$ 20.00 ☐
Multimedia, IEEE Transactions on:	Electronic	$ 42.00 ☐	$ 21.00 ☐
IEEE MultiMedia Magazine:	Electronic and Digital	$ 39.00 ☐	$ 19.50 ☐
	Print	$149.00 ☐	$ 74.50 ☐
Network Science and Engrg., IEEE Trans. on:	Electronic	$ 33.00 ☐	$ 16.50 ☐
IEEE Reviews in Biomedical Engineering:	Print	$ 25.00 ☐	$ 12.50 ☐
	Electronic	$ 25.00 ☐	$ 12.50 ☐
	Print & Electronic	$ 40.00 ☐	$ 20.00 ☐
IEEE Security and Privacy Magazine:	Electronic and Digital	$ 39.00 ☐	$ 19.50 ☐
	Print	$149.00 ☐	$ 74.50 ☐
IEEE Sensors Journal:	Print	$150.00 ☐	$ 75.00 ☐
	Electronic	$ 50.00 ☐	$ 25.00 ☐
Smart Grid, IEEE Transactions on:	Print	$100.00 ☐	$ 50.00 ☐
	Electronic	$ 40.00 ☐	$ 20.00 ☐
	Print & Electronic	$120.00 ☐	$ 60.00 ☐
Wireless Communications, IEEE Transactions on:	Print	$120.00 ☐	$ 60.00 ☐
	Electronic	$ 48.00 ☐	$ 24.00 ☐
	Print & Electronic	$120.00 ☐	$ 60.00 ☐
IEEE Wireless Communications Letters:	Print	$ 80.00 ☐	$ 40.00 ☐
	Electronic	$ 18.00 ☐	$ 9.00 ☐
	Print & Electronic	$ 95.00 ☐	$ 47.50 ☐
***New!* IEEE Life Sciences Letters** (Open Access Pub)	Electronic	No Fee	

*IEEE membership required or requested
Affiliate application to join SP Society only.

Amount Paid $_____

9.

IEEE Membership Affiliate Fee (See pricing in Section 8) $_____

Signal Processing Society Fees $_____

Canadian residents pay 5% GST or 13% HST
Reg. No. 125634188 on Society payment(s) & pubs only Tax $_____

AMOUNT PAID WITH APPLICATION TOTAL $_____

Prices subject to change without notice.

☐ **Check or money order enclosed Payable to IEEE on a U.S. Bank**
☐ **American Express** ☐ **VISA** ☐ **MasterCard**
☐ **Diners Club**

Exp. Date/ Mo./Yr.

Cardholder Zip Code Billing Statement Address/USA Only

Full signature of applicant using credit card Date

10. WERE YOU REFERRED?

☐ Yes ☐ No If yes, please provide the follow information:
Member Recruiter Name:_____
IEEE Recruiter's Member Number (Required):_____

2015 IEEE SIGNAL PROCESSING SOCIETY STUDENT MEMBERSHIP APPLICATION

(Current and reinstating IEEE members joining SPS complete areas 1, 2, 8, 9.)

Mail to: IEEE OPERATIONS CENTER, ATTN: Louis Curcio, Member and Geographic Activities, 445 Hoes Lane, Piscataway, New Jersey 08854 USA
or Fax to (732) 981-0225 (credit card payments only.)
For info call (732) 981-0060 or 1 (800) 678-IEEE or E-mail: new.membership@ieee.org

IEEE

1. PERSONAL INFORMATION

NAME AS IT SHOULD APPEAR ON IEEE MAILINGS: SEND MAIL TO: ☐ **Home Address** OR ☐ **Business/School Address**
If not indicated, mail will be sent to home address. Note: Enter your name as you wish it to appear on membership card and all correspondence.
PLEASE PRINT Do not exceed 40 characters or spaces per line. Abbreviate as needed. Please circle your last/surname as a key identifier for the IEEE database.

TITLE | FIRST OR GIVEN NAME | MIDDLE NAME | SURNAME/LAST NAME

HOME ADDRESS

CITY | STATE/PROVINCE | POSTAL CODE | COUNTRY

2.

Are you now or were you ever a member of IEEE? ☐ **Yes** ☐ **No**
If yes, please provide, if known:

MEMBERSHIP NUMBER_____

Grade_____**Year Membership Expired:**_____

3. BUSINESS/PROFESSIONAL INFORMATION

Company Name

Department/Division

Title/Position | Years in Current Position

Years in the Profession Since Graduation | ☐ PE State/Province

Street Address

City | **State/Province** | **Postal Code** | **Country**

4. EDUCATION

A baccalaureate degree from an IEEE recognized educational program assures assignment of "Member" grade. For others, additional information and references may be necessary for grade assignment.

