IEEE COMMUNICATIONS MAGAZINE

November 2017, Vol. 55, No. 11

- Green Communications and Computing Networks
- Human-Driven Edge Computing and Communication
- Network Services Chaining in the 5G Vision
- Communications Education and Training

IEEE ComSoc IEEE Communications Society

A Publication of the IEEE Communications Society

www.comsoc.org

NATIONAL INSTRUMENTS

INNOVATE FASTER

WITH FIELD-DEPLOYED 5G PROOF-OF-CONCEPT SYSTEMS

In the race to design next-generation wireless technologies, research teams must rely on platforms and tools that accelerate productivity. And whether you're working in the lab or deploying solutions for field trial test, NI software defined radio hardware and LabVIEW Communications software can help you innovate faster and build 5G proof-of-concept systems to demonstrate new technologies first.

Accelerate your innovation at ni.com/5g.

©2017 National Instruments. All rights reserved. LabVIEW, National Instruments, NI, and ni.com are trademarks of National Instruments. Other product and company names listed are trademarks or trade names of their respective companies.

Director of Magazines
Raouf Boutaba, University of Waterloo (Canada)
Editor-in-Chief
Osman S. Gebizlioglu, Huawei Tech. Co., Ltd. (USA)
Associate Editor-in-Chief
Tarek El-Bawab, Jackson State University (USA)
Senior Technical Editors
Nim Cheung, ASTRI (China)
Nelson Fonseca, State Univ. of Campinas (Brazil)
Steve Gorshe, PMC-Sierra, Inc (USA)
Sean Moore, Centripetal Networks (USA)
Peter T. S. Yum, The Chinese U. Hong Kong (China)
Technical Editors
Mohammed Atiquzzaman, Univ. of Oklahoma (USA)
Guillermo Atkin, Illinois Institute of Technology (USA)
Mischa Dohler, King's College London (UK)
Frank Effenberger, Huawei Technologies Co.,Ltd. (USA)
Tarek El-Bawab, Jackson State University (USA)
Xiaoming Fu, Univ. of Goettingen (Germany)
Stefano Galli, ASSIA, Inc. (USA)
Admela Jukan, Tech. Univ. Carolo-Wilhelmina zu Braunschweig (Germany)
Vimal Kumar Khanna, mCalibre Technologies (India)
Yoichi Maeda, Telecommun. Tech. Committee (Japan)
Nader F. Mir, San Jose State Univ. (USA)
Seshradi Mohan, University of Arkansas (USA)
Mohamed Moustafa, Egyptian Russian Univ. (Egypt)
Tom Oh, Rochester Institute of Tech. (USA)
Glenn Parsons, Ericsson Canada (Canada)
Joel Rodrigues, Univ. of Beira Interior (Portugal)
Jungwoo Ryoo, The Penn. State Univ.-Altoona (USA)
Antonio Sánchez Esguevillas, Telefonica (Spain)
Mostafa Hashem Sherif, AT&T (USA)
Tom Starr, AT&T (USA)
Ravi Subrahmanyan, InVisage (USA)
Danny Tsang, Hong Kong U. of Sci. & Tech. (China)
Hsiao-Chun Wu, Louisiana State University (USA)
Alexander M. Wyglinski, Worcester Poly. Institute (USA)
Jun Zheng, Nat'l. Mobile Commun. Research Lab (China)
Series Editors
Ad Hoc and Sensor Networks
Edoardo Biagioni, University of Hawaii, Manoa (USA)
Ciprian Dobre, Univ. Politehnica of Bucharest (Romania)
Silvia Giordano, University of App. Sci. (Switzerland)
Automotive Networking and Applications
Wai Chen, Telcordia Technologies, Inc (USA)
Luca Delgrossi, Mercedes-Benz R&D N.A. (USA)
Timo Kosch, BMW Group (Germany)
Tadao Saito, Toyota Information Technology Center (Japan)
Consumer Communicatons and Networking
Ali Begen, Ozyegin Univ. and Networked Media (Turkey)
Mario Kolberg, University of Stirling (UK)
Madjid Merabti, Liverpool John Moores U. (UK)
Design & Implementation
Vijay K. Gurbani, Bell Labs/Alcatel Lucent (USA)
Salvatore Loreto, Ericsson Research (Finland)
Ravi Subrahmanyan, Invisage (USA)
Green Communicatons and Computing Networks
Song Guo, The Hong Kong Polytechnic Univ. (China)
RangaRao V. Prasad, Delft Univ. of Tech. (The Netherlands)
John Thompson, Univ. of Edinburgh (UK)
Jinsong Wu, Alcatel-Lucent (China)
Honggang Zhang, Zhejiang Univ. (China)
Integrated Circuits for Communications
Charles Chien, CreoNex Systems (USA)
Zhiwei Xu, HRL Laboratories (USA)
Network and Service Management
George Pavlou, U. College London (UK)
Juergen Schoenwaelder, Jacobs University (Germany)
Networking Testing and Analytics
Irena Atov, Microsoft (USA)
Erica Johnson, University of New Hampshire (USA)
Ying-Dar Lin, National Chiao Tung University (Taiwan)
Optical Communications
Admela Jukan, Tech. Univ. Braunschweig, Germany (USA)
Xiang Liu, Huawai Technologies (USA)
Radio Communications
Thomas Alexander, Ixia Inc. (USA)
Amitabh Mishra, University of Delaware (USA)
Columns
Book Reviews
Piotr Cholda, AGH U. of Sci. & Tech. (Poland)
History of Communications
Steve Weinsten (USA)
Technology Leaders' Forum
Steve Weinstein (USA)
Publications Staff
Joseph Milizzo, Assistant Publisher
Susan Lange, Online Production Manager
Jennifer Porcello, Production Specialist
Catherine Kemelmacher, Associate Editor

IEEE COMMUNICATIONS MAGAZINE

NOVEMBER 2017, vol. 55, no. 11

www.comsoc.org/commag

2017 IEEE Communications Society Elected Officers
Harvey A. Freeman, *President*
Khaled B. Letaief, *President-Elect*
Luigi Fratta, *VP–Technical Activities*
Guoliang Xue, *VP–Conferences*
Stefano Bregni, *VP–Member Relations*
Nelson Fonseca, *VP–Publications*
Robert S. Fish, *VP-Industry and Standards Activities*

Members-at-Large

Class of 2017
Gerhard Fettweis, Araceli Garciá Gómez
Steve Gorshe, James Hong

Class of 2018
Leonard J. Cimini, Tom Hou
Robert Schober, Qian Zhang

Class of 2019
Lajos Hanzo, Wanjiun Liao
David Michelson, Ricardo Veiga

2017 IEEE Officers
Karen Bartleson, *President*
James A. Jeffries, *President-Elect*
William P. Walsh, *Secretary*
John W. Walz, *Treasurer*
Barry L. Shoop, *Past-President*
E. James Prendergast, *Executive Director*
Vijay K. Bhargava, *Director, Division III*

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

ANNUAL SUBSCRIPTION: $71: print, digital, and electronic. $33: digital and electronic. $1001: non-member print.

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

COPYRIGHT AND REPRINT PERMISSIONS: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limits of U.S. Copyright law for private use of patrons: those post-1977 articles that carry a code on the bottom of the first page provided the per copy fee indicated in the code is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For other copying, reprint, or republication permission, write to Director, Publishing Services, at IEEE Headquarters. All rights reserved. Copyright © 2017 by The Institute of Electrical and Electronics Engineers, Inc.

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

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

ADVERTISING: Advertising is accepted at the discretion of the publisher. Address correspondence to: Advertising Manager, *IEEE Communications Magazine*, IEEE, 445 Hoes Lane, Piscataway, NJ 08855-1331.

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

BPA WORLDWIDE

IEEE ComSoc™
IEEE Communications Society

Join our Community!

IEEE

Networking • Conference Discounts • Technical Publications • Volunteer

Member Benefits and Discounts

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

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

*2015 Journal Citation Reports (JCR)

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

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

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

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

Join today at www.comsoc.org

THE IEEE COMMUNICATIONS SOCIETY'S NEW WEBSITE

Harvey Freeman

In the July issue of IEEE Communications Magazine, Stan Moyer, ComSoc's Chief Marketing Officer, outlined several different marketing efforts that are taking place this year. These included social media audit and consolidation, brand identity toolkit, marketing collateral, marketing strategy, and website redesign. By far the most important and consequently the most expensive endeavor was the website redesign. Having agreed to move forward with these efforts, the lead for facilitating ComSoc's marketing activities fell to Daphne Bartlett and the ComSoc Marketing staff.

Daphne Bartlett has been ComSoc's Marketing Director since July 2016 and leads the development of marketing and communications strategies for the organization. Her previous experiences include marketing and brand management roles at The College Board, The American Society of Mechanical Engineers (ASME), Beauté Prestige International, a Division of Shiseido Cosmetics, The Beanstalk Group, Produce for Better Health Foundation, and Exelon Corporation.

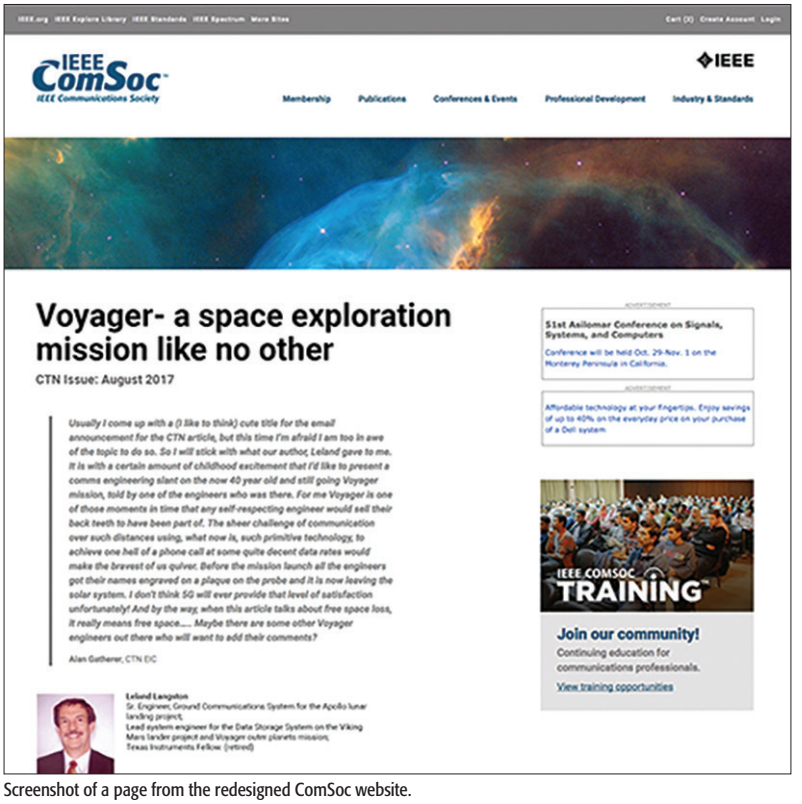
In terms of the website, ComSoc leadership and staff agreed that the current website, created in 2014, no longer serves our community. It does not help ComSoc meet its strategic objectives nor does it reflect our desired brand image. Additionally, the navigation and search is difficult for members and visitors to find what they want and need. Since a website redesign will have the largest impact on ComSoc, as a website is typically the main window for viewing, finding, and purchasing ComSoc products and services, it is important to get it right.

After an extensive RFP process, ComSoc engaged with an outside web design agency, Four Kitchens, to assist with the redesign and ensure the features of the site support our overall objectives:

- Advance communications technology for the betterment of humanity.
- Engage, connect, and support diverse communities.
- Clearly communicate ComSoc's mission and offerings.
- Make the website easy to use regardless of skill level, preference, or device.
- Increase membership.

We conducted member interviews and surveys, as well as a series of intensive content, competitive, and technical audits. This work took about four months to complete, and it helped us to identify key technology and user experience issues, success criteria, and priorities for the new website. This comprehensive exercise was the foundation for a project plan that served as the roadmap for the rest of the redesign.

Based on all of the information gleaned from the research and conversations with members, we will launch a new website with our member and non-member audience in mind. We will improve the overall design and structure of our content, and have a more responsive interface with faster load times, no matter which device is being used. The optimized navigation structure will make it easier to find the needed content. In addition, with seamless single-sign on, member-only ComSoc exclusive content will be accessible. There are additional small but impactful changes that will make for a much easier and user-friendly ComSoc web experience:



Screenshot of a page from the redesigned ComSoc website.

Better Page Structure and Visuals: A clear visual hierarchy will be established so that key actions and information can be found immediately — reducing page complexity and avoiding unnecessary text. Long pages can work for a website, but they need to be able to be scanned by modern web readers. The new ComSoc website will take advantage of visual elements that wrap content into easy-to-read portions, and in-page navigation will make it easier to stay oriented on long pages.

Comprehensive Search: Using SOLR with keyword search, the new website will deliver faster, more targeted search results.

Improved Messaging: We have a lot of con-

tent, but the ComSoc story and member benefits should be clearer to visitors. We have improved key pages like the Homepage, Membership, and About pages, which will make it clear who we are and what we have to offer.

We want the new website to better articulate our mission and core values by having our key messages be more visible, including adding high-priority content to the Homepage. The website will better promote the value of membership by identifying and highlighting which content is most useful for members.

Standardized Sub-navigation: Sub-navigation can help users stay oriented in a content-heavy site and manage longer pages. The new website will use sub-navigation consistently to help users understand where they are, and where else they can go while providing in-page navigation to help navigate longer pages.

Related, Dynamic Content: The new ComSoc website should provide pathways for users to continue to explore and engage with site content. We will surface popular or key content, provide "tags" that will allow related content to be displayed, and position related content at key points on the pages to provoke interest and avoid dead ends.

Consistent Branding and Tone of Voice: The current ComSoc website has a lot of variance in its overall look and feel. The new site will be visually consistent from page to page. We will develop tone of voice and an editorial guideline for content contributors to ensure that content is written with ComSoc audiences in mind.

We are excited to share the new ComSoc website, set to launch at the end of this year. We hope that you will find the new site modern and fresh, and easier to use and navigate. We will provide more information about the new website as we get closer to launch.

If you have any questions, comments, or suggestions, please contact Daphne Bartlett, the Marketing Director, at daphne.bartlett@comsoc.org, or Stan Moyer, the Chief Marketing Officer, at smoyer@comsoc.org.

CALL FOR PAPERS

IEEE COMMUNICATIONS MAGAZINE

EDUCATION & TRAINING: HUMANITARIAN ENGINEERING AND COMMUNITY ENGAGEMENT IN EDUCATION

BACKGROUND

Humanitarian engineering is research and design that aims to directly improve the well-being of poor, marginalized, or under-served communities. A fundamental tenet of humanitarian engineering is the need to engage the community in need and seek their active participation in the development and assessment of appropriate solutions rather than simply provide technology and techniques. In recent years, several pioneering institutions have begun to incorporate Humanitarian Engineering and Community Engagement principles and concepts into their teaching and research agenda. In so doing, they are helping their students become more globally aware; more conversant with the intersection between economics, sociology, politics and engineering fundamentals; and more engaged with communities in need. This FT is intended to hasten the incorporation of Humanitarian Engineering and Community Engagement principles and concepts into communications engineering curricula by providing educators, researchers and standards professionals with an opportunity to share their experience, best practices and case studies.

SUBMISSIONS

Authors from industry, government and academia are invited to submit papers for this FT (Feature Topic) of IEEE Communications Magazine on Humanitarian Engineering and Community Engagement in Education. The FT scope includes, but is not limited to, the following topics of interest:

- Case studies of the incorporation of Humanitarian Engineering and Community Engagement principles and concepts into communications engineering curricula.
- Best practices for incorporating Humanitarian Engineering and Community Engagement principles and concepts into communications engineering curricula.
- Case studies of the incorporation of Humanitarian Engineering and Community Engagement principles and concepts into professional training.
- Best practices for incorporating Humanitarian Engineering and Community Engagement principles and concepts into professional training.
- Development of tools for use in learning about Humanitarian Engineering and Community Engagement principles and concepts and their impact on design and development.
- Development of humanitarian communication technologies and their use in education.

Articles should be tutorial in nature, with the intended audience being all members of the communications technology community. They should be written in a tutorial style comprehensible to readers outside the specialty of the article. Mathematical equations should not be used (in justified cases up to three simple equations are allowed). Articles should not exceed 4500 words (from introduction through conclusions, excluding figures, tables and captions). Figures and tables should be limited to a combined total of six. The number of archival references is recommended not to exceed 15. In some rare cases, more mathematical equations, figures, and tables may be allowed, if well-justified. In general, however, mathematics should be avoided; instead, references to papers containing the relevant mathematics should be provided. Complete guidelines for preparation of the manuscripts are posted at http://www.comsoc.org/commag/paper-submission-guidelines. Please submit a PDF (preferred) or MS Word formatted paper via Manuscript Central (http://mc.manuscriptcentral.com/commag-ieee). Register or log in, and go to Author Center. Follow the instructions there. Select "May 2018/Humanitarian Engineering and Community Engagement in Education" as the Feature Topic category for your submission.

IMPORTANT DATES

- Manuscript Submission Deadline December 1, 2017
- Decision Notification: January 15, 2018
- Final Manuscript Due Date: February 15, 2018
- Publication Date: May 2018

GUEST EDITORS

David G Michelson (Lead)
University of British Columbia, Canada
davem@ece.ubc.ca

Kevin M Passino
Ohio State University, USA
passino.1@osu.edu

Joe Decuir
University of Washington – Bothell, USA
jdecuir@uw.edu

UPDATED ON THE COMMUNICATIONS SOCIETY'S WEB SITE

www.comsoc.org/conferences

2017

NOVEMBER

WINCOM 2017 — Int'l. Conference on Wireless Networks and Mobile Communications, 1–4 Nov.
Rabat, Morocco
http://www.wincom-conf.org/?p=welcome

IEEE NFV-SDN 2017 — IEEE Conference on Network Function Virtualization and Software Defined Networks, 6–8 Nov.
Berlin, Germany
http://nfvsdn2017.ieee-nfvsdn.org/

FRUCT21 2017 — Conference of Open Innovations Association (FRUCT) 2017, 6–10 Nov.
Helsinki, Finland
http://fruct.org/conference21

IEEE LATINCOM 2017 — 9th Latin-American Conference on Communications, 8–10 Nov.
Guatemala City, Guatemala
http://latincom2017.ieee-comsoc-latincom.org/

IEEE COMCAS 2017 — Int'l. Conference on Microwaves, Communications, Antennas and Electronic Systems, 13–15 Nov.
Tel Aviv, Israel
http://www.comcas.org/

IEEE ICSOS 2017 — IEEE Int'l. Conference on Space Optical Systems and Applications, 14–16 Nov.
Naha, Japan
http://icsos2017.nict.go.jp/

ISWSN 2017 — Int'l. Symposium on Wireless Systems and Networks, 19–22 Nov.
Lahore, Pakistan
http://sites.uol.edu.pk/iswsn17/

NOF 2017 — Int'l. Conference on the Network of the Future, 22–24 Nov.
London, United Kingdom
http://www.network-of-the-future.org/

CNSM 2017 — Int'l. Conference on Network and Service Management, 26–30 Nov.
Tokyo, Japan
http://www.cnsm-conf.org/2017/

IEEE VNC 2017 — IEEE Vehicular Networking Conference, 27–29 Nov.
Torino, Italy
http://www.ieee-vnc.org/

ITU-K 2017 — ITU Kaleidoscope: Challenges for a Data-Driven Society, 27–29 Nov.
Nanjing, China
http://www.itu.int/en/ITU-T/academia/kaleidoscope/2017/Pages/default.aspx

PEMWN 2017 — Int'l. Conference on Performance Evaluation and Modeling in Wired and Wireless Networks, 28–30 Nov.
Paris, France
https://sites.google.com/site/pemwn2017/

DECEMBER

IEEE GLOBECOM 2017 — IEEE Global Communications Conference, 4–8 Dec.
Singapore
http://globecom2017.ieee-globecom.org/

ICT-DM 2017 — Int'l. Conference on Information and Communication Technologies for Disaster Management, 11–13 Dec.
Münster, Germany
http://ict-dm2017.ercis.org/

ICSPCS 2017 — Int'l. Conference on Signal Processing and Communication Systems, 13–15 Dec.
Surfers Paradise, Australia
http://www.dspcs-witsp.com/icspcs_2017/index.html

IEEE ANTS 2017 — IEEE Int'l. Conference on Advanced Networks and Telecommunications Systems, 17–20 Dec.
Bhubaneswar, India
http://ants2017.ieee-comsoc-ants.org/

JANUARY

IEEE CCNC 2018 — IEEE Consumer Communications and Networking Conference, 12–15 Jan.
Las Vegas, NV
http://ccnc2018.ieee-ccnc.org/

FEBRUARY

IEEE WF-IOT 2018 — IEEE World Forum on Internet of Things, 5–8 Feb.
Singapore
http://wfiot2018.iot.ieee.org/

ICIN 2018 — Conference on Innovations in Clouds, Internet and Networks, 20–22 Feb.
Paris, France
http://www.icin-conference.org/

MARCH

ICNC 2018 — Int'l. Conference on Computing, Networking and Communications, 5–8 Mar.
Maui, HI
http://www.conf-icnc.org/2018/

OFC 2018 — The Optical Networking and Communication Conference & Exhibition, 11–15 Mar.
San Diego, CA
http://www.ofcconference.org/en-us/home/about/

APRIL

IEEE ISPLC 2018 — IEEE Int'l. Symposium on Power Line Communications and Its Applications, 8–11 Apr.
Manchester, United Kingdom
http://isplc2018.ieee-isplc.org/

IEEE WCNC 2018 — IEEE Wireless Communications and Networking Conference, 15–18 Apr.
Barcelona, Spain
http://wcnc2018.ieee-wcnc.org/

WTS 2018 — Wireless Telecommunications Symposium, 18–20 Apr.
Phoenix, AZ
http://www.cpp.edu/~wtsi/

–Communications Society portfolio events appear in bold colored print.
–Communications Society technically co-sponsored conferences appear in black italic print.

GCN
GLOBAL
COMMUNICATIONS
NEWSLETTER

November 2017
ISSN 2374-1082

CHAPTER ACTIVITIES

IEEE ComSoc Austin Chapter, USA: Winner of the 2017 IEEE ComSoc Chapter-of-the-Year Award and North-America Chapter Achievement Award

By Fawzi Behmann, Chair, IEEE ComSoc and Signal Processing Joint Chapter in Austin

Since our last CAA and CoY awards in 2015, The Austin Chapter strived to a greater level in serving our membership and the local community, and extending our reach within the region, to North America and, when opportunities opened, internationally.

Through focused vision, high energy and passion, the core team led by Fawzi Behmann exhibited high capacity in executing many activities and programs. Such efforts were not only recognized by the Society but also by the region where the Chapter Chair received an outstanding member Award for three consecutive years 2013-2014 and 2015.

Our core focus is fulfilling the mission of IEEE and tailoring it to the societies we are privileged to serve. The ComSoc Chapter had the opportunity as a joint chapter with the Signal Processing Chapter to bring intriguing topics and speakers to broaden the advancement in technology, research and use cases in different vertical markets.

In 2016, the ComSoc (Austin) chapter carried out core activities that included conducting 19 sessions (11 technical and eight non-technical/professional/administrative), capitalizing on the DLT/DSP program, engaging with student activities, reaching out to the community for membership development, participating and supporting technical events, and supporting one member to a senior level.

Among the interesting talks was one related to drones given by Robert Youens on the topic, "New FAA Drone Regulations."

Among the Distinguished Lecturers we hosted were: Dr. Anthony Chan, Huawei, Dallas, TX; Prof. Riku Jäntti, Head of the Department of Communications and Networking at Aalto University, Finland; and Prof. Todd Humphreys, UT, Austin, TX.

Lecture by Prof. Jäntti, held at the same time we celebrated IEEE Day on October 4. Left to right: Leslie Martinich, CTS Section Chair, Fawzi Behmann, ComSOc & SP Join Chapter chair and Prof. Riku Jäntti, Aalto University, Finland.

The core value offering was enhanced by executing a number of steps:

•Conducting feedback survey after ach session to capture attendees' evaluation for improvement, suggestions for future topics and presenters, as well as other ideas such as volunteering to assist in support activities.

•Conducting a leadership meeting for strategic planning and budgeting, setting the theme of the year and topics to be pursued, and assessing on a quarterly basis.

•Conducting a year-end assessment of achievements and lessons learned.

•Being creative and innovative in preparing personalized gifts (plaques with the speaker's name, title and date) for speakers as a token of appreciation of their time and efforts and sharing their experience.

•Having introductory slides that were presented at the beginning of each session promoting the IEEE brand, membership, most recent news and important chapter activities.

•Continuing to build the relationship with AT&T as the facility sponsor and ensuring that acknowledgments are made every time we hold the meeting at their facilities.

•Reaching out and collaborating with other Chapters, including the Computer Society, EMBS, Consulting Networks, and others.

•Preparing a process for re-election, sending information to members, and tabulating and announcing the results.

We have taken a proactive approach in managing Chapter activities on a quarterly basis. This approach helped define tangible steps in subsequent quarters to plan more events and attract higher attendance. As the subjects change, we see many new faces, helping us to ensure that the message is reaching a wider audience.

The Austin Communications Chapter sponsored, supported and participated in several local events and activities. Some examples include IEEE/NI 5G Day (August 3), NI Week (August 1-4), IEEE N3XT (September 17), and the Texas Wireless Summit (October 18).

Fawzi Behmann, Chapter Chair, participated at the R5 annual meeting and received the Outstanding Member Award at the banquet. He also supported the NA ComSoc region as a board member and vice chair in the planning, coordination and execution of the RCCC Regional Chapter Chair Conference held in December 2016 in Washington. Behmann also made a presentation at the RCCC as an exemplary Chapter with best practices.

The Chapter reached out and supported the student community by being part of the Advisory Board for the Westwood High School Ambassador Academy for Health and Science. The Chapter invitated students and teachers to the Chapter meeting. The Chapter also extended support to Texas State University by participating in Senior Design Day, when senior students demonstrated their projects. During lunch, Behmann was a keynote speaker for the event.

REGIONAL REPORT

Asia-Pacific IEEE Comsoc Summer School 2017 in Lahore, Pakistan

By Kashif Bashir, Chair IEEE ComSoc Lahore Section Pakistan

The IEEE ComSoc Lahore Section, in collaboration with the Al-Khawarizmi Institute of Computer Science (KICS), University of Engineering and Technology Lahore and Fast National University Lahore, organized the four-day "Asia Pacific IEEE ComSoc Summer School 2017."

Other collaborators are HEC Pakistan, IEEE UET and IEEE FAST. It was very inspirational for all who attended, particularly for the young engineers in Pakistan who very actively participated in the event. The summer school program was designed for new graduates, M.S. and Ph.D. students so they could advance their knowledge and research activities. The IEEE ComSoc Summer School has its historical background as well, as it first started two years ago in 2015 in Italy.

The opening ceremony of the IEEE ComSoc Summer School 2017 took place at FAST NU Lahore from 11–14 July 2017. The event began with a video message by Prof. Dr. Vincenzo Piuri, who is a full professor at the University of Milan, Italy. Dr. Piuri said it was his honor to address the participants of the IEEE ComSoc Summer School 2017. He mentioned that the event was a unique opportunity to learn new trends in communication technology since it provided a sharing experience for all. Dr. Vincenzo then said that events like this allowed all of us to improve and facilitate the citizens. He revealed that technology empowerment was benefiting not only Pakistan but also other countries. Dr. Piuri in the end congratulated the IEEE ComSoc Lahore Section, KICS-UET and FAST-NU for conducting this event in Pakistan.

Prof. Fazal Ahmad Khalid, Vice Chancellor of the University of Engineering and Technology Lahore, was the honorable chief guest. Prof. Khalid said that IEEE ComSoc was a good initiative and would surely bring lucrative results. He also mentioned that it would be beneficial for both industry and academia. He appreciated the technological advancements in the country and emphasized the efforts and contributions of our students who were active participants in smart world and smart technologies. Other guests included Prof. Dr. Waqar Mahmood, Director KICS UET and Sultan Qaboos IT Chair, Prof. Dr. Noor Muhammad Shaikh, GC University Lahore, Dr. Sayyed Aaun Abbas and Dr. Tariq Jadoon, LUMS, and the Chair of the IEEE ComSoc Lahore Section, Mr. Kashif Bashir.

Group photo of All Participants with COO Mr. Akbar Nasir Khan, Kashif Bashir Chair IEEE COMSOC Lahore Section, Dr. Amjad Hussain during Technical Tour to Punjab Safe City Authority Lahore.

COMSOC Speakers on Technical Visit to Punjab Safe City Authority. From left to right: Dr Mahtab Alam, Mr. Kashif Bashir, Mr. Akbar Nasir Khan, Mr. Faisal K Sheikh, Dr. Amjad Hussain, Hussain Mahdi, Mr. Mehran Memon, Dr. Ali Hammad Akbar.

Prof. Dr. Waqar Mahmood Director KICS UET receiving souvenir form Vice Chancellor Prof. Dr. Fazal Ahmad Khalid. From Left: Prof. Dr Waqar Mahmood, Prof. Dr. Fazal Ahmad Khalid, Dr. Amjad Hussain, Mr. Kashif Bashir.

This year the IEEE ComSoc Summer School 2017 attracted 65 research participants from nine universities in Pakistan. There were 14 speakers, including five from abroad and nine from national universities. International speakers were from China, Saudi Arabia, Estonia, Ireland and Malaysia; national speakers were from LUMS, COMSATS Lahore, UET Lahore and Mehran University Jamshoro.

A number of talks coupled with speeches by the guests were seen at this year's summer school. Featured speakers were Dr.

(Continued on Newsletter page 4)

Group photo at the Opening Ceremony Day 1: all guests with Participants of COMSOC Summer School 2017.

REGIONAL ACTIVITIES

Toward Development of IEEE ComSoc Activities in Kazakhstan

By Oleg Stukach, Tomsk Chapter Chair

The policy of energy independence is a strategic target of many states. Kazakhstan is the eighth largest country in the world in terms of territory, but not population. The only explored deposits of coal extracted from an open-cast mine surface is enough for 300 years. Nevertheless wind power engineering is widely developed in Kazakhstan. A possible reason is the high cost of constructing power lines to far away places. Also, the strong wind in the steppe blows constantly and mainly in one direction, in contrast with the weak breezes in Europe. However, it is necessary to monitor the remote power objects and to provide control and safety. It is effective only by wireless communications.

The second advantage of Kazakhstan is the organic agriculture. It is a very technological part of the economy where each plant growing is under control. Because of the distances that need to covered on the steppe, control is possible only by wireless communication.

Hence, energy and agriculture development inevitably requires the design of wireless control and communication facilities. But IEEE activity has not been present in Kazakhstan until now. More correctly, Kazakhstan is a white spot on the IEEE map. It seems that ComSoc should not miss the opportunities in this market before it quickly becomes overcrowded.

To attract attention to these problems, the first Euroasian Conference on Future Energy, together with the International Siberian Conference on Control and Communications, was held in June in the capital of Kazakhstan. Another purpose was the propagation of IEEE development through the formation of Chapters and Sections. We are convinced that in the large industrial and academic cities of Kazakhstan (Astana, Almaty, Karaganda, etc.) it is necessary to have Chapters for the basic directions of strategic

(Continued on Newsletter page 4)

Participants of the SIBCON 2017 Conference, Astana.

CONFERENCE REPORT

A Series of Cyber Security Workshops at INTERPOL World 2017

By Leo Hwa Chiang and Ewell Tan, IEEE APO Office, Singapore

Today, we are living in a connected society and the world has become borderless. Law enforcement officials are therefore facing a challenging environment, as technology has created new opportunities for criminals. The INTERPOL World conference, which was held at the Suntec Singapore Convention and Exhibition Centre 4–7 July 2017, brought together security professionals, experts and police forces from around the world, to gain knowledge, share experiences and address security challenges, to further safeguard the cities we are living in and manage cyber threats.

Leo Hwa Chiang, Director of IEEE Asia Business Development, curated three cyber security workshops with the topics "Surviving and Thriving through and after a Cybersecurity Incident," "Man of Mode: Fashion, Hacking Tool Reuse and Copycat Crime in Cyberspace," and "Forensic Computing" on the 5–6 July at the INTERPOL World Congress, on behalf of the IEEE Communications Society. More than 60 participants attended the workshop. Participants at these workshops gained knowledge on improving the risk management process to rapidly recover from cyber-attacks and minimize the impact to the organization. They also discovered the increasingly sophisticated criminal ecosystem and the Darknet. Through the workshops, participants should be able to identify internal and external threats, and enforce security resilience.

From left to right: Regional Sales Manager, Ira Tan; Client Service Manager, Alex Liu; Project Manager, Ewell Tan and the Director of IEEE Asia Business Development, Leo Hwa Chiang.

Leo Hwa Chiang (right) received a certificate of appreciation from Mr. Noboru Nakatani, the Executive Director of INTERPOL Global Complex for Innovation.

In addition to these informative workshops, IEEE ComSoc had a booth to promote IEEE to the security professionals, industry leaders and government agencies. Through the combined efforts of InfoHost (the IEEE dealer), the IEEE Marketing and Sales Department team (Ira and Alex), and the IEEE Singapore office staff (Leo and Ewell), visitors' enquiries were answered at the exhibition booth.

Last but not least, the Executive Director of the INTERPOL Global Complex for Innovation, Mr. Noboru Nakatani, presented a certificate of appreciation to the IEEE Communications Society, for organizing the three cyber security workshops successfully at this Congress.

SUMMER SCHOOL 2017/*Continued from page 2*

Mohamed-Slim Alouini, Dr. Momin Uppal, Dr. Zartash Uzmi, Dr. Ghalib Shah, Dr. M. Majid Butt, Dr. Faisal K. Shaikh, Dr. Ubaid ullah Fiaz, Dr. Muhammad Tahir, Dr. Sobia Baig, Dr. Hussain Mahdi, Dr. Waqar Baig, Mr. He Gang and Dr. Muhammad Mahtab Alam. The topics included, among others, Internet Censorship Research, USAID Fulbright and other scholarship opportunities abroad, Smart Grid Communication technology, FM Radio Broadcast Using RTL Software, Internet of Things, Polar Codes, Shared Spectrum Access, 5G Wireless Communication Networks, Wireless Acoustic Sensor Network. Paper presentations were also presented. A panel discussion on the topic of how to elevate research in Pakistan was also a major highlight of the event.

Two visits were also arranged. The first visit was to the best eye-candy locales in the city of Lahore. Second, the participants were taken on a technical tour of Punjab Safe City Headquarters, where they were briefed by Mr. Akbar Nasir Khan, Chief Operating Officer and Project Director PPIC3. He also showed how the city was monitored by hundreds of cameras and how all the emergency services are integrated into a single hotline.

The chief guest of the closing ceremony was Engr. Jawed Salim Qureshi, Chairman of the Pakistan Engineering Council. He said that the Pakistan Engineering Council was closely working with the government and industry for the betterment and future of the students who were appointed as interns. Mr. Jawed Salim Qureshi also said that Pakistani engineers had been doing great not only in Pakistan but also across the world to prove their mettle. He said that the formula of success was being a champion. Mr. Akbar Nasir Khan, Chief Operating Officer of the Punjab Safe City Authority, said that what he learned from his education was leadership and it was what took you to the heights of success. Mr. Edward Zhang, Vice President Huawei Technologies China, spoke about the CPEC and Huawei's ICT contribution in Pakistan. Prof. Dr. Fazal Ahmed Khalid, Vice Chancellor UET Lahore, expressed appreciation for the efforts of the Chair of the IEEE ComSoc Lahore Section (Kashif Bashir) and volunteers for organizing this mega event between KICS-UET and FAST NU successfully.

Other guests were Prof. Dr. Waqar Mahmood, Director of KICS UET Lahore; Prof. Dr. Suhail Aftab Qureshi, Dean of EE at UET Lahore; Prof. Dr. Amjad Hussain, Director of FAST NU Lahore; Mr. Kashif Bashir, Chair of the IEEE ComSoc Lahore Section, and Mr. Usman Munawar, Secretary of the IEEE Lahore Section. The guests were given souvenirs and shields, while the participants and the event organizers were given certificates.

GCN GLOBAL COMMUNICATIONS NEWSLETTER

STEFANO BREGNI
Editor-in-Chief
Politecnico di Milano – Dept. of Electronics and Information
Piazza Leonardo da Vinci 32, 20133 MILANO MI, Italy
Tel: +39-02-2399.3503 – Fax: +39-02-2399.3413
Email: bregni@elet.polimi.it, s.bregni@ieee.org

FABRIZIO GRANELLI
Associate Editor-in-Chief
University of Trento
Email: fabrizio.granelli@unitn.it

IEEE COMMUNICATIONS SOCIETY

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

REGIONAL CORRESPONDENTS WHO CONTRIBUTED TO THIS ISSUE

EWELL TAN, SINGAPORE (ewell.tan@ieee.org)

IEEE ComSoc IEEE Communications Society

www.comsoc.org/gcn
ISSN 2374-1082

ACTIVITIES IN KAZAKHSTAN/*Continued from page 3*

development: communication, power energy, control, and future agriculture.

The conference supports interdisciplinary discussions and interaction among scientists and engineers, establishing international cooperation through participation in the activities of the professional communities of IEEE.

The Conference was held at Saken Seifullin Kazakh Agrotechnical University in Astana. The University has a strong Power Energy faculty with a Radio Engineering, Electronics and Telecommunications Department. Financial support has been provided by the administration, namely, the Rector Akhylbek Kurishbayev, the Vice-Rector on strategic development Sergey Mogilnyy, and the dean of faculty Sultanbek Issenov. Employees of the Radio Engineering Department volunteered as organizers of the Conference, including Arman Mirmanov and Botagoz Khamzina. The Vice-President of the creation and operation of space systems, the Kazakhstan cosmonaut A. Aimbetov, and professor of the School of Engineering of the Nazarbaev University, IEEE Member Mehdi Bagheri, provided invited welcome lectures. Ordinary reports were presented within two days at 32 sessions in eight tracks, including Communications, Control, and Power Electronic Devices. The Proceedings of SIBCON-2017 are published in IEEE Xplore in the form of full papers.

National Instruments Russia R&D is the old partner of the SIBCON conference. Representatives of the company organized the following hands-on classes: "Basics of LabView," "MyRIO Platform for quick support of the engineering projects," and "New generation of the graph design of applications on LabView Software Technology Preview." In the special session "Computer Measurement Technologies" some questions on the introduction of innovative technologies for decision making for engineering problems in communications, automation of manufacturing, modeling, and processing of experiments were discussed. KATU is completely equipped with NI tools, and students fulfill class projects on communication systems.

The social program included a banquet, technical excursions and a sight seeing tour. The International Exhibition ASTANA EXPO-2017 devoted to future energy technologies was open in the capital of Kazakhstan during the Conference, and all participants could visit the exhibition. Astana is the new capital of Kazakhstan, and it is a historic but very modern city. The rich history, huge cultural heritage, and modern technologies harmoniously intertwine here. It is practically the second Singapore, and maybe the first. The capital is dazzlingly beautiful. Walking around the city, you can admire unusual modern installations, which give Astana its special charm: Bayterek, the symbol of Astana, the Green Water Boulevard "Nurzhol" recreational pedestrian zone with the avenue of singing fountains, the Independence Palace, the Palace of Peace and Concent, the Palace of Arts, the National Museum, the Astana Opera, etc.

ComSoc is actively present in Europe, but it is necessary to expand into new territories and new markets. We invite to Astana all who are interested in extending the development of ComSoc activities. Above all, we suggest locating a future ICC in Astana. We are sure you will not be disappointed.

IEEE Access®

The **journal** for rapid **open access** publishing

Become a published author in 4 to 6 weeks.

Get on the fast track to publication with the multidisciplinary open access journal worthy of the IEEE.

IEEE journals are trusted, respected, and rank among the most highly cited publications in the industry. IEEE *Access* is no exception with a typical **one-third** acceptance rate. Even though it provides authors faster publication time, every submitted article still undergoes extensive peer review to ensure originality, technical correctness, and interest among readers.

Published only online, IEEE *Access* is ideal for authors who want to quickly announce recent developments, methods, or new products to a global audience.

Publishing in IEEE *Access* allows you to:

- Submit multidisciplinary articles that do not fit neatly in traditional journals
- Reach millions of global users through the IEEE *Xplore*® digital library with free access to all
- Integrate multimedia with articles
- Connect with your readers through commenting
- Track usage and citation data for each published article
- Publish without a page limit for **only $1,750** per article

Learn more about this award-winning journal at:
www.ieee.org/ieee-access

IEEE
Advancing Technology for Humanity

14-PUB-196 11/15

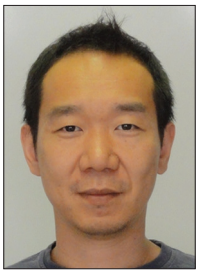
GREEN COMMUNICATIONS AND COMPUTING NETWORKS


Jinsong Wu


John Thompson


Honggang Zhang


RangaRao Venkatesha Prasad


Song Guo

Green communications and computing (GCC) has been an active and will be a long lasting important field for technological research and development, although, in different stages, GCC may have different focuses and characteristics. In the near future, there will be three different, correlated, and important relevant directions and focuses: green Internet of Things (IoT) [1], green fifth generation (5G) wireless systems, and big data meeting green challenges [2]. The IoT refers to internetwork physical sensors, devices, and entities to enable collecting and exchanging data and signals. There are a number of concepts relevant to green IoT in different contexts, such as green cyber-physical systems, green machine-to-machine communications, and green device-to-device communications.

GCC stands for a subset of global green efforts. As two leading global green efforts of the United Nations Environment Programme (UNEP), the 23rd Climate Change Conference (COP23) [3] and the third UN Environment Assembly (UNEA 3) [4] are being held in November and December 2017. COP23 will take place in Bonn, Germany, from 6 to 17 November this year [3]. Under the Presidency of the Republic of Fiji, delegates from 197 countries will negotiate the implementation of the Paris Agreement [5]. More than 20,000 delegates are expected to attend, and more than 830 organizations want to showcase their actions toward climate change [3]. The highest-level world decision-making body on the environment, UNEA 3, with the universal membership of all 193 UN Member States and the full involvement of UN organizations, specialized agencies, inter-governmental organizations, civil society, and the private sector, will be held in Nairobi, Kenya, from 4–6 December 2017 under the clear theme of pollution issues, which are expected to achieve a number of tangible commitments to terminate the pollution of our air, land, waterways, and oceans, and to safely manage our chemicals and waste [4].

All of the above efforts continue to inspire the IEEE Series on Green Communications and Computing Networks to support green efforts. The seventh, November 2017, issue of this Green Series includes relevant articles.

There are two articles in this issue related to big data meeting green challenges. The article "Toward Big Data in Green City" first reviews the latest work concerning big data and sensor-cloud, respectively, and then introduces three types of sensor-cloud for green city. The article "Big Data Analytics for Electric Vehicle Integration in Green Smart Cities" first makes a survey of the data analytics techniques for smart grid and EVs, and provides an overview of the data analytics landscape on the EV integration to green smart cities.

There is one article related to wireless power transfer. "Simultaneous Wireless Information and Power Transfer: Technologies, Applications, and Research Challenges" surveys the current architectures and enabling technologies for SWIPT, and illustrates their importance.

There are five articles related to advanced techniques in green wireless networks.

The article "Green Heterogeneous Cloud Radio Access Networks: Potential Techniques, Performance Trade-offs, and Challenges" first proposes some potential techniques to energy-efficiently operate H-CRANs, and then elaborates on initial ideas of modeling three fundamental trade-offs.

"Fully Exploiting Cloud Computing to Achieve a Green and Flexible C-RAN" reviews the recent advances of exploiting cloud computing to form a green and flexible C-RAN.

The article "Enhancing Energy Efficiency via Cooperative MIMO in Wireless Sensor Networks: State of the Art and Future Research Directions" surveys several cooperative multiple-input multiple-output (CMIMO) models for different scenarios and discusses the implementations relevant to green communications.

"Energy-Sustainable Traffic Steering for 5G Mobile Networks" proposes an energy-sustainable traffic steering framework to match energy distributions in both the spatial and temporal domains.

The article "A Software-Defined Green Framework for Hybrid EV-Charging Networks" discusses how to jointly use these two types of charging methods whose advantages are complementary to each other, and then propose a software-defined green framework for hybrid EV-charging networks.

ACKNOWLEDGMENTS

We would like to acknowledge the great support from Osman S. Gebizlioglu, the current Editor-in-Chief of *IEEE Communications Magazine*, Peggy Kang, the Managing Editor of IEEE Communications Society Magazines, Jennifer Porcello, Production Specialist, and Joseph Milizzo, Assistant Publisher, and

the other IEEE Communications Society publication staff. We also highlight the great support of this Green Series from the members of the IEEE Technical Committee on Green Communications and Computing (TCGCC) as well as the IEEE Environmental Engineering Initiative (EEI) under IEEE Technical Activities Board (TAB).

REFERENCES

[1] H. Chao *et al.*, "Power Saving for Machine to Machine Communications in Cellular Networks," Proc. 2011 IEEE GLOBECOM Wsps., Dec 2011
[2] J. Wu, et. al., "Big Data Meet Green Challenges: Big Data Toward Green Applications," *IEEE Systems J.*, vol. 10, no. 3, Sept. 2016.
[3] UN Climate Change Conference, Bonn, Germany, 6–17 November 2017; http://newsroom.unfccc.int/cop-23-bonn/, accessed Aug. 25, 2017.
[4] UN Environment Assembly, Nairobi, Kenya, 4-6 December 2017; http://www.unep.org/environmentassembly/un-environment-assembly, accessed Aug. 25, 2017.
[5] J. Wu *et. al.*, Green Communications and Computing Networks Series Editorial, *IEEE Commun. Mag.*, vol. 54, no. 5, May 2016.

BIOGRAPHIES

JINSONG WU [SM] (wujs@ieee.org) is the elected Vice-Chair of Technical Activities, IEEE Environmental Engineering Initiative, a pan-IEEE effort under IEEE Technical Activities Board (TAB). He is the founder and founding Chair of IEEE Technical Committee on Green Communications and Computing. He has been selected as the winner of the 2017 *IEEE Systems Journal* Best Paper Award. He was the leading editor and co-author of the comprehensive book *Green Communications: Theoretical Fundamentals, Algorithms, and Applications* (CRC Press, 2012).

JOHN THOMPSON [F] (john.thompson@ed.ac.uk) currently holds a personal chair in signal processing and communications, University of Edinburgh, United Kingdom. He was deputy academic coordinator for the Mobile Virtual Centre of Excellence Green Radio project, and now leads the U.K. SERAN project, which studies spectrum issues for 5G wireless, and leads the European Marie Curie Training Network ADVANTAGE, which trains 13 Ph.D. students in smart grid technology. He was also a Distinguished Lecturer on green topics for ComSoc in 2014–2015.

HONGGANG ZHANG [SM] (honggangzhang@zju.edu.cn) is a professor at Zhejiang University, China, and was International Chair Professor of Excellence for UEB and Supélec, France (2012–2014). He was Chair, IEEE Technical Committee on Cognitive Networks (2011–2012). He was Lead Guest Editor of *IEEE Communications Magazine* Feature Topics on Green Communications. He was General Co-Chair of 2010 IEEE GreenCom and Co-Chair of IEEE Online GreenComm 2015. He is the book co-editor/co-author of *Green Communications: Theoretical Fundamentals, Algorithms, and Applications* (CRC Press, 2012).

RANGARAO VENKATESHA PRASAD [SM] (R.R.VenkateshaPrasad@tudelft.nl) received his Ph.D. from IISc, Bangalore, India, when he designed a scalable VoIP conferencing platform. Part of his thesis led to a startup venture, Esqube Communication Solutions. In 2005, he joined TUDelft. He has worked on personal networks (PNs), IoT, CPS, and energy harvesting networks. His work at TUDelft has resulted in 180+ publications. He is a Senior Member of ACM.

SONG GUO [SM] (song.guo@polyu.edu.hk) is a full professor at the Department of Computing, Hong Kong Polytechnic University. He has published over 300 papers in referreed journals/conferences and received multiple IEEE/ACM best paper awards. He is an Editor of *IEEE Transactions on Green Communications and Networking* and the Secretary of the IEEE Technical Subcommittee on Big Data. He is a Senior Member of the ACM and an IEEE Communications Society Distinguished Lecturer.

Toward Big Data in Green City

Chunsheng Zhu, Huan Zhou, Victor C. M. Leung, Kun Wang, Yan Zhang, and Laurence T. Yang

Toward big data in green city, the authors review the latest work concerning big data and sensor-cloud, respectively. They introduce three types of sensor-cloud (i.e., PSC, ASC, and SSC) for green city. Specifically, about PSC, participatory sensing is incorporated into sensor-cloud for sensing big data. In terms of ASC, an agent is incorporated into sensor-cloud for transmitting big data. For SSC, a social network is incorporated into sensor-cloud for sharing big data.

ABSTRACT

Integrating sensors and cloud computing, sensor-cloud is a very powerful system for users to obtain big data in green city. In this article, toward big data in green city, we first review the latest work concerning big data and sensor-cloud, respectively. Further, we introduce three types of sensor-cloud (i.e., PSC, ASC, and SSC) for green city. Specifically, about PSC, participatory sensing is incorporated into sensor-cloud for sensing big data. In terms of ASC, an agent is incorporated into sensor-cloud for transmitting big data. For SSC, a social network is incorporated into sensor-cloud for sharing big data. Finally, the open research issues with respect to big data and sensor-cloud are discussed, respectively. We hope this article can serve as enlightening guidance for future research regarding big data in green city.

INTRODUCTION

Recently, with the rapid advances in green sensing, computing, and communication technologies, our city is evolving into a green city with numerous green applications (e.g., green homes, green grids, green factories). In particular, with the fast growth of urban population and fast urban development, a huge amount of data from various sources is being generated by green city. This explosive data growth in green city is making big data [1] a hot topic for the fulfillment of green city.

Incorporating sensors and cloud computing, sensor-cloud [2] is a very powerful system for users to achieve big data in green city. Particularly, as shown in Fig. 1, on one hand, various big data in our city is sensed and gathered by a variety of ubiquitous sensors (e.g., temperature sensors, humidity sensors, pressure sensors, light sensors, video sensors) and further transmitted to the cloud via a sink or sinks. On the other hand, the received big data is stored and processed by the powerful data centers in the cloud and further delivered to users on demand. Thus, with sensor-cloud, users are able to have access to the desirable big data anytime and anywhere if there is any network connection.

In this article, toward big data in green city, recent work about big data and sensor-cloud is reviewed first. Further, three types of sensor-cloud — participatory sensor-cloud (PSC), agent-based sensor-cloud (ASC), and social-sensor-cloud (SSC) — are introduced for green city. Particularly, with respect to PSC, participatory sensing [3] is incorporated into sensor-cloud for sensing big data. In ASC, an agent [4] is incorporated into sensor-cloud for transmitting big data. In terms of SSC, a social network [5] is incorporated into sensor-cloud for sharing big data. Eventually, the open research issues concerning big data and sensor-cloud are investigated. This article aims to serve as inspiring guidance for future research with respect to big data in green city.

The rest of this article is organized as follows. The following two sections review the latest work about big data and sensor-cloud, respectively. Then we present PSC, ASC, and SSC. Following that, we investigate the open research issues on big data and sensor-cloud, respectively. The final section concludes this article.

LATEST WORK ON BIG DATA

Focusing on the network infrastructure of big data, the unique challenges in establishing a network infrastructure for geologically distributed and rapidly generated big data are discussed in [6]. Specifically, it first investigates the challenges in terms of every segment in the network highway, such as the access networks connecting data sources, the Internet backbone bridging them to remote data centers, and the dedicated networks within a data center and among data centers. Moreover, it presents two case studies empowered by networking, and highlights the promising and interesting future directions based on those case studies.

Discussing the framework of big data, a novel framework to deliver big data over content-centric mobile social networks with a satisfactory quality of experience is proposed in [7]. The characteristics and challenges of mobile big data are presented first. Then a content-centric network architecture to deliver mobile big data in mobile social networks is proposed. With that, how to choose the relay node for forwarding data packets is introduced by defining priorities of data packets. Lastly, simulation results are demonstrated on the performance of the proposed framework.

Investigating the cost minimization of big data, the authors in [8] try to investigate the cost minimization problem by a joint optimization of three factors (i.e., task assignment, data placement, and data movement) for big data services in geo-distributed data centers. Specifically, to describe the task completion time considering both data transmission and computation, a 2D Markov chain is proposed, and the average task completion time

Chunsheng Zhu and Victor C. M. Leung are with the University of British Columbia; Huan Zhou is with China Three Gorges University and also with the University of British Columbia; Kun Wang is with Nanjing University of Posts and Telecommunications; Yan Zhang is with the University of Oslo; Laurence T. Yang is with St. Francis Xavier University.

Digital Object Identifier:
10.1109/MCOM.2017.1700142

0163-6804/17/$25.00 © 2017 IEEE

is derived in closed form. Then the problem is modeled as mixed-integer nonlinear programming, and an efficient solution is proposed to linearize it. Finally, extensive simulation results validate the efficiency of the proposed proposal.

Studying the role of big data, the authors in [9] first show the state-of-the-art communication technologies and smart-based applications in the context of smart cities. Then visions of big data analytics for supporting smart cities are presented by discussing the ways that big data fundamentally alters urban populations at different levels. After that, a future business model of big data for smart cities is introduced. Eventually, the business and technological challenges of the proposed business model are observed.

Considering the application of big data, a real-time big data analytical architecture for remote sensing satellite application is proposed in [10]. Particularly, taking into account of both real time and offline data processing, the introduced architecture consists of three main units: a remote sensing big data acquisition unit; a data processing unit; and a data analysis decision unit. Further, a detailed analysis regarding the proposed analytical architecture is performed using Hadoop, on the remotely sensed big data for land and sea area using an Earth observatory system.

Latest Work about Sensor-Cloud

Concerning the pricing of sensor-cloud, five sensor-cloud pricing models are proposed in [11]. Specifically, the following elements are considered by each sensor-cloud pricing model to charge the user: 1) user's lease period; 2) sensor-cloud's working time; 3) the sensor-cloud resources used by the user; 4) the volume of data achieved by the user; 5) the sensor-cloud path that transmits data from the sensor network to the user. With that, the characteristics and case studies of the proposed sensor-cloud pricing models are investigated, followed by a review regarding a user behavior study.

Regarding the security of sensor-cloud, a risk assessment framework utilizing attack graphs for sensor networks in a sensor-cloud is proposed in [12]. Particularly, the impacts of attacks on sensor networks are reviewed by the proposed framework first. Then reasonable timeframes that predict the degradation of the sensor network's security performance (e.g., confidentiality, integrity, and availability) are estimated by the proposed framework. With that, the assessment results of the proposed framework are validated in various simulated attack scenarios.

About the quality of service of sensor-cloud, [13] focuses on scheduling a particular cloud data center that congregates data from various virtual sensors and further delivers the data to users. Specifically, the scheduling of the cloud data center is conducted considering several network constraints (e.g., data migration cost, data delivery cost, and service delay). Then an optimal decision rule for selecting a particular cloud data center is determined to satisfy the quality of service. Experimental results about the proposed scheduling mechanism in real-time sensor-cloud scenarios are shown.

With respect to the modeling of sensor-cloud, [14] pays attention to theoretical characterization and analysis of sensor-cloud. Particularly, a mathematical formulation of sensor-cloud is presented. Then a detailed analysis is performed, taking into account the performance metrics (e.g., energy consumption, fault tolerance, and lifetime of a sensor). Further, the evaluation of the cost effectiveness of sensor-cloud is conducted. Analytical results with the proposed formulation are also demonstrated, in terms of a sensor's average lifetime, a sensor's average energy consumption, and a user's average expenditure.

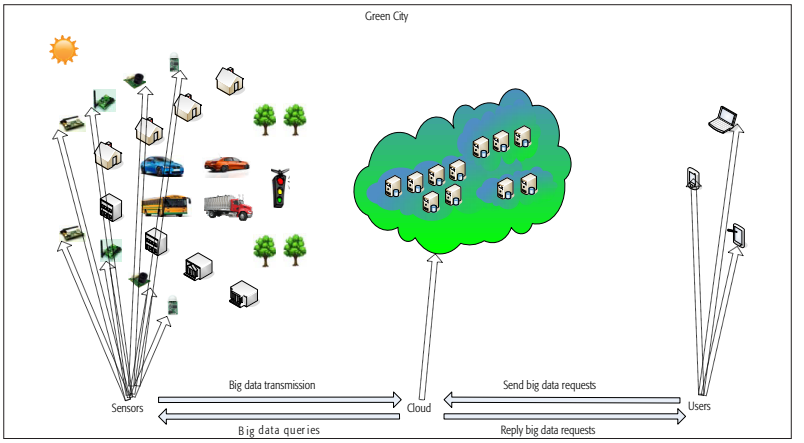


Figure 1. An example of Sensor-Cloud.

For the application of sensor-cloud, [15] discusses target tracking with sensor-cloud. Specifically, how to schedule sensors to track targets in the presence of overlapping coverage is studied. A dynamic mapping algorithm is proposed, based on the Theory of Social Choice for achieving an unbiased mapping of sensors to targets. The uniformity of the dynamic mapping algorithm is demonstrated, while every target is covered by the sensors.

Three Types of Sensor-Cloud

PSC

As an emerging sensing and computing paradigm, participatory sensing [3] tasks everyday mobile devices (e.g., cellular phones) to form interactive and participatory sensor networks, which enable public and professional users to gather and analyze as well as share local knowledge. In other words, participatory sensing combines the ubiquity of mobile devices with the sensing capabilities of sensor networks, targeting the seamless collection of data from a large number of user-carried devices.

Induced by the participatory sensing concept, as shown in Fig. 2a about PSC, users play the role of sensors by sensing and gathering data with mobile devices and further transmitting the data to the cloud. After the cloud stores and processes the received data, the processed data is delivered to users on demand. Moreover, clouds share the data with each other by communications between them.

In such a way, when offering big data to users, the following benefits can be achieved with PSC:

- The number of sensors needed to be deployed to sense and gather data can be reduced greatly, since a lot of data is sensed and gathered with user-carried devices.

- The amount of data that needs to be gathered can be decreased substantially due to a certain number of data being shared through cloud-to-cloud communications.

ASC

In the field of artificial intelligence, agent-based system technology [4] is a very powerful paradigm to conceptualize, design, and implement software systems. Specifically, characterized by acting autonomously on behalf of their users, agents are computer programs that address a growing number of complex problems across distributed and open environments.

Motivated by the agent concept, as illustrated in Fig. 2b concerning ASC, there are agents near the sensors. The agents are in charge of transmitting the data to the cloud, while the sensors sense, gather, and transmit the data to the agents. After the cloud stores and processes the received data, the processed data is delivered to users on demand. In addition, the clouds share the data with each other via communications between them.

In such a manner, when providing big data to users, the following advantages can be obtained with ASC:

- The energy consumption of sensors can be reduced substantially, because the multihop communications from the sensors to the sink/sinks are eliminated with the agents.
- The speed at which the cloud receives the sensory data can be faster, since the data sensed by the sensors is directly transmitted to the cloud via the nearby agents rather than indirectly transmitted to the cloud via a faraway sink or sinks.
- The volume of data that needs to be collected can be decreased substantially, because a certain number of data is shared via cloud-to-cloud communications.

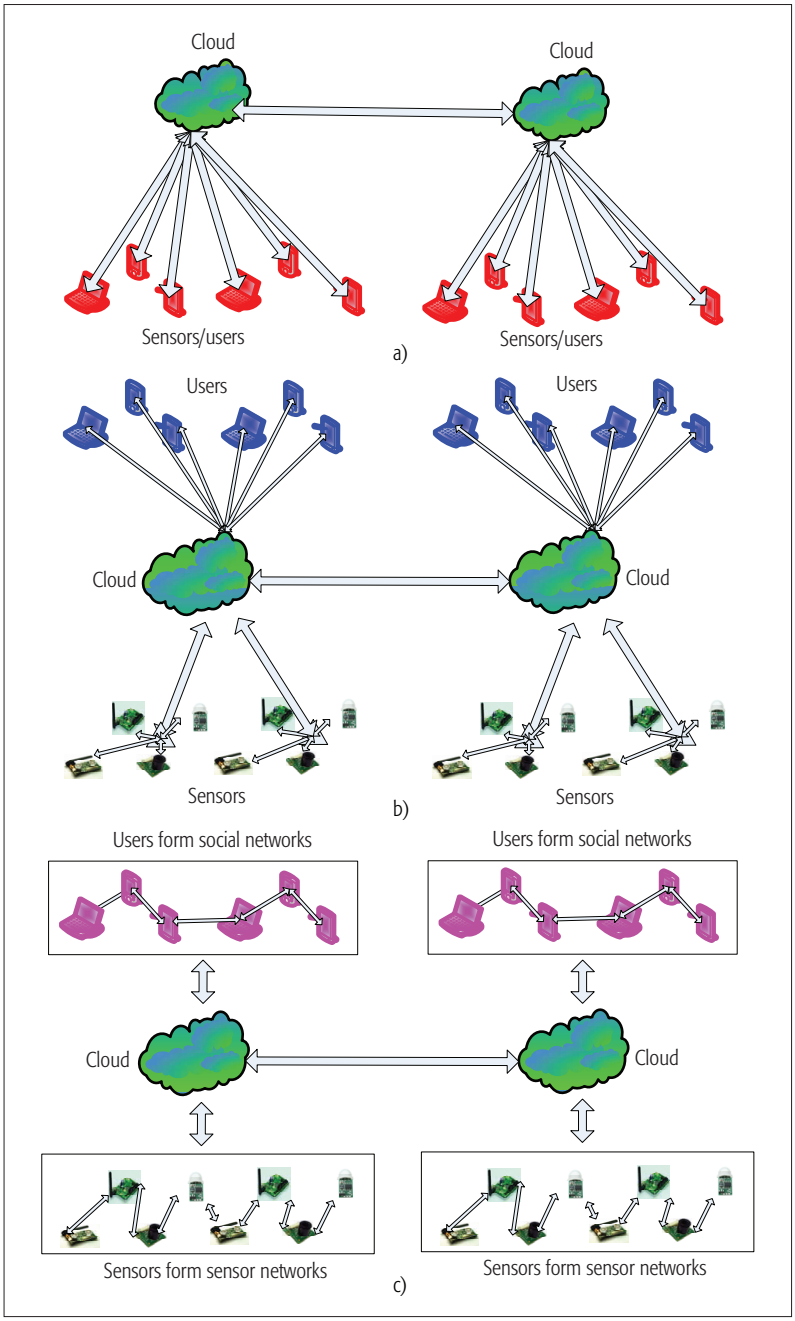


Figure2. a) PSC: b) ASC; c) SSC.

SSC

A social network [5] is a network in which social members (e.g., individuals, organizations) interact with each other. Particularly, in a social network, the resources and services can be shared among social members with established relationships.

Based on the social network concept, as demonstrated in Fig. 2c regarding SSC, the users form social networks and the sensors form sensor networks. The data sensed and gathered by the sensors in the sensor networks is transmitted to the cloud via the sink or sinks in the sensor networks. After the cloud stores and processes the received data, the processed data is delivered to the social networks on demand. Then the users in the social networks share the received data. Moreover, the clouds shares the data with each other through communications between them.

In such a way, when offering big data to users, the following benefits can be obtained with SSC:

- The energy consumption of cloud can be decreased greatly, since the desirable data is delivered from the cloud to the social networks once rather than from the cloud to various users multiple times.
- The amount of data that needs to be gathered can be reduced substantially, due to a certain number of data being shared through cloud-to-cloud communications.

Case Study

As presented in Fig. 3, in green city, various entities (shopping malls, markets, retail stores, restaurants, playgrounds, parks, traffic routes, factories, plants, universities, schools, companies, etc.) exist. In some places (e.g., shopping malls, markets, retail stores, restaurants, playgrounds, parks, traffic routes) where there is high diversity and mobility of users, PSCs can be applied to sense the big data. For some sites (e.g., factories, plants) where there is high stability in terms of the type of users and the expected performance of sensors, ASCs can be deployed to transmit the big data. Regarding some places (e.g., universities, schools, companies) in which there is a high similarity of users leading to an easily shared network, SSCs can be formed to share the big data.

Open Research Issues on Big Data

Integration of Big Data: Big data is in various data formats from various sources. For big data, one of the most important issues is integrating such a large amount of data with different formats from multiple sources. Therefore, integration of big data is one of the key challenges to be addressed, especially if some data are incomplete, incorrect, or in the wrong format.

Processing Platforms of Big Data: Substantial big data applications are based on big data analysis, which usually requires huge processing capability. As a result, the hardware and software platforms for big data should offer communication and computing capabilities, with high and stable performance. In addition, the software platforms should be compatible with different kinds of hardware platforms.

Management of Big Data: Big data that have been created at ever increasing rates are characterized by volume, variety, and velocity. Thus, there is great potential in terms of obtaining valuable insights from big data. For proper and efficient utilization of big data, it is critical to design appropriate and effective big data management techniques based on various applications.

Open Standards of Big Data: Open standards are a significant element for ensuring that different systems can interoperate with each other. For big data in multiple systems, open-standards-based technologies should be explored. With open-standards-based technologies, various big data systems can exchange and utilize information cohesively.

Security and Privacy of Big Data: One essential issue for big data is security and privacy. Different data need different levels of security policies to protect them against malicious attacks. Furthermore, various data require different levels of privacy policies for preventing unauthorized access. In addition, for different big data transmissions over various types of networks, different levels of security and privacy policies are demanded.

Quality of Service of Big Data: In terms of big data, the quality of service of big data is always a fundamental concern for users. Quality of service is affected by various factors in the specific application scenarios, while satisfactory quality of service is different for various users. Therefore, it might be promising to investigate techniques that can predict and control the quality of service of big data for various application scenarios and different users.

Open Research Issues about Sensor-Cloud

Framework of Sensor-Cloud: Sensor-cloud is a system with a number of interesting applications (e.g., vital signs monitoring, water quality monitoring, fault diagnostics). For various applications of sensor-cloud, the frameworks of sensor-cloud are generally different. Thus, it is worthwhile to explore the framework of sensor-cloud, considering the specific application requirement (throughput, response time, etc.).

Energy Efficiency of Sensor-Cloud: Energy efficiency is always a crucial factor for sensor-cloud, since most sensors are generally equipped with non-rechargeable batteries, and the implementation of cloud inherently involves tremendous energy consumption. Therefore, mechanisms (e.g., duty scheduling for sensor networks, energy-efficient medium access control for sensor networks, energy-efficient data transmission from sensor networks to cloud, energy-efficient job scheduling in cloud) to save the energy consumption of sensor-cloud are in urgent need.

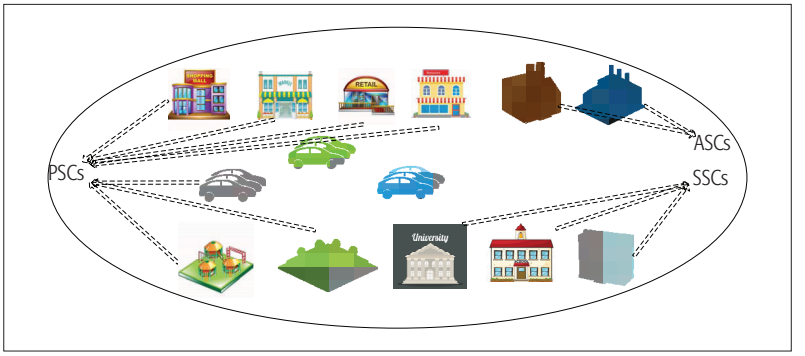


Figure 3. Case study of PSC, ASC, and SSC.

Resource Management of Sensor-Cloud: There are various resources (e.g., sensing resources, storage resources, computing resources) in sensor-cloud. Maintaining the resources, utilizing the resources, and eventually offering desirable data to users is a challenging task. In this regard, resource management strategies (e.g., resource allocation, resource optimization, resource coordination) for sensor-cloud are very important.

Localization of Sensor-Cloud: Localization is a necessary technique for sensor-cloud, since the data sensed and gathered by the sensors will be useless if the sensors have no idea of their geographical locations. Particularly, for a sensor-cloud in which there are a large number of sensors, a localization process that finds the position of sensors is needed because it will be very expensive to use GPS for many sensors. As a result, localization mechanisms for sensor-cloud should be investigated.

Mobility of Sensor-Cloud: Mobility is a fundamental attribute that should be considered for sensors. Mobile sensor networks have attracted a lot of attention from both the academic and industrial communities, and they have demonstrated several benefits (e.g., better energy efficiency, improved coverage, enhanced target tracking) over static sensor networks. Similarly, taking into account the mobility of sensors in sensor-cloud, the performance of mobile sensor-cloud is worthy of study, in contrast to static sensor-cloud.

Testbed of Sensor-Cloud: Regardless of the size of sensor-cloud, there is considerable cost with respect to the deployment, operation, and maintenance of sensor-cloud. To better use the installed capacity of sensor-cloud to reduce the underutilization of sensor-cloud to the barest minimum, it is very valuable to explore a testbed of sensor-cloud, on which the performance of sensor-cloud can be tested.

Conclusion

Targeted to serve as enlightening guidance for future research regarding big data in green city, this article has investigated sensor-cloud, which is a very powerful system for users in green city

There is considerable cost with respect to the deployment, operation, and maintenance of sensor-cloud. To better use the installed capacity of sensor-cloud to reduce the underutilization of sensor-cloud to the barest minimum, it is very valuable to explore a testbed of sensor-cloud, on which performance of sensor-cloud can be tested.

to achieve big data. Particularly, toward big data in green city, this article has reviewed the recent work about big data and sensor-cloud. Moreover, this article has presented three types of sensor-cloud (i.e., PSC, ASC, and SSC) for green city. In terms of PSC, participatory sensing is incorporated for sensing big data. Regarding ASC, the agent is incorporated for transmitting big data. Concerning SSC, a social network is incorporated for sharing big data. Finally, this article has outlined open research issues regarding big data and sensor-cloud.

Acknowledgment

This work was partially supported by funding from the Natural Sciences and Engineering Research Council of Canada, the ICICS/TELUS People & Planet Friendly Home Initiative at the University of British Columbia, TELUS, and other industry partners. This work was partially supported by the National Natural Science Foundation of China (NSFC) under Grant 61602272 and Grant 61572262, and the NSF of Jiangsu Province under Grant BK20141427. This work was partially supported by the project IoTSec – Security in IoT for Smart Grids, with number 248113/O70 part of the IKTPLUSS program funded by the Norwegian Research Council. This research is partially supported by the projects 240079/F20 funded by the Research Council of Norway. H. Zhou is the corresponding author.

References

[1] J. Wu *et al.*, "Big Data Meet Green Challenges: Big Data Toward Green Applications," *IEEE Systems J.*, vol. 10, no. 3, Sept. 2016, pp. 888–900.
[2] C. Zhu *et al.*, "Collaborative Location-Based Sleep Scheduling for Wireless Sensor Networks Integrated with Mobile Cloud Computing," *IEEE Trans. Comp.*, vol. 64, no. 7, July 2015, pp. 1844–56.
[3] H. Gao *et al.*, "A Survey of Incentive Mechanisms for Participatory Sensing," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 2, 2nd qtr. 2015, pp. 918–43.
[4] D. Ye, M. Zhang, and A. V. Vasilakos, "A Survey of Self-Organization Mechanisms in Multiagent Systems," *IEEE Trans. Sys., Man, Cybern.*, vol. 47, no. 3, Mar. 2017, pp. 441–61.
[5] Y. Jiang and J. C. Jiang, "Understanding Social Networks from a Multiagent Perspective," *IEEE Trans. Parallel Distrib. Sys.*, vol. 25, no. 10, Oct. 2014, pp. 2743–59.
[6] X. Yi *et al.*, "Building a Network Highway for Big Data: Architecture and Challenges," *IEEE Network*, vol. 28, no. 4, July-Aug. 2014, pp. 5–13.
[7] Z. Su, Q. Xu, and Q. Qi, "Big Data in Mobile Social Networks: A QoE-Oriented Framework," *IEEE Network*, vol. 30, no. 1, Jan.-Feb. 2016, pp. 52–57.
[8] L. Gu *et al.*, "Cost Minimization for Big Data Processing in Geo-Distributed Data Centers," *IEEE Trans. Emerging Topics Comp.*, vol. 2, no. 3, Sept. 2014, pp. 314–23.
[9] I. A. T. Hashema *et al.*, "The Role of Big Data In Smart City," *Int. J. Info. Mgmt.*, vol. 36, no. 5, Oct. 2016, pp. 748–58.
[10] M. Rathore *et al.*, "Real-Time Big Data Analytical Architecture for Remote Sensing Application," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 10, Oct. 2015, pp. 4610–21.
[11] C. Zhu *et al.*, "Towards Pricing for Sensor-Cloud," *IEEE Trans. Cloud Comp.*, vol. PP, no. 99, Jan. 2017.
[12] A. Sen and S. Madria, "Risk Assessment in a Sensor Cloud Framework Using Attack Graphs," *IEEE Trans. Serv. Comp.*, vol. PP, no. 99, Mar. 2016.
[13] S. Chatterjee, S. Misra, and S. Khan, "Optimal Data Center Scheduling for Quality of Service Management In Sensor-Cloud," *IEEE Trans. Cloud Comp.*, vol. PP, no. 99, Oct. 2015.
[14] S. Misra, S. Chatterjee, and M. S. Obaidat, "On Theoretical Modeling of Sensor Cloud: A Paradigm Shift from Wireless Sensor Network," *IEEE Sys. J.*, vol. PP, no. 99, Nov. 2014, pp. 1–10.
[15] S. Chatterjee and S. Misra, "Target Tracking Using Sensor-Cloud: Sensor Target Mapping in Presence of Overlapping Coverage," *IEEE Commun. Lett.*, vol. 18, no. 8, Aug. 2014, pp. 1435–38.

Biographies

Chunsheng Zhu [S'12, M'16] (chunsheng.tom.zhu@gmail.com) is a postdoctoral research fellow in the Department of Electrical and Computer Engineering, University of British Columbia, Canada. He received his Ph.D. degree in electrical and computer engineering from the University of British Columbia in 2016. His current research interests mainly include wireless sensor networks, cloud computing, the Internet of Things, big data, social networks, and security.

Huan Zhou is a professor in the College of Computer and Information Technology, China Three Gorges University. He is also a visiting scholar at the University of British Columbia. He received his Ph.D. degree from the Department of Control Science and Engineering, Zhejiang University, China. His research interests include mobile social networks, opportunistic mobile networks, and wireless sensor networks.

Victor C. M. Leung [S'75, M'89, SM'97, F'03] is a professor in the Department of Electrical and Computer Engineering and holder of the TELUS Mobility Research Chair, University of British Columbia. His research is in the areas of wireless networks and mobile systems. He is a Fellow of the Royal Society of Canada, a Fellow of the Canadian Academy of Engineering, and a Fellow of the Engineering Institute of Canada.

Kun Wang [M'13] is an associate professor in the School of the Internet of Things, Nanjing University of Posts and Telecommunications, China. He received his Ph.D. degree from the School of Computing, Nanjing University of Posts and Telecommunications in 2009. In 2016, he was a research fellow with the School of Computer Science and Engineering, University of Aizu, Fukushima, Japan. His current research interests include information security, ubiquitous computing, and wireless communications technologies.

Yan Zhang [SM'10] is a professor in the Department of Informatics, University of Oslo, Norway. He is also a chief research scientist at Simula Research Laboratory, Norway. He is an Associate Technical Editor of *IEEE Communications Magazine*, an Editor of *IEEE Transactions on Green Communications and Networking*, and an Editor of *IEEE Communications Surveys & Tutorials*. His current research interests include next-generation wireless networks leading to 5G, green, and secure cyber-physical systems.

Laurence T. Yang [M'97, SM'15] is a professor in the Department of Computer Science, St. Francis Xavier University, Canada. His research interests include parallel and distributed computing, embedded and ubiquitous/pervasive computing, and big data. He has published more than 220 papers in various refereed journals (around 40 percent in top IEEE/ACM transactions and journals). His research has been supported by the National Sciences and Engineering Research Council and the Canada Foundation for Innovation.

Big Data Analytics for Electric Vehicle Integration in Green Smart Cities

Boyang Li, Mithat C. Kisacikoglu, Chen Liu, Navjot Singh, and Melike Erol-Kantarci

ABSTRACT

The huge amount of data generated by devices, vehicles, buildings, the power grid, and many other connected things, coupled with increased rates of data transmission, constitute the big data challenge. Among many areas associated with the Internet of Things, smart grid and electric vehilces have their share of this challenge by being both producers and consumers (ie., prosumers) of big data. Electric vehicls can significantly help smart cities to become greener by reducing emissions of the transportation sector and play an important role in green smart cities. In this article, we first survey the data analytics techniques used for handling the big data of smart grid and electric vehicles. The data generated by electric vehicles come from sources that vary from sensors to trip logs. Once this vast amount of data are analyzed using big data techniques, they can be used to develop policies for siting charging stations, developing smart charging algorithms, solving energy efficiency issues, evaluating the capacity of power distribution systems to handle extra charging loads, and finally, determining the market value for the services provided by electric vehicles (i.e., vehicle-to-grid opportunities). This article provides a comprehensive overview of the data analytics landscape on the electric vehicle integration to green smart cities. It serves as a roadmap to the future data analytics needs and solutions for electric vehicle integration to smart cities.

INTRODUCTION

According to International Data Corporation's (IDC's) visionary presentation on "The Digital Universe of Opportunities," the overall created and copied data volume worldwide was 4.4 zettabytes (ZB) in 2013. The volume of data is doubling every two years, and by 2020 the total volume will exceed 44 ZB (44 trillion GB). Besides the volume, the velocity of the data is growing as a result of the advances in communication technologies and the Internet of Things (IoT). Such enormous datasets with high velocity, veracity, and variety are expressed as the big data phenomenon.

Smart grid and electric vehicles (EVs) are among the main drivers of IoT, as they form a large connected network of things, such as vehicles, charging stations, smart meters, intelligent electronic devices (IEDs), and phasor measurement units (PMUs). They are also anticipated to be the drivers of green smart cities by enabling efficient integration of renewable energy and lower emissions. The green smart city vision anticipates almost all flat surfaces, including roads, covered by solar panels to maximize the utilization of solar energy [1]. EVs carry dozens of sensors that provide data including user driving behaviors, battery security via a battery management system (BMS), and grid charge management via charging stations. Drivers, as well, carry smart devices and wearables that contribute to the data generated on roads. With smart, autonomous, self-driving cars, those data will be continuously moving from cars to servers and cars to cars. In the case of EV grid integration (EVGI), their charging and discharging pattern is tightly coupled with the operation, security, and efficiency of the smart grid. In that sense, data analytics play a critical role in EVGI, green smart cities, and other green infrastructure as presented in [2]. In particular, charging planning and harmonization of EVs for selling power back to the grid (i.e., vehicle to grid, V2G) require fast and reliable data analytics techniques.

In this article, we provide a comprehensive survey of existing techniques, and provide a roadmap for future technologies in data analytics for EVGI applications in green smart cities. We start with a brief overview of smart grid and EVs to present the applications and potential challenges. We discuss the sources of big data generation in detail. Then we continue with a survey on data analytics tools that are used in this domain. The article aims to introduce the existing data analytics studies on EVs and smart grid. Hadoop-based cloud platforms, prediction methods, and decision support tools are among the surveyed data analytics studies. The article provides a requirement analysis for future data analytics tools and aims to serve as a roadmap for researchers in this area.

Figure 1 provides an overview of EV integration with V2G, G2V, power and information exchange, heterogeneous communication technologies, data flow, cloud integration, applications, and big data analytics tools. As shown in the figure, EV-EVSE-grid communication can be implemented by power line communications (PLC) and wireless networks. Note that EVSE is the EV supply equipment and is used interchangeably throughout this article. EVs can be charged and

The authors provide a comprehensive overview of the data analytics landscape on the EV integration to green smart cities. It serves as a roadmap to the future data analytics needs and solutions for EV integration to smart cities.

Boyang Li is with Illinois Institute of Technology; Mithat C. Kisacikoglu is with the University of Alabama; Chen Liu is with Clarkson University; Navjot Singh and Melike Erol-Kantarci are with the University of Ottawa.

Digital Object Identifier: 10.1109/MCOM.2017.1700133

0163-6804/17/$25.00 © 2017 IEEE

Despite the advantages, there are still challenges for the widespread adoption of EVs. Limited driving range and associated driver range anxiety, long duration for charging, and non-ubiquity of charging stations are the critical barriers to the penetration of EVs.

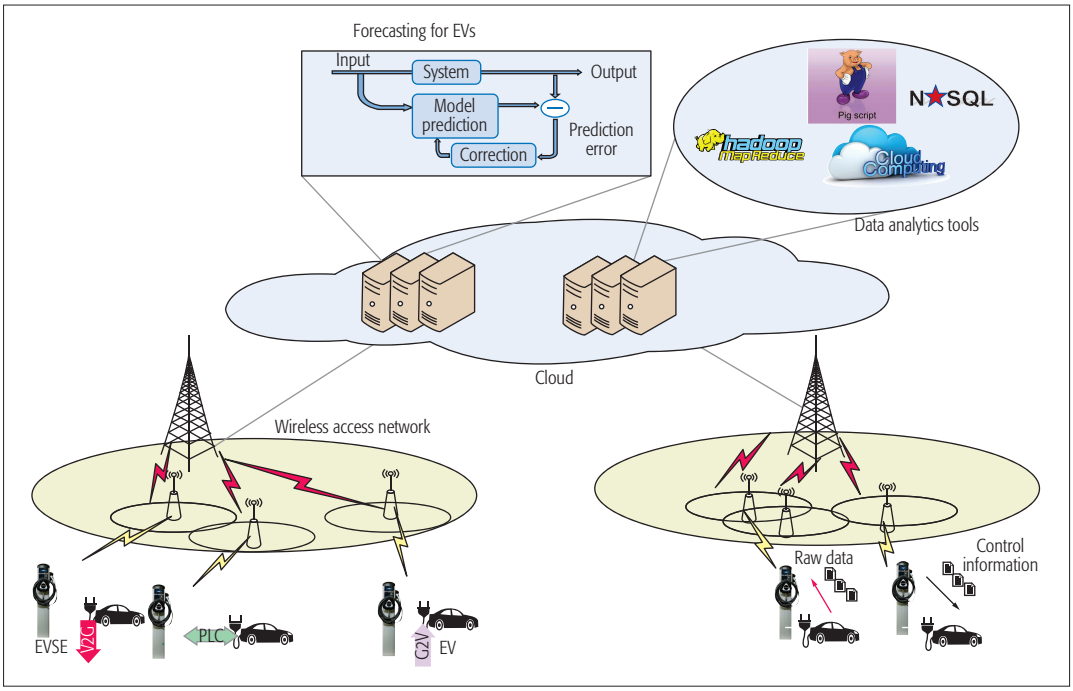


Figure 1. Implementation diagram of EVGI.

discharged in a coordinated fashion. The big data from EVs and all other entities are stored and processed over the cloud for various application purposes including optimized charging. Some widely used big data tools are also plotted in the figure.

The rest of the article is organized as follows. The following section provides an overview of the integration of EVs to the smart grid. Then we describe the sources of big data in EVs and G2V/V2G applications. Following that, we provide a detailed survey on the available big data approaches and platforms for processing EV data. We then discuss the open issues, requirements for data analytics tools tailored for EVs, and some future concepts that require more research. In the final section, we conclude the article.

An Overview of Smart Grid and Electric Vehicle Integration

Smart grid is the modernized electrical grid that integrates advanced sensing, communication, and control functionality for the purpose of enhanced efficiency, reliability, and security in the operation of the utility grid. EVs are either completely or partially powered by their onboard batteries, which are charged by the power grid. Plug-in EVs use less fossil fuels and emit much less CO_2 compared to conventional vehicles and therefore are incentivized by many governments to improve air quality and reduce greenhouse gas emissions. Despite the advantages, there are still challenges for the widespread adoption of EVs. Limited driving range and associated driver range anxiety, long duration for charging, and non-ubiquity of charging stations are the critical barriers to the penetration of EVs. In addition, with a growing EV market, the impact of EVs on the power grid is a matter of concern, especially at the distribution level. This may result in adverse effects such as peak loading, increased losses, voltage unbalance/deviations, and need for additional network reinforcements. Data management in EVGI plays a vital role for healthy integration of these new technology vehicles into the future green smart cities.

EV batteries are charged using onboard chargers and EVSE, also known as charging stations. EVSEs can be located at residential premises, parking lots of commercial buildings, and any roadside charging facility. The EV integration framework enables EVs to be controlled by the smart grid or aggregators via the communication between vehicles and the grid. Communication between EVs and the smart grid can be a mix of wireless and wired technologies including PLC, Zigbee, WiFi, LTE, and fifth generation (5G) wireless networks. A hardware description of EV grid interaction and the EVGI system with bidirectional power and communication architecture are shown in Fig. 2. The communication between an EV and the smart grid includes two concepts: basic signaling and high-level communication. Basic signaling refers to EV charging control methods utilizing the control pilot signal of the charging plug for basic G2V charging control, as shown in Fig. 2. This control is realized by modifying the duty cycle ratio of the control pilot signal, which is already available in all of the plug-in EVs in the market. Communication functions via the current control pilot pin have a fairly simple structure and cannot provide the bidirectional information required between EV and grid. On the other hand, V2G control is achieved via high-level communication that uses PLC superimposed on the control pilot signal. In the case of PLC, V2G communications is overseen by the EV communication controllers (EVCCs) and supply equipment communication controllers (SECCs). While the EVCC and SECC are primary actors, grid operators, charging aggregators, and electricity providers are the secondary actors of the charging communication system. EVs and EVSEs can also communicate through wireless networks for sharing data that is useful for trip planning, real-time pricing, and so on.

In G2V charging, the challenges mostly arise from the increased stress on distribution systems of the smart grid. Due to increased power consumption on the network during peak hours, off-peak hours are preferable for EV charging. In addition, if the night valley in the 24-hour electricity demand profile is filled with EV loads, the ramping up/down costs that occur in the morning/evening can be avoided. Meanwhile, charging during daytime, especially during peak hours of electricity demand, requires extra planning. Moreover, the distribution system suffers from overloading if several EV batteries are fed from the same transformer.

The charging management mainly relies on certain information being available to local (distributed) or global (centralized) controllers. Information exchanged between vehicle and controllers include user departure time, state of charge (SOC) of the battery, charging active/reactive power reference, and user-specific information such as charging preference, vehicle vendor, onboard charger power, and battery capacity. Compared to distributed control, centralized control achieves better utilization of EVs for grid support due to having more information and achieving optimum results. Therefore, central power optimization is one of the most explored analysis methods in EV charging networks to solve congestion related problems. On the other hand, the distributed strategy allows each EV to determine its own charging profile, which may not always result in an optimal aggregated charging regime. However, the distributed approach has gained more attraction in the literature because of its higher flexibility for the EV user, higher reliability, and easier field implementation [3]. In this case, data communication is much lower, and private information is mostly kept in vehicle.

Despite the large number of studies on EV charging, less has been explored on how this data will be attained, how it will be processed for larger penetration of EVs, and how data from other sources could be coupled with data from EVs to predict the behavior of a driver for charging or discharging the battery. In particular, data that are potentially available while the EV is on the move have been underutilized. Obviously, more data give more chance to derive useful insights, but decision making from big data of EVs requires finding the right information in near real time. In this context, data analytics techniques can increase the efficiency of EV data.

During V2G, when EVs act as distributed generators, data analytics become more of a concern and are more needed. When an EV is allowed to sell electricity, information on where the vehicle will be in the next time frame, how much energy will be left in the battery when it is reconnected, how much of this energy will be reserved for trading, load on the utility when the EV is plugged in, and similar information need to be available. Part of this data could be voluntarily made available by the driver, some could be predicted, and the rest could be collected from sensors. In any case, the amount and speed of data flow are quite big, and robust data analytics tools are needed to make effective and timely decisions.

In the following section, before we survey the data analytics tools, we analyze the potential data

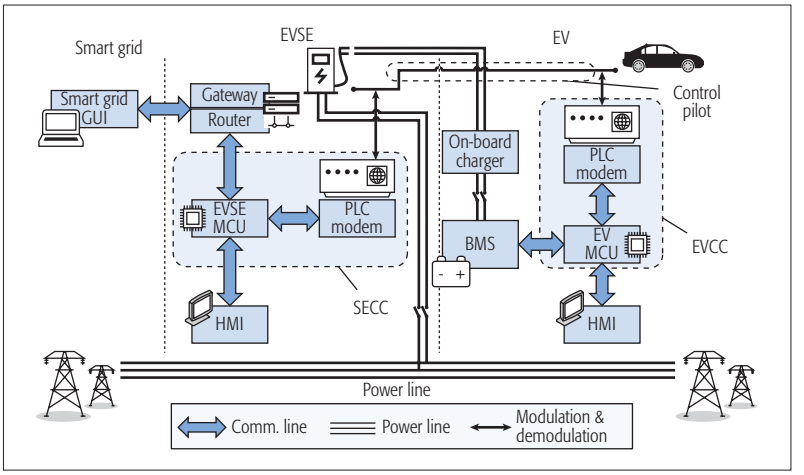


Figure 2. Overview of EV integration with V2G, G2V, heterogeneous communication technologies, data flow, cloud integration, applications, and big data analytics tools.

sources related to EVs. The data from EVs can be vehicle, driver, charging station, or even smart city related such as traffic condition on roads. The vehicle data can come from various sources such as batteries, onboard chargers, and trip logs. In addition, wearables on drivers contribute to the big data of cars. The utility grid power consumption data stream is also important to determine which charging/discharging scenario should be employed for the specific geographical location.

Big Data of Electric Vehicles

New generation autonomous self-driving cars, whether electric or not, are equipped with hundreds of sensors and surrounded by smart technologies. Furthermore, road infrastructure is also underway with large deployment of connected technologies (i.e., traffic lights, signs, and road cameras). The advances in wireless and vehicular communications enable these smart cars to be able to communicate with the infrastructure and other smart cars. Autonomous connected vehicles and their interaction with smart cities will increase the amount of data that is generated and shared. In addition, drivers carry a number of sensors on their smartphones and wearable devices. In general, IoT, and in particular the Internet of Vehicles (IoV) and Internet of Energy (also known as Energy Internet), benefit from cloud services [4]. Onboard and on-body devices have limited storage and processing capabilities. Meanwhile, their communication capability opens the door to accessing powerful cloud servers. The data from EVs, drivers, charging stations, and infrastructure constitute the big data of EVs, which requires data analytics tools running on the cloud.

Many automobile manufactures allow drivers to check the status of their EVs and remotely control their charging through mobile apps. These applications collect vehicle and trip data. EV data mostly come from onboard electronic control unit (ECUs) and battery management systems (BMSs). SOC of EV batteries is a key parameter for most charging and discharging decisions. BMS logs show SOC information and how an EV battery is performing. Malfunctioning battery cells, and heating and cooling details can be observed by these logs. Based on BMS logs, state of health

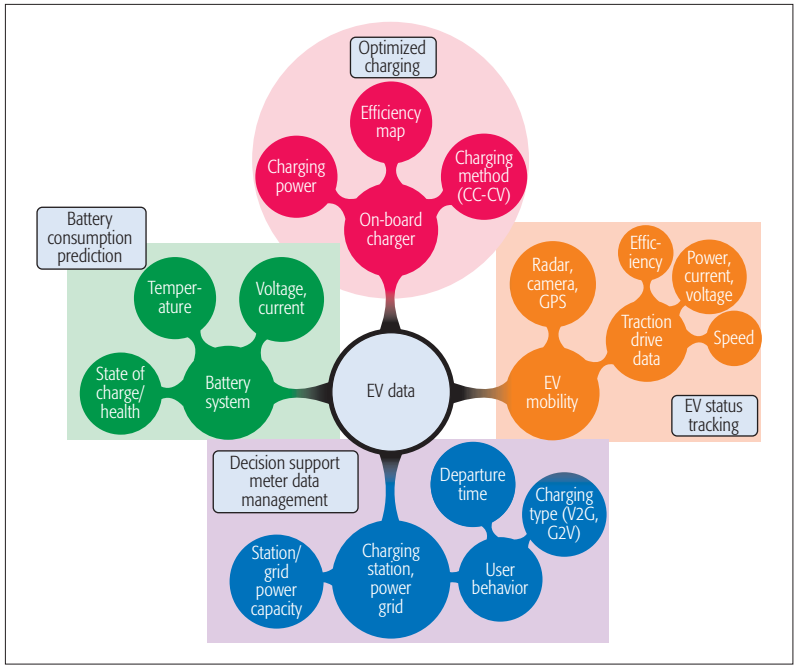


Figure 3. Summary of EV related data sources.

(SOH) information can be obtained, and the impact of V2G services on battery life can be accurately observed.

In addition to the data directly collected from EVs, drivers can voluntarily share information about their driving patterns and charging habits. Trip information including start and end times of journeys, connect and disconnect times of chargers, and the battery SOC can easily be collected. Advanced systems can record details like how much air conditioning is used, or how a driver accelerates or breaks. All these various kinds of data can be used for decision making through data analytics tools.

An important parameter for EV performance is the driving range. For a long time, market acceptance of EVs was low due to range anxiety, which is the worry that the EV battery will run out of power before the destination or a suitable charging point is reached. Big data is frequently used to estimate the driving range, which is an efficient way to diminish the range anxiety. In [5], the authors proposed a classification method related to driving range estimation. The data are classified as standard, historical, and real-time data. Those are defined as follows.

Standard Data: This includes the data obtained from official sources such as scheduled tours and activities from websites, the usual driving time to the destination according to Google Maps, or climatic conditions such as hurricane season or dry season.

Historical Data: It refers to the indirect data resulting from the feedback of other drivers' experience. For example, recent miles per gallon equivalent (MPGe) of a car can be used to predict the refueling stops on the road. Websites such as tripadvisor provide reviews from previous travelers who share similar trips. Yelp provides information on accommodation and food stops. These are examples of historical data.

Real-Time Data: This kind of data is closely related to emergency issues. Real-time traffic conditions are monitored by the GARMIN app, including examples of sudden rain or snow and unplanned road closures [5].

In [6], the authors proposed a framework to explore drivers' behavioral patterns and estimate the driving range. They have collected data from an EV in Taiwan over one year. They used the Growing Hierarchical Self-Organizing Maps (GHSOM) algorithm to categorize the driving pattern.

Besides range estimation, big data from EVs can be used by municipalities to make decisions on siting public charging stations. In this respect, the key factor is the evaluation of charging demand. Various kinds of data have been employed such as road traffic density, distribution of gas stations, and vehicle ownership. There are also several studies that use travel patterns of taxi fleets in order to derive optimal charging station siting. In [7], the authors proposed a way to site public EV charging stations using big-data-informed travel patterns of a taxi fleet. Using Beijing as a case study, they examined a large-scale data set containing 11,880 taxis for a month. Meanwhile, in [8] information from over 30,000 personal trip records in Seattle, Washington, gathered from the Puget Sound Regional Council's 2006 household travel survey, were used to determine public EV parking locations and durations. Regression methods have been used to predict parking demand variables, including total vehicle hours per zone, neighborhood and parked time per vehicle trip, and so on, as a function of site accessibility, local jobs, population densities, and trip attributes. As cities become smarter, such data will have vast volume, and mining them along with EV data will provide more opportunities for planning. A summary of the data related to EVGI explained in this section is presented in Fig 3. In the next section, we survey the data analytics tools that make use of the data described above.

Big Data Analytics Platforms and EV Integration

The data collected from EVs are various, and the volume is huge; therefore, traditional statistical ways to build a model may not work very well. Big data analytics have been useful for EV integration in a variety of ways such as optimized charging, battery management, and EV status tracking. In this section, we group the existing studies based on the platform used. The first subsection focuses on Hadoop-based techniques that allow parallel processing of EV big data. The second subsection presents a study that uses the Weka data mining tool.

Hadoop-Based Approaches

Optimized Charging: In [9], *Wei et al.* developed an optimized charging model, the multi-level feedback queue. Their model uses grid demand data, charging station data, EV battery data, user data, and data from a local distribution system. In order to handle the large amount of data from multiple sources, they proposed processing data in parallel using MapReduce over the Hadoop framework. The authors store their data using HBase, which is a NoSQL-based database used for big data storage on the cloud computing platform.

Scheme	Data source	Purpose	Platforms and tools
Optimized charging [9]	Demand information, charging station parameters, car battery data and user data	Optimizing the charging via job scheduling	Hadoop and HBase
Battery consumption prediction [10]	EVs data collected from Jeju Island testbed	Improving the accuracy of battery consumption model	Hadoop and R statistical package
Charging meter data management [11]	EVSE data collected from Jeju Island testbed	Improving the interoperability of heterogeneous chargers	Hadoop, Pig script, MySQL
EV status tracking [12]	Sensor data collected from EVs in Indianapolis	Extracting raw data and transforming into classified buckets	Hadoop and HBase
Weka-based decision support scheme [13]	New York City (NYC) demand data	Building decision support engine for power system operators	J48 and M5 algorithms from Weka platform

Table 1. Comparison of big data analytics tools used for EVs.

The presented approach is promising; however, its performance has not been evaluated with real data. Further studies are needed to show how charging can be optimized using Hadoop and big data from EVs.

Battery Consumption Prediction: In [10], the authors proposed spatio-temporal analysis of EV data using Hadoop and the R statistical package. The data were collected through a battery monitoring device that accumulated SOC records of EVs along major roads in Jeju Island, Korea. The authors used Pig scripts, which are high-level programming scripts designed for Hadoop, to filter the necessary fields from the raw data heap. Then they used the R package to conduct time series analysis and provided prediction of EVs' battery consumption.

Charging Meter Management: In [11], the authors used a similar framework to [10] to implement meter management over streaming EV data. They implemented a data analysis framework, which, after retrieving the temporal stream records from the Master Data Management Software (MDMS), used Hadoop Pig scripts to filter the raw data. Then the Hadoop Pig script results were converted to SQL commands to insert data to MySQL. At the final step, a neural network library was used to forecast future EV connections. Similar to [10], the authors used data collected from EVs in Jeju Island. The island aims for all its vehicles to be electric by 2030 as part of its becoming a carbon-fee city (http://spectrum.ieee.org/energywise/transportation/ efficiency/korean-island-plans-for-all-electric-vehicles-by-2030). The research in [10, 11] represent those efforts toward green cities and demonstrate how EV data analytics techniques can be used for this purpose.

EV Status Tracking: In [12], the authors addressed the unstructured nature of big EV data. They also used Hadoop and MapReduce. The authors first employed a preprocessing stage to remove inconsistencies and duplicates in the data to ensure optimum storage. Then the data was imported to HBase. In this study, raw EV data was extracted and transformed into classified buckets. The data were collected through the Think City project in Indianapolis. As a result, the authors observed more than 10 features for over 200 EVs. They tracked analytics on SOC, maximum vehicle speed, location, temperature, maximum voltage, and current of charging.

WEKA-Based Approach

Decision Support Tool: In [13], Ranganathan *et al.* proposed using decision tree algorithms provided in the Weka data mining platform to analyze smart grid and EV data, and form a decision support tool for grid operators. The authors used NY Independent System Operators (NYISO) demand data that is publicly available. The proposed decision support system has two phases: data preprocessing and data classification. The data preprocessing stage removes irrelevant data and noise, while classification is used to reach a decision and is based on a decision tree with predefined rules. For classification, the authors use the J48 ad M5 algorithms, which are readily available in the Weka platform. Although the authors worked on large datasets from the power grid, these data sets are offline and their size is still manageable compared to the big data that will be flowing from millions of EVs. The proposed Weka-based decision support scheme needs to be further evaluated over streaming data from EVs.

A comparison of the surveyed big data analytics approaches is given in Table 1. The first three studies of this section [9–11] focus on big EV data in order to provide input for EV applications. The research in [12] studies EV data processing only and aims to structure the data for general EV applications. The final surveyed scheme, [13], is a decision support tool for power system operators. It is suitable for large datasets; however, when streaming big data from EVs are analyzed, Weka-Hadoop-based platforms may be considered. In addition, surveyed studies have focused on HBase, while there are other NoSQL databases such as Cassandra and MongoDB.

Requirement Analysis and Future Research Directions

Overview of the Landscape

The existing literature on applying big data tools on EV and smart grid data is limited. One of the major challenges is lack of publicly available real-world data. Several of the surveyed studies have worked on data collected in related projects; however, these are still not large-scale when compared to the anticipated higher penetration of EVs and increased data flow. The research in [14] generates synthesized EV data where EV characteristics have been superimposed on real

The value of big data analytic tools would be better evaluated as they transition into decisions for system operators. Nevertheless, there are many future opportunities to explore in this area.

traces of taxis in the San Francisco, California, area. This study provides a useful public dataset for EV integration; however, it is not suitable to evaluate big data methods as the size of data is still manageable with traditional data analytic techniques. In addition, in the real world, heterogeneous, unstructured EV data will be streaming from EVs in real time, and it is hard to mimic the challenges and evaluate the true performance of big data approaches with such static datasets. Electric vehicles and the new generation of autonomous vehicles can generate data on the order of several hundreds of gigabytes to thousands of gigabytes during mobility where the amount of the data depends on the variety of sensors used for autonomy (http://www.networkworld.com/article/3147892/internet/one-autonomous-car-will-use-4000-gb-of-dataday.html). During V2G/G2V operation of electric vehicles, the data generation rate and speed are expected to be lower compared to mobility, but the varied sensors collecting data from power grid, vehicle, charging station, and driver will need solutions from the big data domain.

Our article categorizes the surveyed approaches based on their platform selection. Most of the studies work on distributed Hadoop clusters and benefit from parallelization with MapReduce. They either use the data for a specific EV application such as charging, battery, or charger management, or store data in HBase to provide for future applications. There is also work on utilizing Weka, which is a well-known data mining platform.

The research on big data analytics for EVs is in its infancy. The performance of NoSQL databases such as Cassandra and MongoDB is unexplored. Furthermore, the analyzed data has not been transformed into decision making in many of the studies. The value of big data analytic tools would be better evaluated as they transition into decisions for system operators. Nevertheless, there are many future opportunities to explore in this area. In the following subsection, we focus on various applications in the EV and smart grid domain that can benefit from big data analytics.

Requirement Analysis and Applications for EVs

Analysis of grid integration of EVs includes different subsystems that operate in various time domains from microseconds to hours. These subsystems include transportation mobility and grid service requirements, EVSE and user behavior models, and onboard power electronics and battery system modeling for charging operation. In order for the utility to be spared the impact of the large number of EV connections and to utilize already available mass energy storage capacity in EVs, communication design of the EVGI framework and decision making through big data analytics play important roles.

Mobility needs of drivers can usually be captured with data tracking devices, which in turn helps to understand energy consumption profiles of drivers. Along with modeling the battery-to-wheel energy efficiency of different EV vendors, it may be possible to generate custom charging requirements for EVs. Data analytics is also important on the utility side requirements when controlling charging. The utility will eventually decide which services are needed by the EVs via analyzing its daily demand data stream. Data analytics is expected to become more important when EVs and intermittent renewable energy generators are integrated with daily demands of utility customers.

In addition to planning on the utility side, decision making tools need to account for user convenience such as each EV having satisfactory SOC at morning departure, as well as the emergency driving range that would be offered anytime. Besides the charging aspect, EVs can provide power to the grid through V2G applications. They can support ancillary services such as voltage support, reactive power compensation, active harmonic filtering, and power factor regulation, as well as load balancing, peak shaving, and renewable energy tracking. Additionally, in the case of a power outage, an EV can be used as an emergency backup source for the home, which is often called vehicle-to-home (V2H). An advanced charger can also provide vehicle-to-vehicle (V2V) charging to increase the charging availability of the EV, even when the EV is out of charge without a nearby charging station. This can be accomplished via wireless charging and by utilizing Uber-like social networking applications. All of these future applications of EVs will call for strong data analytics tools that fully integrate EV and smart grid data. As a result, data analytics techniques will need to work on more heterogenous and unstructured data from multiple sources flowing at higher speeds, while the decision making timeframe will need to be relatively smaller than for today's applications.

Future Directions

As mentioned above, the decision making timeframe will need to be reduced from minutes to seconds for integration with most smart grid applications. All of the surveyed approaches in this article consider Hadoop clusters in the cloud. However, the delay for accessing the cloud is a major concern for real-time applications. In this case, mobile edge computing using Hadoop-like parallelization can reduce the response time of decision making. This approach could parallelize computing tasks using MapReduce on the EVs. In fact, EVs have more computational power than other mobile devices such as smartphones or wearables; therefore, a group of EVs can be tasked with running data analytics in order to reduce latency.

On one hand, big data has enormous benefits in the economy, society, and the environment. On the other hand, there is concern about data security protection and privacy [15]. If data are excessively protected due to an individual's privacy, information would be significantly curtailed. As a result, much valuable information might be lost. Thus, there needs to be a balance between consumers' privacy and the benefit of sharing data. The utility-privacy trade-off has been explored in several studies, but there are open issues on how much uncertainty can be handled for EV integration to smart grid and green smart cities.

In summary, fast and effective data analytics approaches are required for real-time interaction of the EVs with the smart grid and smart city. Those approaches can benefit from the advances in mobile edge computing. However, security and privacy concerns escalate with distributed processing of EV data by other EVs. The nexus of processing capacity, delay, security, and privacy is

an open issue that has yet to be addressed in the domain of big data analytics for EVs.

Conclusion

In this article, we review the state-of-the-art data analytics tools for electric vehicle integration to smart grids and thereon to green smart cities as well as the big data that are generated by cars and drivers. We first provide an overview of smart grid and electric vehicle integration. We present the challenges of EV integration, and discuss how these challenges can be addressed by data analytics. Then we discuss the sources of big data including EVs, drivers, EV batteries, chargers, and EVSEs. We provide a comprehensive survey on data analytics tools that are used in this domain. We conclude the article with a summary, a requirement analysis for data analytics tools for EV related applications, and finally, with a discussion of future directions.

Acknowledgment

Dr. Erol-Kantarci's research is supported by the U.S. National Science Foundation (NSF) under Grant Number CNS-1647135 and the Natural Sciences and Engineering Research Council of Canada (NSERC) under RGPIN-2017-03995.

References

[1] R. Northfield, "Greening the Smart City," *IEEE Engineering & Technology*, vol. 11, no. 5, June 2016, pp. 38–41.
[2] J. Wu *et al.*, "Big Data Meet Green Challenges: Big Data Toward Green Applications," *IEEE Systems J.*, vol. 10, no. 3, Sept. 2016, pp. 888–900.
[3] M. C. Kisacikoglu, F. Erden, and N. Erdogan, "Distributed Control of PEV Charging Based on Energy Demand Forecast," *IEEE Trans. Industrial Informatics*, vol. PP, no. 99, 2017, pp. 1–1.
[4] K. Wang *et al.*, "A Survey on Energy Internet: Architecture, Approach, and Emerging Technologies," *IEEE Systems J.*, vol. PP, no. 99, 2017, pp. 1–1.
[5] H. Rahimi-Eichi and M.-Y. Chow, "Big-Data Framework for Electric Vehicle Range Estimation," *Proc. Annual Conf. IEEE Industrial Electronics Society*, 2014, pp. 5628–34.
[6] C.-H. Lee and C.-H. Wu, "Collecting and Mining Big Data for Electric Vehicle Systems Using Battery Modeling Data," *Proc. 12th Int'l. Conf. Info. Technology — New Generations*, 2015, pp. 626–31.
[7] H. Cai *et al.*, "Siting Public Electric Vehicle Charging Stations in Beijing Using Big-Data Informed Travel Patterns of the Taxi Fleet," *Transportation Research Part D: Transport and Environment*, vol. 33, 2014, pp. 39–46.
[8] T. Chen, K. Kockelman, and M. Khan, "Locating Electric Vehicle Charging Stations: Parking-Based Assignment Method for Seattle, Washington," *Transportation Research Record: J. Transportation Research Board*, vol. 2385, 2014, pp. 28–36.
[9] W. Wei *et al.*, "Multi-Level Feeder Queue Optimization Charging Model of Electric Vehicle and its Implementation of MR Algorithm," *Int'l. J. u- and e-Service, Science and Technology*, vol. 9, no. 3, 2016, pp. 199–208.
[10] J. Lee *et al.*, "Spatio-Temporal Analysis of State-of-Charge Streams for Electric Vehicles," *Proc. 14th ACM Int'l. Conf. Info. Processing in Sensor Networks*, 2015, pp. 368–69.
[11] J. Lee and G.-L. Park, "Electric Vehicle Charger Management System for Interoperable Charging Facilities," *J. Teknologi*, vol. 78, no. 5–8, 2016.
[12] V. Bolly, J. Springer, and E. Dietz, "Using Open Source NOSQL Technologies in Designing Systems for Delivering Electric Vehicle Data Analytics," *Proc. 121st ASEE Annual Conf.*, Indianapolis, IN, 2014, pp. 24.1339.1–339.9.
[13] P. Ranganathan and K. Nygard, "Smart Grid Data Analytics for Decision Support," *Proc. IEEE Electrical Power and Energy Conf.*, 2011, pp. 315–21.
[14] H. Akhavan-Hejazi, H. Mohsenian-Rad, and A. Nejat, "Developing a Test Data Set for Electric Vehicle Applications in Smart Grid Research," *Proc. IEEE VTC-Fall 2014*, Vancouver, BC, Canada, 2014.
[15] P. D. Diamantoulakis, V. M. Kapinas, and G. K. Karagiannidis, "Big Data Analytics for Dynamic Energy Management in Smart Grids," *Big Data Research*, vol. 2, no. 3, Sept. 2015, pp. 94–101.

Biographies

Boyang Li obtained his B.E degree from Tianjin University in 2010 and his M.S. degree from Clarkson University in 2016, respectively. He is currently working toward a Ph.D. degree in computer science at Illinois Institute of Technology. His former research interests lie in big data and anomaly detection. His current research interests include HPC data analytics and job scheduling in large-scale systems.