A.
Baccalaureate Degree Received | Program/Course of Study

College/University | Campus

State/Province | Country | Mo./Yr. Degree Received

B.
Highest Technical Degree Received | Program/Course of Study

College/University | Campus

State/Province | Country | Mo./Yr. Degree Received

5.

Full signature of applicant

6. DEMOGRAPHIC INFORMATION – ALL APPLICANTS -

Date Of Birth _____
Day Month Year

☐ Male ☐ Female

7. CONTACT INFORMATION

Office Phone/Office Fax | Home Phone/Home Fax

Office E-Mail | Home E-Mail

8. 2015 IEEE STUDENT MEMBER RATES

IEEE DUES Residence	16 Aug 14-28 Feb 15 Pay Full Year	1 Mar -15 Aug 15 Pay Half Year**
United States	$32.00 ☐	$16.00 ☐
Canada (incl. GST)	$33.60 ☐	$16.80 ☐
Canada (incl. HST for NB, NF, and ON)	$36.16 ☐	$18.08 ☐
Canada (incl. HST for Nova Scotia)	$36.80 ☐	$18.40 ☐
Canada (incl. HST for PEI)	$36.48 ☐	$18.24 ☐
Canada (incl. GST and QST Quebec)	$36.79 ☐	$18.40 ☐
Africa, Europe, Middle East, Latin America, Asia, Pacific	$27.00 ☐	$13.50 ☐

Canadian Taxes (GST/HST): All supplies, which include dues, Society membership fees, online products and publications (except CD-ROM and DVD media), shipped to locations within Canada are subject to the GST of 5% or the HST of 13%,14% or 15%, depending on the Province to which the materials are shipped. GST and HST do not apply to Regional Assessments. (IEEE Canadian Business Number 12563 4188 RT0001)

Value Added Tax (VAT) in the European Union: In accordance with the European Union Council Directives 2002/38/EC and 77/388/EEC amended by Council Regulation (EC)792/2002, IEEE is required to charge and collect VAT on electronic/digitized products sold to private consumers that reside in the European Union. The VAT rate applied is the EU member country standard rate where the consumer is resident. (IEEE's VAT registration number is EU826000081)

U.S. Sales Taxes: Please add applicable state and local sales and use tax on orders shipped to **Alabama, Arizona, California, Colorado, District of Columbia, Florida, Georgia, Illinois, Indiana, Kentucky, Massachusetts, Maryland, Michigan, Minnesota, Missouri, New Jersey, New Mexico, New York, North Carolina, Ohio, Oklahoma, West Virginia, Wisconsin.** Customers claiming a tax exemption must include an appropriate and properly completed tax-exemption certificate with their first order.