Mithat C. Kisacikoglu received his Ph.D. degree from the University of Tennessee in 2013 in electrical engineering. He joined Hacettepe University, Ankara, Turkey, as an assistant professor in 2014. He then worked at the National Renewable Energy Laboratory, Golden, Colorado, as a research engineer between 2015 and 2016. He is currently an assistant professor in the Electrical and Computer Engineering Department at the University of Alabama, Tuscaloosa. His research interests include electric vehicles, grid integration, and power electronics.

Chen Liu currently is an assistant professor in the Department of Electrical and Computer Engineering at Clarkson University, Potsdam, New York. He received his M.S. in electrical engineering in 2002 from the University of California, Riverside, and his Ph.D. in Electrical and Computer Engineering in 2008 from University of California, Irvine, respectively. His research interests include computer architecture, the interaction between system software and microarchitecture, power-aware computing, and hardware acceleration techniques.

Navjot Singh is pursuing his M.A.Sc. degree in electrical and computer engineering at the University of Ottawa. He received his B.Tech. degree in computer science and engineering from Punjab Technical University, India, in 2016. His research interests include big data and wireless networks.

Melike Erol-Kantarci (melike.erolkantarci@uottawa.ca) is an assistant professor at the School of Electrical Engineering and Computer Science, University of Ottawa, and a courtesy assistant professor at Clarkson University. She received her Ph.D. and M.Sc. degrees in computer engineering from Istanbul Technical University in 2009 and 2004, respectively. Her main research interests are 5G and beyond wireless networks, smart grid, electric vehicles, and the Internet of Things.

Security and privacy concerns escalate with distributed processing of EV data by other EVs. The nexus of processing capacity, delay, security and privacy are open issues that are yet to be addressed in the domain of big data analytics for EVs.

Simultaneous Wireless Information and Power Transfer: Technologies, Applications, and Research Challenges

Jun Huang, Cong-Cong Xing, and Chonggang Wang

The authors survey the current architectures and enabling technologies for SWIPT and identify technical challenges to implement SWIPT. Following an overview of enabling technologies for SWIPT and SWIPT-assisted wireless systems, they showcase a novel SWIPT-supported power allocation mechanism for D2D communications to illustrate the importance of the application of SWIPT.

ABSTRACT

Energy efficiency will play a crucial role in future communication systems and has become a main design target for all 5G radio access networks. The high operational costs and impossibility of replacing or recharging wireless device batteries in multiple scenarios, such as wireless medical sensors inside the human body, call for a new technology by which wireless devices can harvest energy from the environment via capturing ambient RF signals. SWIPT has emerged as a powerful means to address this issue. In this article, we survey the current architectures and enabling technologies for SWIPT and identify technical challenges to implement SWIPT. Following an overview of enabling technologies for SWIPT and SWIPT-assisted wireless systems, we showcase a novel SWIPT-supported power allocation mechanism for D2D communications to illustrate the importance of the application of SWIPT. As an ending note, we point out some future research directions to encourage and motivate more research efforts on SWIPT.

INTRODUCTION

Energy-constrained wireless devices are typically powered by batteries that suffer from limited lifetime. While replacing or recharging wireless device batteries can sustain the network's normal operations, it either incurs high operational costs or is even impossible in some scenarios.

This situation calls for a new technology by which wireless devices can constantly replenish themselves with energy from the ambient environment. Simultaneous wireless information and power transfer (SWIPT), which exploits the same emitted electromagnetic wave field to deliver both energy and information, has emerged as a powerful means to address the above issue. SWIPT promises three types of gain. First, wireless devices with SWIPT support are able to scavenge energy when receiving data, thereby prolonging their lifetime. Second, compared to the conventional time-division multiplexing mechanism, where the transmissions of power and information are separate, the transmission efficiency under SWIPT is improved. Third, with SWIPT, the interference to the communications is kept under control and can even be beneficial for energy harvesting (EH).

SWIPT is applicable to various wireless systems. One of its notable applications is RF identification (RFID) tags, which are passive devices and battery-free. Although RFID technology has been extensively investigated, its potential is not fully exploited due to the small reading range of RFID readers and constrained power. Another SWIPT application can be found in relay-assisted wireless communication systems where energy is transferred to remote terminals via one or more intermediate relays. In this situation, with SWIPT, the high path loss rate of energy-bearing signals can be mitigated.

The purpose of this article is to provide a brief overview of the current SWIPT architectures, enabling technologies, and applications, and some technical challenges in realizing SWIPT. Specifically, multiple technologies whose integration is needed to enable SWIPT in wireless systems are surveyed. As a case study to illustrate the importance of the application of SWIPT, a novel SWIPT power allocation mechanism for device-to-device (D2D) communications is presented. Also, future SWIPT research directions are comprehensively discussed in this article.

SWIPT ARCHITECTURES

A significant advance in wireless power transfer is the rectifying-antenna (rectenna), which is a diode-based circuit that converts RF signals into direct current voltage. Early research efforts on wireless power transfer focused on the design of compact and efficient rectennas or similar energy harvesters.

With the ever increasing amount of wireless devices in today's society, substantial research efforts in both academia and industry have recently focused on SWIPT due to its dual roles in transferring information and energy. Typical SWIPT architectures are shown in Fig. 1, where the time switching refers to the architecture in which each receiving antenna periodically switches between the energy harvester and the information decoder, and the power splitting refers to the architecture, where the received signal is divided into two separate signal streams, with one being sent to the energy harvester and the other to the information decoder. To realize SWIPT in practical wireless systems, sophisticated receivers have been designed based on these two architectures. In

Digital Object Identifier:
10.1109/MCOM.2017.1600806

Jun Huang is with Chongqing University of Posts and Telecommunications; Cong-Cong Xing is with Nicholls State University; Chonggang Wang is with InterDigital Communications.

0163-6804/17/$25.00 © 2017 IEEE

particular, *antenna switching* [1] is devised where the receiving antennas are separated into two groups with one group dedicated to information decoding and the other to EH. Nevertheless, this design is essentially a special case of the power splitting architecture with binary splitting power ratios. Another notable practical receiver design is *dynamic power splitting* [2] where an adjustable power ratio for EH and information decoding is designed.

Given these architectures, corresponding technologies need to be developed to successfully implement SWIPT in wireless systems. These technologies are summarized in the next section.

Enabling Technologies for SWIPT

Enabling SWIPT in wireless systems requires the integration of multiple technologies such as multi-antenna transmission, EH, resource allocation, and signal processing. The state of the art of these technologies is briefly discussed in this section.

Multi-Antenna Transmission

Among the efforts searching for a solution to reducing the transmission range of SWIPT, the idea of installing multiple antennas on devices appears to be sound and feasible, as multiple antennas not only can increase the antenna aperture, but also can attain higher gain. In order to arrange multiple antennas into a small pocket-sized device, a higher communication frequency would be useful for SWIPT systems. Incidentally, equipping a SWIPT system with multiple antennas enables two different signal processing techniques: analog domain beamforming and digital domain precoding. The former can be realized by a complex weighting phase shifter, and the latter can be specially designed to satisfy some predefined power or rate conditions/constraints.

An issue that arises and needs to be addressed in multi-antenna transmissions is co-channel interference, because the system consists of multiple users. In this regard, various existing interference alleviation techniques (e.g., block diagonalization precoding [3], which sends information to interference-free receivers and energy to others) may be applied in SWIPT systems to deal with this issue.

Efficient Energy Harvesting

Toward achieving a green and self-sustaining system that requires less energy from fixed sources, efficient EH methods and techniques need to be considered in SWIPT systems. Unlike traditional EH sources such as solar power, wind, and tide, the location of a transceiver has a great impact on the EH performance in SWIPT. A SWIPT-enabled transmitter can work in either a periodic manner or a time-varying manner. When most of the nodes in the system have a strong power level, SWIPT may be turned off by the system for reduced overhead. On the other hand, when most of the nodes in the system suffer from a lack of power, SWIPT may be turned on to power the nodes.

EH for SWIPT has been explored in opportunistic and cooperative ways. An optimal time switching rule for a point-to-point wireless link over the flat-fading channel subject to the time-varying co-channel interference was derived in [4], where the receiver was assumed to have no fixed power supplies, and thus needs to replenish energy from the unintended interference and/or the intended signal sent by the transmitter. Relay selection for achieving a trade-off between the efficiency of the information transmission to the receiver and the amount of energy transferred to the energy harvesters has been studied recently, and a relay selection policy that yields the optimal trade-off was proposed in [5]. The latest progress on efficient EH techniques in SWIPT implementation can be found in [6].

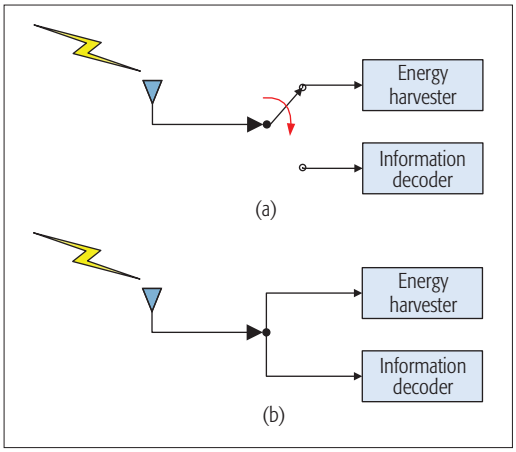


Figure 1. SWIPT architectures at the receiving antenna: a) time switching; b) power splitting..

Resource Allocation

Resource allocation in SWIPT systems primarily refers to the optimization of the utilization of various limited resources in the system, such as energy, bandwidth, time, and space. Of course, any required or predefined constraints with respect to relevant parameters must be satisfied.

Due to the dual identities of RF signals, transmitting information and power simultaneously calls for joint consideration of resource allocation with power control and user scheduling, which entails the following two things. First, opportunistic power control can be used to improve the energy and information transfer efficiency by exploiting the channel fading feature. Second, idle users who have high channel gains can be scheduled for power transfer to prolong the network lifetime. It has been discovered that with optimal power control consideration, both the system capacity and the harvested energy increase significantly, and the average harvested energy can be improved as well.

Moreover, resource allocation is an effective way to mitigate the interference in wireless systems. With SWIPT, harmful interference to a system can be turned into useful energy for the system. An interference-based resource allocation mechanism can gather the interference and direct it to specific energy-hungry devices, thereby enhancing the system's performance.

Signal Processing

A main concern for SWIPT is the growing decay in energy transfer efficiency, caused by the propagation of path loss when the transmission distance increases. Beamforming, as an advanced signal

Resource allocation in SWIPT systems refers to, primarily, the optimization of the utilization of various limited resources in the system, such as energy, bandwidth, time, and space. Of course, any required or predefined constraints with respect to relevant parameters must be satisfied.

SWIPT can be expected to provide an ultimate solution to the power issue of WSNs or the IoT, since it can prolong the network's lifetime without wired battery recharge or battery replacement due to the fact that sensors can harvest power through SWIPT as they communicate with one another.

processing technique, can be applied to SWIPT to improve its power transfer efficiency without additional bandwidth or increased transmitting power. Indeed, beamforming has been deemed a primary technique for feasible implementations of SWIPT.

In addition to beamforming, the transmitted power in a wireless power transfer system varies over time provided that the average power delivered to the receiver is above a certain required threshold. Because of this, information can be encoded in the energy signal by varying its power levels over time, thereby achieving continuous information transfer without degrading the power transfer efficiency. To emphasize this dual use of energy signal in both wireless power transfer and wireless information transfer, the resulting modulation scheme is called energy modulation.

A new general modulation mechanism that is influenced by the spatial modulation was proposed in [7]. This new modulation uses multiple antennas, is suitable for SWIPT, and, notably, carries information via energy patterns. With regard to pulse position modulation (PPM) and pulse amplitude modulation (PAM), the following two energy patterns can be conceived [8]:

(1) A PPM-resembling position-based energy pattern that is used in the spatial domain
(2) A PAM-resembling intensity-based energy pattern that completely depends on positive values

Potentially, some other modulation techniques may also be of possible inspirational value in devising additional energy patterns for energy-pattern-based transmissions in integrated receivers.

Wireless Systems with SWIPT

As SWIPT-enabled devices can function dually, that is, they can decode information and harvest energy from the same stream of signals simultaneously, the SWIPT technology can be applied to various wireless systems. In this section, typical wireless networks with SWIPT are summarized by organizing them into the following five types: wireless sensor networks (WSNs), relay networks, coordinated multipoint (CoMP) networks, collaborative mobile clouds (CMCs), and cognitive radio networks.

Wireless Sensor Networks

One of the essential features of the Internet of Things (IoT) is that all (small) physical objects around us will be connected via some WSN in which a small, inexpensive, and low-power sensor is attached to each physical object to collect information from the immediate surrounding environment and transmit it to other nodes in the network, thereby enabling future smart homes, smart cities, smart hospitals, and so forth. For a long time, the notable bottleneck for any WSN, which consists of a large number of sensors as the "skin" of the IoT structure, is the (short) period of time in which the network can be functional. Since sensors are typically powered by batteries, those batteries need to be recharged or replaced on a regular basis. Even worse, it is sometimes very difficult or impossible to replace sensor batteries [9].

SWIPT can be expected to provide an ultimate solution to the power issue of WSNs or the IoT, since it can prolong the network's lifetime without wired battery recharge or battery replacement due to the fact that sensors can harvest power through SWIPT as they communicate with one another.

Relay Networks

Relay networks are devised to improve the efficiency and reliability of signal/data transmissions by setting up intermediate network nodes and requiring them to relay signals/data in a cooperative manner to reduce the fading and attenuation of signals. Relay networks present a particularly appropriate scenario to which the SWIPT technology can be applied, in the sense that the relay network itself can benefit from the relayed and cooperative transmissions in terms of saving energy, and the harvested energy can be used to charge relay nodes as compensation for their role of data forwarding [10]. Generally speaking, current research on relay-assisted systems in the literature studies the following two scenarios: the SWIPT scenario and the multihop energy transfer scenario. The former refers to the scheme where the employed relay and the source terminal seek energy from each other's radiated signals, and the latter refers to the situation where energy needs to be transferred to some remote terminals through multiple relays.

Also, studies in the literature on SWIPT relay networks primarily focus on performance enhancement of the physical layer, the medium access control layer, and the network layer. For instance, there are issues related to relay operation, power allocation, and relay selection. With regard to relay selection, the mechanism of SWIPT in fact poses a new challenge for the design of relay networks, since the demanded relay for data transmission may not be the same relay that has the most powerful channel for EH. As such, a trade-off between the efficiencies of information transfer and energy transfer must be considered when devising relay selections.

CoMP Networks

At present, CoMP networks can be categorized into the following two types: joint transmission (JT), where multiple base stations serve an edge user simultaneously and the edge user's data is shared commonly, and coordinated beamforming (CB) where, in contrast to JT, an edge user is served only by one base station, and its data is locally owned. Pragmatically, CB-CoMP networks are more preferable than JT-CoMP networks since they require much less data exchanging overhead.

It is an emerging practice to investigate the benefits and challenges of integrating CoMP networks with the SWIPT technology, with the proven fact that full-scale integration/cooperation can reduce the total amount of required transmitting power in a CoMP network [11]. Note, however, that an enormous backhaul capacity will be needed for a CoMP network integrated with SWIPT if all base stations participate in energy transfer and information sharing with respect to all EH terminals and information-receiving terminals.

Collaborative Mobile Clouds

A CMC is a cooperative content distribution architecture in mobile computing where users share acquired multimedia information in a collaborative and complementary manner [12]. A CMC,

unlike conventional cloud computing or cloud radio access networks, constitutes a set of mobile terminals that, in a way similar to the peer-to-peer mechanism, completes the desired tasks distributively, cooperatively, and collaboratively.

A CMC constituted by a set of SWIPT mobile terminals can be expected to be more energy-efficient than a CMC of regular mobile terminals due to the fact that the SWIPT technology allows a terminal to harvest energy and receive information simultaneously. Furthermore, integrating SWIPT with CMCs not only can result in more efficient energy utilizations, but can restructure the formation of the CMC as well. Considering that the data transmission performed by a CMC member can cost that member/user/terminal a considerable amount of energy, the selfish nature of a user/terminal may consequently prevent it from being willing to join a CMC. Thus, applying the SWIPT technology to CMCs may act as an incentive for users to join CMCs, as doing so may benefit users with extra energy.

Cognitive Radio Networks

Cognitive radio networks refer to systems of spectrum sharing between primary users (PUs) and secondary users (SUs), where the PUs share their underutilized licensed spectrum with the SUs under the condition that the SUs' communications do not generate harmful interference to PUs' communications. Specifically, the following spectrum sharing methods are available: interweave, overlay, and underlay. Interweave spectrum sharing requires the SU transmit data over the silent intervals of PU's communications; overlay spectrum sharing enables the SU to exchange available spectrum with the PU; and underlay spectrum sharing allows the SU and the PU to coexist on the spectrum band by requesting the SU to restrict its transmitting power to a certain extent.

Because of the ways in which cognitive radio networks and the SWIPT technology work, making a cognitive radio network SWIPT can be expected to enhance the spectrum utilization and the EH efficiency [8]. To see this, consider a SWIPT-enabled PU network that shares its licensed spectrum with an SU network. While the spectrum utilization efficiency in the PU network can be improved, the energy transfer efficiency in the PU network can be augmented due to the extra energy source from the SU.

Power Allocation for D2D with SWIPT

The most challenging issue in successfully implementing D2D communications underlaying cellular networks lies in the mutual interference between D2D and cellular communications.

Power allocation plays a vital role in addressing such an issue. While numerous research studies have been conducted on D2D power allocation, most of them have assumed that the interference is harmful, and little work has been done for D2D power allocation in the presence of SWIPT. In this section, we address the D2D power allocation problem with SWIPT by using game theory.

System Model

Consider a single cell in cellular networks that D2D communications underlay. Within the cell, there are one base station (BS), one cellular user, and n D2D links. As D2D communications are implemented as an underlay to cellular communications, D2D links nonorthogonally reuse cellular uplink resources, and thus they will cause interference to cellular users and among themselves as well.

Each D2D user is installed with a SWIPT power splitting unit, which allows simultaneous information and power transfer. Hence, the D2D device is able to harvest energy from signals, interference, and noise. As noted earlier, a D2D pair will be interfered not only by the cellular user, but also by the other D2D links; that is, a D2D receiver can scavenge energy from the associated D2D transmitter and cellular users as well as other D2D links.

It is true that higher transmitting power of a D2D link will lead to an increased signal-to-interference-plus-noise ratio (SINR), but this increase will certainly cause stronger interference to other D2D links. Meanwhile, increased power will result in more energy harvested by D2D receivers and more power consumption. The objective is to determine the appropriate amount of power for all D2D pairs in the presence of SWIPT.

Utility Formulation

Game theory is leveraged to model the power allocation among D2D communication transmitters. Specifically, we define the utility of the D2D link i as

$$U_i = \omega_1 \cdot SINR_i + \omega_2 \cdot E_i - \omega_3 \cdot F_i, \quad (1)$$

where E_i is the total amount of energy harvested by D2D link i, F_i signifies the negative impacts of link i including the interference (caused by i) to other D2D links and its own power consumption, and ω_1, ω_2, and ω_3 are (positive) weights reflecting the importance of each term, and $\sum_{i=1}^{3} \omega_i = 1$.

Note that the power of a D2D transmitter, p_i, should not be so high that it would consume more power and generate harmful interference to other D2D links. Figure 2 illustrates the correlation between U_i and p_i, and between $SINR_i$ and p_i. As can be seen, the utility would not increase unlimitedly with the increase of power, which shows the rationality of the formulation of utility.

Distributed Power Allocation Mechanism

To solve the power allocation problem, we assume that any D2D link only knows its own power and link status and has no way of knowing those of any other D2D links. However, the interference and the harvested energy from the interference are known by this link, and U_i is concave with respect to p_i based on the previous discussion. Here, we give an iterative algorithm for computing the Nash equilibrium where the computation of the Nash-equilibrium-leading power $p_i[t+1]$ for link i at the $(t+1)$th iteration amounts to finding the optimal power for link i. That is,

$$p_i[t+1] = \arg \max_{p_i[t] \in [0,\infty]} \{(U_i[t])\}, \quad (2)$$

where the right side can be obtained by solving the equation

$$\frac{\partial U_i[t]}{\partial p_i[t]} = 0.$$

The major steps of the algorithm for computing the desired Nash equilibrium power p_i^* for each

CMC, unlike the conventional cloud computing or cloud radio access networks, constitutes a set of mobile terminals which, in a way similar to the peer-to-peer mechanism, completes the desired tasks distributively, cooperatively, and collaboratively.

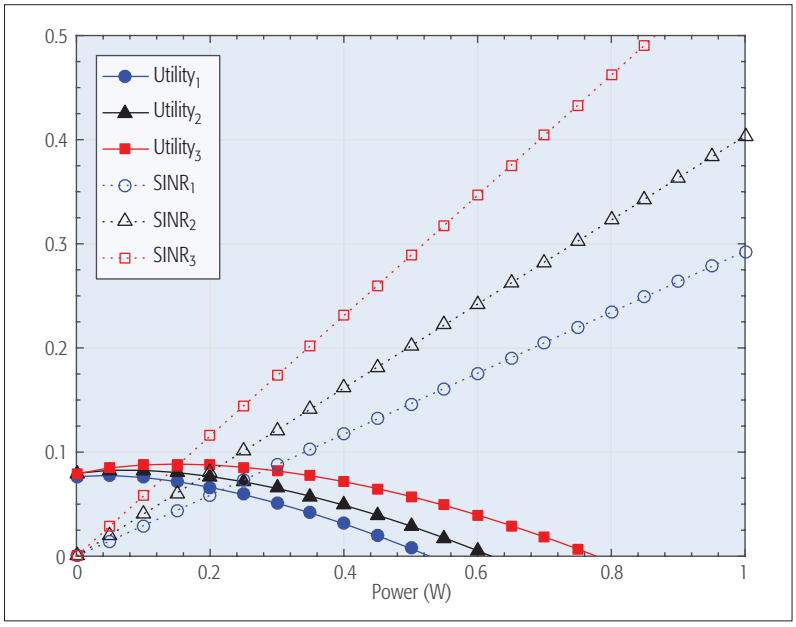


Figure 2. Correlations between U_i and p_i, and $SINR_i$ and p.

D2D link i are shown as follows:

- Each D2D link takes an initial value for its power $p_i[0]$, where $i = 1, 2, ..., n$.
- Each D2D link computes its subsequent powers by Eq. 2, and updates its power value. We use $p_i[t]$ to denote the tth updated power value for link i.
- If $|U_i[t] - U_i[t-1]| \le \varepsilon$ for a predefined threshold ε at the tth round; then the computation terminates and $p_i^* = p_i[t]$; otherwise, the computation continues by starting the next round.

It can be proven that this approach converges to the Nash equilibrium strategy. The detailed proof is omitted here due to space limitation.

Simulations

To validate the proposed power allocation mechanism, comparisons for the mechanism with/without SWIPT and node mobility are thoroughly investigated in this section. We assume that three D2D links reuse the same spectrum with one cellular user in the cell. $p_i \in [0, 2]$ W, and the *Rician* fading model is employed, the weights for each terms of Eq. 1 are set to be the same (i.e., 1/3), meaning that they are equally important, all the experiments are run in MATLAB, and the simulation results are statistically collected by averaging 1000 trials.

With vs. Without SWIPT: Comparisons are performed between the power allocation with SWIPT and that without SWIPT. For the latter case, the utility function of each D2D link i is defined as $U_i = SINR_i - F_i$, where

$$\frac{(1-\delta)p_i h_{ii}}{(1-\theta)(N_0+I)+(1-\eta)\sum_{j\neq i} p_j h_{ji}}$$

and $F_i = p_i^{1.5}$ [13]. It is obvious to see that this definition has almost the same form as the previous utility except for the EH term

$$E_i = \delta \cdot p_i h_{ii} + \eta \cdot \sum_{j\neq i} p_j h_{ji} + \theta \cdot (N_0 + I),$$

where δ, η, and θ are the power conversion efficiencies, which are all set to 0.2. Moreover, $N_0 + I$ is set to 0.4 throughout all the simulations.

The comparison results in terms of power and utility are shown in Fig. 3. As expected, we can see from Fig. 3a that the power of each D2D link with SWIPT is at least 11 percent greater than that without SWIPT. In parallel, the utility of each D2D link with SWIPT is higher than that without SWIPT, as evidenced by Fig. 3b. Thus, we claim that SWIPT plays an important role in improving the system performance, and thus our proposed power allocation mechanism is of practical significance.

With vs. Without Mobility: Using the same parameter settings used in the previous experiments, comparisons are also conducted between power (and utility, respectively) with node mobility and without mobility. The error scaling factors for

$$J[t-1] = \sum_{j\neq i} p_j[t-1]h_{ji}$$

are employed to reflect the mobility. Specifically, the perceived interference from other D2D links at the tth iteration is assumed to be a random variable that is evenly distributed in the interval from $0.8 \cdot J[t-1]$ to $1.3 \cdot J[t-1]$ instead of exact $J[t-1]$ due to mobility.

The results are shown in Fig. 4, where for the case with mobility, the mean value of 1000 trials and standard deviation at each iteration are given. As can be observed from this figure, the gap between results of without mobility and with mobility is small, indicating that mobility may have little impact on the algorithm if it is guaranteed to converge. In addition, we can see that the algorithm can converge very fast even with mobility, and it has the same convergence speed as in the case without mobility. This insight further validates the correctness of the proposed mechanism.

Conclusion and Future Research Directions

In this article, with respect to the emerging complex mechanism of SWIPT, we have surveyed its current architectures and enabling technologies, and have identified some technical challenges in implementing this mechanism. Following an overview of enabling technologies for SWIPT and SWIPT-enabled wireless systems, we have showcased a novel SWIPT-supported power allocation mechanism for D2D communications to illustrate the importance of the application of SWIPT. Toward encouraging more research efforts on SWIPT to forthcome, we point out the following issues that may be worth investigations as future research topics on SWIPT.

Mobility of SWIPT: While the notion that mobility would be a desired feature of any SWIPT systems, as is readily seen by considering the fact that information transmission and EH as well as network status all have a dynamic and time-varying nature, note that the mobility itself will largely affect the availability of channel state information of the network, and therefore it is not an easy task to obtain the precise channel state information in SWIPT systems. Given this situation, devising robust beamformers to cope with the mobility issue in systems with SWIPT is a both needed and challenging task.

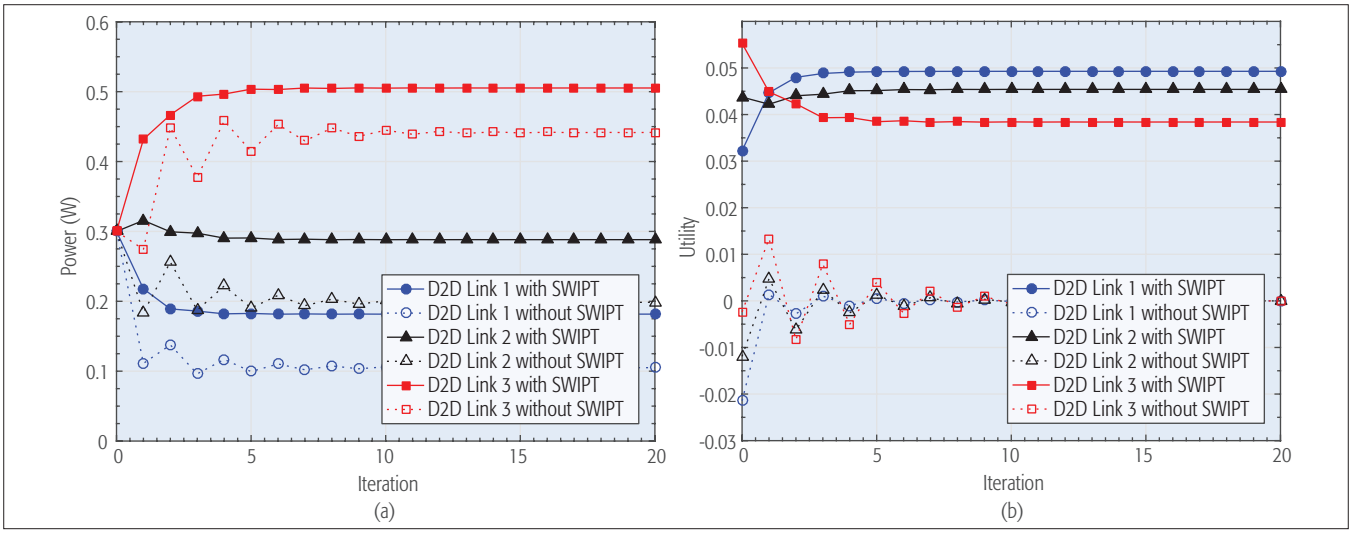


Figure 3. Performance comparisons with/without SWIPT: a) power comparison; b) utility comparison.

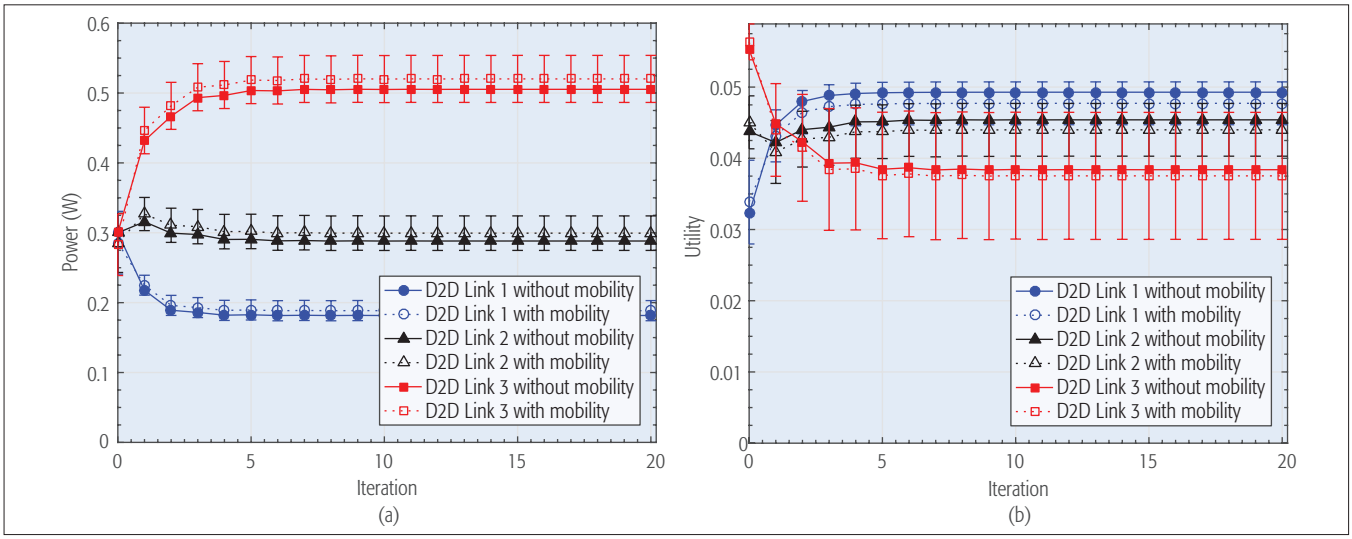


Figure 4. Performance comparisons with/without mobility; a) power comparison; b) utility comparison.

Security of SWIPT: The dual effects of increasing the transmitting power of signals — doing so not only can enhance the desired energy transfer from a transmitter (Alice) to a legitimate receiver (Bob), but can escalate the undesired risk of information stealth by an eavesdropper (Eve) as well — present a security concern in SWIPT systems. Noticing the recognized effectiveness of physical layer security measures, which essentially work by optimizing the secrecy rate, it would be a release to the aforementioned security concern if we are able to find an effective way to increase the signal strength on the legitimate channel and reduce the signal strength on the wiretap channel at the same time.

SWIPT in Multihop Networks: The trade-off issue between the efficiencies of information transmission and energy transfer in SWIPT is manifested when SWIPT is integrated into a multihop relay network, since a relay node that is most suitable for information transmission may not be most suitable for energy transfer. Hence, how to appropriately select relay nodes in terms of the efficiencies of information transmission and energy transfer needs to be further investigated. Considering that network coding is known for its capability to increase the amount of transmitted information per time slot and for allowing receivers to receive messages simultaneously, which now means an increased amount of surrounding energy, it would be an interesting approach to tie network coding, SWIPT, and multihop networks together in this context.

Acknowledgment

This research is supported by NSFC under grant number 61671093.

References

[1] R. Zhang and C. K. Ho, "MIMO Broadcasting for Simultaneous Wireless Information and Power Transfer," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, May 2013, pp. 1989–2001.
[2] X. Zhou, R. Zhang, and C. K. Ho, "Wireless Information and Power Transfer: Architecture Design and Rate-Energy Tradeoff," *Proc. 2012 IEEE GLOBECOM*, Dec. 2012, pp. 3982–87.
[3] Z. Ding *et al.*, "Application of Smart Antenna Technologies In Simultaneous Wireless Information and Power Transfer," *IEEE Commun. Mag.*, vol. 53, no. 4, April 2015, pp. 86–93.
[4] L. Liu, R. Zhang, and K. C. Chua, "Wireless Information Transfer with Opportunistic Energy Harvesting," *Proc. 2012 IEEE Int'l. Symp. Info. Theory*, July 2012, pp. 950–54.

[5] D. S. Michalopoulos, H. A. Suraweera, and R. Schober, "Relay Selection for Simultaneous Information Transmission and Wireless Energy Transfer: A Tradeoff Perspective," *IEEE JSAC*, vol. 33, no. 8, Aug 2015, pp. 1578–94.
[6] X. Lu *et al.*, "Wireless Charging Technologies: Fundamentals, Standards, and Network Applications," *IEEE Commun. Surveys & Tutorials*, vol. 18, no. 2, 2nd qtr. 2016, pp. 1413–52.
[7] R. Zhang, L. L. Yang, and L. Hanzo, "Energy Pattern Aided Simultaneous Wireless Information and Power Transfer," *IEEE JSAC*, vol. 33, no. 8, Aug. 2015, pp. 1492–1504.
[8] S. Bi, Y. Zeng, and R. Zhang, "Wireless Powered Communication Networks: An Overview," *IEEE Wireless Commun.*, vol. 23, no. 2, Apr. 2016, pp. 10–18.
[9] K. Huang, C. Zhong, and G. Zhu, "Some New Research Trends in Wirelessly Powered Communications," *IEEE Wireless Commun.*, vol. 23, no. 2, Apr. 2016, pp. 19–27.
[10] S. Guo *et al.*, "Energy-Efficient Cooperative Transmission for Simultaneous Wireless Information and Power Transfer in Clustered Wireless Sensor Networks," *IEEE Trans. Commun.*, vol. 63, no. 11, Nov. 2015, pp. 4405–17.
[11] W. N. S. F. W. Ariffin, X. Zhang, and M. R. Nakhai, "Sparse Beamforming for Real-Time Energy Trading in Comp-Swipt Networks," *Proc. 2016 IEEE ICC*, May 2016, pp. 1–6.
[12] Z. Chang *et al.*, "Energy Efficient Resource Allocation and User Scheduling for Collaborative Mobile Clouds with Hybrid Receivers," *IEEE Trans. Vehic. Tech.*, vol. PP, no. 99, 2016, pp. 1–1.
[13] Q. Wang, M. Hempstead, and W. Yang, "A Realistic Power Consumption Model for Wireless Sensor Network Devices," *Proc. 2006 3rd Annual IEEE Commun. Society Sensor and Ad Hoc Commun. and Networks*, vol. 1, Sept 2006, pp. 286–95.

Biographies

Jun Huang [M'12, SM'16] (xiaoniuadmin@gmail.com) is currently a full professor at the Institute of Electronic Information and Networking, Chongqing University of Posts and Telecommunications. He has authored 80+ publications, and several of them are in prestigious journals and conference proceedings. His research interests include IoT and D2D/M2M.

Cong-cong Xing (cong-cong.xing@nicholls.edu) is currently a full professor of computer science at Nicholls State University, Thibodaux, Louisiana. He received his Ph.D. in computer science and engineering from Tulane University, New Orleans, Louisiana, joining the Nicholls State University faculty in 2001. His research interests include theoretical foundations of programming languages, category theory, and computer networking analysis. He is active in research in these areas.

Chonggang Wang [F'16] (cgwang@ieee.org) received his Ph.D. degree from Beijing University of Posts and Telecommunications, China, in 2002. He is currently a member of technical staff at InterDigital Communications, with a focus on Internet of Things R&D activities. He is the founding Editor-in-Chief of the *IEEE Internet of Things Journal*, and is on the Editorial Boards of several journals. He is an IEEE ComSoc Distinguished Lecturer (2015-2016).

Green Heterogeneous Cloud Radio Access Networks: Potential Techniques, Performance Trade-offs, and Challenges

Yuzhou Li, Tao Jiang, Kai Luo, and Shiwen Mao

ABSTRACT

As a flexible and scalable architecture, heterogeneous cloud radio access networks (H-CRANs) inject strong vigor into the green evolution of current wireless networks. But the brutal truth is that EE improves at the cost of other indices such as SE, fairness, and delay. It is thus important to investigate performance trade-offs for striking flexible balances between energy-efficient transmission and excellent QoS guarantees under this new architecture. In this article, we first propose some potential techniques to energy-efficiently operate H-CRANs by exploiting their features. We then elaborate the initial ideas of modeling three fundamental trade-offs, namely EE-SE, EE-fairness, and EE-delay trade-offs, when applying these green techniques, and present open issues and challenges for future investigation. These related results are expected to shed light on green operation of H-CRANs from adaptive resource allocation, intelligent network control, and scalable network planning.

INTRODUCTION

BACKGROUND AND MOTIVATION

The dramatic increase in the number of smartphones and tablets with ubiquitous broadband connectivity has triggered an explosive growth in mobile data traffic [1]. Cisco forecasts that the amount of global mobile data traffic will increase 7-fold from 2016 to 2021, the majority of which are generated by energy-hungry applications such as mobile video [1]. This is also referred to as the well-known 1000× data challenge in cellular networks. Meanwhile, the number of devices connected to the global mobile communication networks will reach 100 billion in the future, and that of mobile terminals will surpass 10 billion by 2020 [2].

Although unprecedented opportunities for the development of wireless networks are created by the massive traffic amount and connected devices, a concomitant crux is that this growth skyrockets the energy consumption (EC) and greenhouse gas emissions in the meantime. From statistical data, the information and communication technology (ICT) industry is responsible for 2 percent of worldwide CO_2 emissions and 2–10 percent of global EC, of which more than 60 percent is directly attributed to radio access networks (RANs) [3]. In this regard, 5G wireless communication networks are anticipated to provide spectral efficiency (SE) and energy efficiency (EE) growth by a factor of at least 10 and 10 times longer battery life of connected devices [2].

CONCEPT OF H-CRANS

To meet the 1000× data challenge, heterogeneous networks (HetNets), composed of a diverse set of small cells (e.g., microcells, picocells, and femtocells) overlaying the conventional macrocells, have been introduced as one of the most promising solutions [2]. However, the ubiquitous deployment of HetNets is accompanied by the following shackles:

- Severe interference: The spectrum reuse among cells incurs severe mutual interference, which may significantly reduce the expected system SE and also decrease the network EE.
- Unsatisfactory EE: The densely deployed small cells lead to escalated EC and thus reduced EE, and also increases capital expenditures (CAPEX) and operational expenditures (OPEX).
- No computing-enhanced coordination centers: There are no centralized units with strong computing abilities to globally coordinate multi-tier interference and execute cross-RAN optimization, which dramatically limits cooperative gains among cells.
- Inflexibility and unscalability: Fragmented base stations (BSs) result in inflexible and unscalable network control and operations, thus leading to redundant network planning and inconvenient network upgrade.

To overcome these challenges faced by HetNets, cloud RANs (C-RANs), new centralized cellular architectures armed with powerful cloud computing and virtualization techniques, have been put forward in parallel to coordinate interference and manage resources across cells and RANs [4]. In C-RANs, a large number of low-cost low-power remote radio heads (RRHs), connecting to the baseband unit (BBU) pool through the fronthaul links, are randomly deployed to enhance the wireless capacity in hotspots. Consequently, the combination of HetNets and C-RANs, known

The authors first propose some potential techniques to energy-efficiently operate H-CRANs by exploiting their features. They then elaborate the initial ideas of modeling three fundamental trade-offs, namely EE-SE, EE-fairness, and EE-delay trade-offs, when applying these green techniques, and present open issues and challenges for future investigation.

Yuzhou Li, Tao Jiang, and Kai Luo are with Huazhong University of Science and Technology; Shiwen Mao is with Auburn University.

Digital Object Identifier: 10.1109/MCOM.2017.1600807

0163-6804/17/$25.00 © 2017 IEEE

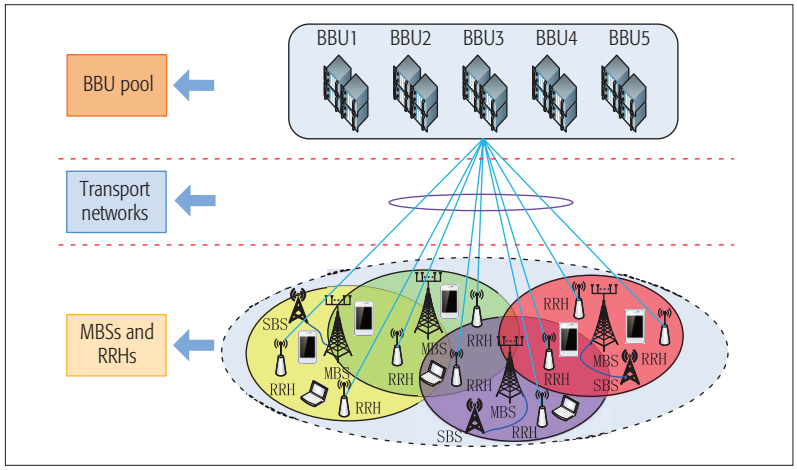


Figure 1. The architecture of H-CRANs.

as heterogeneous C-RANs (H-CRANs), becomes a potential solution to support both spectral- and energy-efficient transmission.

Green H-CRANs

As mentioned above, one of the main missions of H-CRANs from their birth is to construct eco-friendly and cost-efficient wireless communication systems. Benefiting from H-CRANs' global coordination ability, many promising techniques, such as joint processing/allocation, traffic load offloading, energy balance, self-organization, and adaptive network deployment, can be applied in these scenarios for energy-efficient transmissions. Unfortunately, the network EE improves usually at the cost of the performance of other technique metrics, such as SE, fairness, and delay, all of which, however, are equally important as EE to guarantee users' quality of service (QoS). That is, there are EE-SE, EE-fairness, and EE-delay trade-offs. It is thus interesting to investigate these performance trade-offs in H-CRANs for establishing rules to flexibly balance the network EE and users' QoS demands when greening H-CRANs.

Compared to existing works (e.g., [5]) on the system architecture or radio resource management (RRM), mainly in terms of EE and SE, this article focuses on the green evolution of H-CRANs, and particularly investigates it from the perspective of EE-SE, EE-fairness, and EE-delay trade-offs instead of the indices themselves. To reach our targets, we organize the remainder of this article as follows. In the following section, we first simply review the architecture of H-CRANs and then exploit their features to propose three potential techniques for green H-CRANs. Then we introduce the possible methods to depict these trade-offs and also provide corresponding challenges and open problems when applying these proposed techniques. We conclude the article in the final section.

Architecture of H-CRANs and Potential Green Techniques

In C-RANs, the idea of dividing conventional cellular BSs into two parts, BBUs and RRHs, is introduced. BBUs are then integrated into centralized BBU pools, where cloud computing and virtualization techniques are implemented to enhance computational ability and to virtualize network function. BBUs are responsible for resource control and signal processing, and RRHs for information radiation and reception, with their interconnection via dedicated transport networks. Thus, the cloud-computing-enhanced centralized BBU pools facilitate cross-cell and cross-RAN information sharing, which paves the path for global resource optimization adapting to network conditions (e.g., channel conditions, interference strength, traffic loads). H-CRANs absorb this architecture in C-RANs and maintain macro BSs (MBSs) and small cell BSs (SBSs) in HetNets to support both global control and seamless communications.

Architecture of H-CRANs

As shown in Fig. 1, H-CRANs are composed of three functional modules.

Real-Time Virtualized and Cloud-Enhanced BBU Pool: Equipped with powerful virtualization techniques and strong real-time cloud computing ability, BBU pools integrate independent BBUs scattered in cells.

High-Reliability Transport Networks: RRHs are connected to BBUs in the BBU pool via high-bandwidth low-latency fronthaul links such as optical transport networks. The data and control interfaces between the BBU pool and MBSs are S1 and X2, respectively [6].

MBSs, SBSs, and RRHs: In H-CRANs, multiple access points (APs), for example, MBSs, SBSs, and RRHs, coexist. MBSs are deployed mainly for network control and mobility performance improvement, for example, decreasing handover times to avoid ping-pong effects for high-mobility users. SBSs and RRHs are geographically distributed within cells close to users to increase capacity and decrease transmit power in the meantime.

In H-CRANs, the function separation between BBUs and RRHs, the decoupling between control and data planes, and the cloud-computing-enhanced centralized integration of BBUs facilitate efficient management of densely deployed mobile networks. For example, the operators only need to install new RRHs and connect them to the BBU pool to expand network coverage and improve network capacity. Moreover, flexible software solutions can easily be implemented under this architecture. For instance, operators can upgrade RANs and support multi-standard operations only through software update by deploying software defined radio (SDR).

Potential Techniques for Green H-CRANs

The four revolutionary changes, that is, function separation, control-data decoupling, centralized architecture, and cloud-computing-enhanced processing, make H-CRANs significantly different from existing second generation (2G), 3G, and 4G wireless networks. By exploiting these features, it is possible to construct H-CRANs that are flexible in network management, adaptive in network control, and scalable in network planning. As a result, energy-efficient operation of H-CRANs without significant loss in other indices such as SE, fairness, and delay can be achieved.

Joint Resource Optimization across RRHs and RANs: In H-CRANs, each BBU first collects its individual network conditions and then shares this information within the BBU pool. As a result, this distributed-collection centralized-control archi-

tecture, further enhanced by virtualization techniques and cloud computing, enables efficient transmission/reception cooperation across RRHs and convenient global control across RANs. Consequently, the existing cooperative techniques, such as coordinated multipoint (CoMP) transmission, enhanced inter-cell interference coordination (eICIC), and interference alignment (IA), can readily be implemented in H-CRANs. All these techniques are self-contained in theory but have rarely been applied to conventional cellular networks because of difficulties in sharing and handling global network information.

As introduced above, multi-RANs and multi-APs with different coverage and functions are deployed in H-CRANs. As a result, unlike traditional single-mode terminals communicating only through a RAN's AP, multi-mode terminals could send and receive data concurrently through multiples of them. This indicates H-CRANs with a new characteristic of network diversity, which can be exploited to design user association strategies. By this, traffic load distributions among RANs and APs can be well balanced, which in turn affects the working states of RANs and resource optimization, and thus affects network interference and EE.

Moreover, under this new centralized architecture, the network EE can be further improved by incorporating more resource allocation dimensions (e.g., power allocation, subcarrier assignment, user association, and RRH operation) into the formulations. Figure 2 shows that joint optimization of RRH operation and power allocation improves EE by up to 84 percent compared to the power-allocation-only algorithm in downlink H-CRANs. Thus, through the aforementioned joint resource optimization and network-diversity-aware user association, significant improvement in EE and reduction in EC can be achieved.

Large-Scale MBS and SBS Deployment: Compared to the transmit power, the overall static power consumption by MBSs and SBSs, composed of cooling and circuit power, are usually much larger [7]. For example, a typical Universal Mobile Telecommunications System (UMTS) BS consumes 800–1500 W with RF output power of 20–40 W. As a result, under the constraints of basic coverage requirements, the deployment of MBSs and SBSs, characterized by the distance between two MBS sites and the number of SBSs per site, affects the area power consumption (APC) and the area SE (ASE) significantly in H-CRANs. The general purpose of large-scale MBS and SBS deployment is to macroscopically plan an appropriate number of BSs to support users' demands for energy saving by avoiding the static power consumption.

Intuitively, the APC will sharply decrease if we reduce the number of MBSs (i.e., increase the inter-site distance). Meanwhile, the ASE will also decrease, because the increased inter-site distance reduces the spectrum reuse. Similarly, the number of SBSs deployed in each site will also affect the APC and the ASE. As an example, Fig. 3 clearly shows the significant impacts of the configuration of MBSs and SBSs on the APC and ASE under practical parameter settings. Therefore, we need careful network planning from a large-scale perspective to flexibly balance these two metrics and to conveniently upgrade the system.

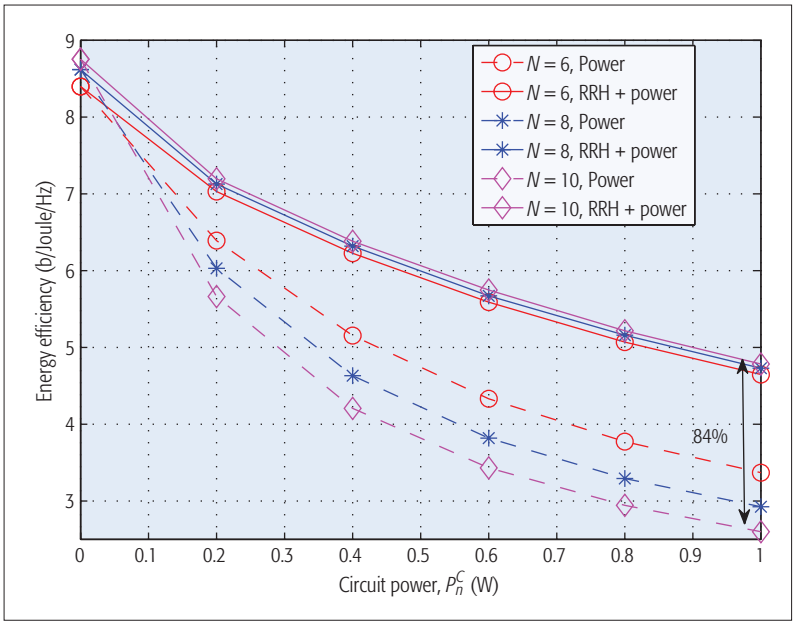


Figure 2. An example: EE variation with the circuit power of each RRH, denoted by P_n^c, in downlink H-CRANs, where an MBS, N RRHs, and 16 users are included. In this example, we maximize the network EE by optimizing RRH operation and power allocation subject to constraints of users' minimum rate requirements of R^{req} = 2 b/Hz.

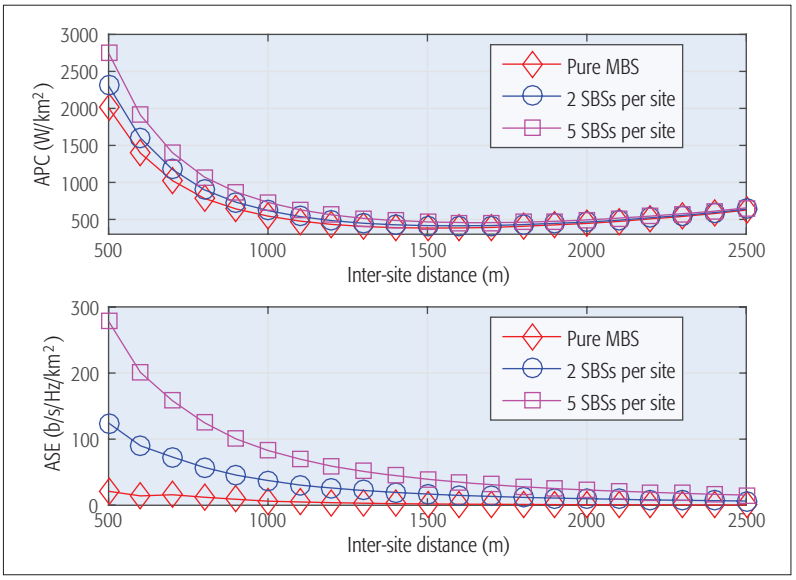


Figure 3. An example: The APC and ASE vs. the inter-site distance subject to a 95 percent coverage constraint. In the figure, we adopt the practical models for the BS power consumption given by $P^{tot} = aP^{tx} + b$, where a_{MBS} = 22.6, b_{MBS} = 412.4 W, a_{SBS} = 5.5, and b_{SBS} = 32 W (note that SBSs refer to micro BSs in the figure) [15].

Load-Aware RRH Operations: The so-called worst case network planning philosophy has been widely adopted to guarantee users' QoS even during peak traffic periods in conventional cellular networks. However, mobile traffic loads usually vary in both spatial and temporal domains, which is referred to as the tidal phenomenon. Specifically, the fraction of time when the traffic is below 10 percent of the peak during a day is about 30 percent on weekdays and 45 percent on week-

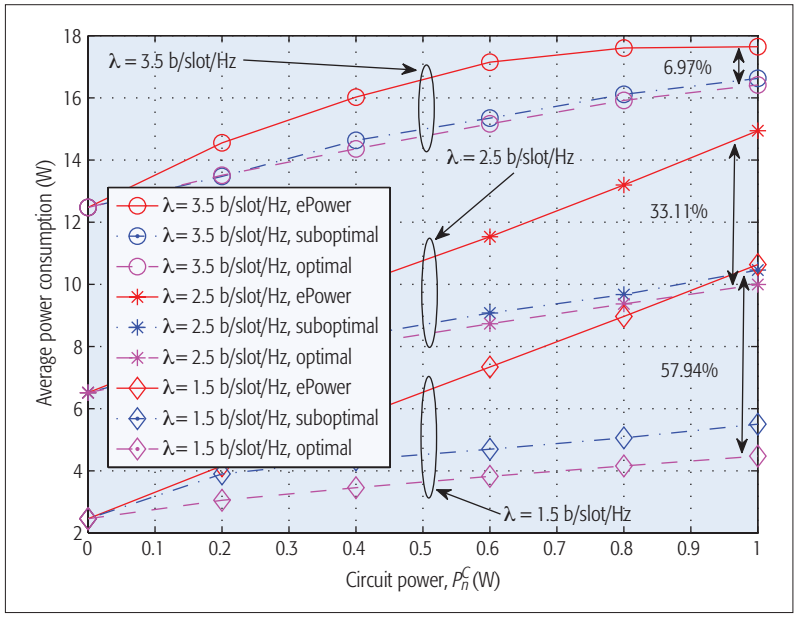


Figure 4. An example: average power consumption with the circuit power of each RRH, denoted by P_n^C, under different traffic arrival rates λ in downlink H-CRANs, where a MBS, 8 RRHs, and 12 users are included. In this example, we formulate a network EE maximization problem that enfolds stochas tic and time-varying traffic arrivals to jointly optimize RRH operation and power allocation.

ends [8]. As a result, a large number of RRHs are extremely underutilized in cases of dense deployment in H-CRANs during off-peak periods, but RRHs still consume circuit power even with little or no activity. Consequently, a significant waste of EC and a sharp decrease in EE will be caused if RRHs are underutilized but still activated. Thus, apart from the aforementioned spatial deployment, energy conservation can also be achieved by exploiting temporal traffic variations. For the fixed deployment, we can adopt load-aware network control in H-CRANs to perform on/off operations of RRHs adapting to spatial and temporal traffic amounts to improve EE.

As an example, we consider a downlink H-CRAN to show the impacts of load-aware RRH on/off operations on energy expenditure. Specifically, we jointly optimize RRH operation and power allocation to maximize the network EE with stochastic and time-varying traffic arrivals taken into account. Two algorithms, optimal and suboptimal, are developed to solve the problem. Figure 4 shows that the proposed algorithms can dramatically reduce the energy consumption compared to the algorithm without RRH operation (i.e., only optimizing power allocation), denoted by ePower, especially in light and middle traffic states (up to a 58 percent gain in light traffic states when the traffic arrival rate λ = 1.5 b/slot/Hz).

Performance Trade-offs and Challenges for Green H-CRANs

Leveraging the proposed potential green techniques in H-CRANs, it is then of importance to explore the key theories that support ubiquitously energy-efficient transmission and meanwhile provide satisfactory QoS for users. Among them, performance trade-offs deserve significant consideration [9].

Apart from the widely studied deployment efficiency-EE, EE-SE, bandwidth-power, and delay-power trade-offs [9], there are two additional fundamental trade-offs: EE-fairness and EE-delay trade-offs. This section elaborates the ideas of modeling these two trade-offs, analyzes challenges and open problems, and provides some possible solutions. Since H-CRANs originally are designed to enhance the network SE and thus the wireless capacity as well, we also review the key concepts and present challenges associated with the EE-SE trade-off under this new architecture.

EE-SE Trade-off

Vast existing research falls into this area due to the following reasons. The traditional indices EC and SE measure how small the amount of energy is needed to satisfy users' QoS and how efficiently limited spectrum is utilized, respectively. However, both of them fail to quantify how efficiently the energy is consumed (i.e.,, EE). Moreover, the optimality of EE and EC and that of EE and SE are not always achieved simultaneously and may even conflict with each other [9]. As a consequence, the existing results from the EC minimization or the SE maximization usually can hardly provide insights into EE-SE trade-off problems.

The general idea of modeling the EE-SE trade-off is that the system maximizes the network EE [10] or a weighted EE-SE trade-off index [11] under the constraints of users' QoS and resource allocation (e.g., power allocation and RRH operation). As a common feature, these works usually assume infinite backlog, that is, there is always data for transmission in the buffer. Under this view, formulations are presented and algorithms are developed only based on the observation time, where the network EE is defined as the ratio of the instantaneous achievable sum rate R_{tot} to the corresponding total power consumption P_{tot} [10, Eq. 5]. Note that P_{tot} is usually modeled to include both transmit and circuit energy consumption, which is affected by the power amplifier inefficiency, transmit power, and circuit power. In this article, we call these formulations short-term (i.e., snapshot-based) models, since only short-term system performance is considered. Accordingly, we denote the network EE of this kind of definition by $EE_{short\text{-}term}$ for simplicity.

Although there have been a large number of works addressing the EE-SE trade-off based on the short-term models, lots of problems remain open in complex H-CRANs. First, jointly considering multi-dimensional resource optimization and multi-available signal processing techniques, it is challenging to formulate EE-SE trade-off problems with network conditions and users' requirements both taken into account in H-CRANs. Furthermore, due to the nonconvexity of $EE_{short\text{-}term}$ [10, Eq. 5; 11, Eq. 26], EE-SE trade-off problems are usually difficult to solve even if we only optimize power allocation in spectrum-sharing H-CRANs. As a result, these problems become much more complicated once we extend from one-dimensional to multi-dimensional resource optimization. Thus, how to develop joint resource allocation algorithms that reach the theoretical limits of the network EE and thus serve as benchmarks to evaluate performance of other heuristic algorithms is

another challenge. Moreover, it is also necessary to develop cost-efficient and easy-to-implement algorithms with acceptable performance levels to solve these problems for practical applications.

EE-Fairness Trade-Off

The widely studied EE-optimal problems (NEPs) in H-CRANs emphasize the network EE maximization without considering EE fairness (i.e., ignoring the EE of individual links). By purely benefiting the links in good network conditions (e.g., excellent wireless channel, little interference, low traffic loads, or all), the NEPs improve the network EE at the cost of the EE of the links in poor conditions. As a result, the NEPs would inevitably lead to severe unfairness among links in terms of EE. However, as traditional concerns on individual links' SE or EC, it is also important to guarantee the EE of each link in the users' perception. It is therefore of interest to investigate the EE-fairness trade-off in H-CRANs, but to the best of our knowledge, studies on this issue have so far been very scarce.

To intuitively show the EE-fairness trade-off, we take the max-min EE fairness in an uplink orthogonal frequency-division multiple access (OFDMA)-based cellular network (it can be seen as a special case of single-cell H-CRANs) as an example. Specifically, we maximize the EE of the worst case link subject to subcarrier assignment and power allocation constraints to ensure the max-min EE fairness among links, which is referred to as the max-min EE-optimal problem (MEP). In Fig. 5, we compare the statistical performance between the NEP and the MEP from three aspects: the EE of the network, the best link, and the worst link. Observe that the EE of the best and worst links in the NEP differs significantly, while the EE, whether of the network, the best link, or the worst link in the MEP, is well balanced. This is because the NEP maximizes the network EE at the cost of the EE fairness among links, but reversely, the MEP sacrifices the network EE to guarantee the max-min EE fairness.

Figure 5 exhibits the phenomenon of the EE-fairness trade-off, but we are still at a very primary stage of revealing and tuning this trade-off, limited by the following two challenges:

- Unified frameworks to quantify and formulate the EE-fairness trade-off are currently not available.
- General techniques or analytical methods to tackle the EE-fairness trade-off problems are still open.

It should be pointed out that the utility theory, originally used to investigate the rate-fairness trade-off [12], is a possible method to demystify the quantitative EE-fairness trade-off.

EE-Delay Trade-off

As far as we know, the concept of the EE-delay trade-off was first proposed by H. V. Poor *et al.* in 2009 [13], where the authors showed that the delay constraints would lead to a loss in EE at equilibrium by a game-theoretical approach. However, to date, how to quantify and control the EE-delay trade-off is still unresolved.

In our view, one possible reason that prevents the existing works including [13] from obtaining a quantitative trade-off is the choice of adopting short-term models with the full buffer assumption, where $EE_{short\text{-}term}$ is used to characterize the network EE. However, different from the full buffer assumption, practical H-CRANs operate in the presence of time-varying wireless channels and stochastic traffic arrivals, both of which significantly affect the EE and delay, and thus the EE-delay trade-off. Hence, short-term formulations in general cannot reflect the delay due to their independence of time and without considering traffic arrivals. As a result, it is unlikely for such models to show the explicit EE-delay relationships.

We further illustrate the principles behind the EE-delay trade-off with two extreme cases. Regarding stochastic traffic arrivals, in the case of aggressive emphasis on the EE, transmission decisions should be triggered only when network conditions are good enough, by which the delay performance degrades inevitably. Alternatively, to ensure small delay, the network has to transmit data at the cost of energy expenditure even when network conditions are very poor, which undoubtedly decreases the EE. Thus, to model the EE-delay trade-off, the following two issues need to be considered:

- How to decide whether to transmit data or defer a transmission in each slot in terms of the EE and delay and how to optimize resource allocation such as power allocation, subcarrier assignment, and RRH operation if transmission is chosen
- How to ensure that deferring transmissions to anticipate more advantageous network conditions becoming available in the future would not result in an uncontrollable delay because of time-variant, stochastic, and unpredicted network conditions

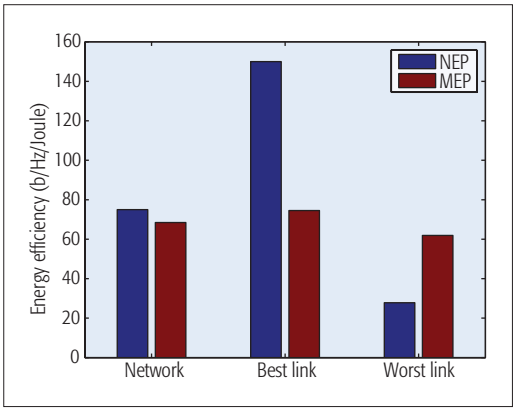


Figure 5. Illustration of the EE-fairness trade-off. In this example, we consider an uplink OFDMA-based cellular network and formulate an optimization problem that maximizes the EE of its worst case link subject to subcarrier assignment and power allocation constraints. In the figure, the number of users $K = 16$, number of subcarriers $N = 128$, power amplifier inefficiency factor $\xi_k = 18$, terminal's circuit power $P_k^C = 0.4$ W, user's rate requirement $R_k^{req} = 15$ b/s/Hz, and maximum transmit power $P_k^{max} = 0.2$ W for all k. Note that the EE of the best/worst link is obtained by saving the EE of the link who has the highest/lowest EE in each sample and then taking an average on 5000 of them.

How to develop joint resource allocation algorithms that reach the theoretical limits of the network EE and thus serve as benchmarks to evaluate performance of other heuristic algorithms is another challenge. Moreover, it is also necessary to develop cost-efficient and easy-to-implement algorithms with acceptable performance levels to solve these problems for practical applications.

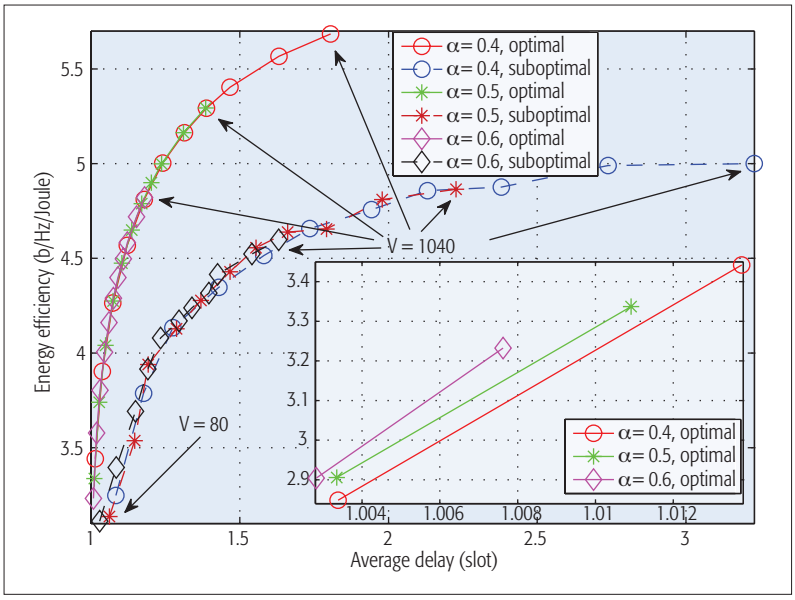


Figure 6. Illustration of the EE-delay trade-off. In this example, we consider a downlink single-MBS H-CRAN and maximize its network $EE_{long\text{-}term}$ subject to a queue length control constraint by jointly optimizing RRH operation and power allocation. In the figure, the traffic arrival rate λ = 2.5 b/slot/Hz, RRH's circuit power P_n^c = 0.4 W, number of RRHs N = 8, and number of users M = 12. In particular, $V \geq 0$ and $\alpha \in [0,1]$ are two control parameters introduced to adjust the EE-delay trade-off.

In what follows, we present a possible method to model and reveal the quantitative EE-delay trade-off.

To formulate EE and delay in a framework, we first need to shift from previously short-term to long-term models. In long-term formulations, random traffic arrivals can be enfolded to obtain a dynamic arrival-departure queue for each user, given as $Q_i(t+1) = \max[Q_i(t) - R_i(t), 0] + A_i(t)$, $\forall i$ [14]. Here, $A_i(t)$ and $Q_i(t)$ denote the amount of newly arrived data and queue length of user i at slot t, respectively. Note that the average delay can be characterized by queue length, as it is proportional to the queue length for a given traffic arrival rate from Little's Theorem.

Furthermore, it is also necessary to inject the concept of time into the EE definition $EE_{short\text{-}term}$ in order to bridge the EE and delay. One possible way to achieve this is to define the EE from a long-term average perspective, given by the ratio of the long-term aggregate data delivered to the corresponding long-term total power consumption in [14, Eq. 10]. For simplicity, we denote this kind of network EE definition by $EE_{long\text{-}term}$. From [10, 14], we know that $EE_{long\text{-}term}$ can also be seen as an extension of $EE_{short\text{-}term}$, because it degenerates to $EE_{short\text{-}term}$ if there are no time averages and expectations in $EE_{long\text{-}term}$. Then, by integrating the queue length control (i.e., delay control) and EE maximization into a framework, we can depict the EE and average delay simultaneously.

We utilize the above ideas to display the EE-delay trade-off in H-CRANs by formulating a stochastic optimization problem that maximizes the network $EE_{long\text{-}term}$ subject to a queue length control constraint through joint optimization of RRH operation and power allocation. Two algorithms, referred to as optimal and suboptimal, are developed to solve this problem. Figure 6 intuitively shows the EE-delay trade-off, where $V \geq 0$ and $\alpha \in [0,1]$ are two control parameters introduced in the model to adjust the EE-delay trade-off. Specifically, from Fig. 6, for the same V, the smaller α is, the better the EE, and the larger the average delay. In addition, for the same α, the bigger V is, the better the EE, and the larger the average delay. These observations together exhibit the EE-delay trade-off, which can be explicitly balanced by V and α. Hence, the long-term model can be used to tune the EE-delay trade-off via adjusting V and α. More clearly, α is used to confine the trade-off range between the EE and average delay (a small α gives a large range and vice versa) and V to tune the trade-off point between the EE and average delay (a small V yields a small delay but low EE and vice versa).

Although [13] found the EE-delay trade-off and [14] obtained an EE-delay trade-off of $[O(1/V), O(V)]$, the optimal EE-delay trade-off, that is, the optimal order for the average delay in V when the EE increases to the optimal by the law of $O(1/V)$, is still unknown. Moreover, [13, 14] focused on the average delay, and thus the obtained results therein are valid only for non-real-time traffic such as web browsing and file transfers. However, there are some other real-time applications, for example, voice and mobile video, in H-CRANs that impose hard-deadline (or maximum delay) constraints. It is thus worthwhile to study how to provision deterministic delay guarantees and improve the EE in the meantime. Moreover, in more realistic H-CRANs with both non-real-time and real-time traffic, it is also well worth investigating how to flexibly balance the EE-delay performance for each kind of traffic from a perspective of systematic design and further devise control algorithms. Potential techniques that can be used to settle these unresolved issues are stochastic optimization, dynamic programming, Markov decision process, queue theory, and stochastic analysis.

Conclusions

Under the triple drives of capacity enhancement, EE improvement, and communication ubiquity, H-CRANs have emerged as a promising architecture for future wireless network design. In this article, we have first exploited the features of H-CRANs to propose three green techniques and then particularly have focused on three fundamental trade-offs, namely EE-SE, EE-fairness, and EE-delay trade-offs. We have introduced the methods to model and analyze these trade-offs, presented open issues and challenges, and also provided some potential solutions. However, we are still at a very primary stage in these studies, and thus further investigations on exploitation of the high-dimension, flexible, and scalable architecture of H-CRANs are eagerly needed for a green future.

Acknowledgment

This work was supported in part by the National Science Foundation of China under Grants 61601192, 61601193, 61631015, and 61471163; the U.S. NSF under Grant CNS-1320664; the Major Program of the National Natural Science Foundation of Hubei in China under Grant 2016CFA009; and the Fundamental Research Funds for the Central Universities under Grant 2016YXMS298.

References

[1] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021, Cisco, Feb. 2017.
[2] IMT-2020 (5G) Promotion Group, "5G Vision and Requirements," white paper, May 2014.
[3] A. Fehske *et al.*, "The Global Footprint of Mobile Communications: The Ecological and Economic Perspective," *IEEE Commun. Mag.*, vol. 49, no. 8, Aug. 2011, pp. 55–62.
[4] China Mobile Research Institute, "C-RAN: The Road Towards Green RAN," white paper, Oct. 2011.
[5] M. Peng *et al.*, "Recent Advances in Cloud Radio Access Networks: System Architectures, Key Techniques, and Open Issues," *IEEE Commun. Surveys & Tutorials*, vol. 18, no. 3, 3rd qtr. 2016, pp. 2282–2308.
[6] M. Peng *et al.*, "Heterogeneous Cloud Radio Access Networks: A New Perspective for Enhancing Spectral and Energy Efficiencies," *IEEE Wireless Commun.*, vol. 21, no. 6, Dec. 2014, pp. 126–35.
[7] J. Wu, "Green Wireless Communications: From Concept to Reality," *IEEE Wireless Commun.*, vol. 19, no. 4, Aug. 2012, pp. 4–5.
[8] E. Oh *et al.*, "Toward Dynamic Energy-Efficient Operation of Cellular Network Infrastructure," *IEEE Commun. Mag.*, vol. 49, no. 6, June 2011, pp. 56–61.
[9] Y. Chen *et al.*, "Fundamental Trade-offs on Green Wireless Networks," *IEEE Commun. Mag.*, vol. 49, no. 6, June 2011, pp. 30–37.
[10] C. Xiong et al., "Energy- and Spectral-Efficiency Tradeoff in Downlink OFDMA Networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 11, Nov. 2011, pp. 3874–86.
[11] J. Tang *et al.*, "Resource Efficiency: A New Paradigm on Energy Efficiency and Spectral Efficiency Tradeoff," *IEEE Trans. Wireless Commun.*, vol. 13, no. 8, Aug. 2014, pp. 4656–69.
[12] C. Joe-Wong et al., "Multiresource Allocation: Fairness-Efficiency Tradeoffs in a Unifying Framework," *IEEE/ACM Trans. Net.*, vol. 21, no. 6, Dec. 2013, pp. 1785–98.
[13] F. Meshkati, H. Poor, and S. Schwartz, "Energy Efficiency-Delay Tradeoffs in CDMA Networks: A Game-Theoretic Approach," *IEEE Trans. Info. Theory*, vol. 55, no. 7, July 2009, pp. 3220–28.
[14] M. Sheng *et al.*, "Energy Efficiency and Delay Tradeoff in Device-to-Device Communications Underlaying Cellular Networks," *IEEE JSAC*, vol. 34, no. 1, Jan. 2016, pp. 92–106.
[15] O. Arnold *et al.*, "Power Consumption Modeling of Different Base Station Types in Heterogeneous Cellular Networks," *Future Networks and Mobile Summit*, June 2010, pp. 1–8.

Biographies

YUZHOU LI [M'14] (yuzhouli@hust.edu.cn) received his Ph.D. degree in communications and information systems from the School of Telecommunications Engineering, Xidian University, Xi'an, China, in December 2015. Since then, he has been with the School of Electronic Information and Communications (EIC), Huazhong University of Science and Technology (HUST), Wuhan, China, where he is currently an assistant professor. His research interests include 5G wireless networks, marine object detection and recognition, and undersea localization.

TAO JIANG [M'06, SM'10] (taojiang@hust.edu.cn) is currently a Distinguished Professor with the School of Electronic Information and Communications, HUST. He has authored or co-authored over 300 technical papers and five books in the areas of wireless communications and networks. He is the Associate Editor-in-Chief of *China Communications* and on the Editorial Boards of *IEEE Transactions on Signal Processing* and *IEEE Transactions on Vehicular Technology*, among others.

KAI LUO (kluo@hust.edu.cn) received his B.Eng. degree from the School of EIC, HUST, in 2006. Then he received his Ph.D. degree in electrical engineering from Imperial College London in 2013. In 2013, he joined the Institute of Electronics, Chinese Academy of Sciences. Since 2014, he has been an assistant professor with the School of EIC, HUST. His research interests are signal processing, MIMO communications, and heterogeneous networks.

SHIWEN MAO [S'99, M'04, SM'09] (shiwen.mao@gmail.com) received his Ph.D. in electrical and computer engineering from Polytechnic University, Brooklyn, New York, in 2004. He is the Samuel Ginn Distinguished Professor and director of the Wireless Engineering Research and Education Center at Auburn University, Alabama. His research interests include wireless networks and multimedia communications. He is a Distinguished Lecturer of the IEEE Vehicular Technology Society and on the Editorial Boards of *IEEE Transactions on Multimedia* and *IEEE Multimedia*, among others.

In more realistic H-CRANs with both non-real-time and real-time traffic, it is also well worth investigating how to flexibly balance the EE-delay performance for each kind of traffic from a perspective of systematic design and further devise control algorithms.

Fully Exploiting Cloud Computing to Achieve a Green and Flexible C-RAN

Jianhua Tang, Ruihan Wen, Tony Q. S. Quek, and Mugen Peng

The authors review the recent advances of exploiting cloud computing to form a green and flexible C-RAN from two cloud-based properties: centralized processing and the software-defined environment. For the centralized processing property, we include coordinated multipoint and limited fronthaul capacity, multicasting, and CSI issues in C-RAN. For the software-defined environment property, we summarize elastic service scaling, functionality splitting, and functionality extension.

ABSTRACT

By merging cloud computing into the RAN, C-RAN has been foreseen as a prospective 5G wireless systems architecture. Due to the innovative move of migrating the baseband processing functionalities to the centralized cloud baseband unit pool, C-RAN is anticipated to reduce energy consumption significantly to be a green RAN. Moreover, with the cloud-based architecture, lots of new functionalities and RAN designs are ready to be incorporated, which redefines the RAN as a flexible RAN. In this article, we review the recent advances of exploiting cloud computing to form a green and flexible C-RAN from two cloud-based properties: centralized processing and the software-defined environment. For the centralized processing property, we include coordinated multipoint and limited fronthaul capacity, multicasting, and CSI issues in C-RAN. For the software-defined environment property, we summarize elastic service scaling, functionality splitting, and functionality extension. We also include some of our recent research results and discuss several open challenges.

INTRODUCTION

Looking back on the mobile communications system's evolution, that is, from first generation (1G) (analog) through to 4G (LTE), the main efforts have been committed to obtaining faster data rates and lower latency. However, during this evolution, a side-effect is causing the system to look grim: the energy consumption. For instance, in 2012, more than 200 GW annual average power consumption in the information and communications technologies (ICT) industry was monitored, one quarter of which are from telecoms infrastructure and devices. To cut down on carbon emissions and have a sustainable future for the ICT industry, 90 percent reduction in energy consumption must be accomplished in the 5G era.

To reduce energy consumption, lots of research works have emerged from both the technique and infrastructure categories. To be more specific:

- In the technique category, people mainly focus on developing energy-efficient algorithms, including base station (BS) sleeping, cell zooming, multiplexing, beamforming, and so on.
- For the infrastructure category, most research attention is on producing energy-efficient hardware, including renewable-source-supported access points, energy-efficient radio heads, millimeter-wave backhaul, and so on.

Instead of upgrading or evolving from these two categories separately, the cloud radio access network (C-RAN) was proposed as a competitive 5G structure that combines the advances from both of them. There are three main components of a C-RAN: remote radio heads (RRHs), fronthaul links, and a baseband unit (BBU) pool. The key innovation of C-RAN is decoupling baseband processing functionalities from the RRHs and migrating these functionalities to the centralized cloud BBU pool, which consists of many general-purpose servers. Hence, the functionality at RRHs can be just as slim as basic signal transceiving.

What gives the C-RAN a big advantage over the conventional RAN in energy reduction is the centralized cloud BBU pool. Specifically, by cloud computing, the RAN manages the transitions not only from many distributed BSs to a centralized BBU pool, but also from a hardware-defined infrastructure to a software-defined environment. As a result, on one hand, cloud computing technology in the centralized BBU pool pushes the RAN to be more energy-efficient. For example, the cooling system is deployed for each BS in the conventional RAN; however, in C-RAN, the cooling system is deployed for the whole centralized BBU pool,[1] and the cooling power can be adaptive to the number of active servers, which is dynamically adjusted in cloud computing. On the other hand, the software-defined environment of the BBU pool facilitates more flexibility for the C-RAN. For instance, many new functionalities can be added onto a C-RAN by just upgrading the software, since the processing, controlling, and management in a C-RAN are all software-defined.

As a result, the energy consumption reduction objective in 5G turns out to be easier to fulfill. That is, much higher multiplexing gain can be obtained by the centralized cloud BBU pool, and renewable sources (e.g., solar power) are also appropriate for supporting RRHs [1]. In addition, compared to the the conventional RAN, the C-RAN offers more flexibility due to the software-defined cloud environment. All this means that a green and flexible C-RAN can be achieved

[1] Normally, there is no cooling system at the RRHs.

Digital Object Identifier:
10.1109/MCOM.2017.1600922

Jianhua Tang is with Chongqing University of Posts and Telecommunications and Seoul National University; Ruihan Wen is with the University of Electronic Science and Technology of China; Tony Q. S. Quek is with Singapore University of Technology and Design and Kyung Hee University; Mugen Peng is with Beijing University of Posts and Telecommunications. The corresponding author is Ruihan Wen.

0163-6804/17/$25.00 © 2017 IEEE

by exploiting cloud computing. Over the past several years, the research on cloud computing in C-RAN has been gone through two stages.

The First Stage: People make use of the centralized processing property brought by cloud computing, in which the cloud BBU pool is regarded as a central super node. That means this node has the global information of the whole system, and can manage and process most of the tasks. However, there is no detailed and in-depth study about the mechanisms inside this node.

The Second Stage: People begin to pay attention to another property, that is, the software-defined environment. At this stage, the detailed mechanisms of cloud computing in the BBU pool and how these mechanisms help improve the C-RAN's efficiency are investigated.

In this article, we discuss the recent advances on fully exploiting the cloud BBU pool to achieve a greener and more flexible RAN, from the above two stages (aspects): centralized processing and the software-defined environment. We also present some of our recent results and highlight some open challenges.

A comprehensive survey on C-RAN was recently conducted in [2]. Compared to [2], the contributions and significance of this article can be considered as follows:

- We emphasize how cloud computing works in C-RAN. This means that we look at C-RAN from the aspect of cloud computing instead of mobile communications. This viewpoint can provide more insights for C-RAN researchers.
- In addition, we unify lots of prospective applications that can be incorporated in C-RAN into the cloud-based service model: X as a service (XaaS).
- We include some recent academic results from our own research (e.g., elastic service scaling) and also some industry progress (e.g., big data as a service), neither of which is covered by [2].
- We try to systematize the RAN and core network in 5G by leveraging cloud computing. To achieve this, we discuss the decentralized core network as a service and C-RAN interacting with network slicing.

Centralized Processing

Centralized processing in the BBU pool offers a natural place to implement many theoretically mature techniques, and brings along new side-effects as well. In this section, we outline the problems that people have studied in C-RAN by leveraging the centralized processing property.

Coordinated Multipoint and Limited Fronthaul Capacity

Over the past decade, coordinated multipoint (CoMP) techniques have attracted comprehensive research attention from both academia and industry, which aim to enhance the system throughput, especially for the cell edge users. However, there are some challenges to be overcome to gain the high performance of CoMP, such as clustering and synchronization.

The emergence of C-RAN provides an ideal environment to implement CoMP, since most challenges in CoMP are eliminated by the centralized processing property of C-RAN (e.g., synchronization). In return, the system power consumption of C-RAN can be further reduced by leveraging CoMP. For instance, the set of active RRHs can be adjusted by the clustering algorithms in CoMP. Hence, the inactive RRHs can be switched into sleep mode to save power consumption. One typical CoMP technique is joint transmission, which duplicates users' desired data to multiple coordinated BSs and transmits the data to each user from multiple coordinated BSs simultaneously. Applying joint transmission in C-RAN means each user's data has to be shared among all the coordinated RRHs and their connected fronthauls. This causes a side-effect in C-RAN: the fronthaul's capacity becomes demanding. Therefore, considering the limited fronthaul capacity is imperative.

To reduce the data transmission rate in the fronthaul, another approach is performing compression in the fronthaul. For instance, the authors in [3] explore compression in C-RAN fronthaul uplinks. By leveraging the correlation between RRHs, they propose a joint decompression and demodulation algorithm in C-RAN fronthaul uplinks, that is, jointly conducting decompression and detection in a single step, to minimize the transmission rate on the fronthaul, and guarantee an acceptable distortion of the decompressed signal in the meantime.

Multicasting

The current RAN is designed to deliver information to specified individuals based on unicasting. However, in 5G, unicasting may not be a good choice anymore for some communication scenarios (e.g., live video streaming), from both the energy efficiency and spectrum efficiency aspects, especially when the fronthaul capacity is scarce.

Due to the centralized processing property, multicasting has been proposed as a promising replacement for unicasting in C-RAN for some communication scenarios. As a natural result, the transmit power consumption in C-RAN is foreseen to be reduced by utilizing multicasting, attributed to the multiplexing gain. For example, in [4], the authors investigate multicast beamforming for different user groups in C-RAN to minimize the weighted sum of backhaul cost and transmit power. Their results show a significant performance advantage by utilizing multicasting rather than unicasting.

CSI Issues

Due to the large number of RRHs and users' equipments (UEs), the channel state information (CSI) is becoming huge in C-RAN. Thus, the so-called curse of dimensionality effect will be a potential obstacle to enhance the performance of centralized processing. There are some works that look at reducing CSI overhead in C-RAN. For example, the authors in [5] propose a compressive CSI acquisition method, which only acquires the instantaneous channel coefficients for a subset of channel links and uses statistical CSI for the others. From the power consumption perspective, reducing CSI overhead means the power consumption to exchange CSI between RRHs and UEs can be cut down as well.

Owing to the centralized processing property, multicasting has been proposed as a promising replacement for unicasting in C-RAN for some communication scenarios. As a natural result, the transmit power consumption in C-RAN is foreseen a reduction by utilizing multicasting, attributing to the multiplexing gain.

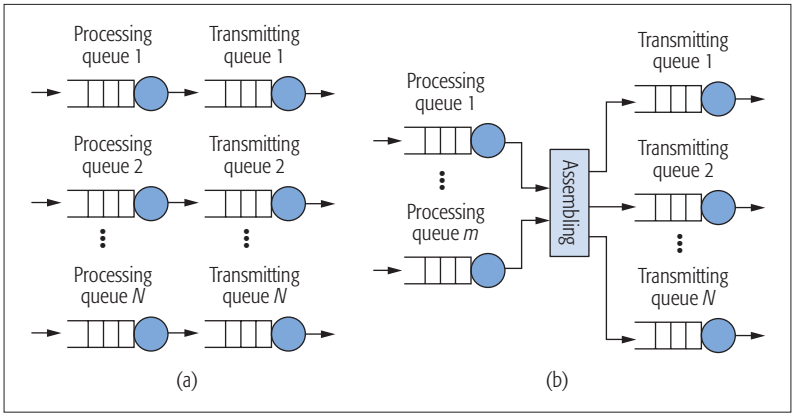


Figure 1. Two different elastic service scaling models: a) an idealized model.; b) a practical model.

Tackling imperfect CSI is always a big concern in wireless communications, and it becomes more significant for the centralized processing tasks in C-RAN, since even a slight imperfection in CSI between the estimated and true channel coefficients may lead to system-wide suboptimal operation. Recently, the authors in [6] make use of the stochastic optimization framework to deal with the noisy and delayed CSI in C-RAN. The results show that the impact of imperfect CSI can be reduced by their proposed approach.

Software-Defined Environment

Besides the centralized processing property, an innovative transition of C-RAN is that by introducing cloud computing, the BBU pool transfers from a hardware-defined infrastructure to a software-defined environment. In this section, we summarize some recent advances in C-RAN by utilizing the software-defined environment.

Elastic Service Scaling

Cloud computing has the famous five essential characteristics, that is, on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service, which have promoted it as a popular and successful computing paradigm over the past decade. By introducing cloud computing in the BBU pool, C-RAN also inherits these characteristics. For example, the rapid elasticity characteristic in C-RAN can be interpreted as meaning that, in the BBU pool, the operator can dynamically scale up and down the required computation resources to support baseband processing to improve the resource utilization. We also call this procedure in the BBU pool elastic service scaling.

Recently, we studied two different elastic service scaling models: an idealized model in [7] and a practical model in [8]. We extract the two main functionalities (i.e., baseband processing and signal transmitting) in C-RAN as two types of queues: the processing queue and transmitting queue, respectively.

The idealized model has the following highlights:

- One-to-one mapping, that is, one UE's incoming traffic is served by only one virtual machine (VM) in the BBU pool, and one VM only serves one UE's traffic. Hence, the number of VMs is equivalent to the number of UEs.
- To capture the elasticity, here, each VM's computation capacity can be dynamically adjusted according to incoming traffic rate, CSI, QoS requirement, and so on. This means that each VM's computation capacity is a variable to be determined.

The corresponding queueing model is represented in Fig. 1a, where the service rate of each processing queue is the computation capacity of each VM, and N is the number of UEs.

However, in a real system, it is not possible to implement the one-to-one mapping and adjust VMs' computation capacity for each UE individually in the BBU pool. Alternatively, in [8], we studied a more practical model, which has the following highlights:

- Multiple-to-multiple mapping, that is, one UE's incoming traffic can be served by multiple VMs in the BBU pool, and one VM can serve multiple UEs' traffic. Thus, the number of VMs is no longer equivalent to the number of UEs.
- Each VM's computation capacity is predefined and fixed, while the optimal number of active VMs is dynamically adjusted, that is, a variable to be optimized, according to the system status. This is how elasticity is captured here.

The practical model reflects the popular commercial cloud service models (e.g., Amazon EC2). The corresponding queueing system model is represented in Fig. 1b, where m is the optimal number of active VMs in the BBU pool.

For both models, there is a transmitting queue for each UE. The service rate of the transmitting queue is the wireless achievable rate to each UE, and the wireless achievable rate is also a variable to be identified. With these queueing system models, we are able to achieve a holistic design for C-RAN, with the following system delay constraint to couple BBU pool and RRHs:

delay in the processing queue + delay in the transmitting queue ≤ system delay threshold.

This constraint is applicable for both the idealized and practical models, but the delay terms have different mathematical expressions in different models.

An interesting trade-off is captured by the system delay constraint. Intuitively and qualitatively, more computation resource allocated in the BBU pool leads to lower delay in the processing queue, while resulting in higher power consumption in the BBU pool. However, thanks to the lower delay in the processing queue, a higher delay in the transmitting queue is hence acceptable based on the system delay constraint, which yields lower transmitting power consumption at the RRHs in return. This trade-off can be optimized mathematically to promote a greener C-RAN. More details and results are given in [7, 8].

Functionality Splitting

The most common architecture of C-RAN is a fully centralized system: processing, control, and management functionalities are located in the BBU pool, and the RRHs just keep basic RF functionality for signal transmission and reception. However, due to some practical constraints (e.g.,

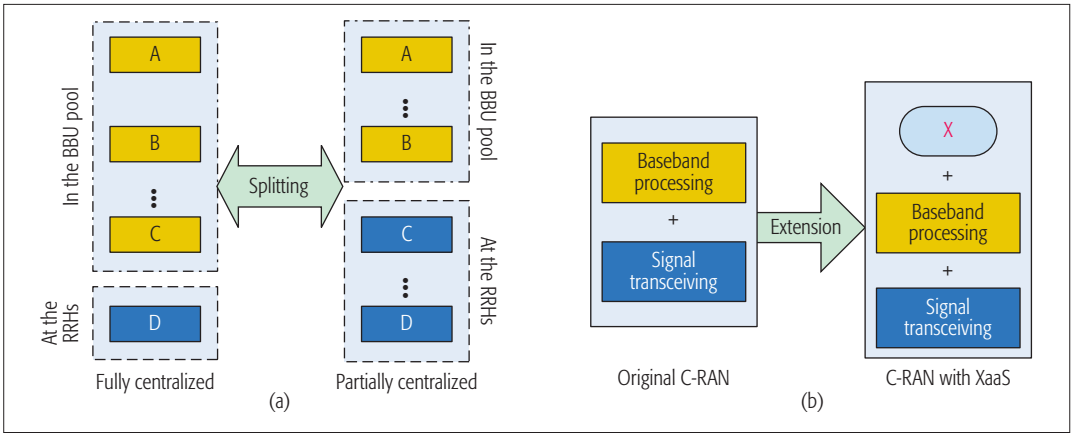


Figure 2. Cloud-assisted functionality flexibility in C-RAN: a) functionality splitting in C-RAN; b) functionality extension for C-RAN.

With the software-defined environment, the C-RAN operator can easily implement a functionality splitting system, instead of a fully centralized system. That means the operator is able to decide each function module to be realized in either the BBU pool or RRHs dynamically, based on different application scenarios.

limited fronthaul capacity), a fully centralized system may not be optimal in some scenarios.

With the software-defined environment, the C-RAN operator can easily implement a functionality splitting system instead of a fully centralized system. This means the operator is able to decide each function module to be realized in either the BBU pool or RRHs dynamically, based on different application scenarios (Fig. 2a).

As an example, Fig. 2a can be regarded as the functionality splitting of C-RAN's radio protocol stack [9]. The left side depicts a fully centralized scenario, in which modules {A, B, ..., C} denote {network management (NM), admission/congestion control (A/CC), radio resource management (RRM), media access control (MAC), and physical layer (PHY)}, and module D stands for the RF module. To reduce the fronthaul traffic amount, some modules can be migrated to the RRH side, as shown on the right side, where {A, ..., B} represents {NM, A/CC}, and {C, ..., D} represents {RRM, MAC, PHY, RF}. This results in a partially centralized C-RAN.

Functionality Extension

There are two main logical functionalities in original C-RAN: baseband processing and signal transceiving, where the baseband processing functionality is executed in the BBU pool, and the signal transceiving functionality is placed at the RRHs.

As the cloud-based BBU pool consists of many general-purpose servers, to fully utilize the software-defined environment of the BBU pool, it is reasonable and inevitable to incorporate more functionalities into it. As depicted in Fig. 2b, a new functionality, X, is added into C-RAN. We call this functionality extension C-RAN with XaaS. We list some recent progress and potential applications in XaaS in the following.

Mobile Cloud Computing as a Service (MCCaaS): Mobile cloud computing (MCC) is proposed to extend the computation ability and prolong the battery life of mobile devices by offloading the computation-intensive tasks to the cloud data center for processing. Although the cloud data center is much more powerful than a mobile device, conventional cloud data centers are always logically and physically far away from mobile devices, which lessens the benefits of MCC. Due to the proximity feature, the cloud-based BBU pool provides a new way to enjoy the advantages of MCC. That is, the computation-intensive tasks can be offloaded to the cloud-based BBU pool for processing instead of the conventional cloud data center in the remote end. In recent progress, the authors in [10] jointly study offloading, computation provisioning, and beamforming in C-RAN with MCCaaS.

Big Data as a Service (BDaaS): Different from MCCaaS, which aims to serve the individual user, BDaaS is proposed to serve the enterprise. For instance, Cazena, a startup company whose mission is to radically simplify enterprise big data processing in the cloud, is offering BDaaS. Under BDaaS, an enterprise's data infrastructure can be built, maintained, and upgraded nearly instantaneously, instead of spending months to set up an on-premises one. This is a very agile and cost-effective way of big data processing. By introducing BDaaS in C-RAN, many big-data-processing-based wireless communications applications can be executed in the BBU pool. An example is social-aware processing, where each user may belong to many different social groups, and each social group is classified by its interest, location, activities, and so on. Accurate and dynamic group classification is the basis to providing and improving specified wireless services for a targeted social group. This group classification problem is a typical application of big data processing, and resolving it in the BBU pool can greatly reduce the latency and backhaul traffic amount.

Decentralized Core Network as a Service: The core network plays the role of the brain of the whole network. It contains the data forwarding functionality in its user plane and the network controlling functionality in its control plane. Furthermore, the user plane and control plane are always coupled in the conventional core network. This core network is always logically located in the center of the whole network, which means all user traffic must go through it to access the Internet. However, in the 5G era, this centralized architecture will be overwhelmed by user data with much higher rate and lower latency requirements. Recently, SK Telecom proposed the concept of a "distributed core network," which separates the user plane

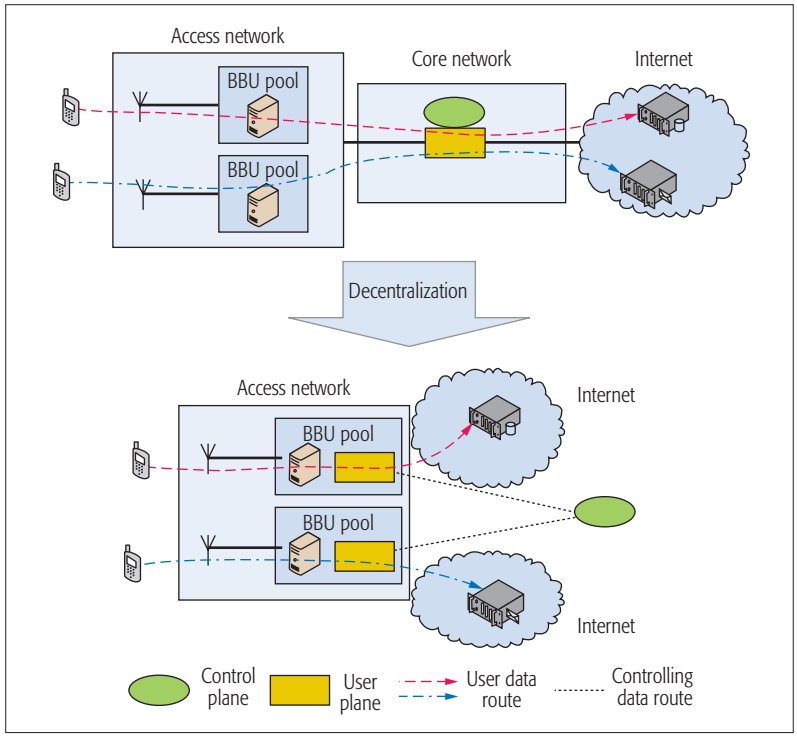


Figure 3. Evolution to the decentralized core network.

and control plane and partitions the physical core network into multiple virtual core networks, leveraging the software-defined network (SDN) and virtualization technology. In Fig. 3, we show the approach to alleviate the core network's burden by decentralizing the core network, with the assistance of C-RAN. Specifically, by migrating the user plane into the software-defined cloud BBU pool, users' data traffic through the physical core network can be greatly reduced.

Caching as a Service: Due to the limited space here, more details are elaborated in our previous work [11].

Thanks to the software defined environment, to realize XaaS in C-RAN is also not complicated. Most of the X can be accomplished by virtualization, without any additional hardware. For instance, MCCaaS can be achieved by generating some specified VMs (in the BBU pool, which support the applications running on the UE side. Then these VMs can help process the application tasks just as clones of the UE.

To sum up the last two sections, we qualitatively plot Fig. 4 to illustrate the interaction between greenness and flexibility. This means that although the two cloud-based properties (i.e., centralized processing and the software-defined environment) each has its own emphasis on greenness and flexibility, respectively, there is no hard bound between greenness and flexibility. In other words, some techniques can contribute to both greenness and flexibility in C-RAN (e.g., elastic service scaling).

Open Challenges

The study of fully exploiting cloud computing in C-RAN is the prerequisite to reach a greener and more flexible RAN. However, the state of the art is still not comprehensive and mature enough. We list some open challenges in this section.

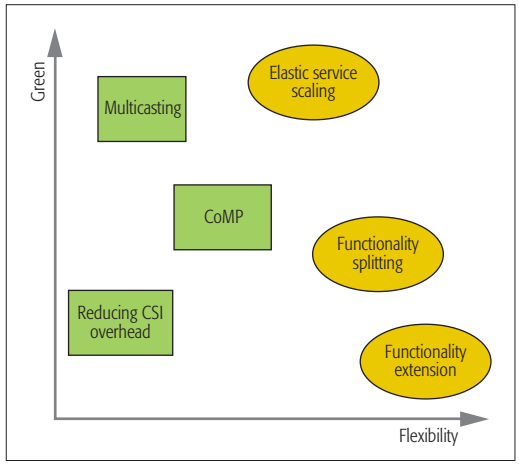


Figure 4. The interaction between green and flexibility.

The Two-Timescale Problem

In the last two sections, we summarize two main perspectives, centralized processing and the software-defined environment, that people have leveraged to exploit the cloud BBU pool. A straightforward extension is to utilize these two properties simultaneously. For example, in [12], we jointly study CoMP (due to centralized processing) and caching as a service (due to the software-defined environment).

However, these types of joint problems always have a common practical limitation: the two-timescale issue. Specifically, resource allocation in the BBU pool is always on the timescale of minutes to hours, while wireless resource allocation or beamforming is on the timescale of only milliseconds due to the wireless channel nature. As shown in Fig. 5, one slow timescale slot (e.g., 30 minutes)consists of T fast timescale slots (e.g., 1 ms). At the beginning of each slow timescale slot, we reallocate the resources in the BBU pool (e.g, activating new VMs), and at the beginning of each fast timescale slot, we redesign the beamformers based on the current CSI.

Problems under these two timescales are always coupled and complicated. For a typical example, the resources allocated in the BBU pool have to meet the wireless QoS requirements over a long time, that is, an entire slow timescale span in Fig. 5. However, this goal is not easy to achieve. This is because in order to achieve the goal at the time we perform resource allocation in the BBU pool, that is, the beginning each slow timescale slot, we have to consider every CSI in T fast timescale slots ahead, which is not possible (since we are unable to know the exact CSI in the future at that time). This becomes a main challenge to be resolved in our future work.

Reducing Latency

The main intention of making use of C-RAN is to improve the energy efficiency of conventional RANs. However, there is a well-known trade-off between energy consumption and latency in communication systems. Generally speaking, to reduce latency means to use higher-power transmitters and processors, which results in higher energy consumption. For example, this trade-off has been recently examined in edge cloud systems [13] and

device-to-device communications systems [14]. However, in C-RAN, there are extensive works aimed at improving energy efficiency, while only very few works pay attention to reducing latency.

In the 5G era, it is expected to have a much more stringent end-to-end latency requirement (i.e., 1 ms). As a potential 5G structure, C-RAN can contribute to reducing latency through the following potential approaches.

Functionality Splitting: In addition to reducing fronthaul traffic, functionality splitting is also applicable in reducing latency. To be more specific, Fig. 2a also illustrates the placement of atomic functions in the baseband processing structure. In particular, for some stringent latency requirement protocols, a fully centralized system may not be satisfiable. For instance, in LTE, the hybrid automatic retransmission request (HARQ) feedback only has 3 ms to be sent out once the corresponding frame is received. Therefore, relegating some atomic functions from the BBU pool to RRHs can effectively reduce the latency. Therefore, if the left hand side of Fig. 2a, {A, B, ..., C} denotes {coding, modulation, MIMO TX, IFFT}, and module D denotes the radio TX, after functionality splitting, the right side of Fig. 2a, {A, ..., B} can represent {coding, modulation, MIMO TX}, and {C, ..., D} can stand for {IFFT, radio TX} to reduce the latency (IFFT: inverse fast Fourier transform).

Fog RAN (F-RAN): To alleviate the traffic on C-RAN fronthaul and reduce latency, F-RAN is proposed as an evolution and complement of C-RAN. For example, for the Internet of Things (IoT) applications in 5G, some applications supported by C-RAN may not be able to meet their quality of service (QoS) requirements, such as low latency, high mobility, and location awareness, since the BBU pool is still relatively far from these "things." To overcome this limitation, one more tier of architecture needs to be deployed between UEs and RRHs. This tier is called fog, and consists of many edge devices. Hence, the RAN becomes a fog RAN. In F-RAN, edge devices, including RRHs and UEs, are also equipped with computation and storage capacities to perform radio signal processing, radio resource management, and caching. In this way, it is possible to transmit the entire torrent of data to the BBU pool via the fronthauls, and some of them can be just processed at the edge devices. As a result, the latency can be reduced accordingly. Nevertheless, F-RAN is not intended to replace C-RAN; It works cooperatively with C-RAN to extend the capability of C-RAN.

Caching as a Service: More details can be found in our previous work [11].

However, the investigation of leveraging C-RAN to achieve the 1 ms end-to-end latency requirement is still open.

Interacting with Network Slicing

Network slicing is proposed as a technique to construct virtual dedicated networks by logically separating the set of network functionalities and resources. These virtual dedicated networks are tailored by some special (technical or commercial) requirements. That means a virtual dedicated network can allow a group of UEs to exclusively access and use a part of the network functionalities and resources [15].

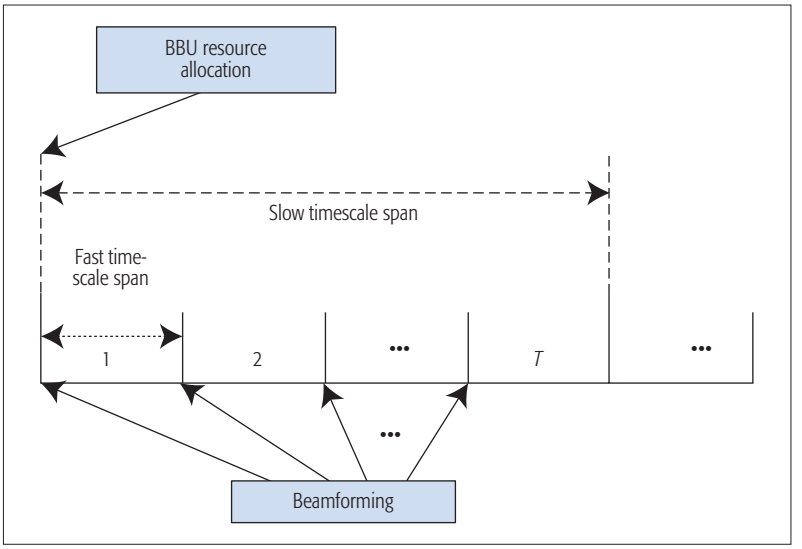


Figure 5. The two-timescale issue.

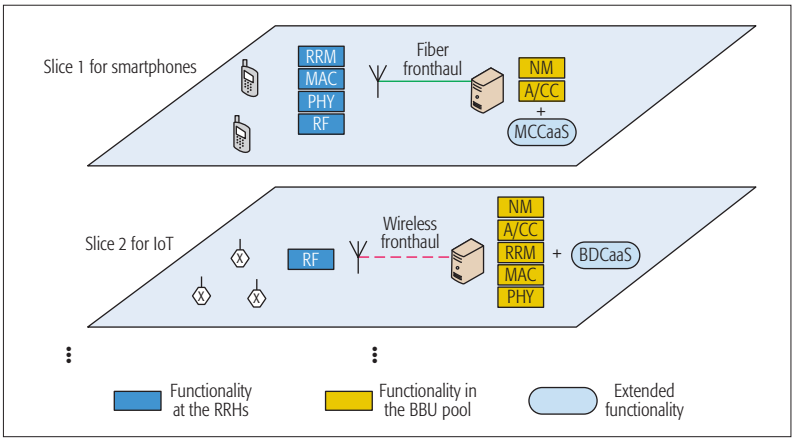


Figure 6. Network slicing for C-RAN.

Based on our discussion in the aforementioned sections, C-RAN is also ready to interact with network slicing. For example, in Fig. 6, we show two typical network slices in the coming 5G era. Slice 1 is for smartphones, which may contain some computation-intensive applications. Slice 2 is for IoT, where the sensors' data rate and latency requirements may not be high, but the features inside the data are very important. Based on these two different scenarios, the operator can produce two different slices in C-RAN accordingly.

Slice 1: To save the fronthaul overhead, we adopt the partially centralized structure for the radio protocol stack. As shown in the figure, only NM and AC/C are kept in the BBU pool; the rest are sunk to the RRHs. In addition, the fiber links are adopted as the fronthaul to further support the high data rate. Furthermore, we incorporate MCCaaS in the BBU pool to help execute the computation-intensive applications.

Slice 2: We still keep the fully centralized structure to coordinate the sensors better. Moreover, to save fronthaul cost and also keep the fronthaul data rate satisfactory, adopting the wireless fronthaul here can be a good choice. In addition, we include BDaaS in the BBU pool to analyze the features inside the data.

With the urgent need to reduce energy consumption in the ICT industry in the 5G era, C-RAN, a solution candidate, has been attracting a broad range of research attention. Due to its cloud-based architecture, C-RAN is going to be not only green but also flexible.

Therefore, together with network slicing in the core and aggregation network, we are able to accomplish the "network on demand" concept (proposed by AT&T). A fundamental of network slicing is that one slice should not be affected by the behavior of other slices. Nevertheless, this rule is difficult to accomplish in the sliced C-RAN due to "inter-slice" interference. As the investigation on network slicing is at a rudimentary level, to design, realize, and manage a C-RAN with multiple isolated slices, and to holistically slice the RAN, aggregation network, and core network are still big challenges.

Conclusion

With the urgent need to reduce energy consumption in the ICT industry in the 5G era, C-RAN, a solution candidate, has been attracting a broad range of research attention. Due to its cloud-based architecture, C-RAN is going to be not only green but also flexible. In this article, we dissect and review the benefits of cloud computing in C-RAN from two aspects, that is, centralized processing and the software-defined environment, which lay the foundation of a green and flexible C-RAN. We summarize some recently studied problems with respect to each aspect and introduce our latest results. Some potential research directions are also explored.

Acknowledgment

This work was supported in part by the Startup Funds of Chongqing University of Posts and Telecommunications under Grant A2016-114, the National Natural Science Foundation of China (NSFC) under Grant 61601071, the Open Foundation of State Key Lab of Integrated Services Networks of Xidian University under Grant ISN17-01, the SUTD-ZJU Research Collaboration under Grant SUTD-ZJU/RES/01/2014, and the MOE ARF Tier 2 under Grant MOE2014-T2-2-002.

References

[1] A. Alameer and A. Sezgin, "Joint Beamforming and Network Topology Optimization of Green Cloud Radio Access Networks," *Proc. Int'l. Symp. Turbo Codes Iterative Info. Processing*, Brest, France, Sept. 2016, pp. 375–79.
[2] M. Peng *et al.*, "Recent Advances in Cloud Radio Access Networks: System Architectures, Key Techniques, and Open Issues," *IEEE Commun. Surveys & Tutorials*, vol. 18, no. 3, pp. 2282–2308.
[3] T. X. Vu, H. D. Nguyen, and T. Q. S. Quek, "Adaptive Compression and Joint Detection for Fronthaul Uplinks in Cloud Radio Access Networks," *IEEE Trans. Commun.*, vol. 63, no. 11, Nov. 2015, pp. 4565–75.
[4] M. Tao *et al.*, "Content-Centric Sparse Multicast Beamforming for Cache-Enabled Cloud RAN," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, Sept. 2016, pp. 6118–31.
[5] Y. Shi, J. Zhang, and K. B. Letaief, "CSI Overhead Reduction with Stochastic Beamforming for Cloud Radio Access Networks," *Proc. IEEE ICC*, Sydney, Australia, June 2014, pp. 5165–70.
[6] Y. Cai, F. R. Yu, and S. Bu, "Dynamic Operations of Cloud Radio Access Networks (C-RAN) for Mobile Cloud Computing Systems," *IEEE Trans. Vehic. Tech.*, vol. 65, no. 3, Mar. 2016, pp. 1536–48.
[7] J. Tang, W. P. Tay, and T. Q. S. Quek, "Cross-Layer Resource Allocation with Elastic Service Scaling in Cloud Radio Access Network," *IEEE Trans. Wireless Commun.*, vol. 14, no. 9, Sept. 2015, pp. 5068–81.
[8] J. Tang *et al.*, "System Cost Minimization in Cloud RAN with Limited Fronthaul Capacity," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, May 2017, pp. 3371–84.
[9] P. Rost *et al.*, "Cloud Technologies for Flexible 5G Radio Access Networks," *IEEE Commun. Mag.*, vol. 52, no. 5, May 2014, pp. 68–76.
[10] J. Cheng *et al.*, "Computation Offloading in Cloud-RAN Based Mobile Cloud Computing System," *Proc. IEEE ICC*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.
[11] J. Tang and T. Q. S. Quek, "The Role of Cloud Computing in Content-Centric Mobile Networking," *IEEE Commun. Mag.*, vol. 54, no. 8, Aug. 2016, pp. 52–59.
[12] J. Tang, T. Q. S. Quek, and W. P. Tay, "Joint Resource Segmentation And Transmission Rate Adaptation in Cloud RAN with Caching as a Service," *Proc. IEEE SPAWC*, Edinburgh, U.K., July 2016, pp. 1–6.
[13] X. Guo *et al.*, "An Index Based Task Assignment Policy for Achieving Optimal Power-Delay Tradeoff in Edge Cloud Systems," *Proc. IEEE ICC*, Kuala Lumpur, Malaysia, May 2016, pp. 1–7.
[14] M. Sheng *et al.*, "Energy Efficiency and Delay Tradeoff in Device-To-Device Communications Underlaying Cellular Networks," *IEEE JSAC*, vol. 34, no. 1, Jan. 2016, pp. 92–106.
[15] X. Zhou et al., "Network Slicing as a Service: Enabling Enterprises' Own Software-Defined Cellular Networks," *IEEE Commun. Mag.*, vol. 54, no. 7, July 2016, pp. 146–53.