IEEE Signal Processing Society

2015 SPS STUDENT MEMBER RATES

		16 Aug-28 Feb Pay Full Year	1 Mar-15 Aug Pay Half Year
Signal Processing Society Membership Fee*		$10.00 ☐	$ 5.00 ☐
Fee includes: **IEEE Signal Processing Magazine** (electronic and digital), **Inside Signal Processing eNewsletter** (electronic) and **IEEE Signal Processing Society Content Gazette** (electronic).			
Add $8 to enhance SPS Membership and also receive: **IEEE Signal Processing Society Magazine** (print) and **SPS Digital Library**: online access to Signal Processing Magazine, Signal Processing Letters, Journal of Selected Topics in Signal Processing, Trans. on Audio, Speech, and Language Processing, Trans. on Image Processing, Trans. on Information Forensics and Security and Trans. on Signal Processing.		$ 8.00 ☐	$ 4.00 ☐
Publications available only with SPS membership:			
Signal Processing, IEEE Transactions on:	Print	$ 95.00 ☐	$ 47.50 ☐
Audio, Speech, and Lang. Proc., IEEE/ACM Trans. on:	Print	$ 73.00 ☐	$ 36.50 ☐
Image Processing, IEEE Transactions on:	Print	$ 94.00 ☐	$ 47.00 ☐
Information Forensics and Security, IEEE Trans. on:	Print	$ 82.00 ☐	$ 41.00 ☐
IEEE Journal of Selected Topics in Signal Processing:	Print	$ 80.00 ☐	$ 40.00 ☐
Affective Computing, IEEE Transactions on:	Electronic	$ 18.00 ☐	$ 9.00 ☐
Biomedical and Health Informatics, IEEE Journal of:	Print	$ 28.00 ☐	$ 14.00 ☐
	Electronic	$ 20.00 ☐	$ 10.00 ☐
	Print & Electronic	$ 65.00 ☐	$ 32.50 ☐
IEEE Cloud Computing	Electronic and Digital	$ 20.00 ☐	$ 10.00 ☐
***New!* IEEE Trans. on Cognitive Comm. &Networking**	Electronic	$ 13.00 ☐	$ 6.50 ☐
***New!* IEEE Trans. on Computational Imaging**	Electronic	$ 14.00 ☐	$ 7.00 ☐
***New!* IEEE Trans. on Big Data**	Electronic	$ 13.00 ☐	$ 6.50 ☐
***New!* IEEE Trans. on Molecular, Biological, & Multi-Scale Communications**	Electronic	$ 12.00 ☐	$ 6.00 ☐
IEEE Internet of Things Journal	Electronic	$ 13.00 ☐	$ 6.50 ☐
IEEE Trans. on Cloud Computing	Electronic	$ 21.00 ☐	$ 10.50 ☐
IEEE Trans. on Computational Social Systems	Electronic	$ 15.00 ☐	$ 7.50 ☐
***New!* IEEE Trans. on Signal & Info Proc. Over Networks**	Electronic	$ 14.00 ☐	$ 7.00 ☐
IEEE Biometrics Compendium:	Online	$ 15.00 ☐	$ 7.50 ☐
Computing in Science & Engrg. Mag.:	Electronic and Digital	$ 20.00 ☐	$ 10.00 ☐
	Print	$ 75.00 ☐	$ 37.50 ☐
Medical Imaging, IEEE Transactions on:	Print	$ 37.00 ☐	$ 18.50 ☐
	Electronic	$ 27.00 ☐	$ 13.50 ☐
	Print & Electronic	$ 45.00 ☐	$ 22.50 ☐
Mobile Computing, IEEE Transactions on:	ELE/Print Abstract/CD-ROM	$ 20.00 ☐	$ 10.00 ☐
Multimedia, IEEE Transactions on:	Electronic	$ 21.00 ☐	$ 10.50 ☐
IEEE MultiMedia Magazine:	Electronic and Digital	$ 20.00 ☐	$ 10.00 ☐
	Print	$ 75.00 ☐	$ 37.50 ☐
Network Science and Engrg., IEEE Trans. on:	Electronic	$ 17.00 ☐	$ 8.50 ☐
IEEE Reviews in Biomedical Engineering:	Print	$ 13.00 ☐	$ 6.50 ☐
	Electronic	$ 13.00 ☐	$ 6.50 ☐
	Print & Electronic	$ 20.00 ☐	$ 10.00 ☐
IEEE Security and Privacy Magazine:	Electronic and Digital	$ 20.00 ☐	$ 10.00 ☐
	Print	$ 75.00 ☐	$ 37.50 ☐
IEEE Sensors Journal:	Print	$150.00 ☐	$ 75.00 ☐
	Electronic	$ 28.00 ☐	$ 14.00 ☐
Smart Grid, IEEE Transactions on:	Print	$ 50.00 ☐	$ 25.00 ☐
	Electronic	$ 20.00 ☐	$ 10.00 ☐
	Print & Electronic	$ 60.00 ☐	$ 30.00 ☐
Wireless Communications, IEEE Transactions on:	Print	$ 60.00 ☐	$ 30.00 ☐
	Electronic	$ 24.00 ☐	$ 12.00 ☐
	Print & Electronic	$ 60.00 ☐	$ 30.00 ☐
IEEE Wireless Communications Letters:	Print	$ 40.00 ☐	$ 20.00 ☐
	Electronic	$ 9.00 ☐	$ 4.50 ☐
	Print & Electronic	$ 48.00 ☐	$ 24.00 ☐
***New!* IEEE Life Sciences Letters** (Open Access Pub)	Electronic	No Fee	

*IEEE membership required or requested
Affiliate application to join SP Society only.

9.

Amount Paid $_____

IEEE Membership Fee (See pricing in Section 8) $_____

Signal Processing Society Fees $_____

Canadian residents pay 5% GST or 13% HST
Reg. No. 125634188 on Society payment(s) & pubs only Tax $_____

AMOUNT PAID WITH APPLICATION TOTAL $_____
Prices subject to change without notice.

☐ **Check or money order enclosed Payable to IEEE on a U.S. Bank**
☐ **American Express** ☐ **VISA** ☐ **MasterCard** ☐ **Diners Club**

Exp. Date/ Mo./Yr.

Cardholder Zip Code Billing Statement Address/USA Only

Full signature of applicant using credit card Date

10. WERE YOU REFERRED?

☐ Yes ☐ No If yes, please provide the follow information:
Member Recruiter Name:_____
IEEE Recruiter's Member Number (Required): _____