Biographies

Jianhua Tang [S'11, M'15] received his B.E. degree in communication engineering from Northeastern University, China, in 2010, and his Ph.D. degree in electrical and electronic engineering from Nanyang Technological University, Singapore, in 2015. He is with the School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, China. Currently, he is a research assistant professor at Seoul National University, Korea. His research interests include cloud computing, content-centric networks, and cloud RAN.

Ruihan Wen received her B.E. degree in communication engineering from Jilin University in 2010 and her M.E. degree in electronic information science and technology from the University of Electronic Science and Technology of China (UESTC) in 2013, respectively. She is now a Ph.D. student at the National Key Laboratory of Science and Technology on Communications in UESTC. Her research interests include resource management, network virtualization, and network slicing technologies in future networks.

Tony Q. S. Quek [S'98, M'08, SM'12] received his B.E. and M.E. degrees in electrical and electronics engineering from Tokyo Institute of Technology. At MIT, he earned his Ph.D. in electrical engineering and computer science. He is a tenured associate professor with the Singapore University of Technology and Design. He is currently an Editor for *IEEE Transactions on Communications* and was an Executive Editorial Committee Member for *IEEE Transactions on Wireless Communications*.

Mugen Peng [M'05, SM'11] received his B.E. degree in electronics engineering from Nanjing University of Posts and Telecommunications, China, in 2000, and his Ph.D. degree in communication and information systems from Beijing University of Posts and Telecommunications (BUPT), China, in 2005. He is a full professor with the School of Information and Communication Engineering at BUPT. His main research areas include wireless communication theory, radio signal processing, and convex optimizations.

Enhancing Energy Efficiency via Cooperative MIMO in Wireless Sensor Networks: State of the Art and Future Research Directions

Yuyang Peng, Fawaz Al-Hazemi, Raouf Boutaba, Fei Tong, Il-Sun Hwang, and Chan-Hyun Youn

ABSTRACT

CMIMO is an effective approach to increase throughput and energy efficiency through the collaboration of individual antennas working together as a virtual multi-antenna system. Several CMIMO strategies have been propounded as major candidates for achieving green communications in wireless sensor networks. Compared to conventional MIMO, CMIMO provides significant gains in terms of flexibility. Recently, more advanced cooperation strategies have been proposed to improve the performance of CMIMO by using emerging techniques such as spatial modulation and coding. Although some breakthroughs have been made in this area, the problem of how to accurately adopt these emerging techniques to model CMIMO is far from being fully understood. This article surveys several state-of-the-art CMIMO models for different scenarios, including data aggregated, multihop-based, and clustered schemes. Moreover, it discusses the implementation of CMIMO techniques, which are expected to be candidate techniques for green communications in modern applications. In the implementation, the trade-offs between energy efficiency and spectral efficiency, quality of service, fairness, and security are discussed. Several simulation results are given to show how emerging techniques in CMIMO design can lead to energy efficiency enhancement. Finally, some challenges and open issues that present future research directions are discussed.

INTRODUCTION

Recently, energy consumption has become a primary performance factor in communication systems. In response, several green communication techniques have been proposed. Typically, they require the target of energy efficiency improvement, which is a near-term goal of going "green." Multiple-input multiple-output (MIMO) has been proved as a key technology to reduce the energy consumption in communication systems. Higher bit rates and energy gains are promised in MIMO systems compared to single-antenna systems. In order to achieve the objectives of MIMO, each wireless device has to be designed with multiple antennas as transmitters or receivers, where the antennas are packed together with spacing on the order of a wavelength. However, due to the small size of the sensor device, referred to as a node, it can be difficult to embed multiple antennas in such nodes. Due to this limitation, traditional MIMO systems cannot produce the expected performance, and the concept of cooperative MIMO (CMIMO) implemented by cooperation of individual antennas is explored to solve the size limitation problem.

CMIMO, also known as virtual MIMO, utilizes the benefit of MIMO elements from independent fading and is one type of cooperative communications. The major difference between MIMO and CMIMO is that in CMIMO, each node is only equipped with one antenna, and nodes are located in different areas. These distributed nodes form a virtual antenna array in order to achieve higher spatial diversity gain, which is also referred to as cooperative diversity gain [1]. The advantages of CMIMO are due to its ability to improve throughput, coverage, and capacity in a cost-effective manner. Because CMIMO offers these tremendous benefits, the research on CMIMO is an active area, and several CMIMO strategies have already been adopted in major wireless standards.

After the pioneering work of Cui *et al.* [2], significant attention has been paid to the design of CMIMO systems over the years. CMIMO has started to become important candidates for many communication networks such as wireless sensor networks (WSNs), ad hoc networks, and vehicular networks. On the other hand, some emerging technologies, such as spatial modulation (SM), have been explored to combine with CMIMO for energy efficiency improvement, and it has been shown that the combined strategy can provide significant performance in terms of energy efficiency [3]. Despite lots of research activity on CMIMO schemes over the last several years, there still remain many technical challenges in the design of CMIMO schemes. The aim of this article is two-fold:

- To present a comprehensive overview on the current state of the art in this research area
- To define open issues for future research directions

The authors survey several state-of-the-art CMIMO models for different scenarios, including data aggregated, multihop-based, and clustered schemes. They discuss the implementation of CMIMO techniques, which are expected to be candidate techniques for green communications in modern applications, as well as the trade-offs between energy efficiency and spectral efficiency, quality of service, fairness, and security.

Yuyang Peng, Fawaz AL-Hazemi, and Chan-Hyun Youn are with Korea Advanced Institute of Science and Technology; Raouf Boutaba is with the University of Waterloo; Fei Tong is with Zhejiang University; Il-Sun Hwang is with Korea Institute of Science and Technology Information.

Digital Object Identifier: 10.1109/MCOM.2017.1600837

0163-6804/17/$25.00 © 2017 IEEE

Reducing the energy consumption to increase energy efficiency sometimes can result in lower spectral efficiency due to the reduction of the transmission diversity. Usually, the transmission diversity is related to the modulation constellation size and dimension of the transmitted symbol. Therefore, the appropriate techniques such as adaptive modulation and index modulation techniques need to be considered.

The remainder of the article is organized as follows. We first briefly introduce the basic principle of CMIMO and its development. Following that, we discuss the energy efficiency and fundamental factors in CMIMO, and then we review some interesting recent results in CMIMO techniques. Finally, we propose some challenges and future research directions on this topic.

Overview of CMIMO

CMIMO is a novel approach of transmitting information by using collaboration of individual antennas; the idea behind of it can be traced back to the virtual antenna array as the groundbreaking work. In CMIMO schemes, the antennas are self-configured to form a cooperative network without any established infrastructure, as shown in Fig. 1. The communication between the transmitter and receiver proceeds in two phases: information sharing and cooperative transmission. Through the first phase, all the nodes get the information data from the others and enable independent data transmission. In the second phase, all the nodes or selected nodes cooperate together to form a virtual MIMO system through techniques such as distributed space time block coding or repetition. The antennas handle the necessary control and communication tasks by themselves via the use of distributed algorithms without an inherent infrastructure. CMIMO schemes are highly appealing for many reasons. In contrast to conventional MIMO schemes, which relay by packaging multiple antennas in one device, CMIMO schemes break this limitation and work in a flexible way without the performance decreasing in terms of throughput. Also, due to the distributed nature, CMIMO schemes can be rapidly deployed and reconfigured. However, these advantages should not be taken to mean that CMIMO schemes are totally flat. Indeed, many CMIMO schemes require a backbone for use by the cluster head node or assistant node to form cooperative transmissions. The cluster head node and assistant node are usually selected from distributed devices, which makes the implementation of the whole CMIMO system complex. Therefore, exploiting the good design structures of CMIMO without violating the fundamental requirements such as spectral efficiency, quality of service (QoS), fairness, and security has important value. The energy constraint is another vital concern in CMIMO schemes. Most existing applications for CMIMO are implemented by assuming that the individual antennas are embedded in the devices with limited energy, and the devices are dropped into a remote region. Therefore, conserving energy to maximize the lifetime is very important, and motivates the research focusing on energy efficiency. On the other hand, ensuring energy efficiency has an effect on the aforementioned fundamental requirements. Thus, finding the trade-off [4] between energy efficiency and the aforementioned fundamental requirements is critical to design energy-efficient CMIMO. Considering the analysis in CMIMO, we may conclude that CMIMO can be a good candidate for next generation communications if we appropriately utilize its advantages and solve its drawbacks.

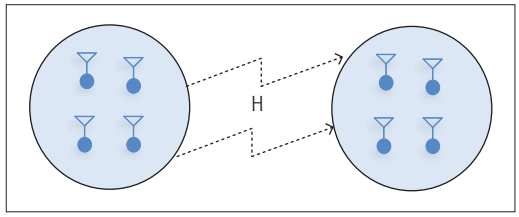


Figure 1. A cooperative MIMO scheme

Energy Efficiency in CMIMO

Since the aim of using CMIMO schemes is to tailor the CMIMO design to the appropriate application for improving energy efficiency, it is useful to discuss the energy efficiency in the CMIMO design. Several fundamental factors need to be considered in energy-efficient CMIMO design. In what follows, we consider the spectral efficiency, QoS, fairness, and security as the fundamental factors and discuss their effects on energy efficiency in CMIMO design. Reducing the energy consumption to increase energy efficiency sometimes can result in lower spectral efficiency due to the reduction of the transmission diversity. Usually, the transmission diversity is related to the modulation constellation size and dimension of the transmitted symbol. Therefore, the appropriate techniques such as adaptive modulation and index modulation techniques need to be considered. In [3, 5], the index modulation technique is taken into account in CMIMO design. The results show that significant energy efficiency is achieved compared to the traditional way under the same spectral efficiency. QoS is a factor interacting with energy efficiency and needs to be guaranteed in most cases. Apparently, the QoS improvement mechanism is contradictory to the energy efficiency requirements because good QoS usually requires big energy consumption. However, for a given QoS, energy efficiency can be achieved by adapting the modulation scheme at the cost of increasing the transmission power. Some systems are energy efficiency preferred, whereas some are power preferred. The type of reference depends on the system itself. If the system is energy efficiency preferred, the property of big power can be neglected. In [2, 6], the optimal modulation constellation sizes are derived to achieve energy efficiency under the given bit error ratio (BER) requirement. The results show that by adapting the modulation to choose the optimal modulation constellation size, the energy efficiency is obtained. Fairness for a communication system refers to the degree to which a fair share of system resources is utilized. For instance, in cognitive-radio-based wireless networks, a certain amount of spectrum should be assigned regardless of the ambient environment. In many cases, increasing energy efficiency causes unfair sharing of the system resources. Therefore, it aims to allocate resource as fairly as possible while keeping energy efficiency. The authors in [7] propose a cognitive CMIMO by considering the radio resource being fairly utilized, and the results show good performance in terms of energy efficiency. For CMIMO-based wireless sensor networks, security is usually not mandatory but desired. In a normal security-based environment, more energy is needed during the transmission

due to the additional processing at both the transmitter and receiver. Spending more energy will decrease the energy efficiency. However, in some special environments such as the military environment, secure transmission can improve energy efficiency by avoiding the additional energy due to misdetection and retransmission. Hence, considering security or not to improve energy efficiency in wireless sensor networks is highly dependent on the operating environment. In the following section, we discuss more energy efficiency issues of the recent advanced techniques in CMIMO design.

Recent Advances in CMIMO

Diversity Gain in CMIMO Schemes

The first study on the CMIMO concept in WSNs dates back to 2004, in which Cui *et al.* [2] were the first to propose a CMIMO with Alamouti code in clustered WSNs. By cooperating with the neighboring wireless nodes, CMIMO can efficiently reduce the transmission energy, but this benefit comes at the cost of higher circuit energy consumption. Since the transmission energy of wireless nodes is proportional at least to the square of the distance, the transmission energy dominates the total energy consumption for a long transmission distance. On the other hand, when the transmission distance is short, circuit energy becomes the major contributor in the total energy consumption. Therefore, in the case of long transmission distance, more cooperative nodes should be used to reduce the transmission energy consumption via antenna diversity, while in the case of short transmission distance, fewer cooperative nodes are preferred to reduce the circuit energy. Moreover, the authors also show that there is an optimal modulation constellation size for each transmission distance. By considering this factor, the energy consumption performance of CMIMO can be further improved.

Multiplexing Gain in CMIMO Schemes

Vertical-Bell Labs Layered Space-Time (VBLAST)-virtual MIMO [6] is yet another classical CMIMO, which provides multiplexing gain by allowing a virtual antenna array to transmit N independent data streams. The core technique of this scheme is to point a data gathering node that can cope with more computational complexity than other normal nodes at the receiver. At the transmitter, each of the nodes broadcasts its data to the other nearby nodes by means of a time-division multiple access scheme. After that, each node has data from all the others to transmit through space time coding techniques. At the receiver, the data gathering node receives data from the transmitter, which allows realization of real MIMO capability with only transmitter side local communications. Using this method, significant energy reduction is achieved.

Data Aggregation Gain in CMIMO Schemes

The CMIMO with data aggregation technique is a way in which the correlated data size can be significantly reduced according to the correlation factor [8]. The underlying philosophy is to reduce the amount of redundant data depending on the data similarity at the transmitter. Specifically, the sensor nodes send the information data to their cluster head, and then the cluster head aggregates the collected data and sends them back to all the sensor nodes in that cluster. Thus, all the sensor nodes at that cluster have the same aggregated information data. After that, the sensor nodes transmit the received aggregated data to the sensor nodes that are located in the receiving cluster, and then sensor nodes at the receiving cluster transmit the received data to their cluster head for joint detection. By considering data aggregation, the transmitted data amount is significantly reduced, and so is the total energy.

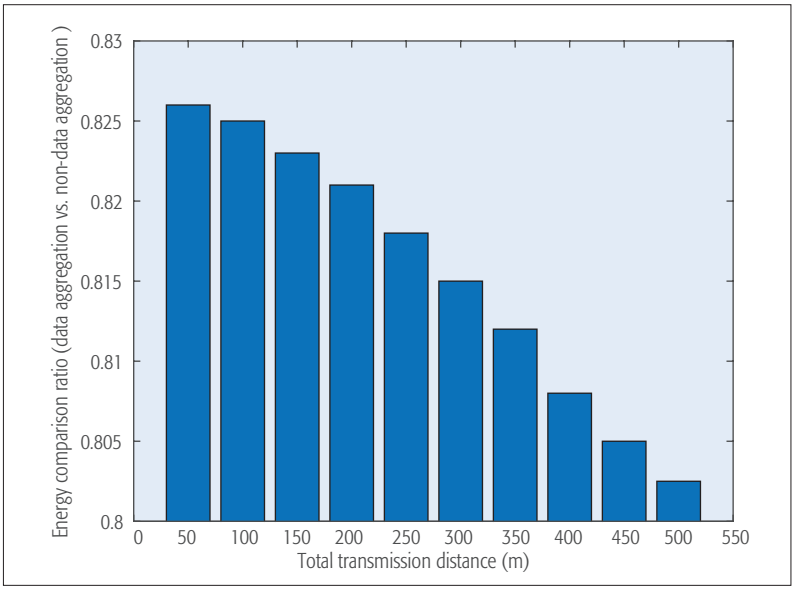


Figure 2. Energy comparison over transmission distance.

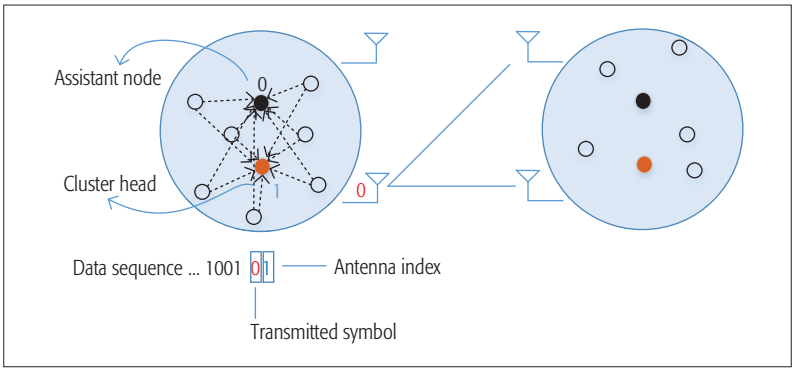


Figure 3. A CMIMO-SMR system.

Figure 2 illustrates the energy performance of data aggregation in CMIMO schemes. The comparisons are carried out under the data-aggregation-based and non-data-aggregation-based CMIMO schemes for a transmission distance of 500 m. The energy comparison ratio shows that data-aggregation-based CMIMO provides significant energy saving over non-data-aggregation-based CMIMO. We observe that although the process of data aggregation requires additional energy, this part of energy has little effect on total energy consumption. In addition, the transmission energy saving due to the small amount of data after data aggregation affects a lot of the total energy consumption. Moreover, the energy

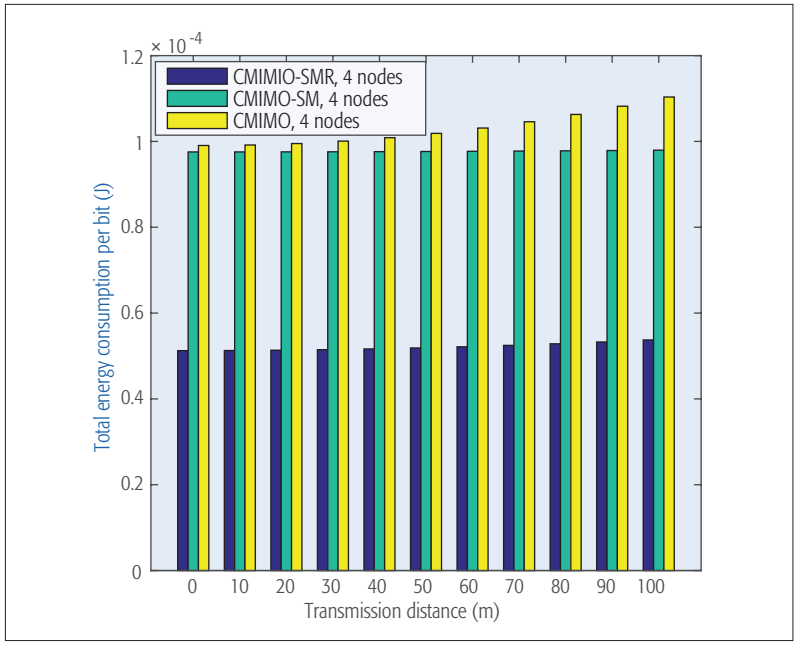


Figure 4. Energy consumption per bit over transmission distance under four-node transmission schemes.

saving performance of data aggregation becomes apparent because the transmission energy dominates the total energy consumption as the transmission distance increases.

Indexing Gain in CMIMO Schemes

CMIMO-spatial modulation (CMIMO-SM) is a novel CMIMO transmission scheme based on the SM technique [3]. The adoption of SM makes CMIMO systems operate without inter-channel interference (ICI). In CMIMO-SM, each node at the transmitter side broadcasts its information data to all the other nodes inside the transmitting cluster using different time slots as the first stage. In the second stage, after each sensor node receives all the other information data, the data sequence is transmitted via the MIMO channel. Note that, for each time instant, the transmitted data sequence is split into two parts: the multiple quadrature amplitude modulation (MQAM)/ multiple phase shift keying (MPSK) modulated symbol part and the antenna index part. Only the modulated symbol part is transmitted, while the antenna index part is reserved for the selection of active transmit antenna and will be detected at the receiver as hidden information. Therefore, for the same spectral efficiency, fewer bits are transmitted in CMIMO-SM compared to CMIMO. Additionally, CMIMO-SM requires a single RF chain, unlike plain CMIMO. Overall, the total energy consumption, including transmission energy and circuit energy, is reduced in CMIMO-SM when compared to that in CMIMO.

CMIMO-SM with randomly distributed nodes (CMIMO-SMR), shown in Fig. 3, is a recently proposed clever modification of CMIMO-SM to improve the flexibility while maintaining its advantages such as ICI-free and energy-efficient transmission [5]. In CMIMO-SMR, the cluster head node and assistant node are jointly set up by means of a cooperative technique in each cluster to obtain diversity. Specifically, the randomly distributed nodes form clusters; and in each cluster there are a cluster head, an assistant node, and several nodes. The cluster head and assistant node have a preassigned index by use of 1 and 0, respectively, to represent them. In the cluster, each node decides if it works as a cluster head for each round according to the rounds in which the node has been a cluster head. After that, the nodes inside the cluster inform the selected cluster head that they will operate as normal nodes or the assistant node by transmitting an extra bit along with the information data. According to the received signal strength (RSS) of the acknowledgment from the other nodes, the cluster head selects the assistant node from the interested candidates. Once the formation of the cluster is done, the nodes only transmit information data to the cluster head and the assistant node. After that, the cluster head and the assistant node transmit the received information data by use of SM. Compared to CMIMO-SM, CMIMO-SMR has less operation inside the cluster and lower circuit energy consumption at the receiver due to the existence of the cluster head and assistant node. Thus, total energy reduction is achieved.

To analyze the energy consumption performance, Fig. 4 shows the total energy consumption per bit and transmission distance for the case of four nodes in one cluster. Three cooperative transmission schemes are considered in Fig. 4: CMIMO, CMIMO-SM, and CMIMO-SMR. It can be seen from Fig. 4 that the energy consumption per bit of each scheme increases as the transmission distance increases. This is because longer transmission distance requires bigger energy consumption. In Fig. 4, the case with CMIMO-SMR achieves the highest performance in terms of energy consumption due to the energy-efficient transmission.

Hop in Multihop-CMIMO Schemes

When the transmission distance is far, CMIMO is often presented in the context of multihop architectures. The authors in [3] derive an optimal hop length expression for multihop-CMIMO-based linear networks, where the sensor nodes form clusters using CMIMO-SM to transmit the information. The optimal length is derived mathematically by considering the transmitted load. Specifically, a cluster close to the destination forwards more load than another cluster far away from the destination; thus, the hop length for the cluster that is close to the destination should be small. The opposite way can be explained for the big hop length. In this way, for each hop length, there is a matched optimal value for it to achieve the minimum energy consumption.

When the intermediate hops act only forwarding information data rather than transmitting their own information data, an optimal number of hops that can be used to minimize the total energy consumption of the whole network can be found by solving the optimization problem. Specifically, the total energy consumption is dependent on the number of hops and can be treated as a convex function. The optimal number of hops can be achieved by taking the first-order derivative of the total energy function with respect to the number of hops and setting it to zero. Because

the number of hops is defined over integer values, the adjacent value, which is with regard to the minimum total energy consumption, can be selected as the optimal number of hops. Figure 5 shows the optimal number of hops and the total energy consumption for a CMIMO-SMR-based multihop network where four different BER situations are considered. For each different BER case, a different optimal number of hops can be found to minimize the total energy consumption when both transmission energy and circuit energy are considered. It can be seen that the optimal number of hops increases as the BER performance increases. This can be explained as the truth that good link quality requires short transmission distance, namely, more hops.

Communication Modes Adaptation in CMIMO Schemes

In cooperative communications, it is possible to form CMIMO mode, cooperative single-input multiple-output (CSIMO) mode, and cooperative multiple-input single-output (CMISO) mode via the collaboration of nodes. In wireless sensor networks, the energy reduction can be achieved by adapting the communication modes for each transmission hop. In [9], a novel communication mode adaptation algorithm is introduced to improve energy efficiency in wireless sensor networks. For each hop, the adaptation of communication modes is considered and determined by optimizing the parameters such as the number of transmitters, receivers, and cooperation nodes to achieve the minimum energy consumption. The experiment results show that significant energy reduction can be obtained by using the communication modes adaptation algorithm.

Radio Resource Management in CMIMO Schemes

Radio resource management is one effective way to reduce energy consumption of wireless systems [10]. The goal of the research on radio resource management is to efficiently utilize the radio resource of the whole network by use of traffic-aware and cognitive radio techniques. In [7], the authors present cognitive CMIMO where the local broadcasting phase and distributed MIMO access phase operate in the cognitive and licensed bands, respectively. In the broadcasting phase, each active node broadcasts its own message with a transmit power, whereas an inactive node becomes a participant of the CMIMO only when it can successfully decode the message transmitted from the nearest active node. It shows that the energy efficiency can be increased by tuning the bandwidth ratio for a given spectral efficiency range.

Sleeping Strategy in CMIMO Schemes

In CMIMO networks, it is not always reasonable to assume that all nodes simultaneously participate in the transmission or are active in certain scenarios. Therefore, sleep mode is an effective tool for energy saving. For example, Li *et al.* [11] consider the design of CMIMO schemes to decide whether a node with a single antenna should be active to become a part of CMIMO under an energy constraint requirement so as to achieve better system performance. In particular, under large-scale CMIMO-based networks, the number of cooperating nodes is high. In this case, if all the nodes participate and keep active in the cooperation stage, the local energy consumption, especially the circuit energy consumption, will increase. On the other hand, the cooperation of large active nodes makes the cooperative transmission consume less transmission energy in the long haul due to the added spatial diversity. Therefore, there is a trade-off between the number of active nodes and total energy consumption. This active nodes selection can be done via making a node sleeping strategy, and then the total energy consumption of CMIMO networks can be optimized by use of the sleeping strategy.

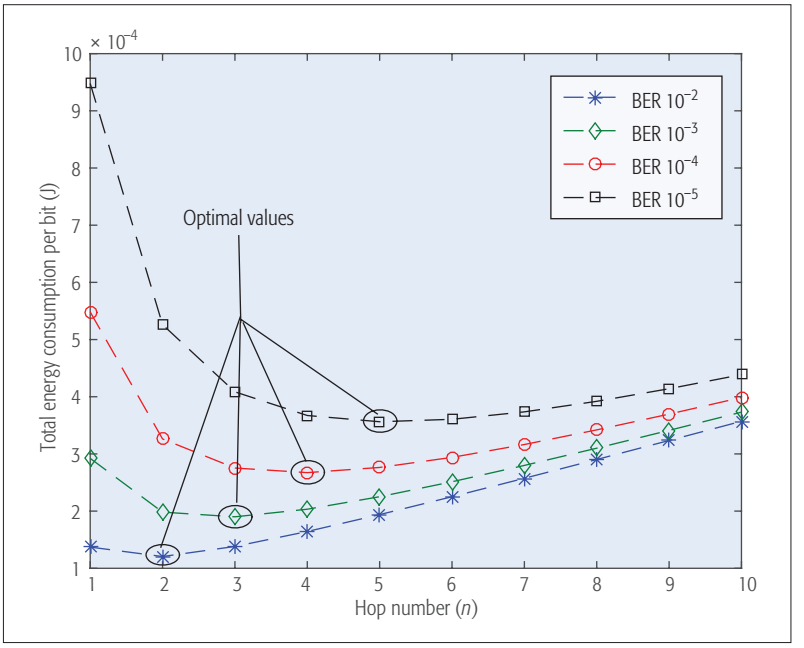


Figure 5. Optimal number of hops vs. energy consumption.

Joint Selecting Gain in CMIMO Schemes

Joint selection is a good candidate to reduce the energy consumption of wireless systems [12]. In multihop-CMIMO networks, the energy consumption is more complex due to the complex transmission environments. CMIMO can exploit the spatial diversity to reduce the transmission energy under a given BER. Therefore, larger hop length should be used to reduce the number of hops. On the other hand, when the hop length is large, the energy consumption of long-haul transmission will dominate the overall energy consumption. Thus, more nodes are required to enlarge the spatial diversity, which will benefit the transmission energy. This analysis means that not only the hop length, but also the number of hops influence the total energy consumption. In [13], a joint selected scheme is proposed in CMIMO systems for energy reduction. In this scheme, the hop length and the number of cooperating nodes are jointly selected under the high node density condition. The simulation results indicate that significant energy saving is obtained by using the joint selected scheme.

In Table 1, a comparison of various energy-efficient CMIMO schemes with the corresponding technologies, fundamental dimensions, and design goals are provided.

To improve the BER performance and optimize the power allocation in CMIMO systems, the channel estimation is indispensable. Although there have been many works in real MIMO channel estimation, the channel estimation in CMIMO is a challenging task because the antenna elements of CMIMO is not integrated.

Schemes	Technologies	Fundamental dimensions	Goals
CMIMO [2]	Diversity	Spectral efficiency, QoS	Reducing energy via cooperation
VBLAST-CMIMO [6]	Multiplexing	Spectral efficiency, QoS	Reducing energy via cooperation and less overhead
CMIMO-DC [8]	Date aggregation	Spectral efficiency, QoS	Reducing energy via data aggregation
CMIMO-SM [3]	Indexing, multihop	Spectral efficiency, QoS	Reducing energy via indexing
CMIMO-SMR [5]	Indexing, multihop, flexibility	Spectral efficiency, QoS	Reducing energy via indexing and flexibility
Scheme in [10]	Radio resource management	Fairness	Reducing energy via radio resource management
Scheme in [11]	Sleeping	Spectral efficiency, QoS	Reducing energy via sleeping mode
Scheme in [13]	Joint selection	Spectral efficiency, QoS	Reducing energy via joint selection

Table 1. Comparison of several CMIMO schemes.

Challenges in Cooperative MIMO

Despite the significant developments that have been achieved in CMIMO, in practical environments, CMIMO still faces several challenges.

Cross-Layer Design

Cross layer design for energy efficiency improvement is important in CMIMO-based transmission. The key design challenge is the performance guarantee when several layers are considered together. There have been many works at different layers in CMIMO networks for energy efficiency. However, their efforts mainly focus on isolated layer design, thus ignoring important interdependencies. Such isolated layer design results in poor performance, especially in real environments when energy and delay are constraints. To overcome this problem, a cross-layer design that supports integration across multiple layers of the protocol is required.

Power Control

Power control is a potent way to improve the performance of CMIMO transmissions. Due to multipath fading, the channel changes randomly. Power control can be used to compensate the random channel, reduce the transmit power, meet the delay constraint, and minimize the probability of link outage. For example, when the CMIMO transmission experiences a deep fading channel, much power should be used to maintain the required signal-to-noise ratio (SNR); when a hard delay constraint requirement is given, the power for transmission of a packet should be increased to improve the probability of successful transmission. For all these purposes, power control in the CMIMO environment needs to be investigated.

Channel Estimation

To improve the BER performance and optimize the power allocation in CMIMO systems, channel estimation is indispensable. Although there have been many works on real MIMO channel estimation, the channel estimation in CMIMO is a challenging task because the antenna elements of CMIMO are not integrated. Moreover, because of the frequent changing of cooperated candidates, the timing requirement of channel estimation is very strict. To achieve accurate channel estimation, there remain many open issues, such as how to utilize only partial channel state information to estimate the whole CMIMO channel, and how to characterize the relation of multiple links at the system level appropriately.

Topology Design in Multihop-CMIMO

Existing topology schemes in CMIMO are mainly developed for linear networks. Specifically, the linear network is categorized into all transmission and single-transmission networks. In these CMIMO-based networks, the optimal values such as the optimal number of hops or optimal hop lengths are calculated for the purpose of total energy saving. Their solutions mainly rely on Lagrange function or derivation. However, in fact, the nodes in CMIMO networks are usually distributed, and the linear topology may not always be suitable. Therefore, the optimal solutions in terms of hops and hop lengths for CMIMO-based random topology are needed. There have been some methods, such as energy awareness optimal relay selection (EAORS) [14], for solving the energy efficiency problem. However, it is not used in the CMIMO scenario yet. Thus, adopting such methods into CMIMO can be a good way to solve the topology problem.

Adaptive Resource Allocation

In CMIMO systems, adaptive resource allocation can provide robust performance while meeting application-specific requirements. The working principle is to achieve better transmission performance via adaptation of the transmission schemes inlcuding constellation size, coding scheme, power level, and so on. In [2], the authors design an adaptive modulation by changing the modulation constellation size for compensating SNR variations. The transmit power can be adapted by changing the modulation constellation size to meet the BER requirements caused by the variations of SNR. However, in real CMIMO systems, how to facilitate and motivate CMIMO to be adaptive is challenging. In other words, mechanisms to facilitate the adaptation need to be investigated.

Service Differentiation

In CMIMO, energy saving should exploit not only the traffic load variations by considering the node sleeping strategy, but also the variations of QoS requirements. Specifically, in CMIMO-based wire-

less networks, some applications require short delay, whereas some applications are delay-tolerant. Therefore, it is reasonable to differentiate the types of traffic and make the energy consumption scale with the corresponding traffic type. Such service-differentiation-based CMIMO can be a potential candidate for energy reduction.

Conclusions and Future Work

CMIMO schemes are regarded as a major innovation that has potential to fundamentally increase the energy efficiency and maintain the advantages of MIMO schemes. Recently, most developments in CMIMO design target a single technique for energy efficiency. The inherent drawback of these solutions is the lack of techniques combination. For each part in CMIMO, there are different techniques for energy efficiency. Appropriate combination of these techniques will enable CMIMO to achieve significant performance improvements.

In this article, we have reviewed and discussed recent advances in CMIMO transmission schemes. Although some significant works in CMIMO schemes have been done, there is still work to be done on energy efficiency improvement. Through analyzing the characteristics of different cooperation schemes, we have given a comprehensive tutorial on CMIMO to provide a guideline for CMIMO design. In particular, we have shown the comparisons among different CMIMO schemes and the key techniques used in these schemes. To solve the energy efficiency problems in CMIMO, we have listed several challenges as future research directions in this area.

Acknowledgment

This work was supported by an Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korean government(MSIT) (no. 2017-0-00294, service mobility support distributed cloud technology).

References

[1] V. Stankovic, A. Host-Madsen, and Z. Xiong, "Cooperative Diversity for Wireless Ad Hoc Networking," *IEEE Signal Processing*, vol. 23, no. 5, Sept. 2006, pp. 37–49.
[2] S. Cui, A. J. Goldsmith, and A. Bahai, "Energy-Efficiency of MIMO and Cooperative MIMO Techniques in Sensor Networks," *IEEE JSAC*, vol. 22, no. 6, 2004, pp. 1089–98.
[3] Y. Peng and J. Choi, "A New Cooperative MIMO Scheme Based on SM for Energy Efficiency Improvement in Wireless Sensor Networks," *Sci. World J.*, vol. 2014, 2014, article no. 975054.
[4] S. Eryigit *et al.*, "Energy Efficiency Is a Subtle Concept: Fundamental Trade-offs for Cognitive Radio Networks," *IEEE Commun. Mag.*, vol. 52, no. 7, July 2014, pp. 30–36.
[5] Y. Peng *et al.*, "Design and Optimization for Energy-Efficient Cooperative MIMO Transmission in Ad Hoc Networks," *IEEE Trans. Vehic. Tech.*, vol. 66, no. 1, 2017, pp. 710–19.
[6] S. Jayaweera, "Virtual MIMO-Based Cooperative Communication for Energy-Constraint Wireless Sensor Networks," *IEEE Trans. Wireless Commun.*, vol. 5, no. 5, 2006, pp. 984–89.
[7] X. Hong *et al.*, "Cognitive Radio in 5G: A Perspective on Energy-Spectral Efficiency Trade-off," *IEEE Commun. Mag.*, vol. 52, no. 7, July 2014, pp. 46–53.
[8] Y. Peng and C.-H. Youn, "An Energy-Efficiency Cooperative MIMO Transmission with Data Compression in Wireless Sensor Networks," *IEEE J. Trans. Elect. Electron. Eng.*, vol. 10, no. 6, 2015, pp. 729–30.
[9] S. Qiu *et al.*, "BER Adaptive Spatial Modulation with Optimized Switching Thresholds for MIMO Systems," *Proc. IEEE WCSP '15*, 2015.
[10] D. Feng *et al.*, "A Survey of Energy-Efficient Wireless Communications," *IEEE Commun. Surveys & Tutorials*, vol. 15, no. 1, 2013, pp. 167–78.
[11] B. Li *et al.*, "Performance Analysis and Optimization for Energy-Efficient Cooperative Transmission in Random Wireless Sensor Networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 9, 2013, pp. 4647–57.
[12] Y. Peng *et al.*, "Joint Selection for Cooperative Spectrum Sensing in Wireless Sensor Networks," *IEEE Sensor J.*, vol. 16, no. 22, 2016, pp. 7837–38.
[13] J. Zhang *et al.*, "Energy-Efficient Multihop Cooperative MISO Transmission with Optimal Hop Distance in Wireless Ad Hoc Networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 10, 2011, pp. 3426–35.
[14] X. Xu *et al.*, "Energy-Efficiency-Based Optimal Relay Selection Scheme with a BER Constraint in Cooperative Cognitive Radio Networks, *IEEE Trans. Vehic. Tech.*, vol. 65, no. 1, 2016, pp. 191–203.

Biography

Yuyang Peng [M'16] received M.S. and Ph.D. degrees in electrical and electronic engineering from Chonbuk National University, Jeonju, Korea, in 2011 and 2014, respectively. He is currently a postdoctoral research fellow with Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea. His research activities lie in the broad area of digital communications, wireless sensor networks, and computing. In particular, his current research interests include cooperative communications, energy optimization, and cloud computing.

Fawaz Al-Hazemi [SM'16] received his B.Sc. degree in computer engineering from King Fahad University of Petroleum and Minerals, Saudi Arabia, in 2004, and his M.Sc. degree in information and communications engineering from KAIST in 2010. He is currently a research scientist at the School of Electrical Engineering in KAIST. His research interests include energy-aware and green computing in data centers. He is a Senior Member of ACM.

Raouf Boutaba [F'12] received M.Sc. and Ph.D. degrees in computer science from the University Pierre & Marie Curie, Paris, in 1990 and 1994, respectively. He is currently a professor of computer science at the University of Waterloo. His research interests include resource and service management in networks and distributed systems. He is the founding Editor-in-Chief of *IEEE Transactions on Network and Service Management* (2007–2010). He is a Fellow of the Engineering Institute of Canada and the Canadian Academy of Engineering.

Fei Tong received his M.S. degree in computer engineering from Chonbuk National University, Jeonju, Korea, in 2011, and his Ph.D. degree in computer science from University of Victoria, Canada, in 2017. He is currently a postdoctoral research fellow in the Department of Control Science and Engineering, Zhejiang University, Hangzhou, China. His research interests include protocol design and performance analysis for advanced wireless communication networks.

Il-Sun Hwang received his Ph.D. degree in computer science from SungKyunKwan University, Korea, in 2007. Since 1981, he has been a chief researcher with the Department of Supercomputing and Advanced Networks, Korea Institute of Science and Technology Information, Korea. He has been an adjunct professor at the School of Electrical Engineering, KAIST, since 2015. His main research interests include network security, grid computing, and high-performance computing.

Chan-Hyun Youn [S'84, M'87] received his Ph.D. degree in electrical and communications engineering from Tohoku University, Japan, in 1994. He is a professor at the School of Electrical Engineering, KAIST. He was also an associate vice-president of the Office of Planning and Budgets and is a director of the Grid Middleware Research Center, KAIST. His research areas are cloud computing, high-performance computing, and deep learning systems and applications. He was an Editor-in-Chief of *KIPS* and an Editor of the *Journal of Healthcare Engineering*, United Kingdom.

In CMIMO based wireless networks, some applications require short delay whereas some applications are delay tolerant. Therefore, it is reasonable to differentiate the types of traffic and make the energy consumption scale with the corresponding traffic type. Such service differentiation based CMIMO can be a potential candidate for energy reduction.

Energy-Sustainable Traffic Steering for 5G Mobile Networks

Shan Zhang, Ning Zhang, Sheng Zhou, Jie Gong, Zhisheng Niu, and Xuemin (Sherman) Shen

The authors propose an energy-sustainable traffic steering framework, where the traffic load is dynamically adjusted to match with energy distributions in both the spatial and temporal domains by means of inter- and intra-tier steering, caching, and pushing.

ABSTRACT

Renewable EH technology is expected to be pervasively utilized in 5G mobile networks to support sustainable network developments and operations. However, the renewable energy supply is inherently random and intermittent, which could lead to energy outage, energy overflow, QoS degradation, and so on. Accordingly, how to enhance renewable energy sustainability is a critical issue for green networking. To this end, an energy-sustainable traffic steering framework is proposed in this article, where the traffic load is dynamically adjusted to match energy distributions in both the spatial and temporal domains by means of inter- and intra-tier steering, caching, and pushing. Case studies are carried out, which demonstrate that the proposed framework can reduce on-grid energy demand while satisfying QoS requirements. Research topics and challenges of energy-sustainable traffic steering are also discussed.

INTRODUCTION

The fifth generation (5G) mobile networks are expected to connect trillions of devices and provide 1000-fold network capacity by 2020 compared to that in 2010. Network densification (i.e., deploying more small cell base stations [SBSs]) can effectively improve the network capacity through spectrum reuse, and thus is considered as the key cornerstone for 5G. However, network densification may lead to huge energy consumption, causing heavy burdens on network operators. To tackle the cumbersome energy consumption, energy harvesting (EH) technology can be leveraged. Particularly, EH enabled base stations (EH-BSs) can exploit renewable energy as supplementary or alternative power sources to reduce operational expenditures (OPEX). In addition, EH-BSs can be deployed more flexibly without the constraint of power lines. By 2011, over 10,000 EH-BSs were deployed globally, and this figure will increase to more than 400,000 by 2020 [1].

Despite the potential advantages, the inherent randomness of renewable energy poses significant technical challenges to network operations. Specifically, the mismatch between harvested energy and traffic distributions may result in energy outage and/or energy overflow, degrading the quality of service (QoS) and energy utilization. Thus, in addition to energy efficiency, a new performance measure, "energy sustainability," should be introduced to keep the energy outage and energy overflow probabilities as low as possible [2]. To this end, we propose a traffic steering framework to enhance energy sustainability in networks with EH-BSs.

Traffic steering goes beyond traffic offloading, which proactively adjusts traffic distribution to match with and better utilize network resources, aiming to enhance network performance or providing better QoS [3].

The proposed framework encompasses three approaches: inter-tier steering, intra-tier steering, and content caching and pushing. Specifically, inter-tier steering adjusts the traffic load of each tier according to the variation of renewable energy arrival rate in a large timescale. In addition, intra-tier steering shifts traffic among neighboring BSs to further break down traffic load in the spatial domain, while content caching and pushing reshape temporal traffic load to overcome small timescale randomness, based on the instant energy status. Together, these three approaches can match traffic demand to renewable energy supply in both the spatial and temporal domains, reducing the probabilities of energy outage and overflow. Therefore, the proposed framework provides two-fold benefits of greenness and QoS provisioning, achieving energy-sustainable networking.

The remainder of this article is organized as follows. An overview of 5G networks with EH is first presented in the following section. Then the energy-sustainable traffic steering framework is introduced, including detailed methods, research topics, and challenges. Case studies are conducted to reveal the effectiveness of energy-sustainable traffic steering, followed by the conclusions.

HETEROGENEOUS NETWORKS WITH ENERGY HARVESTING

In this section, we introduce EH-enabled networks, including architecture, challenges, existing solutions, and limitations.

EH-ENABLED 5G NETWORK ARCHITECTURE

With the promising EH technologies leveraged, the resulting 5G network architecture is shown in Fig. 1. Particularly, SBSs can be further classified into three types based on the functions and power sources:

Digital Object Identifier:
10.1109/MCOM.2017.1700022

Shan Zhang and Xuemin (Sherman) Shen are with the University of Waterloo; Ning Zhang is with Texas A&M University-Corpus Christi; Sheng Zhou and Zhisheng Niu are with Tsinghua University; Jie Gong is with Sun Yat-Sen University

0163-6804/17/$25.00 © 2017 IEEE

- Off-grid EH-SBSs, powered solely by renewable energy without access to power grid
- On-grid EH-SBSs, powered jointly by power grid and renewable energy
- Conventional SBSs, powered only by power grid

Notice that the three types of SBSs have distinct features. Off-grid EH-SBSs enable the most flexible deployment, whereas the QoS cannot be guaranteed well due to the unstable energy supply. Thus, off-grid EH-SBSs can be deployed for opportunistic traffic offloading for macro BSs (MBSs). On-grid EH-SBSs provide reliable services with on-grid power as backup sources, which can also use the renewable energy to reduce the OPEX. However, they bring the highest capital expenditure due to the EH modules and wired connections to power grid. Conventional SBSs are the moderate option, which guarantee QoS requirements but with the highest on-grid power consumption.

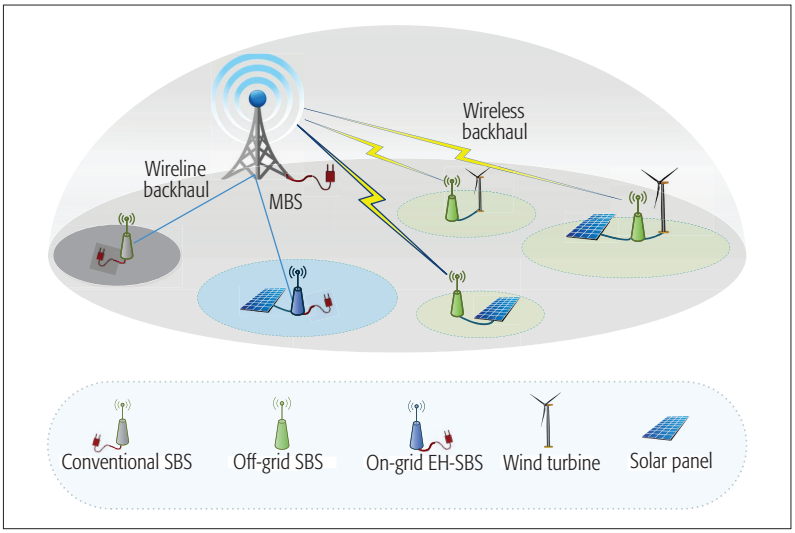


Figure 1. Network architecture with renewable energy harvesting.

Mismatched Traffic and Energy

As the cellular systems are expected to provide reliable services with guaranteed QoS, the power supply and demand of each BS should be balanced. For conventional BSs, the on-grid power supply can be dynamically adjusted based on the traffic variations. However, it is much more challenging for EH-BSs, as the renewable energy arrival is usually mismatched with the traffic demand. For example, the two-day traffic and renewable energy variations are shown in Fig. 2, wherein the traffic profile is obtained from real data measurement in the EARTH project [4], and the renewable power profile is collected by the Elia group [5]. The mismatch of renewable energy and traffic load variations may bring following problems.

Renewable Energy Outage: This happens when the energy arrival rate is lower than the BS power demand, which may cause additional on-grid power consumption or degrade QoS. For example, off-grid EH-BSs even have to be shut down when the renewable energy is insufficient to support their static power need [4]. Although multiplexing diverse renewable energy sources helps to improve the reliability, the randomness still exists, and energy outage cannot be avoided.

Renewable Energy Overflow: This occurs when the renewable energy is oversupplied compared to the traffic demand. To address this problem, batteries can be used to store the redundant energy for future use. However, in practical systems, the battery capacity is usually limited due to high cost, and energy overflow is still inevitable.

Spatial Supply-Demand Imbalance: This is due to the diverse energy sources of different BSs. Furthermore, renewable energy is non-uniformly distributed in the spatial domain, which may be mismatched with the traffic load in that domain. Thus, neighboring BSs can have imbalanced renewable energy supply across the network, leading to inefficient network-wide renewable energy utilization.

Energy-Sustainable Networking

The challenges of renewable energy dictate that the network design criterion should shift from minimizing the total energy consumption toward energy sustainability, that is, to sustain traffic

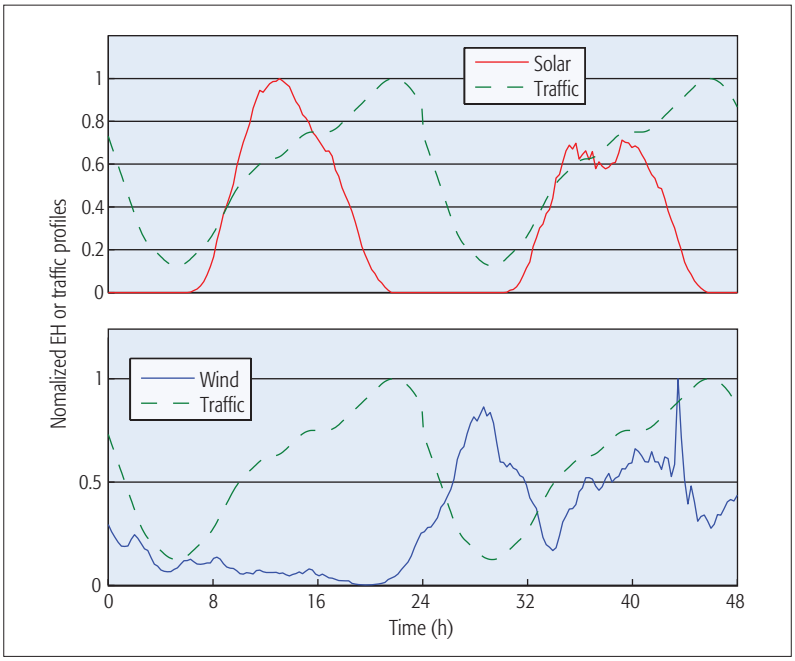


Figure 2. Daily energy harvesting and traffic demand profiles.

while satisfying the QoS requirements with energy dynamics. Specifically, energy-sustainable networking improves the utilization of renewable energy to mitigate energy outage and overflow, which enables better QoS provisioning and reduces on-grid power demand. To achieve this goal, the energy supply and the traffic demand should be matched with each other in both the spatial and temporal domains [6].

Existing research mostly focuses on the energy management perspective, which reshapes energy supply to match the given traffic distribution. The methods can be mainly classified into two categories.

BS-Level Energy Allocation: Through dynamic charging and discharging, the renewable energy can be reallocated in the temporal domain to match the time-varying traffic load. For example, effective online and offline energy scheduling schemes have been proposed to minimize the

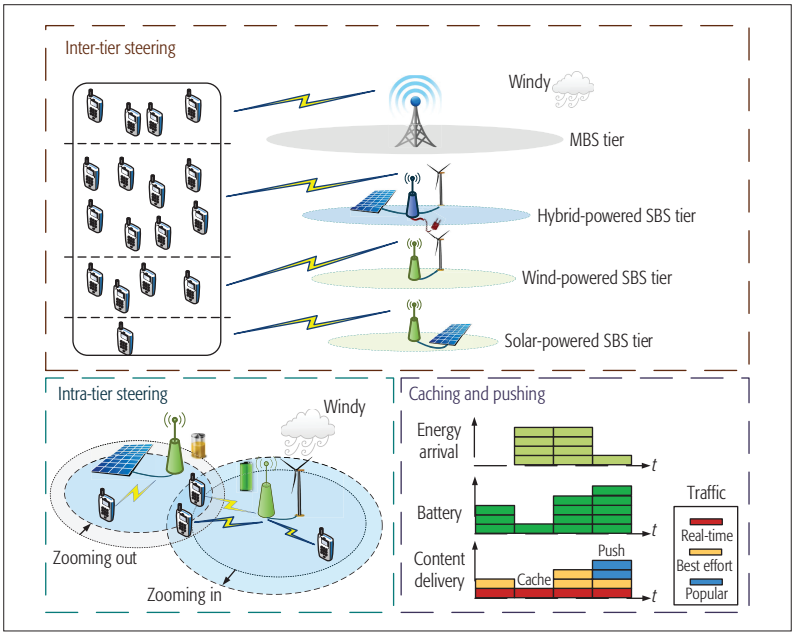


Figure 3. Energy-sustainable traffic steering framework.

on-grid power consumption of a single EH-BS, assuming infinite battery capacity [7]. However, in real systems, the performance of these methods may be degraded due to battery limitations.

Network-Level Energy Cooperation: Through energy transfer among EH-BSs, renewable energy can be redistributed across the network to match the traffic load, which can further reduce renewable energy waste [8]. However, the EH-BSs need to be connected through either dedicated power lines or smart grids to implement this approach [9]. Besides, the two-way power transmission among EH-BSs may cause the loss of renewable energy.

Notice that the existing methods of energy management manifest their limitations in terms of performance and prerequisite power infrastructure; hence, we seek solutions from the traffic steering perspective, which is more flexible and can easily be realized through control signaling without any additional deployment of power infrastructures. Our previous work was the first to adopt EH-SBS traffic offloading to tackle renewable energy dynamics [10]. In this article, we propose a comprehensive traffic steering framework, where traffic is manipulated dynamically in different spatial and temporal scales to match the renewable energy supply.

Energy-Sustainable Traffic Steering

In this section, we first introduce the concept and applications of traffic steering, and then propose an energy-sustainable traffic steering framework to realize traffic-energy matching in heterogeneous networks (HetNets) with EH.

Traffic Steering Concept

The heterogeneous 5G networks call for sufficient utilization of available resources to support the dynamic and differentiated traffic demand. However, the conventional user association method is mainly based on received signal-to-interference-plus-noise ratio (SINR), which can fail to meet this requirement. To deal with this challenge, traffic steering redistributes traffic load across the network based on radio resources to optimize the performance of networks and end users. As user association goes beyond "SINR-based" to "resource-aware," traffic steering can effectively enhance network utility through appropriate traffic-resource matching.

The objectives and policies of traffic steering can be diverse. For example, traffic can be steered from heavily loaded cells to lightly loaded ones for load balancing, aiming to maximize network capacity. Besides, in networks with multiple radio access technologies (RATs), users can be steered to different RATs according to their mobility, such that call dropping probability can be reduced. In addition, traffic steering can also be performed in the temporal domain through transmission scheduling and rate control. For instance, the transmission of best effort traffic can be postponed (i.e., steered to future time slots) when the traffic demand exceeds network capacity.

Traffic Steering with EH

With EH leveraged, the key challenge is that the service capability of a BS can vary dynamically with renewable energy supply, which requires the traffic load to be matched with the corresponding service capability to fully utilize renewable energy. To this end, we propose an energy-sustainable traffic steering framework that encompasses three main approaches, as shown in Fig. 3. Inter-tier steering optimizes the amount of traffic steered to different tiers over a large timescale, based on the statistic information of renewable energy supply. Then intra-tier steering, caching, and pushing dynamically reshape traffic load on smaller scales to achieve fine-grained traffic-energy matching. Specifically, intra-tier steering adjusts the BS-level traffic load to the corresponding energy supply through cooperation among multiple neighboring cells. Meanwhile, caching and pushing schedule content delivery by exploiting the content information and differentiated QoS requirements, which can be conducted independently at each BS to further deal with the small timescale randomness of energy and traffic dynamics.

Inter-Tier Steering: As shown in Fig. 1, BSs can be further divided into multiple tiers based on power sources, cell type, and other system parameters. Each tier has different service capabilities, which can vary dynamically with renewable energy arrival rate. Energy-sustainable inter-tier traffic steering dynamically optimizes the amount of traffic steered to each network tier, based on the information of energy supply and other system parameters. Intuitively, more traffic can be steered to the EH-SBS tiers with sufficient energy for service, which reduces the on-grid power consumption of other tiers. On the contrary, EH-SBSs with insufficient energy supply can reduce transmit power consumption by serving fewer users to maintain the power balance (e.g., solar-powered SBSs on a cloudy day). In addition, EH-SBSs can be deactivated when the power supply is even lower (e.g., solar-powered SBSs at midnight), which can further save the static power consumption (e.g., air conditioning) to enhance renewable energy sustainability.

With intelligent inter-tier traffic steering, the traffic load can be reshaped in both the spatial

and temporal domains simultaneously. In the spatial domain, the traffic load of each tier is optimized based on their service capability, as shown in Fig. 3. From the whole network perspective, the traffic load supported by different energy sources is also dynamically adjusted with on-grid BSs serving as backups, which in fact realize temporal traffic-energy matching.

Intra-Tier Steering: Intra-tier traffic steering further adjusts BS-level traffic load through methods of cell zooming and BS cooperation, which further break up traffic load in the spatial domain based on energy status. With cell zooming, the coverage of a BS can be enlarged (i.e., zoom in) or shrunk (i.e., zoom out) to adjust the corresponding traffic load through transmit power control and antenna tilting. In conventional on-grid networks, traffic is generally steered from heavily loaded cells to lightly loaded ones, leading to load-dependent cell size. With EH technology implemented, EH-BSs should further adjust cell size based on the corresponding energy supply. For example, the EH-BSs with oversupplied energy can zoom in to assist the transmission of neighboring BSs by utilizing the redundant harvested energy. In reward, some EH-BSs encountering energy shortage can shrink their coverage to reduce their traffic load (i.e., zooming out). By doing so, traffic is steered to EH-SBSs with redundant renewable energy, reducing service outage and battery overflow.

In addition to cell zooming, energy-aware cooperative transmission can also be applied. As networks become ultra dense, a mobile user may be covered and served by multiple BSs simultaneously. In this case, the cooperative transmission of these BSs can be optimized based on their energy status in addition to channel condition. For instance, BSs with insufficient renewable energy or large path loss can reduce transmit power or even turn off, while others enlarge transmit power to guarantee QoS.

Caching and Pushing: Caching and pushing aim to schedule content delivery in an optimal manner by exploiting the information of contents, user preference, and differentiated QoS requirements, which can reshape traffic load to deal with the small timescale randomness of renewable energy arrival.

The transmission of non-real-time traffic can be postponed during energy shortage periods by caching corresponding contents at a BS. For example, the packet can be delivered with lower transmission rate to reduce energy consumption as long as the given deadline is satisfied, while the transmission of best effort traffic can be delayed until the energy is sufficient. The main idea of proactive pushing is that the EH-BSs can deliver the popular contents (e.g., videos and news) in a multicast manner before user request, when the energy is sufficient or oversupplied. By storing the contents at end-user devices, the amount of data transmission can be reduced. In fact, the demand of video streaming, currently accounting for over 50 percent of wireless traffic and still increasing, is expected to dominant mobile data service. Surprisingly, 10 percent of the most popular videos receive nearly 80 percent of views. Therefore, proactive pushing can effectively reduce the future traffic load.

Through energy-sustainable caching and pushing, the traffic load is equivalently steered from periods of energy shortage to periods of energy oversupply. This traffic reshaping enables traffic-energy matching in smaller timescales, further enhancing energy utilization.

Research Challenges

Based on the design scales, inter-tier steering, intra-tier steering, caching, and pushing can be conducted at the hour, minute, and second levels, respectively, providing both coarse- and fine-grained network optimization. As such, energy-sustainable traffic steering can match traffic demand to energy supply flexibly in the spatial and temporal domains, providing two-fold benefits of greenness and QoS. To realize the potential advantages, several research challenges to implement energy-sustainable traffic steering are discussed as follows.

Service Capability with EH: Since traffic steering is done mainly to match the traffic load to the corresponding service capability from different spatial-temporal scales, a fundamental problem is to analyze the EH-enabled service capability of each BS and network tier. In the existing literature, the link-level channel capacity has been extensively studied, with an EH-enabled transmitter and a single or multiple receivers [11]. However, network-level analysis can be much more challenging due to factors such as random traffic distribution, user mobility, differentiated QoS requirements, inter-cell interference, and HetNet architecture. Stochastic geometry can be applied to QoS analysis of large-scale networks, based on the information of traffic distribution and network topologies. For conventional on-grid HetNets, network capacity has been investigated with respect to different QoS performance metrics, such as coverage probability and user achievable rate. Future work should revisit these issues considering the influence of renewable energy supply in both large and small timescales [12].

Low-Complexity Operation: Traffic steering in the spatial domain needs the cooperation of multiple cells from the same or different tiers, where the BS activation/deactivation, power control, and user association should be jointly optimized based on the instant information of traffic distribution and renewable energy supply. The problem is to minimize the long-term on-grid energy consumption subject to QoS requirements and power constraints, which can be formulated as a mixed-integer programming problem. The existing literature mainly focus on traffic steering optimization for small-scale networks, whereas the operational complexity can increase exponentially with network scale due to the coupled operation of different cells [13]. As future mobile networks are expected to be ultra-densely deployed with massive device connections, low-complexity steering schemes are critical for practical implementation. To this end, the BS operations can be decoupled in both the temporal and spatial domains. In the temporal domain, the deactivation of BSs can be determined over a large timescale based on the statistic information of energy and traffic arrivals, and then each active BS can further schedule the content delivery over a small timescale based on the instant states. In the spatial domain, the con-

Caching and pushing aim to schedule content delivery in an optimal manner by exploiting the information of contents, user preference, and differentiated QoS requirements, which can reshape traffic load to deal with the small timescale randomness of renewable energy arrival.

Existing studies have designed spatial traffic steering schemes based on a semi-static traffic model, and also proposed dynamic content caching and pushing schemes from a single BS perspective. The joint optimization of spatial and temporal traffic steering offers opportunities to further improve network performance.

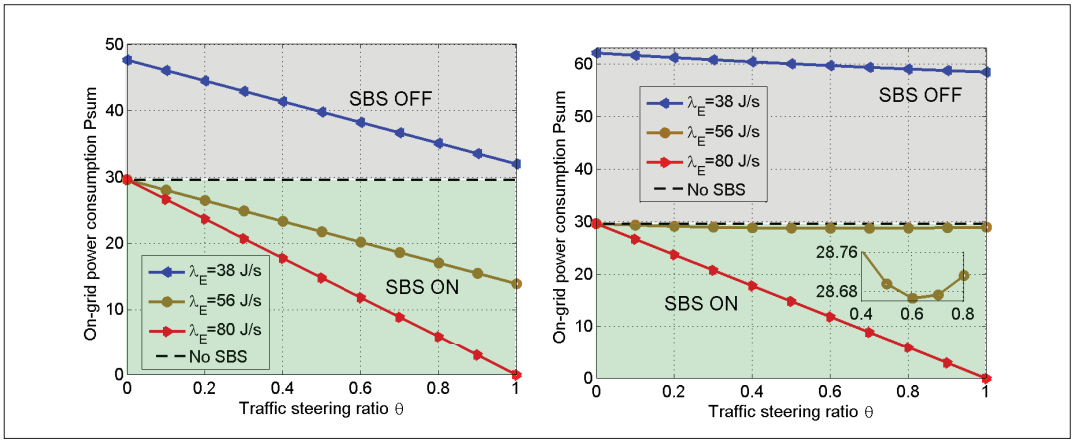


Figure 4. Power consumption with respect to traffic steering ratio: a) on-grid EH-SBS; b) off-grid EH-SBS.

cept of self-organized networking (SON) can be applied to design effective distributed traffic steering schemes, such that BSs can make decisions independently with local information [14].}}

Joint Spatial-Temporal Optimization: Existing studies have designed spatial traffic steering schemes based on a semi-static traffic model [10], and also proposed dynamic content caching and pushing schemes from a single BS perspective. The joint optimization of spatial and temporal traffic steering offers opportunities to further improve network performance. In this case, mobile operators have a higher degree of freedom to reshape traffic load. For instance, the oversupplied energy at a BS can be utilized for:

1. Zooming in to assist neighboring cells
2. Serving more traffic from upper-tier on-grid MBSs
3. Pushing popular contents to users proactively
4. Storing the redundant energy in battery for future use

The Markov decision process (MDP) method can be applied to design optimal steering policies through theoretical analysis of small-scale networks, which can provide a guideline to practical network operation. Furthermore, advanced machine learning technologies can be implemented to devise a spatial-temporal traffic steering method for large-scale networks based on the big data analysis in real systems [15, 16].

Trade-off between System and User Performance: Although traffic steering helps to improve network performance, the interests of end-user devices may be degraded. Some end devices may not be connected with the preferred BS after being steered in the spatial domain, and therefore they have to increase the transmit power to maintain the uplink QoS. Accordingly, the batteries of end devices may be consumed quickly. Moreover, for temporal traffic steering, proactive pushing and caching also consume the resources of end devices, which may be unacceptable from the perspective of mobile users. Existing traffic steering mainly focuses on the performance optimization from the network perspective, while the trade-off between network and user profits has not been well studied. For practical implementation, incentive schemes need to be designed to compensate the performance degradation of steered users, whereby the agreement can be achieved between network operators and mobile users. Such problems can be modeled as two-player or multi-player games, and game theory can be applied to seek equilibria among players.

Case Study: Inter-Tier Steering

In this section, a case study is carried out on inter-tier steering to illustrate practical implementation in detais. We consider a two-tier HetNet consisting of a conventional MBS with radius 1000 m and multiple EH-SBSs with coverage radius 100 m. The available bandwidths for the MBS and SBS are set as 10 MHz and 2 MHz, respectively. The traffic distribution is modeled as a Poisson point process (PPP) with density 1500 /km^2, and the average rate per user should be no smaller than 500 kb/s. Traffic within each small cell can be partially or completely steered from the MBS to the EH-SBS depending on the renewable energy arrival rate. Specifically, we analyze how much traffic should be steered to match the EH-dependent service capability, from the perspective of a typical EH-SBS. Define by θ the steering ratio, that is, the percentage of traffic steered to the EH-SBS within the small cell. For the given renewable energy arrival rate, we demonstrate the on-grid power consumption variation with respect to θ to investigate the influence of energy supply on the BS-level service capability and the optimal traffic steering volume.

Consider the discrete power consumption model, with per unit energy set as 1 J. The energy arrival at an EH-SBS is modeled as Poisson process with rate λ_E for tractable analysis, which is saved at the battery for future use. Both on-grid and off-grid EH-SBSs are considered, which are assumed be to 500 m away from the MBS. For the on-grid EH-SBS, it can consume either renewable or on-grid power, but the on-grid power can only be used as backup when the battery is empty. The off-grid EH-SBS has to shut down when the battery becomes empty, and meanwhile, all traffic is steered to the MBS through a handover procedure, causing additional handover cost. The BS power consumption model is based on the EARTH project [4], and power control can be conducted by partially deactivating the subframes. For the wireless channel model, the path loss exponent, noise, and interference densities are set as 3.5, −105, and −100 dBm/MHz, respectively. Denote by P_{sum} the on-grid power required to serve the traffic located within

the coverage of a small cell. Under different energy arrival rate λ_E, the relationship between P_{sum} and the steering ratio θ is shown in Fig. 4, with the handover cost set as 1.5 J. For comparison, the dashed lines show the power consumption if the EH-SBS is not deployed and all traffic is served by the MBS. Therefore, activating an EH-SBS for traffic steering even increases power consumption in the regions above the dashed lines, and thus the EH-SBSs should be completely deactivated.

The on-grid power consumption P_{sum} is shown to decrease with θ when the on-grid EH-SBS is active, indicating that the active on-grid EH-SBS should serve traffic as much as possible. Furthermore, P_{sum} decreases with λ_E, as more renewable energy can be utilized. However, steering traffic to the on-grid EH-SBS does not always save energy, even under the optimal steering ratio. For example, the minimal on-grid power consumption is around 32 W when λ_E = 38 J/s (with θ = 1), while only 30 W is needed if the EH-SBS is not activated, shown as the dashed lines in Fig. 4a. The reason is due to the trade-off between transmission and static power consumption. On one hand, steering traffic to the SBS helps to reduce transmission power, as the SBS provides higher spectrum efficiency due to the short transmission path. On the other hand, the active SBS also requires static power independent of data transmission [4], which may increase the total on-grid power consumption when renewable energy is insufficient. This trade-off is influenced by the renewable energy arrival rate, which directly determines the network power consumption. Therefore, the on-grid EH-SBSs should be turned off under low energy arrival rate, whereas they should try to serve as much traffic as possible once activated.

As for the off-grid EH-SBS, the optimal ON-OFF states also depend on the renewable energy arrival rate. However, the optimal traffic steering ratio may not be 1. As an example, the energy-optimal steering ratio is 0.6 when the renewable energy arrival rate is 56 W. This is due to the trade-off between the handover cost and the power saved at the MBS. With more traffic steered to EH-SBSs, the power consumption of the MBS is reduced, whereas the renewable energy of the EH-SBSs can be used up more quickly if insufficient, leading to frequent user handover and extra cost. This trade-off determines the optimal steering ratio and ON-OFF status of the off-grid EH-SBS, based on the renewable energy arrival rate. Furthermore, the optimal steering ratio indicates the service capability of EH-SBSs, which increases with renewable energy arrival rate. When the energy arrival rate is high (e.g., λ_E = 80 J/s), the EH-SBS should serve all traffic within coverage (i.e., θ = 1), reflecting high service capability. On the contrary, the EH-SBS should be deactivated and serve no traffic (i.e., θ = 0) under low energy arrival rate (e.g., λ_E = 38 J/s), reflecting zero service capability.

To evaluate the effectiveness of energy-sustainable traffic steering, we illustrate the daily power consumption profiles of an on-grid EH-SBS under different traffic steering schemes. The traffic and solar energy profiles shown in Fig. 2 (the first day) are adopted, with the peak energy arrival rate and traffic load set as 100 λ_E/s and 3000 θ/km^2, respectively. The greedy algorithm keeps the EH-SBS on to serve all the traffic within coverage. The proposed energy-sustainable scheme dynamically adjusts the ON/OFF states and the traffic load of the EH-SBSs to minimize the on-grid power consumption, based on the results of Fig. 4. The temporal variations of on-grid power consumption under the two schemes are demonstrated in Fig. 5. Specifically, it is shown that the proposed energy-sustainable traffic steering scheme can reduce the on-grid power consumption by 48 percent on average, compared to the greedy scheme. Furthermore, the proposed algorithm is more effective during low energy hours.

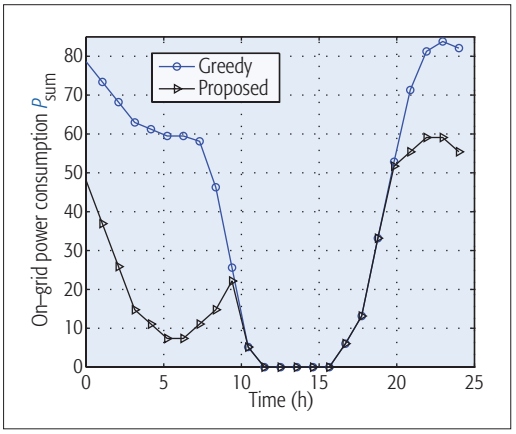


Figure 5. Daily on-grid power consumption under different traffic steering schemes.

Although traffic steering helps to improve network performance, the interests of end user devices may be degraded. Some end devices may not be connected with the preferred BS after steered in spatial domain, and thereby they have to increase the transmit power to maintain the uplink QoS. Accordingly, the battery of end devices may be quickly consumed.

Conclusions

In this article, an energy-sustainable traffic steering framework has been proposed to address the sustainability issue of 5G networks by encompassing three approaches:

- Inter-tier steering
- Intra-tier steering
- Content caching and pushing

The proposed framework can better balance the power demand and supply of individual EH-BSs by matching the traffic load and renewable energy distribution in both the spatial and temporal domains. The case study on inter-tier offloading has demonstrated that the ON-OFF status and the steered traffic load of each EH-SBS should be adapted to the corresponding energy-dependent service capability, such that the on-grid power consumption can be effectively reduced. Future research topics and challenges are also discussed.

Acknowledgment

This work is sponsored in part by the National Nature Science Foundation of China (No. 91638204, No. 61571265, No. 61461136004), Hitachi R&D Headquarters, and Natural Sciences and Engineering Research Council of Canada.

References

[1] R. Martin, "Nearly 400,000 Off-Grid Mobile Telecommunications Base Stations Employing Renewable or Alternative Energy Sources Will Be Deployed from 2012 to 2020," Navigant Research, tech. rep., Feb. 2013; http://www.navigantresearch.com/newsroom/nearly-400000-off-grid-mobile-telecommunications-base-stations-employing-renewable-or-alternative-energy-sources-will-be-deployed-from-2012-to-2020, acessed Nov. 4, 2016.

[2] L. X. Cai *et al.*, "Dimensioning Network Deployment and Resource Management in Green Messh Networks," *IEEE Wireless Commun.*, vol. 18, no. 5, Oct. 2011, pp. 58–65.
[3] Nokia, "Business Aware Traffic Steering," tech. rep., Feb. 2015; http://info.networks.nokia.com/Business-aware-traffic-steering-LP.html, accessed Nov. 4, 2016.
[4] G. Auer *et al.*, "D2.3: Energy Efficiency Analysis of the Reference Systems, Areas of Improvements and Target Breakdown," EARTH Project, tech. rep., Nov. 2010; https://www.ict-earth.eu/publications/deliverables/deliverables.html, accessed Nov. 4, 2016.
[5] Elia, Power Generation; http://www.elia.be/en/grid-data/power-generation, accessed Nov. 2, 2016.
[6] Z. Zheng *et al.*, "Sustainable Communication and Networking in Two-Tier Green Cellular Networks," *IEEE Wireless Commun.*, vol. 21, no. 4, Aug. 2014, pp. 47–53.
[7] D. W. K. Ng, E. S. Lo, and R. Schober, "Energy-Efficient Resource Allocation in OFDM Systems with Hybrid Energy Harvesting Base Station," *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, July 2013, pp. 3412–27.
[8] Y.-K. Chia, S. Sun, and R. Zhang, "Energy Cooperation in Cellular Networks with Renewable Powered Base Stations," *IEEE Trans. Wireless Commun.*, vol. 13, no. 12, Dec. 2014, pp. 6996–7010.
[9] S. Bu *et al.*, "When the Smart Grid Meets Energy-Efficient Communications: Green Wireless Cellular Networks Powered by the Smart Grid," *IEEE Trans. Wireless Commun.*, vol. 11, no. 8, Aug. 2012, pp. 3014–24.
[10] S. Zhang et al., "Energy-Aware Traffic Offloading for Green Heterogeneous Networks," *IEEE JSAC*, vol. 34, no. 5, May 2016, pp. 1116–29.
[11] O. Ozel and S. Ulukus, "Achieving AWGN Capacity Under Stochastic Energy Harvesting," *IEEE Trans. Info. Theory.*, vol. 58, no. 10, Oct. 2012, pp. 6471–83.
[12] H. S. Dhillon *et al.*, "Fundamentals of Heterogeneous Cellular Networks with Energy Harvesting," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, May 2014, pp. 2782–97.
[13] P. Li *et al.*, "On Efficient Resource Allocation for Cognitive and Cooperative Communications," *IEEE JSAC*, vol. 32, no. 2, Feb. 2014, pp. 264–73.
[14] Y. Zhou, and W. Zhuang, "Performance Analysis of Cooperative Communication in Decentralized Wireless Networks with Unsaturated Traffic," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, May 2016, pp. 3518–30.
[15] J. Wu *et al.*, "Big Data Meet Green Challenges: Big Data Toward Green Applications," *IEEE Sys. J.*, vol. 10, no. 3, Sept. 2016, pp. 888–900.
[16] X. Chen *et al.*, "Energy-Efficiency Oriented Traffic Offloading in Wireless Networks: A Brief Survey and a Learning Approach for Heterogeneous Cellular Networks," *IEEE JSAC*, vol. 33, no. 4, Apr. 2015, pp. 627–40.

Biographies

Shan Zhang [M] (s372zhan@uwaterloo.ca) received her Ph.D. degree from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2016. She is currently a postdoctoral fellow in the Department of Electronical and Computer Engineering, University of Waterloo, Ontario, Canada. Her research interests include resource and traffic management for green communication, intelligent vehicular networking, and software defined networking. She received the Best Paper Award at the Asia-Pacific Conference on Communication in 2013.

Ning Zhang [M] (ning.zhang@tamucc.edu) received his Ph.D degree from the University of Waterloo in 2015. He is now an assistant professor in the Department of Computing Science at Texas A&M University-Corpus Christi. Before that, he was a postdoctoral research fellow first at the University of Waterloo and then at the University of Toronto. He was the co-recipient of the Best Paper Awards at IEEE GLOBECOM 2014 and IEEE WCSP 2015. His current research interests include next generation wireless networks, software defined networking, vehicular networks, and physical layer security.

Sheng Zhou [M] (sheng.zhou@tsinghua.edu.cn) received his B.S. and Ph.D. degrees in electronic engineering from Tsinghua University in 2005 and 2011, respectively. He is currently an associate professor in the Electronic Engineering Department, Tsinghua University. From January to June 2010, he was a visiting student at the Wireless System Lab, Electrical Engineering Department, Stanford University, California. His research interests include cross-layer design for multiple-antenna systems, cooperative transmission in cellular systems, and green wireless communications.

Jie Gong [M] (xiaocier@gmail.com) received his B.S. and Ph.D. degrees from Department of Electronic Engineering, Tsinghua University in 2008 and 2013, respectively. He is currently an associate research fellow at the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, Guangdong Province, China. He was a co-recipient of the Best Paper Award from the IEEE Communications Society Asia-Pacific Board in 2013. His research interests include cloud RAN, energy harvesting, and green wireless communications.

Zhisheng Niu [F] (niuzhs@tsinghua.edu.cn) graduated from Beijing Jiaotong University, China, in 1985, and got his M.E. and D.E. degrees from Toyohashi University of Technology, Japan, in 1989 and 1992, respectively. During 1992–1994, he worked for Fujitsu Laboratories Ltd., Japan, and in 1994 joined Tsinghua University, where he is now a professor in the Department of Electronic Engineering. He is also a guest chair professor of Shandong University, China. His major research interests include queueing theory, traffic engineering, mobile Internet, radio resource management of wireless networks, and green communication and networks.

Xuemin (Sherman) Shen [F] (xshen@bbcr.uwaterloo.ca) is a university professor, Department of Electrical and Computer Engineering, University of Waterloo. He is also the associate chair for graduate studies. His research focuses on resource management, wireless network security, social networks, smart grid, and vehicular ad hoc and sensor networks. He was an elected member of the IEEE ComSoc Board of Governors and the Chair of the Distinguished Lecturers Selection Committee. He has served as the Technical Program Committee Chair/Co-Chair for IEEE GLOBECOM '16, IEEE INFOCOM '14, IEEE VTC-Fall '10, and GLOBECOM '07. He received the Excellent Graduate Supervision Award in 2006, and the Outstanding Performance Award in 2004, 2007, 2010, and 2014 from the University of Waterloo. He is a registered professional engineer of Ontario, Canada, an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering Fellow, and a Royal Society of Canada Fellow, and was a Distinguished Lecturer of the IEEE Vehicular Technology Society and the IEEE Communications Society.

Increase Your Knowledge of Technical Standards

The foundation for today's global high-tech success

Prepare yourself to enter the high-tech industry by learning how technical standards develop, grow, and relate to technologies worldwide.

IEEE Standards Education is dedicated to helping engineers and students increase their knowledge of technical standards, applications, and the impact standards have on new product designs.

Begin your journey into the high-tech world: http://trystandards.org

Stay current with access to regularly updated educational programs

Discover learning opportunities through tutorials and case studies

Understand the role of technical standards in the world

For information about IEEE Standards Education, visit http://standardseducation.org

IEEE
Advancing Technology for Humanity

A Software-Defined Green Framework for Hybrid EV-Charging Networks

Yanfei Sun, Xiaoxuan Hu, Xiulong Liu, Xiaoming He, and Kun Wang

The authors propose a software-defined green framework for hybrid EV charging networks, which consists of three planes: the application plane provides customized services for users; the control plane aims to guide both data flow and energy flow to implement an efficient and economic strategy for EV charging; and the physical plane collects the information from a large number of EVs and other infrastructures.

ABSTRACT

EVs provide a promising green solution to reduce dependence on fossil fuel and the emission of greenhouse gas. This article concerns hybrid EV systems that contain both wired charging and wireless charging vehicles. To the best of our knowledge, we are the first to discuss how to jointly use these two types of charging methods whose advantages are complementary to each other; for example, wireless charging has better user friendliness and flexibility, while wired charging can provide higher charging efficiency and bidirectional energy transmission. We believe such a hybrid system can benefit us much more than either a pure wireless or pure wired EV charging system. On the other hand, we also pay attention to a practically important problem with DSM. The existing schemes can hardly ensure efficient and stable charging services in the foreseeable future due to the rapidly increasing number of electric vehicles and the limited capacity of local grids. To address this issue, we propose a software-defined green framework for hybrid EV charging networks, which consists of three planes: the application plane provides customized services for users; the control plane aims to guide both data flow and energy flow to implement an efficient and economic strategy for EV charging; and the physical plane collects the information from a large number of EVs and other infrastructures. Furthermore, we present a DSM approach for EV charging as a case study. Simulation results reveal that our proposals could achieve delightful DSM performance in the proposed software-defined EV charging networks. Although the proposed hybrid framework offers promising opportunities, we still face some technical challenges that solicit further research efforts from both academia and industry, which are discussed in detail at the end of this article.

INTRODUCTION

With the increasing attention to the shortage of fossil fuel and the high carbon dioxide emissions, the electric vehicle (EV) is being considered as one of the most important green solutions to these problems [1]. Compared to conventional vehicles, EVs are more economical and environmentally friendly. In addition, with the popularization of EVs, renewable energy, such as solar, wind energy, and hydro energy, can be fully utilized for charging EVs to further cut cost. However, due to the increasing penetration of EVs, unregulated EV-charging management may lead to unexpected load on power grids and make the grids break down. In order to satisfy the quick charging requirements, the power transmission capacity will be increased, thus raising the infrastructure cost of power grids [2]. To solve this problem, one popular method is to use EVs to support the power supply to the grid when they are parked. EVs can achieve a general balance of power load through peak shaving and valley filling [3]. However, there is still a huge gap between the charging demand and the actual service capability at rush hours, which hinders the users from enjoying stable energy supply in power grids.

Generally, there are two types of charging technologies for EVs: wired charging and wireless charging. Wired charging technologies are widely applied in the EV transportation network, and they can provide high charging efficiency and bidirectional energy transmission. However, the restricted capacity of battery and the number of charging stations limit the driving range of EVs. In contrast, with the development of wireless power transfer technologies, wireless charging becomes an alternative for EV charging in order to extend the driving range [4]. Theodoropoulos *et al.* [5] studied the potential of a feasible demand-side management (DSM) scheme for EV wireless charging. With the advantages of wireless charging, the charging process can be implemented in a both simple and economical way. Compared to traditional wired charging, wireless charging technologies have many benefits. First, the wireless charging technique provides convenience and improves user friendliness. Second, it provides better product durability and enhances flexibility for EVs [6]. Despite these advantages, compared to the wired charging technologies, the wireless charging technologies can only provide unidirectional ancillary services (e.g., frequency regulation) for mitigating negative impacts on the power grid, making the power grid more challenging to implement a stable and efficient energy supply. Clearly, these two types of charging techniques are complementary to each other. Combining the wired charging tech-

Digital Object Identifier:
10.1109/MCOM.2017.1601190

Yanfei Sun, Xiaoxuan Hu, and Xiaoming He are with Nanjing University of Posts and Telecommunications; Xiulong Liu is with Hong Kong Polytechnic University; Kun Wwang is with Nanjing University of Posts and Telecommunications and also with Hong Kong Polytechnic University. Corresponding author: Kun Wang

0163-6804/17/$25.00 © 2017 IEEE

nologies with wireless charging technologies has the potential to make full use of the advantages of both charging technologies, achieving bidirectional energy transmission and large driving range.

Nevertheless, both the real-time status of power grid and the requirements of EVs should be taken into consideration so that the charging operations can be adjusted in time and the traditional peak loads can be avoided at the same time. Recently, advanced communication technologies have been applied to many scenarios, such as power grid and vehicular networks, enabling a large number of entities to exchange information without region restriction. Hence, the integration of both wired charging technologies and wireless charging technologies into power grid needs a comprehensive communication technology to implement DSM for EVs and the stability of power grid at the same time. Software-defined networking (SDN) [7] is such a promising paradigm, which gives hope of satisfying the demands of EV charging operations. Since the emergence of SDN, it has been applied successfully in both power grid and vehicular networks. Wang *et al.* [8] introduced SDN technology to a smart grid communication network for minimizing the total deployment cost of aggregation points (APs), enabling reliable data transmission between smart meters (SMs) and a control center. Li *et al.* [9] considered the balance between the latency requirement and the cost to the vehicular ad hoc network (VANET) and proposed an optimizing strategy applied in the control plane, which is a logical layer of SDN infrastructure. In addition, much research has been conducted on applying SDN in vehicular networks, which are named software-defined vehicular networks (SDVN), in order to deal with various communication challenges. Bozkaya *et al.* [10] addressed the data flow and interference management challenges in SDVN and proposed a power management model for serving more vehicles. Huang *et al.* [11] proposed a software-defined pseudonym system in SDVN for pseudonym management, and a matching theory is formulated to deal with the problem of pseudonym resource scheduling.

In this article, we consider a software-defined green framework for efficient EV charging management. The architecture consists of three planes: the application plane, control plane, and physical plane. Specifically, the application plane provides the EV charging service applications and an intelligent decision making center for customized demands. The control plane is responsible for both data flow control and energy flow control, managed by an information controller and an energy controller, respectively. The physical plane contains the vehicular network and the grid network. A large number of network devices and communication infrastructures are involved in the vehicular network, while various electric entities are located in the grid network. By introducing the SDN framework into the EV charging network to implement the optimization of energy and information control, the grid consumption can be further reduced in two aspects: On one hand, the impacts of EV charging operations on load can be reduced; on the other hand, the energy resources from different local areas can be centrally controlled and managed, contributing to full utilization [12]. In addition, we consider the DSM of EV charging operations as a case study. Finally, simulation results are given to reveal the feasibility of the proposed software-defined charging network architecture.

The remainder of our article is organized as follows. The following section describes the architecture of an EV charging network and points out the existing technical challenges, which then motivates us to propose a green software-defined charging network for EVs. A case study on EVs charging management is then discussed. Following that, we discuss some open issues of the proposed framework for EV charging networks. Finally, we conclude this article.

Electric Vehicle Charging and Supporting Network

The Architecture of the Electric Vehicle Charging Network

The EV charging network structure is shown in Fig. 1, and is divided into two major systems: wired charging systems and wireless charging systems. According to the different charging states of EVs, wired charging systems can be further divided into three subsystems: residential charging, charging stations, and parking lots. These systems are all connected to the smart grid, which communicates with the control center for further management.

Residential Charging System: The residential charging system serves a lot of residents equipped with EV chargers. When EVs are at home during the off-peak period (10 p.m.–7 a.m.) [13], they plug into the household load and charge their batteries. Due to the stationary state of EVs over a long period, EVs at home can be considered flexible electric energy storage devices for balancing the household loads. In each house, the smart meter can be equipped to enable the two-way communication between EVs and grid. EVs upload their own information (i.e., state of charge [SOC]) to the local collector through smart meters, and the charging operations are controlled by the smart meters. If the household load exceeds the expected level, the EVs will stop charging operation to conduct discharging operation until the load is at an ordinary level. If the household load is below the expected level, the electric market will reduce the electric price to encourage more charging activities. However, malfunctions in home appliances can be caused by lower voltage. Lower voltage means lower power quality, which is a sign of an imminent serious issue at any time. Several characteristics measure the power quality, including harmonics, power factor, voltage deviation, frequency shift, and so on. Various indications such as larger amplitude at higher harmonics, a lower power factor, voltage deviation beyond a limit, and excessive frequency drift from a target value can also be considered as lower power quality.

Charging Station System: The charging station system provides fast charging operations in the shortest period for EVs whose batteries cannot afford the rest of a trip. EVs in a certain area communicate with roadside units (RSUs) directly, and the charging requests can be transmitted through

According to the different charging states of electric vehicles, wired charging systems can be further divided into three subsystems: residential charging, charging station, and parking lots. These systems are all connected to the smart grid which communicates the control center for the further management.

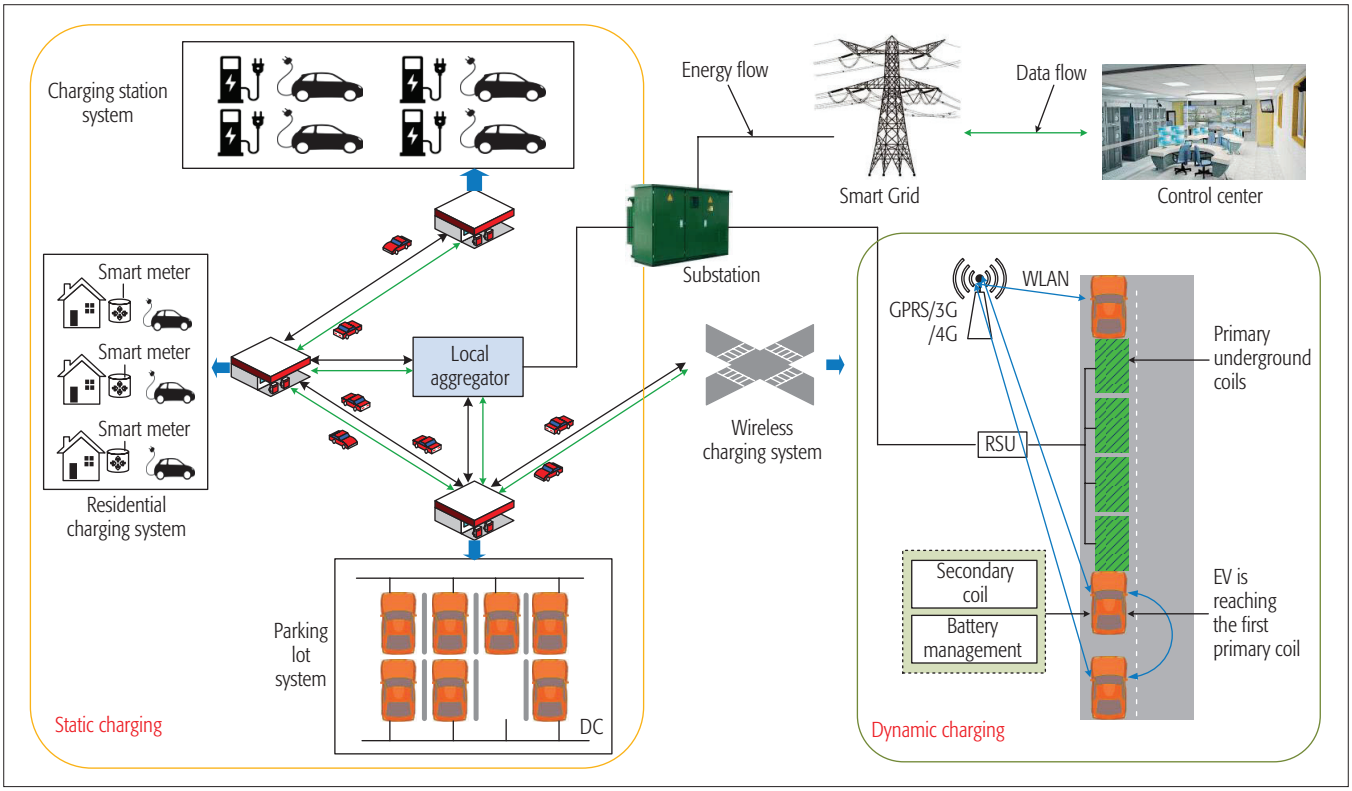


Figure 1. The structure of an EV charging network.

the wireless communication links provided by a VANET [9]. Then the optimal assigned charging stations can be calculated by the RSUs based on the current status of mobile EVs and charging stations, and the optimal scheduling results are sent to EVs to guide the charging operations. However, due to the ubiquitous charging operations, power outage and blackouts will happen, and part of the grid will stop the power supply, leading to instability of the power grid.

Parking Lot System: The EVs in parking lots connect to ac grid via the parking lot dc link. Therefore, EVs can be charged or discharged based on their SOC. The parking time of EVs is much longer than the charging time in a charging station but is shorter than at home. For example, it only takes about half an hour to charge an empty battery of an EV if it uses a quick charger. When the user is shopping or eating, the parking time is usually at least an hour, and it will take eight more hours for an EV to park at home. In addition, the charging operation is not the major task for the EVs, and the EVs are only fully charged just before the next journey. Hence, in order to reduce the extra loads of power grid generated by a large number of EVs' charging operations, the start time of EVs' charging operations can be controlled by the local aggregators (LAGs) [8]. The potential valley stage capacity can be fully used to charge EVs for balancing the load.

Wireless Charging System: Recently, wireless charging technologies are applied in electrified transportation to extend the range of EVs. Comparing the existing conductive charging technologies, wireless charging technologies can hold a promising future for the simplified charging process and economics. There are two major methods of wireless EV charging: static charging and en route charging, which is also called dynamic charging. Due to the high charging efficiency (85–90 percent) [14], in order to be on par with the existing wired charging technologies, the conductive charging technologies will be replaced by the wireless charging technologies to reduce the high cost of large battery packs. In the static charging system, the wireless charging station, including electronics interfaces and safety devices, is on the primary side to produce a high frequency current to the underground coils on the primary side, while the EV equipped with a secondary coil located at the bottom of the EV and a battery management system responsible for the charging or discharging operations of the EV is on the secondary side. Moreover, a closed communication loop controls the charging process through the information transmission between the battery monitoring system and the wireless charging station, as shown by the blue line in Fig. 1. The sensor system is used for identifying the EVs' activities. In the dynamic charging system, the EV battery can be continuously charged when the EV is moving on the roadway. There are several primary coils located successively on the charging road. When an EV reaches the dynamic charging station, the sensor system recognizes it and contacts the control system to enable the first primary coil to be electrified. This wireless charging operation only lasts a few seconds without the need to be parked. After that, when the EV leaves the first primary coil, the first primary will be switched off and the second one is electrified. Each EV can be considered a variable load to the power grid. Much work has been done to balance the electric load, such as vehicle-to-grid (V2G) technologies. However, due to the additional electric load generated from a wireless charging system, traditional V2G technologies cannot satisfy the new

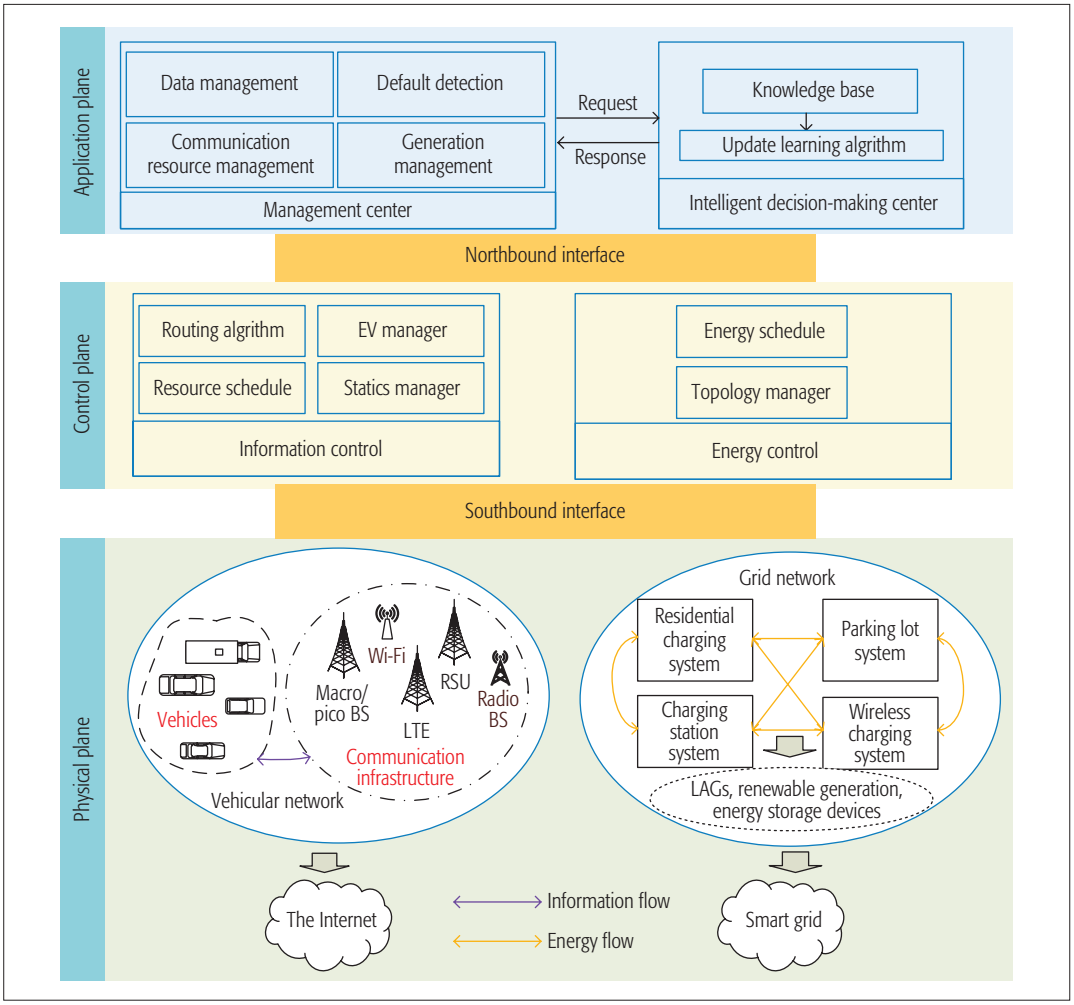


Figure 2. The proposed software-defined EV charging network.

demand of hybrid EV charging networks. In the static charging network, algorithms and mechanisms have been exploited to route EVs. The optimized charging station and routes can be achieved based on the different demands of EVs. However, in the hybrid charging network containing both static charging and dynamic charging, the driving range is expanded, and the traditional solutions cannot adapt to the new needs of EVs.

Software Defined Electric Vehicle Network for Charging Services

Software-defined networking is a promising network architecture that uses the core concept of decoupling the control plane and the data plane. Many outstanding advantages are provided by applying SDN in other network environments such as wired networks, wireless networks, and sensor networks. By integrating SDN with VANETs, SDVN is a centralized paradigm providing routing applications for vehicles. Therefore, the EV network can also be combined with SDN for addressing the aforementioned challenges. First, the energy flow can be controlled programmability through SDN and V2G technologies, and the quality of energy supply can be guaranteed in a global vision. Second, the energy consumption and production of each entity can be monitored and controlled to keep the reliability of the whole power grid and EV network. Last, the topology of the EV charging network changes dynamically. Efficient charging routing and scheduling can be achieved with a full view of the charging network through SDN.

Our Proposed Software-Defined Electric Vehicle Charging Network

Figure 2 shows the proposed architecture of a software-defined EV charging network. The architecture is divided into three parts: the application plane, control plane, and physical plane. The three parts are described in detail below. Then the challenges and opportunities of the proposed architecture are discussed later.

Physical Plane: The physical plane is located at the bottom of the proposed architecture, including the vehicle network and the grid network. In the traditional SDN architecture, the physical plane is also called the data plane and is responsible for data collection. In fact, in our proposed architecture, various kinds of resources are contained in this plane. The vehicular network mainly consists of communication infrastructure and network devices. In the communication infrastructure, the wireless access network connects to the core transmission network through the base station (BS), and the core transmission network connects to the external Internet through the data gateway. There are several access modes: con-

Software-defined networking is a promising network architecture that uses the core concept of decoupling the control plane and the data plane. Many outstanding advantages are provided by applying SDN into other network environments such as wired networks, wireless networks, and sensor networks.

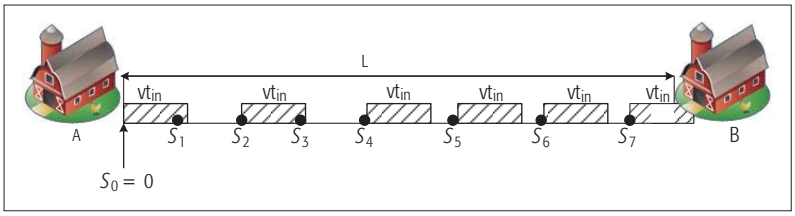


Figure 3. The model of an EV charging network.

ventional BSs for cellular networks (Long Term Evolution [LTE], Universal Mobile Telecommunications System [UMTS], macro BS, and pico BS), radio BSs with large coverage, WiFi access points with small range, and RSUs. In the vehicular network, the involved vehicles act as the end users, which are equipped with onboard units (OBUs). Information can be sent or received between communication infrastructure and the vehicular network. In a different scenario, the users and BS can achieve multiple access methods, such as separation of control plane and data plane, separation of business data downlink, and coordinated multipoint.

The grid network contains the various electric entities located in the residential charging system, charging stations system, parking lot system, and wireless charging system, including a number of LAGs, renewable energy generation, and energy storage devices. Energy flows are generated from bulk generations in smart grid to each LAG and then distributed to the involved charging points. The EVs, especially plug-in hybrid EVs (PHEVs), are also contained in the grid network, acting as the energy storage devices and power lines. Therefore, the vehicular network and the grid network are not separate, but can be connected by information and communications technology (ICT) and smart grid technologies. The information flow and energy flow can enable bidirectional transmission for further control in the upper control plane.

Control Plane: In the control plane, both the information controller and energy controller are connected to the physical plane through southbound interface. This plane is responsible for bridging the users' demands and practical control operations, including data transmission and energy transmission, on the visualized resources in our proposed architecture. Meanwhile, each logical controller connects to the upper application plane though a northbound interface for uploading data based on engines for various functions. To be specific, when the user requests are sent to the control plane, the information controller and energy controller receive the corresponding functional requirements, match out the corresponding service strategies based on their own resources, and then inform the underlying physical plane to achieve the corresponding functions by means of signaling control. For instance, the EV charging network is by nature highly mobile and delay-sensitive. When the requests are sent to the control plane, the information controller is submitted the functional requirements of larger coverage and low latency, and the energy controller is submitted the functional requirements of high reliability and low latency through user authentication and resource adaptation. After receiving the requirements of function, the information controller adopts the separation strategy to allocate a BS to satisfy the uplink connection of EVs. As a result, wide coverage and reduction of the number of handoffs are realized. Broadcasting BSs are assigned to conduct downlink transmission, and the unified transmission can be implemented based on the similar characteristics of user requirements. Similarly, after the energy controller receives the functional requirements, an efficient energy supply scheme is chosen for the EV charging services. In the whole control decision process, the information controller and energy controller can collect the behavior characteristics of information flow and energy flow, respectively, to the intelligent decision making center in the application plane.

Application Plane: The application plane is a collection of EV charging service applications, composed of a management center and an intelligent decision making center. The major functions of the management center are data management, communication and energy resource management, fault detection, and generation management. The application plane provides packed service modules to EV users, receives configuration instructions and requests from EV users, and resolves the instructions and requests to the logical controller for decision making. The intelligent decision making center can maintain a knowledge base for every user and every service, and update it by a learning algorithm. When each controller in the control plane encounters any difficulty in the decision process, it can make a decision request to the intelligent decision center and get a fast response based on the extensible function engine.

Case Study

In this section, we present a case study that is closely related to EV charging management to show the delightful performance of applying the concept of software defined networking into EV charging networks.

The Model of an EV Charging Network

First of all, considering the realistic application of an EV charging network, we have the following assumption: Before the EV drives pass a charging system, it can know the positions of charging systems ahead through the communication infrastructures. The EV has to decide whether to conduct charging operations or not. In addition, the distance between two charging systems cannot be known exactly because the next charging systems cannot be discovered when the condition of driving changes. As a result, how much the battery can be charged should be decided when an EV is charging at a charging system. When an EV is located in the parking lot system, charging the battery for too much time will lead to a long pure waiting time because the charging time exceeds the time for working or shopping. In this section, we consider a model of an EV charging network as shown in Fig. 3 When an EV begins to travel before the first charging operations, it can cover s_0 km initially. S is the total distance of the travel. We have $S_0 < S$, which means the EV must charge its battery for finishing the travel. We also let v be the velocity of an EV moving in transit.

Let d_m be the distance an EV can cover after a full charging operation. Let μ be an increasing distance by charging for one minute of an EV. Let m be the amount of equipment in a charging system. In order to prevent frequent charging, the time interval between two charging operations needs to be at least 10 minutes. Let d_c be the distance between two charging operations we have maintained.

An EV is driving from A to B, and there are seven charging systems located in the path of travel. Let $C_i(i = 1, ..., 7)$ be the coordinates of charging systems following a one-dimensional Poisson process with density λ, where $C_1 < C_2 <$. Let c_i be the minimum value of C_i satisfying $C_i > c_{i-1} + vt_{in}$, where $c_0 = 0$. Let T_w be the total pure waiting time.

The EV charging network is a coupled network consisting of four kinds of charging scenes. These charging scenes have their own characteristics in charging operations. In order to analyze the DSM performance of a software-defined-based EV charging network, three state variables are considered to describe the DSM performance:

- Mean of total pure waiting time (T_w)
- Electricity price (p)
- Electricity demand level (d_e) [15].

Mean of Total Pure Waiting Time (T_w): The mean of total pure waiting time can be denoted as the waiting time at seven charging systems during a trip. There are three kinds of variables affecting the T_w: the time lengths of charging operations (T_{ch}), the shopping or working time at parking lot systems (T_{pa}), and the waiting for other EVs at charging stations (T_{st}).

Electricity Price (p): The electricity price can be changed by the local grid company because the EVs' demand D is changing. Therefore, the electricity price for an EV is dynamically controlled by D. Let C_a be the market capacity and $C_a = D_{max}$.

Electricity Demand Level (d_e): Each charging system is served by a substation, which is responsible for degrading the high-voltage electricity into medium-voltage electricity through a transformer. Hence, the total demand of EVs cannot exceed the capacity of the transformer, which is denoted by maximum demand D_{max}. The real demand from EVs is represented by D, and the electricity demand level is $d_e = D/D_{max}$.

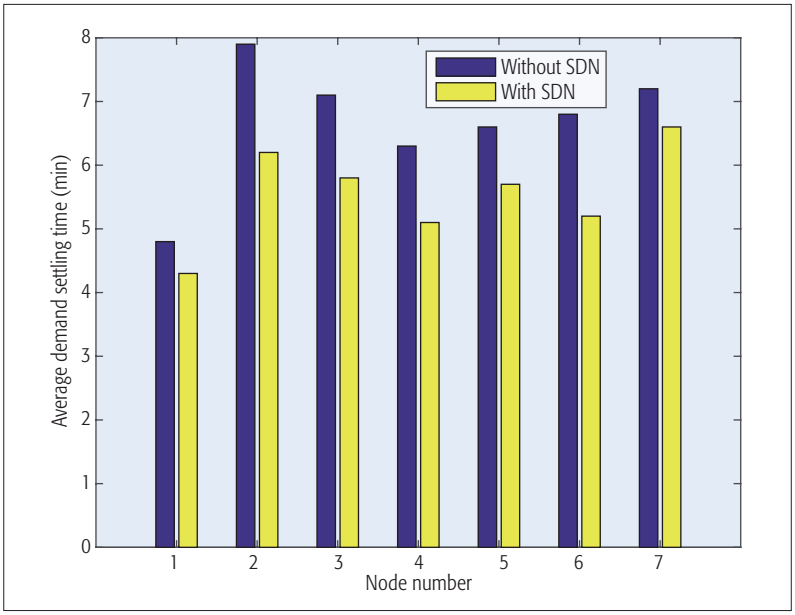


Figure 4. Average demand settling time of the charging nodes.

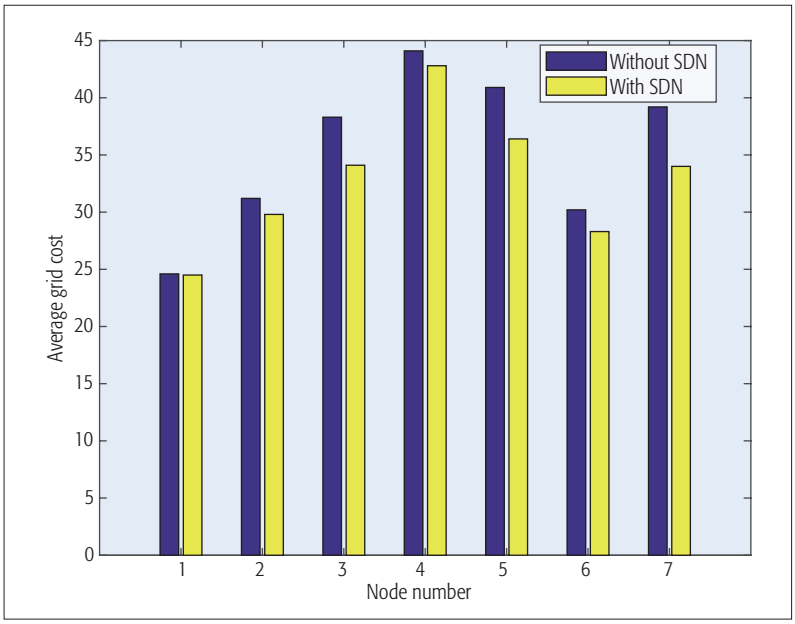


Figure 5. Average grid cost of the charging nodes.

Illustrative Results

The objective of the simulation is to evaluate the performance of a software-definition-based EV charging network. In this section, we select the demand settling time and the grid operation cost as the evaluation of DSM performance, as Figs. 4 and 5 show. The demand settling time is used to denote the duration before the ideal demand level is reached, and the grid operation cost is used to present the operations compensating the difference between real demand levels and intended demand levels. In these figures, S = 150 km, d_m = 160 km, v = 1 km/min, s_0 = 40 km, t_{in} = 20 min, and μ = 0.01 min^{-1}. First, as the s_0 is much smaller than the S, the charging operations can be observed clearly during travel. Next, d_m is the distance an EV can cover after a full charging operation, so $S < d_m$ means the EV must charge its battery to finishing its trip. Last, to prevent the EV from conducting frequent charging operations, we assume the time interval between adjacent charging operations is 20 min, which is the prerequisite of our analysis.

In order to prove the superiority of introducing the SDN framework, we choose the results of two approaches, with and without the SDN framework, as a comparison. Without the SDN framework, the EVs only have the local information about the grid load to make the charging decisions, which is random on the assumption of affording the next trip. Figure 4 shows the demand settling time of the nodes of charging systems. From the result, the demand levels of all the nodes can be controlled to reach the ideal demand levels through the SDN framework. We also find that the initial demand levels and the ideal demand levels in different

To be specific, the decentralized wireless charging operations can reduce the load in a local grid. How to design the rational distribution of wired and wireless charging network to achieve the maximum benefits on both users and grid is an interesting problems to investigate.

charging systems are diverse. Hence, the settling times of charging systems are different. Figure 5 shows the grid operation cost of the nodes of charging systems. From the two figures, we find that both the demand settling time and the grid operation cost can be reduced in all the charging nodes, which indicates that the SDN framework can enhance the DSM performance in an EV charging network.

Open Issues

In this article, we have proposed a hybrid EV charging network to address the major problems that arise in the charging operations of EVs. In addition, we have presented a software-defined green framework for the proposed hybrid EV charging network. In this section, we highlight some open issues about challenges and opportunities arising out of the hybrid EV charging network.

Opportunities

In this subsection, we discuss several emerging opportunities if the software-defined based hybrid EV-charging network can be achieved in the future. We also point out some main technical challenges that may hinder the implementation of such a hybrid EV charging network, which may open some potential directions for future research.

• Similar to the wired communication networks and wireless communication networks that take different responsibilities in the hybrid communication networks, the wireless charging network, especially the mobile wireless charging network, can bring convenience and security to EV charging. Wired charging technologies cannot be replaced by wireless charging technologies within a short time, but the V2G solutions can be implemented to maintain a stable power grid.

• In general, the wireless charging points are decentralized, while the wired charging points are intensive. In the future, EVs can be charging while they are driving on the road, which means the number of charging operations can be reduced over a whole trip.

• The SDN-based hybrid EV charging network also has a good effect on the power grid. To be specific, the decentralized wireless charging operations can reduce the load in a local grid. How to design the rational distribution of a wired and wireless charging network to achieve the maximum benefits for both users and grid is an interesting problem to investigate.

Challenges

Due to the combination of wired charging technologies and wireless charging technologies, a series of new challenges that may appear are discussed as follows.

•Wired charging technologies and wireless charging technologies adopt two completely different sets of charging infrastructure with the large demand of construction cost. Meanwhile, the simultaneous existence of two kinds of devices not only leads to double the operation cost, but also limits the capacity of two sets of equipment.

•In order to implement large-scale application, a unified charging standard is significant. However, the two charging technologies adopt different charging standards. Even in the single wired or wireless charging technology, the charging standards in different areas are diverse.

•For the EVs, both wired charging technologies and wireless charging technologies can be supported for optimal charging operations. However, it may also result in the problem of decision making. For instance, if an EV that supports both the wired and wireless charging technologies needs to charge its battery during a journey, how to choose the optimal way for the next charging operations is also a huge challenge.

•Due to the introduction of wireless charging, especially dynamic wireless charging, the intervals between charging operations get short, and the large-scale charging operations will show the great randomness in the charging network. As a result, the orderly optimization of charging operations will be difficult to achieve.

•There are three major wireless charging technologies: magnetic resonance coupling, inductive coupling, and microwave radiation. The main principle is derived from the phenomenon of magnetic induction and magnetic resonances. Therefore, wireless charging operations may also have an bad impact on the wired charging operations, resulting in additional losses.

Conclusion

In this article, we have presented an architecture of hybrid EV charging networks. Then a green software-defined charging network for EVs has been proposed to address the technical challenges existing in EV charging networks. The proposed software-defined charging network contains three planes, that is, the application plane, control plane, and physical plane. As a case study of the proposed architecture, we have considered the demand-side management of charging operations in the applications of the EV charging network. Finally, the simulation results of the proposed DSM scheme are given to show that our architecture can achieve higher efficiency and lower cost.

Acknowledgment

This work is supported by NSFC (61572262, 61772286, 61533010, 61373135, 61571233, 61532013); the NSF of Jiangsu Province (BK20141427); the China Postdoctoral Science Foundation (2017M610252); and the China Postdoctoral Science Special Foundation (2017T100297).

References

[1] K. Wang *et al.*, "Distributed Energy Management for Vehicle-to-Grid Networks," *IEEE Network*, vol. 31, no. 2, Mar. 2017, pp. 12–18.
[2] K. Wang *et al.*, "A Game Theory Based Energy Management System Using Price Elasticity for Smart Grids," *IEEE Trans. Industrial Informatics*, vol. 11, no. 6, Dec. 2015, pp. 1607–16.
[3] K. Wang *et al.*, "A Survey on Energy Internet: Architecture, Approach and Emerging Technologies," *IEEE Systems J.*, vol. PP, no. 99, Jan. 2017, pp. 1–14.
[4] C. Ou, H. Liang, and W. Zhuang, "Investigating Wireless Charging and Mobility of Electric Vehicles on Electricity Market," *IEEE Trans. Industrial Electronics*, vol. 62, no. 5, May 2015, pp. 3123–33.
[5] T. Theodoropoulos, I. Damousis, and A. Amditis, "Demand-Side Management ICT for Dynamic Wireless EV Charging," *IEEE Trans. Industrial Electronics*, vol. 63, no. 10, Oct. 2016, pp. 6623–30.

[6] X. Lu *et al.*, "Wireless Charging Technologies: Fundamentals, Standards, and Network Applications," *IEEE Commun. Surveys & Tutorials*, vol. 18, no. 2, Apr. 2016, pp. 1413–52.
[7] K. Wang et al., "An SDN-Based Architecture for Next-Generation Wireless Networks," *IEEE Wireless Commun.*, vol. 24, no. 1, Feb. 2017, pp. 25–31.
[8] S. Wang and X. Huang, "Aggregation Points Planning for Software-Defined Network Based Smart Grid Communications," *Proc. IEEE INFOCOM 2016*, Apr. 2016, pp. 1–9.
[9] H. Li, M. Dong, and K. Ota, "Control Plane Optimization in Software-Defined Vehicular Ad Hoc Networks," *IEEE Trans. Vehic. Tech.*, vol. 65, no. 10, Oct. 2016, pp. 7895–7904.
[10] E. Bozkaya and B. Canberk, "QoE-Based Flow Management in Software Defined Vehicular Networks," *Proc. 2015 IEEE GLOBECOM Wksps.*, Dec 2015, pp. 1–6.
[11] X. Huang *et al.*, "Software Defined Networking with Pseudonym Systems for Secure Vehicular Clouds," *IEEE Access*, vol. 4, no. 1, Apr. 2016, pp. 3522–34.
[12] K. Wang *et al.*, "Green Industrial Internet of Things Architecture: An Energy-Efficient Perspective," *IEEE Commun. Mag.*, vol. 54, no. 12, Dec. 2016, pp. 48–54.
[13] S. Yoon *et al.*, "Stackelberg-Game-Based Demand Response for At-Home Electric Vehicle Charging," *IEEE Trans. Vehic. Tech.*, vol. 65, no. 6, June 2016, pp. 4172–84.
[14] F. Lu *et al.*, "A Dynamic Charging System with Reduced Output Power Pulsation for Electric Vehicles," *IEEE Trans. Industrial Electronics*, vol. 63, no. 10, Oct. 2016, pp. 6580–90.
[15] R. Yu *et al.*, "Balancing Power Demand Through EV Mobility in Vehicle-to-Grid Mobile Energy Networks," *IEEE Trans. Industrial Informatics*, vol. 12, no. 1, Feb. 2016, pp. 79–90.

Biographies

YANFEI SUN received his Ph.D. degree in communication and information systems from Nanjing University of Posts and Telecommunications (NJUPT), China, in 2006. He has been a professor with the College of Telecommunication & Information Engineering, Nanjing University of Posts and Telecommunications, since 2006. His main research interests are in the areas of future networks, industrial Internet, big data management and analysis, and intelligent optimization and control.

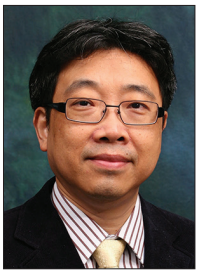
XIAOXUAN HU received her B.Eng. degree in automation from NJUPT in 2014, and is currently pursuing a Ph.D. degree in information acquisition and control at NJUPT. Her research interests include big data management and analysis, vehicle-to-grid networks, and communications networks in the smart grid.

XIULONG LIU received his B.E. and Ph.D. degrees from the School of Software Technology, Dalian University of Technology, China, in 2010 and 2016, respectively. He served as a research assistant at Hong Kong Polytechnic University in 2014, and a visiting scholar at Temple University in 2015. Currently, he is a postdoctoral fellow at Hong Kong Polytechnic University. His research interests include RFID systems and wireless sensor networks.

XIAOMING HE is a postgraduate student in the School of Internet of Things, NJUPT. His current research interests include V2G and deep reinforcement learning.

KUN WANG [SM] received his B.Eng. and Ph.D. degrees from the School of Computers, NJUPT, in 2004 and 2009, respectively. From 2013 to 2015, he was a postdoctoral fellow in the Electrical Engineering Department, University of California, Los Angeles. In 2016, he was a research fellow in the School of Computer Science and Engineering, University of Aizu, Fukushima, Japan. He is currently a research fellow in the Department of Computing, Hong Kong Polytechnic University, China, and also a full professor in the School of Internet of Things, NJUPT. He has published over 50 papers in refereed international conferences and journals. He received the Best Paper Award at IEEE GLOBECOM 2016. He serves as an Associate Editor of *IEEE Access*, the *Journal of Network and Computer Applications*, and *EAI Transactions on Industrial Networks and Intelligent Systems*, and an Editor of the *Journal of Internet Technology*. He was the Symposium Chair/Co-Chair of IEEE IECON '16, IEEE EEEIC '16, IEEE WCSP '16, IEEE CNCC '17, and others. His current research interests are mainly in the area of big data, wireless communications and networking, smart grid, energy Internet, and information security technologies. He is a member of ACM.

HUMAN-DRIVEN EDGE COMPUTING AND COMMUNICATION: PART 1


Jiannong Cao


Aniello Castiglione


Giovanni Motta


Florin Pop


Yanjiang Yang


Wanlei Zhou

The vision of edge computing considers that tasks are not exclusively allocated on centralized cloud platforms, but are distributed toward the edge of the network (as in the Internet of Things and fog computing paradigms), and transferred closer to the business thanks to content delivery networks. The traditional gateway becomes a set-top box machine, with additional computation and storage capabilities, where micro-tasks can be offloaded first, instead of directly to the cloud. Mobile edge computing can also be a more suitable approach to also extract knowledge from privacy sensitive data, which are not to be transferred to third party entities (global cloud operators) for processing. The proliferation of the networking connectivity and the progressive miniaturization of the computing devices have paved the way to sensor networks and their success in the automation of several monitoring and control applications. Such networks are built in an ad hoc manner and deployed in an unsupervised manner, without an a priori design. The consequent availability of long-range communication means at certain nodes of those networks has enabled the possibility of the Internet connection of the sensor network to make use of cloud-based services.

There was overwhelming interest in the Call for Papers of our Feature Topic, and we received a total of 51 submissions co-authored by people belonging to institutions spread over four continents. The submissions went through a rigorous review process, and the result was the acceptance of 12 contributions that will be split over two issues. The first six papers will be included in this issue, while the remaining six papers will be included in the issue of February 2018.

The first contribution, co-authored by Bao *et al.*, is "Follow Me Fog: Toward Seamless Handover Timing Schemes in a Fog Computing Environment," which introduces a new framework named Follow Me Fog. This framework supports a new seamless handover timing scheme among different computation access points.

A new cloud-edge computing framework including the cloud plane and edge plane is presented in the article co-authored by Wang *et al.*, "A Cloud-Edge Computing Framework for Cyber-Physical-Social Services." This framework, based on a tensor-based services model, is used to provide high-quality proactive and personalized services for humans.

In the article "Improving Opportunistic Networks by Leveraging Device-to-Device Communication," co-authored by Marin *et al.*, the authors propose a novel mechanism for connecting peers in opportunistic networks over Wi-Fi Direct in a secure and seamless fashion, without requiring any human intervention throughout the process of pairing. To this aim, the authors present their opportunistic engine implementation over Wi-Fi and show how easily it is able to accommodate the addition of Wi-Fi Direct without impacting the battery life of mobile devices.

Markakis *et al.*, in their article "Efficient Next Generation Emergency Communications over Multi-Access Edge Computing," introduce the challenges that next generation emergency services need to overcome in order to fulfill the requirements for rich-content, real-time, location-specific communications of 5G networks. They also present a vision of how this concept can satisfy the 5G requirements for ultra-low-latency and ultra-reliable emergency communications.

The next article, "Pseudo-Dynamic Testing of Realistic Edge-Fog Cloud Ecosystems," co-authored by Ficco *et al.*, considers how it can be challenging to test a software artefact to be deployed in nodes spread among edge and fog computing architectures.

In the last article, "Vehicular Fog Computing: Architecture, Use Case and Security and Forensic Challenges," Huang *et al.* present an architecture for vehicular fog computing, and discuss the potential benefits, security and forensic challenges, and mitigation strategies using the fog-assisted traffic control system as a use case.

We would like to thank all the authors for their contributions to this Feature Topic and to appreciate the efforts of the reviewers to provide constructive feedback to the authors. We address our personal warm regards to Osman Gebizlioglu, the Editor-in-Chief of *IEEE Communications Magazine*, for his support and advice, as well as to the editorial and managerial publications team for their assistance and excellent cooperative collaboration to produce this valuable scientific work.

BIOGRAPHIES

JIANNONG CAO [M'93, SM'05, F'15] (csjcao@comp.polyu.edu.hk) received his M.Sc. and Ph.D. degrees in computer science from Washington State University. He is currently a Chair Professor with the Department of Computing at Hong Kong Polytechnic University. He is also the director of the Internet and Mobile Computing Lab in the department and the director of the university's Research Facility in Big Data Analytics. His research interests include parallel and distributed computing, mobile computing, and big data analytics.

ANIELLO CASTIGLIONE [S'04, M'08] (castiglione@ieee.org) has a Ph.D. in computer science from the University of Salerno, Italy. He is an adjunct professor at the University of Salerno and at the University of Naples ''Federico II,'' Italy. He is the Managing Editor of two international journals and has served as a Guest Editor of several Special Issues in many high-ranked journals. His research interests include security and privacy on communication networks, information forensics and Security, and applied cryptography.

GIOVANNI MOTTA [S'97, M'05, SM'11] (giovannimotta@google.com) received his Laurea in informatica in 1996 (Summa Cum Laude) from the University of Salerno and a Ph.D. in computer science from Brandeis University in 2002. He is currently

with Google (Terra Bella), where he works on high resolution satellite imagery and analytics. His main interests are in the fields of data compression, coding, and algorithms. He has been granted 13 patents and published two books and more than 50 peer-reviewed papers in journals and conferences.

FLORIN POP (florinpop@ieee.org), Professor, Ph.D., Habil., received his Ph.D. in computer science at the University Politehnica of Bucharest in 2008. His main research interests are large-scale distributed systems, big data, cloud and edge computing, adaptive methods, multi-criteria optimization methods, prediction methods, self-organizing systems, and performance evaluation using modeling and simulation. He is a reviewer and Guest Editor for several journals. He is a scientific researcher at the National Institute for Research and Development in Informatics (ICI), Bucharest.

YANJIANG YANG (yang.yanjiang@huawei.com) is currently a senior researcher at Shield Lab, Huawei, Singapore. Prior to that, he was with the Institute for Infocomm Research, Singapore, as a research scientist between 2008–2015. His research spans a wide spectrum of information security such as applied cryptography, trusted computing, multimedia security, cloud security; his main current research interest is IoT security, especially connected car security.

WANLEI ZHOU [M'92, SM'09] (wanlei.zhou@deakin.edu.au) received his B.Eng. and M.Eng. degrees from Harbin Institute of Technology, China, in 1982 and 1984, respectively, and his Ph.D. degree from the Australian National University, Canberra, in 1991. He is currently the Alfred Deakin Professor, Chair of Information Technology, and Associate Dean of the Faculty of Science, Engineering and Built Environments, Deakin University, Melbourne, Australia. His research interests include distributed systems, network security, bioinformatics, and e-Learning.

CALL FOR PAPERS

IEEE COMMUNICATIONS MAGAZINE

EMERGING TECHNOLOGIES FOR CONNECTED AND SMART VEHICLES

BACKGROUND

Over the past decade, advances in vehicular communications and intelligent transportation systems (ITS) have been aimed at trimming down the fuel consumption by avoiding traffic congestion, enhancement of traffic safety while initiating new application, i.e., mobile infotainment. To address the individual requirements of both safety and non-safety applications in the Connected and Smart Vehicles field, there is a need to build up a new communication technology for the integrated solutions of vehicular communication and smart communications. The Connected Vehicles infrastructure can be of various models such as Vehicle-to-Vehicle (V2V), Vehicle-to-Infrastructure (V2I), and Vehicle-to-Everything (V2E). Due to the rapid growth in the Connected Vehicles, many research constraints need to be addressed, e.g., reliability and latency, appropriate scalable design of MAC and routing protocols, performance and adaptability to the changes in environment (node density and oscillation in network topology), and evaluation and validation of Connected Vehicles' protocols under the umbrella of coherent assumptions using simulation methodologies. In addition, the information shared among Connected Vehicles is of great importance and it is not yet clear what kind of privacy policies will be defined for the ITS networks.

This Feature Topic (FT) aims to emphasize the latest achievements to identify those aspects of Connected Vehicles and ITS networks that are identical to a traditional communications network in the broader spectrum.

In this FT, we would like to try to answer some (or all) of the following questions:

What will be the effect on Policy Making regulations when the connected vehicles will hit the roads? What are the privacy consequences with Connected Vehicles with Smart onboard peripherals? What kind of privacy policies will be defined for the connected vehicles and smart ITS networks? What kind of education and training are required for drivers and passengers of such connected and smart vehicles to take full benefits of the proposed communication architectures? What kind of ground breaking applications can make Connected Smart Vehicles more attractive? What will be the acceptability aspects of Smart Vehicles with networked hardware? In addition, the authors are expected to address state-of-the art research challenges, results, architecture, applications, and other achievements in the following topics, but not limited to:

- Network and system architecture for connected vehicles
- Physical layer and routing protocols
- Internet-of-vehicles, smart sensors (infrastructure and vehicle based)
- Energy efficient vehicular communication
- Quality-of-service for vehicular communication
- Information and content centric networking in connected vehicles networks
- Vehicular cyber-physical systems and smart devices
- MAC protocols and channel management
- Delay tolerant vehicular networks
- Modeling and theory
- Mobility management (traffic models)
- Real-time optimization system
- Intra-vehicle communication
- Future Internet in ITS and networking systems

SUBMISSION GUIDELINES

Articles should be tutorial in nature, with the intended audience being all members of the communications technology community. They should be written in a tutorial style comprehensible to readers outside the specialty of the article. Mathematical equations should not be used (in justified cases up to three simple equations are allowed). Articles should not exceed 4500 words (from introduction through conclusions, excluding figures, tables, and captions). Figures and tables should be limited to a combined total of six. The number of archival references is recommended not to exceed 15. In some rare cases, more mathematical equations, figures, and tables may be allowed, if well justified. In general, however, mathematics should be avoided; instead, references to papers containing the relevant mathematics should be provided. Complete guidelines for preparation of manuscripts are posted at page, http://www.comsoc.org/commag/paper-submission-guidelines. Please submit a PDF (preferred) or MSWORD formatted paper via Manuscript Central (http://mc.manuscriptcentral.com/commag-ieee). Register or log in, and go to Author Center. Follow the instructions there. Select "October 2018/Emerging Technologies for Connected and Smart Vehicles" as the Feature Topic category for your submission.

IMPORTANT DATES

- Manuscript Submissions Due: February 1, 2018
- Decision Notification: June 1, 2018
- Final Manuscript Submission: July 15, 2018
- Publication Date: October 2018

GUEST EDITORS

Syed Hassan Ahmed
Univ. of Central Florida, USA
sh.ahmed@ieee.org

Jalel Ben-Othman
Univ. of Paris 13, France
jalebeno@yahoo.fr

Jaime Lloret
Polytechnic Univ. of Valencia, Spain
jlloret@dcom.upv.es

Wael Guibene
Intel Corp., Santa Clara, CA, USA
wael.guibene@intel.com

Ashfaq A. Khokhar
Iowa State Univ., USA
ashfaq@iastate.edu

Raheem Beyah
Georgia Inst. of Technol., USA
rbeyah@ece.gatech.edu

Antonio Sánchez-Esguevillas
Telefonica, Spain
a.sanchez-esguevillas@ieee.org

Follow Me Fog: Toward Seamless Handover Timing Schemes in a Fog Computing Environment

Wei Bao, Dong Yuan, Zhengjie Yang, Shen Wang, Wei Li, Bing Bing Zhou, and Albert Y. Zomaya

The authors propose Follow Me Fog (FMF), a framework supporting a new seamless handover timing scheme among different computation access points. Intrinsically, FMF supports a job pre-migration mechanism, which pre-migrates computation jobs when the handover is expected to happen. Such expectations can be indicated by constantly monitoring received signal strengths.

ABSTRACT

Equipped with easy-to-access micro computation access points, the fog computing architecture provides low-latency and ubiquitously available computation offloading services to many simple and cheap Internet of Things devices with limited computing and energy resources. One obstacle, however, is how to seamlessly hand over *mobile* IoT devices among different computation access points when computation offloading is in action so that the offloading service is not interrupted — especially for time-sensitive applications. In this article, we propose Follow Me Fog (FMF), a framework supporting a new seamless handover timing scheme among different computation access points. Intrinsically, FMF supports a job pre-migration mechanism, which pre-migrates computation jobs when the handover is expected to happen. Such expectations can be indicated by constantly monitoring received signal strengths. Then we present the design and a prototype implementation of FMF. Our evaluation results demonstrate that FMF can achieve a substantial latency reduction (36.5 percent in our experiment). In conclusion, the FMF design clears a core obstacle, allowing fog computing to provide interruption-resistant services to mobile IoT devices.

INTRODUCTION

The rapid evolution of the Internet of Things (IoT) will soon have a substantial impact on our daily lives through reforming many Internet applications in a variety of domains. IoT devices, however, create a growing need to store and process significant quantities of data, but they themselves have limited computing and energy resources, and hence are not able to perform sophisticated data processing and storage. Therefore, the IoT is perfectly aligned with emerging fog computing: Equipped with easy-to-access micro computation access points (i.e., fog nodes), it provides low-latency and ubiquitously available computation services [1, 2].

Mobility is an intrinsic trait of many IoT applications in a fog computing environment [3], such as smart transportation, smart sport, smart tourism, and geo-based games. Consider an exemplary application, an augmented reality (AR) assisted tour guide: a mobile tourist holds an IoT device that continuously records its current view and streams the multimedia contents to a nearby fog node, while the fog node performs scene recognition and sends back contextual augmentation labels, to be displayed on the IoT device overlaying the actual scene. Many mobile applications like this require not only powerful computing resources but also sufficiently low latency in data transfer and processing to guarantee users' desired quality of experience. However, we observe that IoT mobility inevitably poses significant challenges for realizing reliable and punctual computing services. Since each computation access point provides only limited wireless coverage, movement of IoT devices will call for handovers among different computation access points, causing complicated internetworking issues and interrupted services.

In conventional mobile networks, mobility of a mobile terminal is enabled by horizontal and vertical handover procedures when it changes serving access point (or base station). However, such handover mechanisms are insufficient to support reliable and punctual computation offloading in the environment of fog computing, where the access point additionally processes jobs uploaded from users. When a handover occurs, those pending jobs (e.g., jobs unprocessed, being processed, or processed but not fed back) must be recovered after the handover. Meanwhile, the new access point should allocate new computing resources to resume job processing. Mobile devices, on the other hand, should also be prepared to recover from potential losses and delays.

There is a set of pioneering studies, based on theoretical analyses, simulations, and testbed experiments, to address the mobility issues in a computation offloading environment, such as "whether to conduct handovers" and "where to hand over." However, one critical question remains unanswered: what is the exact timing procedure of a handover to minimize service interruptions? The answer is vital to many delay-sensitive applications. In the aforementioned AR guided tour scenario, the overall latency, from the beginning of uploading the current scene to the end of successfully downloading augmentation labels, must be reasonably

Digital Object Identifier:
10.1109/MCOM.2017.1700363

The authors are with the University of Sydney.

0163-6804/17/$25.00 © 2017 IEEE

small, and this latency requirement should not be impacted by handovers — our ultimate goal is to make the tourist unaware of the existence of handovers. During a handover, important procedures such as disconnection from the old fog node, connection to the new fog node, and job recovery at the new fog node must be conducted in an orderly and timely manner so that the additional latency caused by handover is minimized.

We see a clear gap between the requirements for a seamless handover and the lack of an efficient handover timing scheme in a fog computing environment. Therefore, we are motivated to propose Follow Me Fog (FMF), a new framework supporting seamless handover timing schemes. The objective of FMF is to orchestrate a sequence of procedures at both the mobile IoT devices and fog nodes, to guarantee service continuity and reduce latency during handovers.

The key design feature of FMF is motivated by an important observation on handovers in the fog computing environment. There are two major processes involved in a handover:

1. *Connection redirection*
2. *Service recovery*

In the connection redirection process, the mobile IoT device disconnects from its original fog node and then connects to a new node. In the service recovery process, the pending jobs offloaded to the old fog node will be recovered at the new node. One can consider a straightforward solution that conducts service recovery after connection redirection: As soon as the mobile device reaches the new fog node, it suggests that there are pending jobs in the old fog node. However, such a handover procedure may cause severe service interruptions, since the computation offloading service is completely halted during both connection redirection and service recovery phases, as shown in Fig. 2a.

Fortunately, we can observe that the service recovery can actually be conducted *in advance* to substantially reduce service interruptions. Our proposed FMF *pre-migrates* computation jobs when the handover is about to happen. Such expectation of handover can be indicated by constantly monitoring the received signal strengths (RSSs) from different fog nodes. When the RSS from the current fog node keeps decreasing while the RSS from a neighboring one keeps increasing, the pre-migration of computation jobs is triggered prior to the connection redirection. As a consequence, the service can be resumed in a timely way when the mobile device is redirected to the new fog node, as shown in Fig. 2b. It is worth noting that FMF follows a cross-layer design: both the RSS levels in the physical/medium access control (MAC) layer and the status of computation jobs in the application layer will influence the decision on the timing of connection redirection and job (pre-) migration.

In what follows, we present the background and compare the differences between FMF and the existing frameworks. We present the FMF design. We present a prototype implementation of FMF and show evaluation results of the implementation. Finally, we conclude our article and discuss some future directions.

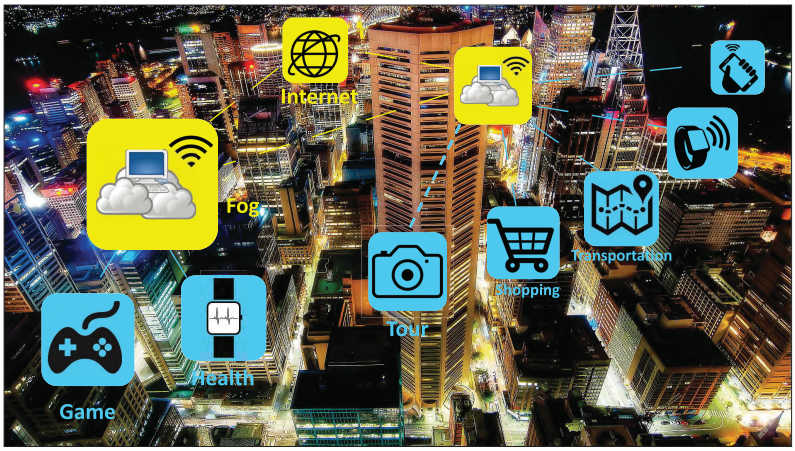


Figure 1. Fog computing environment for IoT applications.

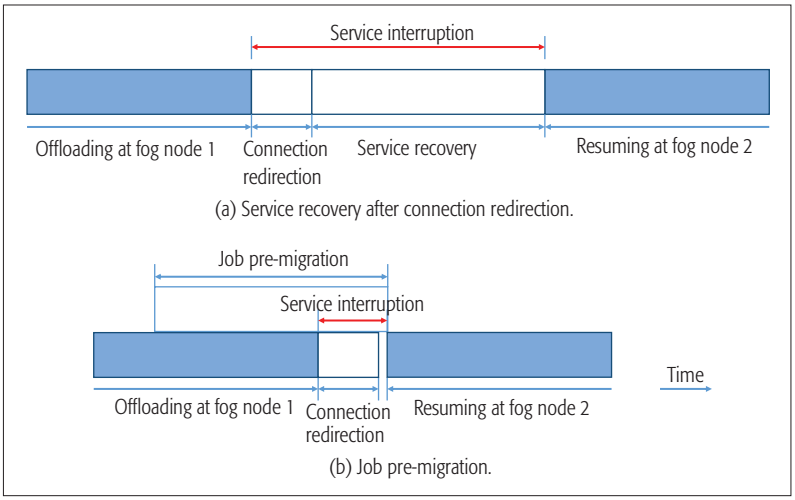


Figure 2. Potential delay reduction of pre-migration.

Background and Related Work

Mobility Management in Wireless Networks

Handover timing algorithms have been intensively investigated in traditional mobile networks. These algorithms aim to transfer mobile terminals to the "best" access point without perceivable interruptions in data transfer. Ping-pong effects should also be avoided. One type of algorithms employ threshold comparison of one or several metrics (e.g., RSS, bandwidth, etc.) [4]. A second type is based on mobility load balancing, where cells suffering congestion can transfer load to other cells [5]. A third type uses dynamic programming or artificial intelligence techniques to improve the effectiveness of handoff procedures (e.g., [6]). However, these algorithms are insufficient in the presence of delay-sensitive computation offloading, since they do not handle promptly and efficiently migrating computation jobs among fog nodes when a handover occurs.

Live Migration

Live migration was first proposed and used in cloud data centers to move running applications between different physical machines without disconnecting the users. It has also been used to hand over computation jobs between fog nodes. Virtual machine (VM) migration [7] and container migration [8] (e.g., Docker [9]) are the two

No existing framework for live migration takes into account the time sensitivity of the applications when they are handed over between fog nodes since delays caused by wireless disconnection and reconnection are out of the scope of live migration.

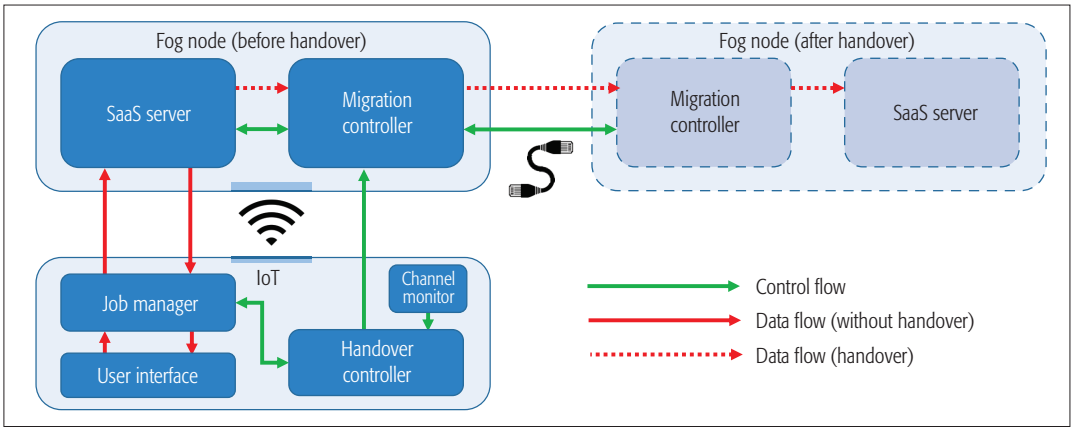


Figure 3. FMF module design.

main solutions. VM migration directly moves the full image of the computing node, including the OS, file system, and all software installed, to the destination node. It is a flexible solution and only requires the destination node to have enough resources to accommodate the VM. Container migration is a lighter-weight solution that can move a single application between computing nodes. It has lower migration cost but requires the container platform to be pre-installed in computing nodes. However, no existing framework for live migration takes into account the time sensitivity of the applications when they are handed over between fog nodes since delays caused by wireless disconnection and reconnection are out of the scope of live migration.

Path Selection

Live migration may not be a convenient solution when a large amount of data are to be migrated among computing nodes — it may introduce large delays and burden backhaul links. An alternative approach is path selection [1], where a more suitable path is found for delivery of computed data from a data center to a mobile terminal. By combining the mechanisms of live migration and path selection, path selection is employed if a reasonably good path can be found, while live migration is performed if even the best available path is not satisfactory. However, like live migration, delays caused by wireless disconnection and reconnection are not considered.

Follow Me Cloud

The concept of Follow Me Cloud (FMC) is to support smooth migration of IP services between a data center and mobile device to another data center with no service disruption [10]. FMC is a service migration mechanism (among data centers) in the wired network, while the wireless handover timing procedure is not redesigned (i.e., it follows a standard procedure in traditional cellular networks). As discussed earlier, in order to promptly hand over delay-sensitive applications, it is more desirable to jointly design the connection redirection and job migration, which is the emphasis of FMF studied in this article.

Analytical Models

There is another category of studies using analytical models to address computation migration problems; examples include the random walk model [11], a Markov decision process [12], combinatorial cost minimization [13], and integer programming [14]. In the scope of the handover timing scheme, decisions are made within small timescales. Analytical models and real-world experiments that can capture variations and make decisions within small timescales are much more desirable.

FMF Design Outline

The Problem

The overall problem is to design a seamless handover timing procedure to avoid service interruptions during handovers in the scenarios where a moving IoT device offloads its delay-sensitive jobs to fog nodes it passes. In this article, we address this issue by developing FMF, a new paradigm for a smooth handover mechanism, and then implement a prototype system.

A Module Design of FMF

The module framework of FMF is shown in Fig. 3 with two parties. IoT devices and fog nodes are connected via wireless links, and the fog nodes are interconnected via a wired network.

The Fog Node: In the framework of FMF, we consider the scenario where each fog node provides software as a service (SaaS) to mobile IoT devices in its *SaaS server*. The SaaS model is suitable for many special-purpose time-sensitive IoT applications (e.g., smart sport and smart tourism) conducted within a given geographical region, and the job processing service for that application is managed by one provider (e.g., administrator of a stadium or museum). The software is deployed in each fog node as a service, which is in the "always-on" status to process offloaded jobs from IoT devices in a timely way and return outcomes. In order to conduct smooth handovers, the *migration controller* is responsible for migrating the jobs to the new fog node to which the IoT device is handed over. Pre-migration is applied so that the migration is triggered prior to connection redirection.

In this article, we focus on the scenarios where an SaaS server is installed at each access point. The designed FMF is also applicable in scenarios where one SaaS server provides computing services to multiple access points. When a mobile IoT device is handed over between two access

points served by two different SaaS servers, both job migration and connection redirection occur, requiring our proposed seamless handover timing scheme.

The IoT Device: At the IoT device, the *job manager* uploads unprocessed jobs and downloads processed jobs (outcomes); the *user interface* interacts with human users; and the *channel monitor* keeps monitoring the channel states from different fog nodes. The key module at the IoT device is the *handover controller*, which is responsible for:

1. Checking the channel states (e.g., RSSs) from different fog nodes and learning their changing tendencies
2. Checking the status of offloaded jobs (e.g., which job is "in flight")
3. Making decisions on job migration and connection redirection
4. Interacting with the migration controllers at fog nodes to advise job pre-migration prior to connection redirection and job recovery after connection redirection

The handover timing scheme is jointly realized by the handover controller at the IoT device and the migration controller at the fog node. The handover controller analyzes the current situation of the IoT device and makes decisions on job pre-migration and connection redirection. When a pre-migration decision is made, the handover controller notifies the migration controller at the fog node, so jobs are migrated to the new fog node. When the IoT device connects to the new fog node, the handover controller contacts the new migration controller to recover the jobs offloaded to the old fog node. The handover timing scheme is further specified as a finite state machine in the following subsection.

Handover Timing Scheme: A Finite State Machine Design

In this subsection, we present the design guideline of the handover timing scheme to be conducted in the FMF framework. The designed handover timing scheme follows a finite state machine, as shown in Fig. 4. There are four states: the connection state, pre-migration state, redirection state, and resuming state. We specify the four states first and then introduce the transitions between the four states.

The States: In the *connection state*, the IoT device continuously offloads computation jobs to its connected fog node and downloads the outcomes. The job offloading is performed in a pipeline fashion on the IoT side: the next job could be uploaded without waiting for the outcome of its previous job, but the maximum number of uploaded but pending jobs is limited by a "window." On the fog side, the jobs are processed in a first-come first-served way. The outcomes of the jobs are sent back to the IoT as soon as processing is completed.

In the *pre-migration state*, an upcoming handover is expected, so pre-migration is conducted. However, job uploading and downloading may continue in this state. On the fog side, the current fog node establishes a connection to the expected fog node and starts to transfer the processed jobs. If a job is not processed, it will be processed first and then migrated. In our current design version, the fog will only migrate the processed jobs (outcomes) rather than unprocessed jobs or jobs being processed. Due to the potential of a "false alarm" in an upcoming handover, it is possible that the IoT will finally cancel the handover even if it enters the pre-migration state. Whenever a "false alarm" happens, migrating unprocessed jobs causes unnecessary job processing at two fog nodes, and migrating jobs being processed causes substantial communication overhead. Please note that "false alarms" are inevitable due to the unpredictability of user mobility and instability of wireless channels. Please further refer to our discussions on "transition from pre-migration to connection state" below.

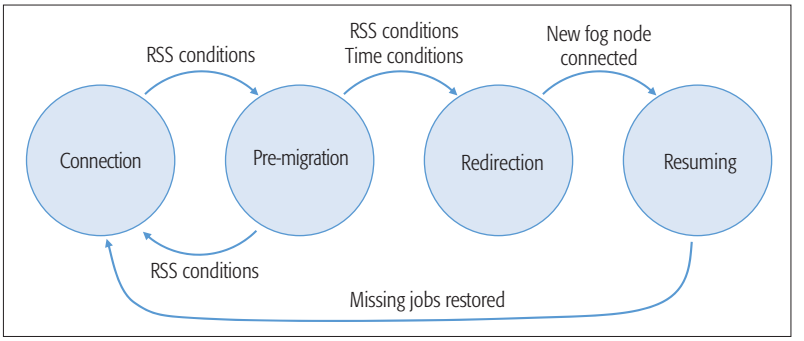


Figure 4. Finite state machine of the handover timing scheme.

In the *redirection state*, the IoT device disconnects from the old fog node and establishes connection to the new one. The old fog node keeps processing and migrating the jobs offloaded prior to reconnection if they are not completely processed and migrated.

In the *resuming state*, the IoT device requests missing jobs (sent to the old fog node before connection redirection) to the new fog node and the new fog node will return the processed jobs. The IoT device is also allowed to send new jobs to the new fog node. The old fog node keeps processing and migrating the jobs prior to connection redirection if they are not completely processed and migrated.

State Transitions: *From connection to pre-migration state:* In the connection state, the IoT device keeps monitoring the RSSs from its current fog node and neighboring ones. The expectation of handover can be predicted by an RSS-based criterion (e.g., relative RSS with hysteresis [4], adaptive lifetime-based [15]). If the criterion is satisfied, the transition from connection to pre-migration state is triggered.

From pre-migration to connection state: During the pre-migration state, the IoT device may notice that the transition to pre-migration state is a false alarm. This can be implied by another RSS based criterion (e.g., the RSS from the current fog node becomes large again). The state is reversed to connection state if such a criterion is satisfied. The pre-migration from the current fog node to the (previously) expected fog node is cancelled.

From pre-migration to redirection state: The transition is triggered if the RSS condition from the new fog node keeps outperforming that from the current fog node for a while. The trigger of the state transition could be reasonably postponed for a small time period if there is a job being transferred (e.g., an unprocessed job being

The handover timing scheme is jointly realized by the handover controller at the IoT device and the migration controller at the fog node. The handover controller analyzes the current situation of the IoT device and makes decisions on job pre-migration and connection redirection.

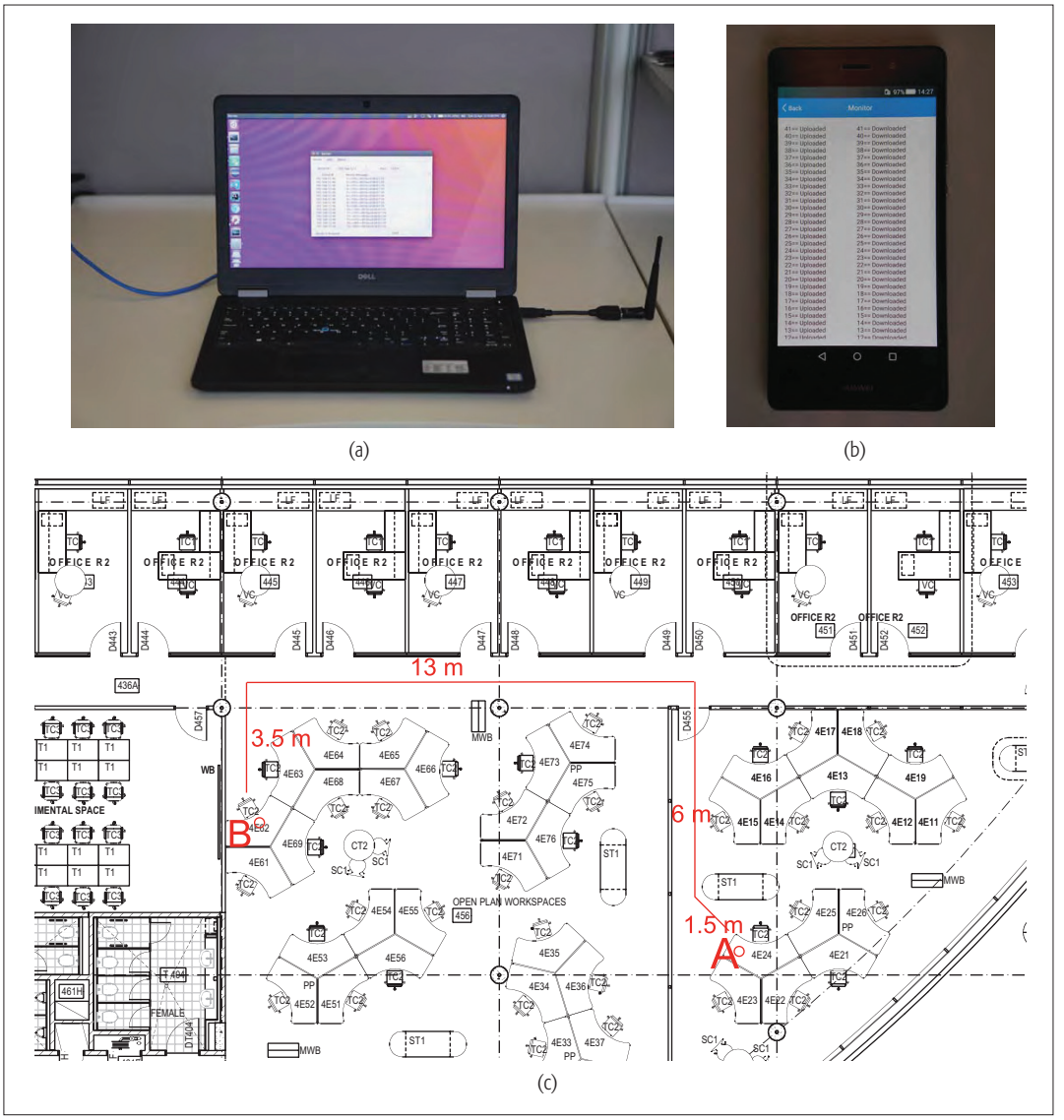


Figure 5. Implementation setup: a) fog node; b) mobile device; c) experiment environment.

uploaded and/or a processed job being downloaded), to wait for the completion of that job transfer.

From redirection to resuming state: The transition happens as soon as the IoT device is connected to the new fog node.

From resuming to connection state: The transition happens as soon as the outcomes of the jobs previously offloaded to the old fog node are successfully sent back to the IoT device via the new fog node.

Prototype and Performance Evaluation

In this section, we present a prototype of FMF that supports job pre-migration. We start with the implementation description of the prototype and then discuss the experiment setup. Finally, we present the evaluation results of the performance.

Implementation Description

To evaluate the proposed FMF framework, we develop an FMF prototype in a real fog computing environment. To create the fog computing environment, we use two laptop computers (Dell Latitude E5570, Core i7-6600U CPU and 8 GB memory) to act as fog nodes. Each fog node has two network adapters:

1. External 300 Mb/s wireless USB adapter (EDiMAX EW-7612UAn V2) that provides local WiFi coverage for the wireless IoT devices
2. Internal Ethernet adapter (Intel I219-LM) that provides connections to other fog nodes, as shown in Fig. 5a

We use smartphones (HUAWEI ALE-L02 P8 Lite) to act as mobile IoT devices and offload jobs to the fog nodes, as shown in Fig. 5b. The IoT devices and fog nodes are connected via WiFi interfaces, and the fog nodes are interconnected via Ethernet interfaces. The FMF application is implemented in both fog nodes and IoT devices. In each fog node, we install the Ubuntu 16.04.2 LTS 64-bit operating system and adopt the QT application framework to implement the application server and migration controller, and deploy them in the SaaS manner. For each IoT device, we adopt the Android platform (Android 5.0.1, API 21) to implement the job manager, user interface, channel monitor, and handover

controller in an all-in-one Android application called "FMF Client." All the communications in the prototype are implemented via sockets programming based on TCP. Between IoT devices and the fog node, three sockets are created with different port numbers for uploading jobs, downloading outcomes, and sending pre-migration signals, respectively. Between every two fog nodes, sockets are created to migrate computation jobs and exchange control signals with each other. WiFi RSS monitoring, connection, and reconnection are realized by Android API class `wifiManager`.

Experiment Setup

Experiment Field: As shown in Fig. 5c, we implement our prototype on the fourth floor of Building J12 in the University of Sydney. We install the two fog nodes at A and B, shown in the figure. In each round of the experiment, the Android phone is held to move from A to B along the red lines, with a total distance of 24 m. The speed of the mobile device is 0.63 m/s. In the experiment, the phone keeps offloading the same job (same communication load and processing load) to its connected fog node. The final performance is averaged on the results of 10 rounds of experiment.

State Transition Conditions: In the prototype implementation, the transition from connection to pre-migration state is triggered if the RSS from the new fog node is at least 10 dB greater than that from the current fog node for 1 s. The transition from pre-migration to connection state is triggered if the RSS from the current fog node is at least 10 dB greater than that from the new fog node for 0.5 s. The transition from pre-migration to redirection state is triggered if pre-migration state lasts for 1 s without transiting back to connection state.

Please note that the system performance can be further improved by carefully designing the transition conditions. However, that may involve complicated theoretical analysis and optimization. It is out of the scope of this article's focus on the framework design of FMF.

Performance Metrics and Benchmark Approach: We evaluate the delay performances of three categories of offloaded jobs. The first category of offloaded jobs do not experience handovers. They are offloaded to one fog node, and their outcomes are sent back from that fog node. The second and third categories of offloaded jobs experience handovers. For the second category, a benchmark re-uploading approach, instead of FMF, is employed: whenever a handover occurs, the mobile device gives up the pending jobs uploaded to the previous fog node and re-uploads these jobs to the new fog node. For the third category, FMF is employed.[1]

Performance Evaluation

The delay performance of the three categories of offloaded jobs are shown in Fig. 6. Each total delay (i.e., from the beginning of uploading one job to the end of successfully downloading its outcome) is further separated into uploading delay, processing delay, downloading delay, and "additional delay." Uploading, processing, and downloading delays correspond to the time durations to upload, process, and download an offloaded job *once*. They are unavoidable for any offloaded job. Additional delay corresponds to total delay minus uploading, processing, and downloading delays. Please note that if a handover happens, jobs in the second category may be re-uploaded or re-processed, and such delays contribute to additional delay; job migration of the third category may also contribute to additional delay.

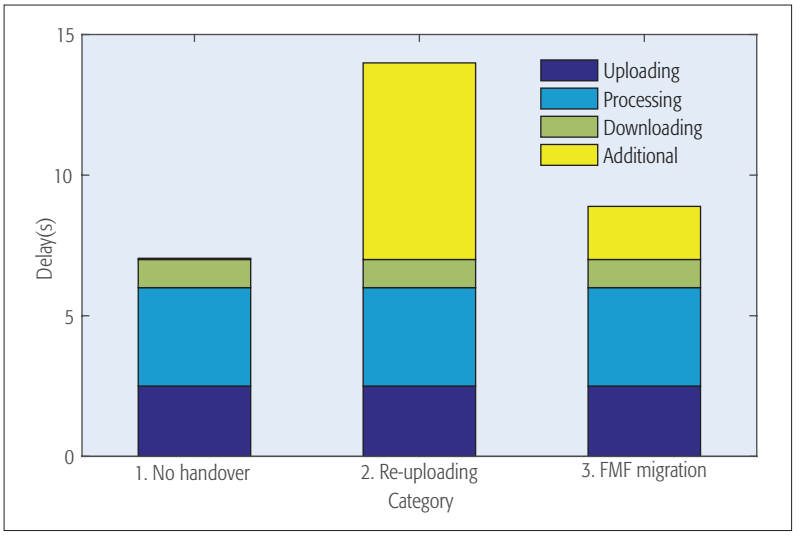


Figure 6. Delay performance by experiment.

From the figure, we notice that for the first category of jobs, the additional delay is quite small. However, for the second category of jobs (FMF is not employed), additional delay is substantially increased since the jobs uploaded to the old fog node are completely wasted, leading to large re-uploading and re-processing delays. For the third category of jobs, since FMF pre-migration is employed, additional delay can be largely avoided. In our experiment, it is 5.1 s smaller than that of the second category, but only 1.9 s greater than that of the first category. When FMF is employed, the additional delay (1.9 s) caused by handover is small compared to unavoidable uploading, processing, and downloading delays (7.0 s in total). The overall delay is reduced by 36.5 percent compared to that of the second category.

Consecutive Handovers

The designed FMF is able to handle the scenario where the IoT device makes two (or more) consecutive handovers within a short period of time (i.e., from fog node 1 to 2 and then from 2 to 3). If there are still missing jobs uploaded to fog node 1 when the second handover occurs, these missing jobs must be recovered at fog node 3. Fog node 1 will send the pending jobs to fog node 2 as usual. Then the jobs will be further forwarded to fog node 3 from fog node 2. The above approach is also applicable when the IoT device makes three or more consecutive handovers.

Using our prototype, we have tested the scenarios where the IoT device moves among three fog nodes (a third laptop is added), making two consecutive handovers within a short period of

[1] Another possible benchmark approach is that job processing remains in the pre-handover serving node. In this case, it is straightforward to show that (1) the delay performance of those jobs uploaded to the old fog node and downloaded from the new fog node is the same as that of FMF, and (2) the delay performance of those jobs uploaded to the new fog node and downloaded from the new fog node is worse than that of FMF, since the new node has to send the unprocessed jobs to the old fog node, and the old fog node has to send the processed jobs to the new fog node.

[2] Due to the space limitation, the experimental results are omitted in this article.

FMF addresses the gap between the requirements for a seamless handover and the lack of an efficient handover timing scheme in a fog computing environment. Our real-world evaluation results showed that FMF can achieve a substantial latency reduction when a mobile device is handed over from one fog node to another.

time. In all test cases, the processed jobs are successfully returned to the IoT device.[2]

Conclusions and Further Discussions

In this article, we propose FMF, a framework supporting a new seamless handover timing scheme among different fog nodes. FMF addresses the gap between the requirements for a seamless handover and the lack of an efficient handover timing scheme in a fog computing environment. Our real-world evaluation results show that FMF can achieve a substantial latency reduction when a mobile device is handed over from one fog node to another.

There can be many future research directions for this study. First, we observe that the conditions (e.g., RSS, job status, and timing) for state transitions influence the delay performance, so our next plan is to further optimize these conditions. Second, in the current design version, the fog will only migrate processed jobs, which may not be suitable for applications with heavy data loads after job processing. The next version of the FMF system will also wisely decide if unprocessed, being processed, or processed jobs should be migrated.

Acknowledgment

Wei Bao would like to acknowledge the support of the University of Sydney DVC Research/Bridging Support Grant. Albert Y. Zomaya would like to acknowledge the support of the Australian Research Council Linkage Grant (LP160100406).

References

[1] P. Mach and Z. Becvar, "Mobile Edge Computing: A Survey on Architecture and Computation Offloading," *IEEE Commun. Surveys & Tutorials*, vol. 19, no. 3, 3rd qtr. 2017, pp. 1628–56.
[2] Y. Mao *et al.*, "Mobile Edge Computing: Survey and Research Outlook," arXiv:1701.01090 [cs.IT], accessed 13 June 2017.
[3] X. Ge, Z. Li, and S. Li, "5G Software Defined Vehicular Networks," *IEEE Commun. Mag.*, vol. 55, no. 7, July 2017, pp. 87–93.
[4] G. Pollini, "Trends in Handover Design," *IEEE Commun. Mag.*, vol. 34, no. 3, Mar. 1996, pp. 82–90.
[5] R. Kwan *et al.*, "On Mobility Load Balancing for LTE Systems," *Proc. IEEE VTC-Fall*, Ottawa, Canada, Sept. 2010, pp. 1–5.
[6] E. Stevens-Navarro, Y. Lin, and V. Wong, "An MDP-Based Vertical Handoff Decision Algorithm for Heterogeneous Wireless Networks," *IEEE Trans. Vehic. Tech.*, vol. 57, no. 2, Mar. 2008, pp. 1243–54.
[7] M. Nelson, B.-H. Lim, and G. Hutchins, "Fast Transparent Migration for Virtual Machines," *Proc. USENIX Annual Technical Conf.*, Anaheim, CA, Apr. 2005, pp. 391–94.
[8] A. Balalaie, A. Heydarnoori, and P. Jamshidi, "Microservices Architecture Enables DevOps: Migration to a Cloud-Native Architecture," *IEEE Software*, vol. 33, no. 3, May–June 2016, pp. 42–52.
[9] T. Harter *et al.*, "Slacker: Fast Distribution with Lazy Docker Containers," *Proc. USENIX Conf. File and Storage Tech.*, Santa Clara, CA, Feb. 2016, pp. 181–95.
[10] T. Taleb and A. Ksentini, "Follow Me Cloud: Interworking Federated Clouds and Distributed Mobile Networks," *IEEE Network*, vol. 27, no. 5, Sept.–Oct. 2013, pp. 12–19.
[11] T. Taleb and A. Ksentini, "An Analytical Model for Follow Me Cloud," *Proc. IEEE GLOBECOM*, Atlanta, GA, Dec. 2013, pp. 1291–96.
[12] S. Wang *et al.*, "Dynamic Service Migration in Mobile Edge-Clouds," *Proc. IFIP Networking*, Toulouse, France, May 2015, pp. 1–9.
[13] S. Wang *et al.*, "Dynamic Service Placement for Mobile Micro-Clouds with Predicted Future Costs," *IEEE Trans. Parallel and Distrib. Systems*, vol. 28, no. 4, Apr. 2017, pp. 1002–16.
[14] X. Sun and N. Ansar, "PRIMAL: Profit Maximization Avatar Placement for Mobile Edge Computing," *Proc. IEEE ICC*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.
[15] A. H. Zahran, B. Liang, and A. Saleh, "Signal Threshold Adaptation for Vertical Handoff in Heterogeneous Wireless Networks," *Mobile Networks and Applications*, vol. 11, no. 4, Aug. 2006, pp. 625–40.

Biographies

Wei Bao received his Ph.D. degree in electrical and computer engineering from the University of Toronto, Canada, in 2016. He is currently a lecturer at the School of Information Technologies, University of Sydney, Australia. His research covers the area of network science, with particular emphasis on 5G systems, the Internet of Things, and mobile computing.

Dong Yuan [M] received his Ph.D. degree from Swinburne University of Technology, Australia, in 2012. He is a lecturer in the School of Electrical and Information Engineering, University of Sydney. His research interests include cloud computing, data management in parallel and distributed systems, scheduling and resource management, the Internet of Things, business process management, and workflow systems.

Zhengjie Yang received his Master of Information Technology in software engineering (research pathway) from the University of Sydney. He is interested in networking, wireless technologies, software development, mobile computing, cloud computing, and big data.

Shen Wang is a Master's student at the School of Information Technologies, University of Sydney. His research interests include gaming, networking, and software.

Wei Li [SM] received his Ph.D. degree from the School of Information Technologies at the University of Sydney. He is currently a research associate in the Centre for Distributed and High Performance Computing, School of Information Technologies, University of Sydney. His current research interests include wireless sensor networks, the Internet of Things, task scheduling, energy-efficient algorithms, and optimization. He is a member of ACM.

Bing Bing Zhou (bing.zhou@sydney.edu.au) received his B.S. degree from Nanjing Institute of Technology, China, and his Ph.D. degree in computer science from Australian National University. He is currently an associate professor at the University of Sydney. His research interests include parallel/distributed computing, cloud computing, parallel algorithms, IoT and bioinformatics. He has a number of publications in leading international journals and conferences. His research has been funded by the Australian Research Council through several Discovery Project grants.

Albert Y. Zomaya [F] is a Chair Professor and director of the Centre for Distributed and High Performance Computing at the University of Sydney. He haspublished more than 500 scientific papers and is an author, co-author, or editor of more than 20 books. He is the Editor-in-Chief of *IEEE Transactions on Sustainable Computing* and serves as an Associate Editor for 22 leading journals. He is a Fellow of AAAS and IET.

CALL FOR PAPERS

IEEE COMMUNICATIONS MAGAZINE

ENABLING COMMUNICATION AND NETWORKING TECHNOLOGIES FOR EDGE COMPUTING

BACKGROUND

With the advancements in Internet of Things (IoTs), billions of devices are expected to generate a huge amount of data that not only inundates the communication networks but may also lead to ineffectiveness of conventional computing paradigms for analysis. The conventional computing approaches require transporting data from source devices over a network, storing it on centralized servers, and then performing analysis. The increase in volume elevates the complexity and costs of transporting and storing the data, and time required to analyze the data. To mitigate these issues, Edge Computing has been introduced that aims to bring the Cloud resources and services closer to data sources. The arrival of Edge Computing facilitates analyzing the IoT data closer to the sources, making applications more responsive to dramatically changing local conditions while avoiding communication bottlenecks of wide area networks. Several new wireless communication technologies are emerging to access the Edge Computing resources and services. These technologies include, but not limited to, Wi-Fi HaLow, LTE, and 5G. Considering the significance of communication and networking technologies for successful deployment of Edge Computing, there exists a need for conducting research to further investigate the standardization efforts and explore different issues/challenges in the communication technologies and Edge Computing networks. The investigation of communication and networking aspects will assist in dramatically increasing IoT capabilities, multiplying practical applications while potentially reducing costs. These new wireless communications networks are vital to support the emerging Edge Computing services in scalable real-time manner.

This Feature Topic (FT) focuses on the crossroads between scientists, industry practitioners, and researchers from different domains in the communication technologies and Edge Computing environments. We envision providing a platform for researchers to further explore the domain and explore the challenges. In this FT, we invite researchers from academia, industry, and government to discuss challenging ideas, novel research contributions, demonstration results, and standardization efforts on enabling wireless communication and networking technologies for Edge Computing.

In this FT, we would like to try to answer some (or all) of the following questions: How can communication and networking technologies improve the performance and services provided by Edge Computing? How can we evaluate the impact of communication and networking technologies on Edge Computing? What are the key communication and networking technological challenges that hinder the success of Edge Computing? How can we standardize the wireless interfaces of devices for communication in Edge computing?

Topics of interest include, but are not limited to:

- Resource and network management in Edge Computing
- Quality of Service mechanisms for wireless Edge Computing networks
- Interactions between the Edge and the Cloud
- Device-to-Device communication and cooperation
- Integration and co-existence of technologies and networks for Edge Computing
- Interoperability between heterogeneous access networks of Edge Computing
- Energy-aware wireless protocols and algorithms for wireless Edge Computing networks
- Wireless communications and networking architecture for Edge Computing systems
- Virtualization in wireless Edge Computing networks
- Software defined networking for the Edge Computing
- Experimental network measurements and characterization for Edge Computing data traffic
- Cross layer design and optimization of Edge Computing networks
- Security and privacy concerns of Edge Computing communication and networking technologies

SUBMISSION GUIDELINES

Articles should be tutorial in nature, with the intended audience being all members of the communications technology community. They should be written in a tutorial style comprehensible to readers outside the specialty of the article. Mathematical equations should not be used (in justified cases up to three simple equations are allowed). Articles should not exceed 4500 words (from introduction through conclusions, excluding figures, tables and captions). Figures and tables should be limited to a combined total of six. The number of archival references is recommended not to exceed 15. In some rare cases, more mathematical equations, figures, and tables may be allowed, if well-justified. In general, however, mathematics should be avoided; instead, references to papers containing the relevant mathematics should be provided. Complete guidelines for preparation of the manuscripts are posted at http://www.comsoc.org/commag/paper-submission-guidelines. Please submit a PDF (preferred) or MSWORD formatted paper via Manuscript Central (http://mc.manuscriptcentral.com/commag-ieee). Register or log in, and go to Author Center. Follow the instructions there. Select "August 2018/Enabling Communication and Networking Technologies for Edge Computing" as the Feature Topic category for your submission.

IMPORTANT DATES

- Manuscript Submission Deadline: December 1, 2017
- Decision Notification: April 1, 2018
- Final Manuscripts Due: May 15, 2018
- Publication Date: August 2018

GUEST EDITORS

Ejaz Ahmed
University of Malaya, Malaysia
imejaz@gmail.com

Ammar Rayes
Cisco Systems, SF, USA
rayes@cisco.com

Muhammad Imran
King Saud University, Saudi Arabia
dr.m.imran@ieee.org

Joel J.P.C. Rodrigues
National Institute of Telecommunications (Inatel), Brazil
Instituto de Telecomunicações, Portugal
joeljr@ieee.org

Albert Zomaya
Sydney University, Australia
albert.zomaya@sydney.edu.au

Wael Guibene
Intel Corporation, Santa Clara, CA, USA
wael.guibene@intel.com

A Cloud-Edge Computing Framework for Cyber-Physical-Social Services

Xiaokang Wang, Laurence T. Yang, Xia Xie, Jirong Jin, and M. Jamal Deen

The authors present a tensor-based cloud-edge computing framework that mainly includes the cloud and edge planes. The cloud plane is used to process large-scale, long-term, global data, which can be used to obtain decision-making information such as the feature, law, or rule sets. The edge plane is used to process small-scale, short-term, local data, which is used to present the real-time situation.

ABSTRACT

Cyber-physical-social systems (CPSSs) represent an emerging paradigm encompassing the cyber world, physical world and social world. One of the main purposes of CPSSs is to provide high-quality, proactive, and personalized services for humans. For CPSSs to realize this purpose, a novel services framework is needed. In this article, we present a tensor-based cloud-edge computing framework that mainly includes the cloud and edge planes. The cloud plane is used to process large-scale, long-term, global data, which can be used to obtain decision making information such as the feature, law, or rule sets. The edge plane is used to process small-scale, short-term, local data, which is used to present the real-time situation. Also, personalized services will be directly provided for humans by the edge plane according to the obtained feature, law, or rule sets and the local high-quality data obtained in the edge plane. Then a tensor-based services model is proposed to match the requirement of users in the local CPSS. Finally, a case study about CPSS services is proposed to demonstrate the application features of the proposed framework.

INTRODUCTION

Over the past few decades, the rapid development of the Internet of Things (IoT), also referred as cyber-physical systems (CPSs), has accelerated the digital revolution and enhanced intelligent human living environments. As an extension of CPS (IoT), cyber-physical-social systems (CPSSs), including the cyber world, physical world, and social world, are proliferating in all aspects of our daily lives. One of the main purposes of CPSSs is to enhance living environments by providing high-quality proactive and personalized services to humans [1–3].

In CPSSs, large-scale data about our daily lives are continually generated from three worlds. Flowing around the three worlds and recording all aspects of our daily lives, data, as the common element, were selected as the starting point of our research [1]. However, large-scale data collected from CPSSs are often redundant, complex, noisy, and low-quality, resulting in unprecedented challenges for providing CPSS services [1]. Furthermore, to provide high-quality services in CPSSs, a comprehensive analysis about big data based on cloud computing is essential.

Cloud computing, with its powerful computational potential, has triggered enormous attention for CPSS big data processing [1, 4]. However, cloud computing has faced increasing computational demands because of the exponential growth of big data, including both large-scale, long-term, global data and small-scale, short-term, local data. Also, the huge computational tasks on the cloud bring others challenging issues related to cost, energy consumption, and quality of the provided services. With the computational capability inherent in many smart devices such as smartphones, innovative research in this challenging area has been motivated by the rise of spatial edge computing [5]. Therefore, a novel computing framework to systematically and comprehensively process those data for providing CPSS services is required.

For example, thousands of cameras have been installed to monitor the traffic conditions in smart cities [6]. With these cameras, large-scale, complex, long-term, global data about city traffic will be collected over a long period. Cloud computing can be used to process the collected data and to obtain the characteristics or rules of the traffic congestion. Then smart devices around cameras can be used to estimate whether the traffic congestion happens according to the obtained characteristics or rules of the traffic congestion and the real-time traffic data. Figure 1 shows the relationship among CPSSs, cloud computing, edge computing, and services. For providing services in this manner, several challenges, including the following three questions, must be addressed.

1. How can we effectively process the big data to obtain high-quality data?
2. How can we analyze the users' corresponding data from the perspective of multi-order attributes?
3. How can we match the requirements, hobbies, and habits according to users' local environments?

In this article, we present a cloud-edge computing framework for providing CPSS services, and the main contributions are as follows. First, we present a novel cloud-edge computing framework for efficiently integrating cloud computing and edge computing. Second, we

Digital Object Identifier:
10.1109/MCOM.2017.1700360

Xiaokang Wang, Xia Xie (corresponding author), and Jirong Jin are with Huazhong University of Science and Technology; Laurence T. Yang is with Huazhong University of Science and Technology and St. Francis Xavier University; M. Jamal Deen is with McMaster University.

0163-6804/17/$25.00 © 2017 IEEE

propose two kinds of tensor integration methods, including the multi-order outer product and multi-order integration strategy, to combine several attributes together. Third, we propose a tensor-based services model used to match the requirements, hobbies, and habits of users. Fourth, we use the multi-order distributed incremental method in this framework to improve the processing efficiency. To describe these contributions, this article is divided into five sections. We provide background information and motivation for our work. We briefly summarize related works including CPSSs, big data, tensor, cloud computing, and edge computing. An overview of the framework and a tensor-based services model are provided. A case study about CPSS services matching is discussed. Finally, conclusions are given.

Related Work

In this section, a concise review of the state of the art in CPSSs, big data, tensor and its high-order singular value decomposition (HOSVD), cloud computing, and edge computing is provided.

Cyber-Physical-Social Systems: Integrating the cyber, physical, and social worlds, CPSSs are considered as a hybrid world where community detection among humans, objects, and cyber actors can be obtained [1]. Many publications about object identification in CPSSs are available to precisely recognize users [7, 8], understand their requirements under different conditions, and even appropriately render high-quality services [8]. However, matching methods, referred to as the basis of precise identification of an object in CPSSs, have been studied and used in many applications such as human motion recognition [9], health [10], human face identification [11], and smart home [6, 12, 13]. Additionally, CPSS big data, as the research starting point of CPSSs, are typically high-order, complex, large-scale, redundant, and noisy. Therefore, systematic methods for representation, integration, and processing of CPSS big data are needed.

Big Data, Tensor, and HOSVD: With the 4V characteristics — volume, variety, veracity, and velocity — CPSS big data bring unprecedented challenges using existing computational methods and models [14]. The tensor, as an extension of a matrix in high-order space, was proven to be an appropriate, reasonable big data representation method [15]. HOSVD, one of the main tensor decomposition methods, was used to process big data for denoising and removing redundant data [1, 15]. To improve the computational efficiency of HOSVD, a tree-based multi-order distributed HOSVD (MDHOSVD), with its incremental computing method in which a tensor can be divided or increased along several or even all orders at the same time, was discussed in [1].

Cloud Computing and Edge Computing: The emergence of cloud computing and its potential to provide computation and storage capabilities for services was propelled by the availability of powerful processing hardware and software. Edge computing is emerging as a novel computational paradigm and is a result of the rapid advances in IoT [5]. Currently, smart devices such as smartphones, laptops, and embedded devices are significantly changing the way we conduct our daily activities, and these smart devices are endowed with increasingly powerful computational capacities. In [5], a detailed survey about mobile edge computing was presented. Also, several applications and key technologies of mobile edge computing were discussed.

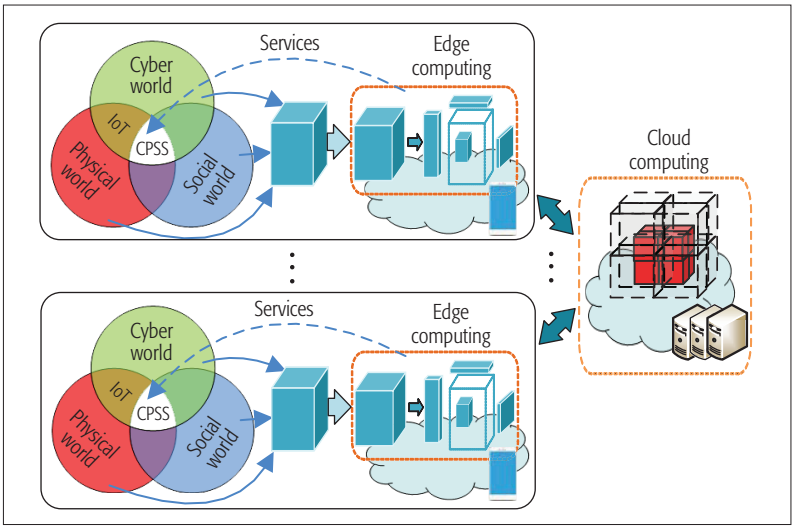


Figure 1. Relationship among a CPSS, cloud computing, edge computing and services.

Overview of the Cloud-Edge Computing Framework

In Fig. 2, as an extension of our previous work in [1], a cloud-edge computing framework including the edge plane, cloud plane, and application plane was studied. Here, the three parts — edge plane, cloud plane, and application plane — with relevant examples are given. The functions of these three parts are now described.

Edge Plane: In this plane, two main tasks of data representation and its initial cleaning, matching, and services providing are performed. A tensor model with its HOSVD is used to represent collected data and extract corresponding high-quality data. Also, matching and services will be implemented and provided directly to users, according to these high-quality data and the obtained feature, law, or rule sets from the cloud plane.

Cloud Plane: In the cloud plane, two important tasks — global data integration and processing, and the acquisition of feature, law, or rule sets — are accomplished. Two methods are used to implement data integration. Also, MDHOSVD is used to obtain the high-quality data of the global data. Furthermore, corresponding methods are implemented to obtain the feature, law, or rule sets in this plane.

Application Plane: In the application plane, high-quality data, as well as the feature, law, or rule sets, are applied for some applications.

The three planes in Fig. 2 complement each other and are used to construct the framework, which forms the basis for CPSS applications and services. In the following subsections, we provide brief descriptions of each plane. In this article, we mainly pay attention to the interconnected operations between the cloud plane and

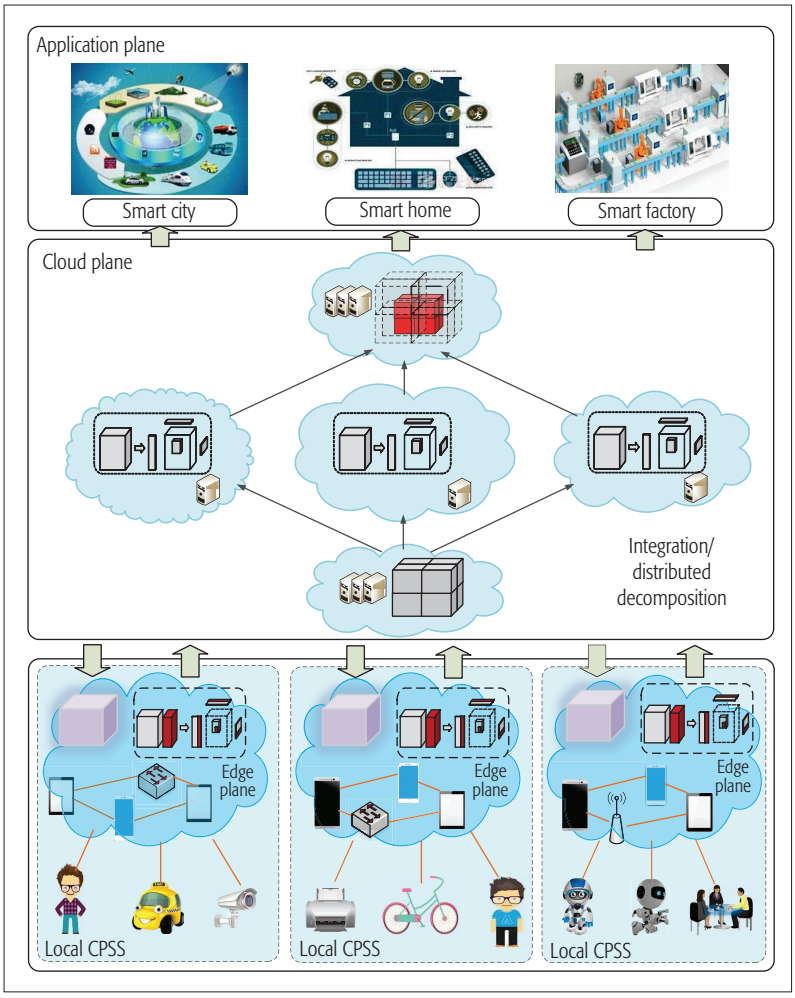


Figure 2. Cloud-edge computing framework.

the edge plane to provide high-quality services for humans.

Data Representation and Reduction in the Edge Plane

In this article, a tensor, a data representation tool, is used to represent the various data collected in the edge plane. In our previous work [15], a tensor-based big data unified representation method was used to represent unstructured data (e.g., video clips), semi-structured data (e.g., XML documents), and structured data (e.g. GPS). Also, HOSVD is employed to process the collected data for obtaining high-quality data. For example, the HOSVD results of an Nth-order tensor $\mathcal{A}$ are computed as

$$\mathcal{S} = \mathcal{A} \times_1 U_1^T \times_2 U_2^T \cdots \times_n U_n^T \cdots \times_N U_N^T, \quad (1)$$

where, $\mathcal{S}$ is the core tensor and matrix U_n, $1 \le n \le N$ is the left singular matrix of the nth order. The core tensor $\mathcal{S}$ and matrix U_n, $1 \le n \le N$ as well as its approximate tensor are referred to as the high-quality data of tensor $\mathcal{A}$ [1, 15].

In every local CPSS, data are continuously generated from the perspective of multi-attributes, represented as the multi-order tensor streaming, which will be processed by multi-order incremental HOSVD (MIHOSVD) as discussed in [1]. Otherwise, in local CPSSs, some simple computational tasks such as matching or ranking will be implemented on the edge plane.

Tensors Integration and Processing in Cloud Plane

In this part, two main tasks — integration and processing of tensors — are discussed. Two methods of tensors integration are proposed to integrate several associated attributes to make sure that they can be analyzed together. Then MDHOSVD is employed to process the integrated tensor and to improve the computational efficiency.

Multi-Order Outer Product: To provide services, the associated collected data should be analyzed from the perspective of multi-attributes to extract the useful information such as the multi-order face feature or multi-order gait feature. Although a tensor is used to represent the multi-attributes data, there is still a challenging question because it cannot completely represent the associated data such as the tensor of face data or that of the same person's gait data. Therefore, a multi-order outer product is proposed to integrate different tensors, which have the same attributes such as persons to uniformly represent these multi-attributes data together.

As shown in Fig. 3, an Nth-order tensor $\mathcal{A}$ and an Mth-order tensor $\mathcal{B}$ have k common orders $I_{n_1} = J_{n_1}$, $I_{n_2} = J_{n_2}$, $\cdots$, $I_{n_k} = J_{n_k}$, referred to as targeted orders. A multi-order outer product to integrate these two tensors and obtain an $(N+ M - k)$ th-order result tensor $\mathcal{C}$, which has all orders from these two tensors that do not have redundant orders, is performed as follows.

1. Tensor $\mathcal{A}$ is fully divided along all k common orders to obtain a series of sub-tensors $\mathcal{A}^{(i_{n_1}, i_{n_2}, \cdots, i_{n_k})}$, $1 \le i_{n_k} \le I_{n_k}$, where the superscript $(i_{n_1}, i_{n_2}, \cdots, i_{n_k})$ is the coordinate of this sub-tensor in tensor $\mathcal{A}$. In the same way, tensor $\mathcal{B}$ is also fully divided into a series of sub-tensors along all the k common orders.
2. Then a sub-tensor $\mathcal{C}^{(i_{n_1}, i_{n_2}, \cdots, i_{n_k})}$ will be computed by the outer product between sub-tensor $\mathcal{A}^{(i_{n_1}, i_{n_2}, \cdots, i_{n_k})}$ and sub-tensor $\mathcal{B}^{(i_{n_1}, i_{n_2}, \cdots, i_{n_k})}$.
3. By reorganizing sub-tensors $\mathcal{C}^{(i_{n_1}, i_{n_2}, \cdots, i_{n_k})}$ according to their coordinates, a resultant tensor $\mathcal{C}$ is obtained. $\mathcal{C}$ has all the orders of both tensors $\mathcal{A}$ and $\mathcal{B}$.

Multi-Order Integration Strategy: The multi-order integration strategy is proposed to integrate the k targeted orders into a hybrid order I_H in a certain tensor such as tensor $\mathcal{A}$. Then a hybrid tensor $\mathcal{A}^h$ will be obtained, in which the dimensionality of the hybrid order is equal to $I_H = I_{n_1} \times I_{n_2} \times \cdots \times I_{n_k}$.

After integrating tensors, we get a large-scale hybrid tensor in the cloud plane. Currently, we prefer to utilize MDHOSVD to process this large-scale hybrid tensor, thus obtaining its high-quality data set.

MDHSOVD: The purpose of MDHOSVD is to realize the HOSVD result of the large-scale tensor according to that of its sub-tensors. Taking the MDHOSVD of an Nth-order $\mathcal{A}$ as an example to explain its computational process, it can be abstracted into a tree with $(N + 1)$ layers and summarized as follows.

1. Each sub-tensor will be sent to a node of the $(N + 1)$th layer, where each sub-tensor will be unfolded along N orders, respectively. Then singular value decomposition will be implemented on these N unfolding matrices in each employed node.

2. Next, the computational result of each unfolding matrix will be sent to a node of the Nth layer. From this step, integrate the sub-tensors along the jth order in the jth layer, where j is selected from N to 1. Then the task is to obtain the singular value decomposition of each unfolding matrix of the integrated tensor according to that of each sub-tensor. After realizing this computational task in the first layer, we will get the left singular value matrix U_i, $1 \le i \le N$. Then the HOSVD of tensor $\mathcal{A}$ will be obtained according to Eq. 1.

A Tensor-Based Services Model

Based on the cloud-edge computing framework, a tensor-based services model is proposed in this article. To clearly express this model, the ServiceTensor is first proposed.

- **ServiceTensor**: The ServiceTensor is used to represent the requirements, interests, and habits in scenarios of local CPSSs. The ServiceTensor of a user is represented as an Nth-order tensor $\mathcal{A} \in R^{I_1 \times I_2 \times \cdots I_{n_1} \times I_{n_2} \cdots \times I_{n_k} \cdots \times I_N}$, where requirements, interests, and habits are represented by k targeted orders under a scenario of a local CPSS represented by the combination of the left $(N - k)$ untargeted attributes.

Next, the working process of the tensor-based services model is summarized as follows.

1. In the cloud plane, ServiceTensor $\mathcal{A}$ will be established by a learning method such as deep learning, according to the long-term, large-scale global data.
2. When a user gets into a local CPSS, sensing devices will collect his/her data in a real-time scenario with multi-attributes. Normally, tensor $\mathcal{B} \in R^{I_1 \times I_2 \times \cdots I_{n_1-1} \times I_{n_k+1} \cdots \times I_N}$ is used to represent the current scenario, in which a certain order is used to represent an attribute. Also, the edge plane process tensor $\mathcal{B}$ by HOSVD is used to obtain its high-quality data. Then the high-quality data of tensor $\mathcal{B}$ will be sent to the cloud plane, where a tensor equation

$$\mathcal{A} \times \begin{matrix} I_{n_1} I_{n_2} \cdots I_{n_k} \\ I_{n_1} I_{n_2} \cdots I_{n_k} \end{matrix} \mathcal{X} = \mathcal{B}, \tag{2}$$

has to be solved to obtain the requirements, interests and habits tensor $\mathcal{X}$. The operation

$$\times \begin{matrix} I_{n_1} I_{n_2} \cdots I_{n_k} \\ I_{n_1} I_{n_2} \cdots I_{n_k} \end{matrix}$$

means multi-order product on the k targeted orders. The main steps of solving this equation are summarized as follows.

a. For tensor $\mathcal{A}$ in Eq. 2, integrate the k targeted orders into a hybrid order I_H and obtain a hybrid tensor $\mathcal{A}^h$.

b. Next, multi-order products on the untargeted orders by $\mathcal{A}^h$ on both tensors $\mathcal{A}^h$ and $\mathcal{B}$ are carried out. After this, we can change the high-order tensor of Eq. 2 into a matrix equation.

c. For the matrix equation in b, we can use the solution of the matrix equation to get the corresponding answer vector. According to the "multi-order integration strategy" and corresponding answer vector, we will get the kth-order tensor $\mathcal{X}$.

3. In the edge plane, fully divide the approximate tensor according to all the untargeted orders. Then we will get a series of sub-tensors $\mathcal{A}_m$. According to [9], a similar matching ratio α_m can also be used to estimate the requirements. Then we will queue a series of matching ratios α_m from large to small values. The requirement tensor $\mathcal{A}_m$ corresponding to the maximal matching ratio is the goal we seek.
4. Finally, the edge plane will inform some smart devices such as smartphones or even a robot to provide services, according to the requirement tensor $\mathcal{A}_m$.

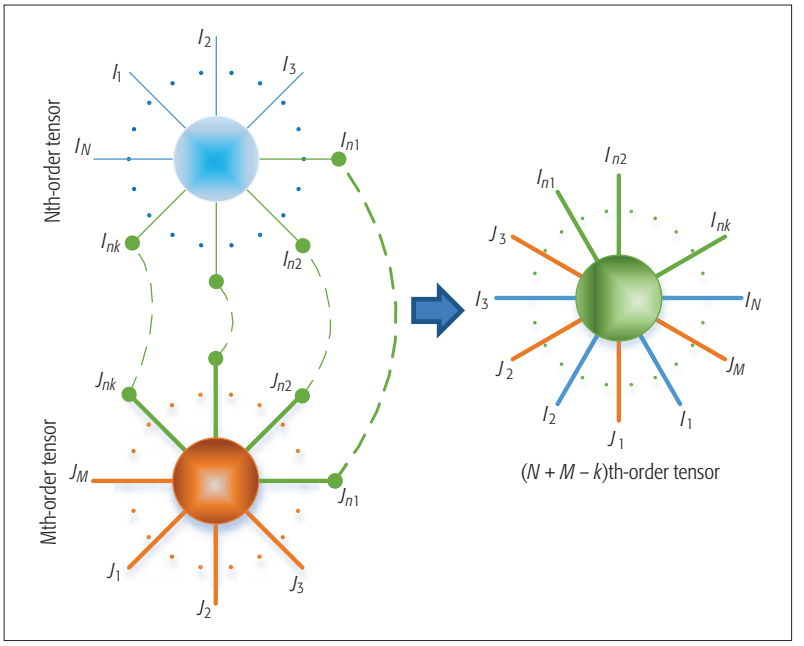


Figure 3. Tensors integration.

Case Study

In this section, a case study about the proactive and personalized services provision is used to illustrate the working process of the tensor-based services model on the cloud-edge computing framework. As shown in Fig. 4, five users, whose ServiceTensors $\mathcal{A}_0$ have been prepared by the cloud plane, get into a local CPSS.

Here, for convenience, we suppose that the corresponding requirements are based on the combination of their face data and gait data in different scenarios. With a similar representation manner to [9, 11], two 4th-order tensors $\mathcal{F} \in R^{I_1 \times I_2 \times I_3 \times I_4}$ and $\mathcal{T} \in R^{J_1 \times J_2 \times J_3 \times J_4}$ are used to represent facial data and gait data, respectively, where $I_1 = J_1 = 5$ means five users; $I_2 = 7$ means seven different expressions including speaking, anger, sadness, shouting, peaceful, happy, and hurried; I_3 and I_4 are used to represent the height and width of the face photo; J_2 means the angle between walking direction and the camera (including 15° to left, 30° to left, 60° to left, 15° to right, 30° to right, 60° to right); J_3 means the number of frames in a walking cycle (also the number of photos in a cycle); and $J_4 = 7$ means the selected joint angle including the front shoulder joint angle, back

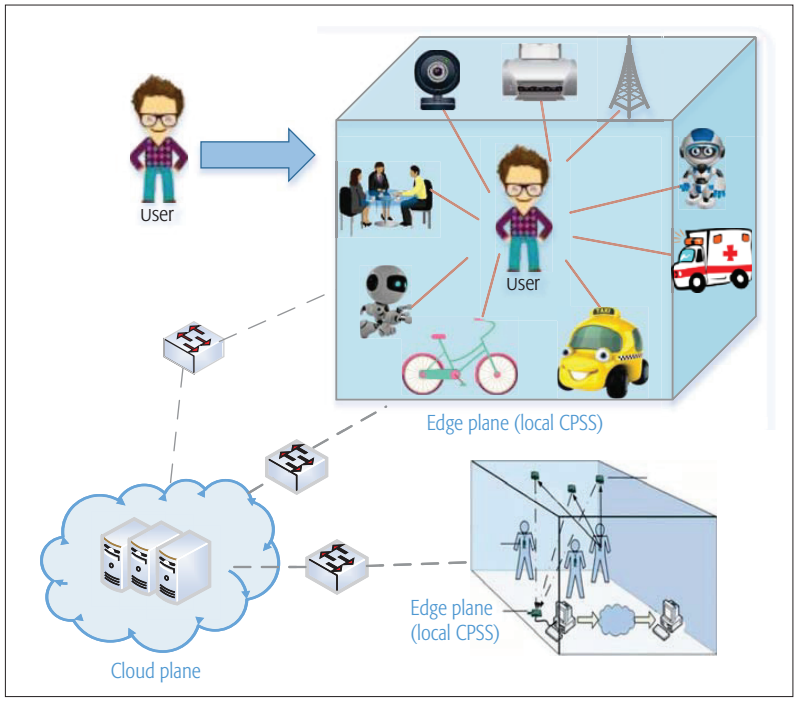


Figure 4. A case study of tensor-based services.

shoulder joint angle, joint angle between two legs, and joint angle of two knees and two ankles.

Also, these requirements were prepared by learning methods in the cloud plane and represented by tensor $\mathcal{A}_0 \in R^{I_1 \times I_2 \times I_3 \times I_4 \times J_2 \times J_3 \times J_4}$. For example, when the third user ($i_1 = 3$) feels very tired and wants to sleep, he will be very sad ($i_2 = 3$) and the walking direction is 15° to the right ($j_2 = 4$). When the fourth user ($i_1 = 4$) is going to a restaurant for dinner, she is always talking with her friend ($i_2 = 1$), and the walking direction is 60° to the left ($j_2 = 3$). When the fourth user ($i_1 = 4$) is hurrying ($i_2 = 7$) and his walking direction is 60° to the right ($j_2 = 6$), he is very busy and needs a vehicle such as a taxi or a shared bicycle.

Now, we take the matching in three scenarios to demonstrate the performance of the tensor-based service model. Suppose a user gets into a local CPSS, and her corresponding data, including the face data $\mathcal{F}^i \in R^{I_1^i \times I_2^i \times I_3 \times I_4}$, $1 \le i \le 3$, $I_1^i = I_2^i = 1$ and gait data $\mathcal{T}^i \in R^{I_1^i \times J_2^i \times J_3 \times J_4}$, $1 \le i \le 3$, $I_1^i = J_2^i = 1$ in a certain scenario, are collected. Then the hybrid tensor $\mathcal{B}^i \in R^{I_1^i \times I_2^i \times I_3 \times I_4 \times J_2^i \times J_3 \times J_4}$, $1 \le i \le 3$, $I_1^i = I_2^i = J_2^i = 1$ is obtained by integrating tensors $\mathcal{F}^i$ and $\mathcal{T}^i$ using the multi-order outer product. $\mathcal{A}_0^j \in R^{I_1^j \times I_2 \times I_3 \times I_4 \times J_2 \times J_3 \times J_4}$, $1 \le j \le 5$ is the sub-tensor divided from the tensor $\mathcal{A}_0$ along the first order. Also, the corresponding high-order tensor equation will obtained according to Eq. 2, which will be processed in the cloud plane. Then the matching ratios will be computed by the edge plane, as shown in Fig. 5.

Finally, the matching results of the three mentioned scenarios are shown in Figs. 5a, 5b, and 5c, respectively. Taking Fig. 5a as an example, the first order is the hybrid order of I_2J_2, which has seven groups corresponding to seven expressions. Each group has six results corresponding to six degrees. The second order means five users, and the third order means the matching ratio. Taking the matching result of the first set scenario as an example, we find that the maximal matching ratio corresponds to ($i_1 = 3$ (the third user), $i_2 = 3$ (the third situation, "sadness"), $j_2 = 4$ (the fourth situation of degree)), which means that the third user is very tired and wants to sleep. Then the proactive and personalized services such as a shower with appropriate temperature will be automatically opened and the window curtain will be automatically closed. Also, the matching results of the second and third scenarios are ($i_1 = 4$, $i_2 = 1$, $j_2 = 3$) and ($i_1 = 4$, $i_2 = 7$, $j_2 = 6$), respectively. Then the fourth user will be told which restaurant that has her favorite food and is convenient with fewer customers in the second scenario. In the third scenario, the fourth user will be informed where the nearest shared bicycle stand is or even inform a nearby taxi to pick him up.

Conclusion

In this article, a cloud-edge computing framework including the cloud plane and edge plane is discussed. This framework, based on a tensor-based services model, is used to provide high-quality proactive and personalized services for humans. In addition, we introduce a tensor-based services model based on ServiceTensor for different scenarios. Also, for even more efficient service provision, several improvements such as optimized methods on computation between cloud plane and edge plane, and improved matching methods will be studied in future.

Acknowledgment

M. Jamal Deen gratefully acknowledges the Canada Research Chair Program for their support of his work.

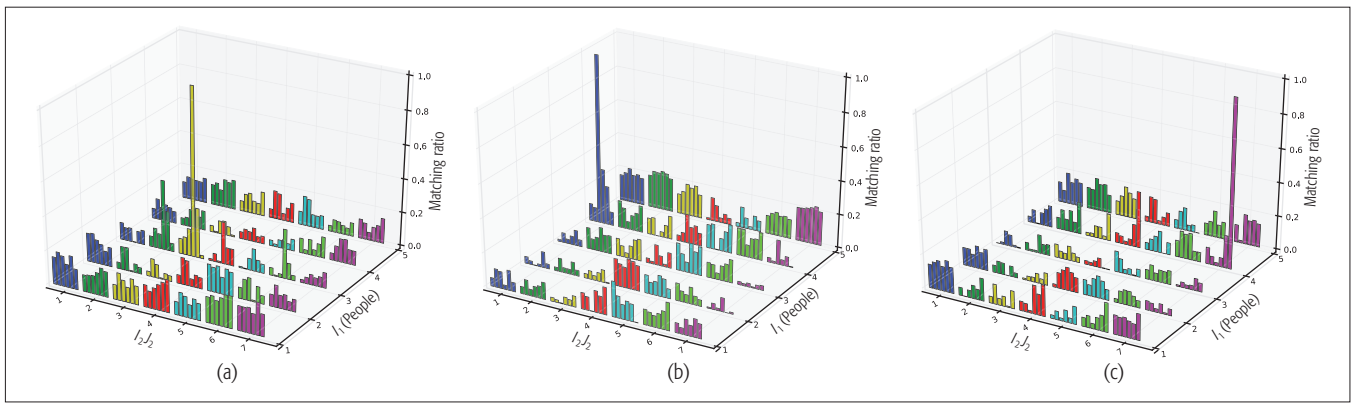


Figure 5. Matching results.

References

[1] X. Wang *et al.*, "A Tensor-Based Big Service Framework for Enhanced Living Environments," *IEEE Cloud Computing Mag.*, vol. 3, no. 6, 2016, pp. 36–43.

[2] J. Zeng *et al.*, "A Systematic Methodology for Augmenting Quality of Experience in Smart Space Design," *IEEE Wireless Commun.*, vol. 22, no. 4, Aug. 2015, pp. 81–87.

[3] H. Li *et al.*, "Mobile Crowdsensing in Software Defined Opportunistic Networks," *IEEE Commun. Mag.*, vol. 55, no. 6, July 2017, pp. 140–45.

[4] V. Marx, "Biology: The Big Challenges of Big Data," *Nature*, vol. 498, 2013, pp. 255–60.

[5] A. Ahmed and E. Ahmed, "A Survey on Mobile Edge Computing," *Proc. 10th IEEE Int'l. Conf. Intelligent Systems and Control*, Coimbatore, India, Jan. 7–8, 2016, pp. 1–8.

[6] S. Majumder, T. Mondal, and M. J. Deen, "Wearable Sensors for Remote Health Monitoring," *Sensors*, vol. 17, no. 1, 2017, pp. 1–45.

[7] I. Kviatkovshy, I. Shimshoni, and E. Rivlin, "Person Identification from Action Styles," *Proc. 2015 IEEE Conf. Computer Vision and Pattern Recognition Wksps.*, Boston, MA, June 7–12, 2015, pp. 84–92.

[8] D. Yao *et al.*, "Human Mobility Synthesis Using Matrix and Tensor Factorizations," *Information Fusion*, vol. 23, no. C, 2014, pp. 25–32.

[9] M.A.O. Vasilescu, "Human Motion Signatures: Analysis, Synthesis, Recognition," *Proc. 2002 Int'l. Conf. Pattern Recognition*, Quebec City, Canada, Aug. 11–15, 2002, pp. 456–60.

[10] P. Mandal, K. Tank, T. Mondal, C. Chen, and M.J. Deen, "Predictive Walking-Age Health Analyzer," *IEEE J. Biomedical and Health Informatics*, vol. PP, no. 99, 2017, pp. 1–1.

[11] M.A.O. Vasilescu and D. Terzopoulos, "Multilinear Analysis of Image Ensembles:TensorFaces," *Proc. 2002 Euro. Conf. Computer Vision*, Copenhagen, Denmark, May 28–31, 2002, pp. 1–7.

[12] M. Tao, K. Ota, M. Dong, "Ontology-Based Data Semantic Management and Application in IoT and Cloud-Enabled Smart Homes," *Future Generation Computer Systems*, vol. 76, 2017, pp. 528–39.

[13] M. J. Deen, "Information and Communications Technologies for Elderly Ubiquitous Healthcare in a Smart Home," *Personal and Ubiquitous Computing*, vol. 19, no. 3–4, 2015, pp. 573–99.

[14] Q. Zhang, L. T. Yang, and Z. Chen, "Deep Computation Model for Unsupervised Feature Learning on Big Data," *IEEE Trans. Services Computing*, vol. 9, no. 1, 2016, pp. 161–71.

[15] L. Kuang *et al.*, "A Tensor-Based Approach for Big Data Representation and Dimensionality Reduction," *IEEE Trans. Emerging Topics in Computing*, vol. 2, no. 4, 2014, pp. 280–91.

Biographies

XIAOKANG WANG (wangxiaokang1002@163.com) is with the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China. His research interests are cyber-physical-social systems, parallel and distributed computing, and big data.

LAURENCE T. YANG [S' 15] (ltyang@gmail.com) is a professor at the School of Computer Science and Technology, Huazhong University of Science and Technology, and in the Department of Computer Science at St. Francis Xavier University, Canada. His research interests include cyber-physical-social systems, parallel and distributed computing, embedded and ubiquitous computing, and big data.

XIA XIE (shelicy@qq.com) received her B.E. degree and Ph.D. degree in computer science and technology from Huazhong University of Science and Technology. She is an associate professor in the School of Computer Science and Technology at Huazhong University of Science and Technology. Her research interests include parallel and distributed computing, data mining and analysis, and big data.

JIRONG JIN (ruanyijinjirong@163.com.) is a staff member at the Haikou Rural Commercial Bank. Her research interest is big data.

M. JAMAL DEEN [F' 02] (jamal@mcmaster.ca) is a Distinguished University Professor and Senior Canada Research Chair in Information Technology at McMaster University, Hamilton, Canada. He is currently the President of the Academy of Science, Royal Society of Canada (RSC). His research interests include nano-/opto-electronics, nanotechnology, and their emerging applications in health and environment.

Cyber-Physical-Social Systems (CPSS) represent an emerging paradigm encompassing the cyber world, physical world and social world. One of the main purposes of CPSS is to provide high quality, proactive and personalized services for humans. For CPSS to realize this purpose, a novel services framework is needed.

Improving Opportunistic Networks by Leveraging Device-to-Device Communication

Radu-Corneliu Marin, Radu-Ioan Ciobanu, and Ciprian Dobre

The authors extend their proposed architecture for covering infrastructure-less environments through a novel mechanism that allows connecting peers through Wi-Fi Direct in a seamless and secure fashion. They emulate real-life user traces to empirically prove that their solution improves hit rate by 13 percent without impacting battery life.

ABSTRACT

Several popular low-level device-to-device techniques, such as Bluetooth and Wi-Fi Direct, are seen today as enablers for 5G mobile networks, as the communication infrastructure for future Internet of Things systems. However, most of these technologies do not support direct over-the-air communication between end users/devices. Opportunistic networks propose the use of delay-tolerant and wireless communication over such technologies, toward routing messages between end users. In our previous studies, we have shown how taking advantage of the existing Wi-Fi infrastructure leads to high hit rates and low latency based on the increased wireless range coupled with high bandwidth. In this article, we extend our proposed architecture for covering infrastructure-less environments through a novel mechanism that allows connecting peers through Wi-Fi Direct in a seamless and secure fashion. Furthermore, we emulate real-life user traces to empirically prove that our solution improves hit rate by 13 percent without impacting battery life.

INTRODUCTION

Opportunistic networks (ONs), originally coined by Pelusi *et al.* [1], are generally considered as a natural evolution of mobile ad hoc networks (MANETs) mostly comprising mobile wireless devices ranging from small wireless-capable sensors to smartphones or tablets, which aid in transmitting data horizontally by taking advantage of the already existent interactions between devices [1–3]. They are generally seen as enablers for fifth generation (5G) networks. For this, ONs employ short-range communication protocols (e.g., IEEE 802.11, Bluetooth) in order to disseminate data and decongest existing back-end protocols.

Communication in ONs is fully decentralized as the composing nodes are only aware of other nodes to which they are in close proximity (based on the range of the transmission media) and have no previous knowledge of the network topology. This has multiple implications. First, no assumptions can be made about the existence of paths between opportunistic nodes, as the network is very dynamic and disconnections are the norm. As such, ONs employ a paradigm entitled store-carry-and-forward (SCF) [1] in which nodes begin by storing local data either generated locally or received from peers; usually being handheld devices, nodes are characterized by a high degree of mobility and move around the network while carrying the stored data until the destination of data items is encountered, in which case a routing process is started for sending the data to their rightful recipients. This leads to the second implication: encountering a data message's destination is not always the case for data forwarding; ONs are based on the altruism of nodes that help in carrying other encountered nodes' data through the networks toward reaching the desired destination in exchange for them bearing the node's messages as well. This is the actual key factor of ONs, which attempts to maximize throughput while reducing latency and leads to our final implication: Understanding human mobility is of paramount importance in designing efficient opportunistic networking protocols [4]. As opposed to mobile edge computing (MEC), where mobile devices connect to a frontier cell tower and therefore can coordinate by communicating over such an infrastructure, all routes in ONs are dynamic, and nodes have no predetermined method of knowing they will interact in any fashion.

The main problem in opportunistic networking is that most of the technologies that can be used do not support direct over-the-air communication between end users/devices, and this is what we aim to address in this article. In our previous work [5, 6], we have designed and implemented a mobile application for tracking interactions between peers in wireless networks, which was used in multiple such tracing experiments and proved that Wi-Fi is more feasible than Bluetooth as opportunistic networking support. In this article, we extend our previous work by implementing an opportunistic communication engine based on the *Interest Spaces* framework [7] that not only can be deployed in Wi-Fi networks, but can also handle non-infrastructure networks. As such, we introduce a novel mechanism in which nodes can act altruistically by hosting Wi-Fi Direct (WFD) access points (APs), which allow peers connecting to them to interact in a seamless and secure fashion. Although this mechanism is largely useful when there is a lack of any pre-existing Wi-Fi infrastructures, it can also be deployed for bridging communities connected to neighboring Wi-Fi networks. Furthermore, we use the traces we previously collected [5] to prove that it improves the performance of ONs — increasing hit rate by up

Digital Object Identifier:
10.1109/MCOM.2017.1700354

Radu-Corneliu Marin and Radu-Ioan Ciobanu are with University Politehnia of Bucharest; Ciprian Dobre is with the National Institute for Research and Development in Informatics.

0163-6804/17/$25.00 © 2017 IEEE

to 13 percent and lowering latency by up to 21 percent, with a minimal impact on battery usage (≈ 7–8 percent). We believe that the techology has evolved sufficiently so that opportunistic networks can start being deployed on a global scale. WFD has become more and more popular (and is starting to be enabled by default, e.g., in Android), and 5G (which will have device-to-device support) is just around the corner.

The remainder of the article is organized as follows. We present a general overview of networking support and provide an in-depth comparison of wireless communication protocols that can be deployed in ONs. We present details on both our opportunistic communication engine and the novel mechanism for extending it over WFD. While we provide an analysis of our experimental results, we conclude the article with a summary of our observations, as well as our thoughts for future work.

Opportunistic Networking Support

The novelty of ONs stems from the use of any available wireless communication media for establishing connections between mobile peers and exchanging data among them [1]. However, there are currently multiple such solutions, and none has been deemed the most feasible to be used in opportunistic networking; multiple criteria must be taken into consideration when choosing a medium to be used in ONs: security, range, power consumption, speed, and infrastructure requirements. Table 1 contains a comparison of the following technologies.

•*Bluetooth* is a low-power and low-cost short-range communication technology for fixed and mobile devices, invented by the Swedish company Ericsson in 1994. It uses short-wavelength UHF radio waves in the industrial, scientific, and medical (ISM) band from 2.4 and 2.485 GHz, and has become ubiquitous in most mobile devices nowadays. In 2010, a new version of Bluetooth was added to the standard by the Bluetooth Special Interest Group, called Bluetooth Smart/Bluetooth Low Energy (BLE) with the purpose of providing considerably lower power consumption while keeping a similar communication range.

•*Near feld communication* (NFC) is a short-range wireless connectivity technology that enables smartphones and other devices to establish radio communications with each other by bringing them in close proximity. It is mostly used for making transactions, exchanging digital content, and connecting electronic devices, being compatible with many existing contactless cards and readers.

•*ZigBee* is a specification for a suite of protocols destined for Internet of Things (IoT) devices with small, low-power digital radios, standardized in 2003 by the ZigBee Alliance. It is based on an IEEE 802.15.4 standard, and can be used to transmit data over long distances by passing them through a mesh network of intermediate devices in an opportunistic fashion. Similar to NFC, ZigBee is generally employed by applications with low data rates that need to have high autonomy.

•*Wi-Fi* is a networking technology that uses the 2.4 GHz UHF and 5 GHz SHF ISM radio bands for offering wireless connection between devices and APs. It follows the IEEE 802.11 standards and is completely ubiquitous nowadays. However, for situations where two devices in close range wish to communicate but have no AP to connect to, there is also WFD, which is a related technology that allows devices to communicate through the wireless interface without needing a fixed AP.

Metric	Bluetooth	BLE	NFC	ZigBee	WFD	WiFi	LoRa
Infrastructure	No	No	No	No	No	Yes	Yes
Max range (m)	100	50	0.2	100	200	100	2,200
Speed (Mb/s)	2.1	1	0.4	0.25	250	600	0.05
Power	1	0.05	0.05	0.33	33	33	0.05
Security	WPA2	AES	N/A	AES	WPA2	WPA2	AES

Table 1. Comparison of wireless and mobile support technologies.

•*LoRa* is a proprietary technology for low-power wide area networking (LPWAN), which offers a long-range protocol for public and private networks with low power consumption. It uses the LoRaWAN protocol to perform communication between remote sensors and gateways connected to the network, and stands at the base of the Internet of Things. Its main goal is to be used for collecting data from small sensors and devices toward static gateways that are able to process and aggregate the data.

The first row in Table 1 shows which of them require an infrastructure and which do not. The ones that do not require the existence of a prior infrastructure can be used for opportunistic communication between devices in range as an alternative to Wi-Fi, which requires an infrastructure. Even though Wi-Fi networks have become nearly ubiquitous, alternative solutions might be required when there are no infrastructure-based solutions, or when they do not function correctly.

Table 1 shows, in the second row, the maximum range in meters for each of the six analyzed technologies. It can easily be seen that NFC has by far the worse range, since it cannot go farther than 20 cm. This happens because communication is done through electromagnetic induction, which cannot be performed if the two communicating devices are farther away. Other than NFC and LoRa, the other technologies offer similar values, with the mention that BLE has a lower range than regular Bluetooth, since it was designed with power saving in mind, thus reducing the maximum range. LoRa has a very high range, but it requires an infrastructure in the shape of gateways. Thus, after analyzing the maximum range for all protocols, it is clear that NFC cannot be used properly for ONs, even in very dense scenarios, because the range is extremely small. It should also be mentioned that ZigBee can have a much higher range since it uses mesh networking to transfer data (similar to opportunistic networking), but this is beyond the scope of this article.

The third row of Table 1 presents the transfer speed (in megabits per second) of each of the analyzed protocols. Wi-Fi clearly has the highest speed, and of the no-infrastructure protocols, WFD has the highest speed. Since NFC, ZigBee, and LoRa were created for short communication, they have very low speeds. The speeds obtained by Bluetooth and BLE are higher by one order

Based on a novel mechanism thaht allows nodes to altruistically become APs for peers in a seamless and secure fashion, our proposed solution can be deployed without any extraneous requirements for the network, and is able to achieve high hit rates and low message delivery latencies without impacting the battery life of mobile devices.

of magnitude. From this analysis, we can conclude that ZigBee would probably be unsuitable for an ON, since the data objects that it needs to exchange are not as small as the data objects exchanged by IoT devices. Nonetheless, Bluetooth speed is acceptable in certain conditions, while WFD is even more feasible.

However, power consumption should also be taken into consideration when choosing a protocol, since opportunistic nodes are battery-powered mobile devices. Thus, we show in the fourth row of Table 1 the power consumed by each of the analyzed technologies. We have taken Bluetooth as a gauge, which is why its power is shown as 1. The BLE protocol consumes about 95 percent less power than Bluetooth, as well as NFC. ZigBee is a higher consumer than NFC, even though the transfer speed is lower, but it has the advantage of a higher communication range, as shown above. Wi-Fi and WFD are the highest consumers, as Bluetooth uses less than 3 percent of the power required by Wi-Fi for the same tasks.

Finally, from a security standpoint (which, as shown in [8], is an important topic in mobile networks, offering new challenges caused by the limited battery life of devices), all protocols, except for NFC, provide more than needed protection: BLE, ZigBee and LoRa use 128-bit AES, while regular Bluetooth, Wi-Fi, and WFD employ WPA2 with 256-bit AES.

Although an infrastructure-based technology, Wi-Fi is a popular communication media in opportunistic networking as it has a sufficiently large range and one of the best throughputs, and with a near-ubiquitous AP presence, it is already considered a norm in modern society. Based on analyzing real-life datasets containing device interactions collected during a three-month-long experiment, we proved that Wi-Fi is more feasible than Bluetooth as opportunistic networking support [5]; due to its increased performance, Wi-Fi manages to establish up to three times more opportunistic contacts, while Bluetooth tends to isolate users into micro-communities. However, only using Wi-Fi is insufficient as it is unable to cover cases in which infrastructure is nonexistent.

Our solution merges infrastructure-based communication technologies (i.e., Wi-Fi) with non-infrastructure based protocols (i.e., WFD) toward extending coverage of the opportunistic engine initially proposed in [6]. Based on a novel mechanism that allows nodes to altruistically become APs for peers in a seamless and secure fashion, our proposed solution can be deployed without any extraneous requirements for the network, and is able to achieve high hit rates and low message delivery latencies without impacting the battery life of mobile devices.

Opportunistic Networks over Wi-Fi Direct

In our previous studies [6], we have designed and implemented a mobile application destined for collecting contextual data, namely the *HYCCUPS Tracer*; the application[1] collects the following information from mobile devices: CPU usage (load, frequency usage, etc.), battery statistics (charge, plug, temperature, etc.), memory availability, screen usage, and sensor data (accelerometer, proximity, etc.). Apart from these metrics, the application also embedded an opportunistic engine capable of tracing interactions over Bluetooth by periodically scanning for beacons from paired devices, and Wi-Fi by using the AllJoyn framework [9].

AllJoyn is an open source software framework that offers a peer-to-peer communication environment for heterogeneous distributed systems across different device classes with emphasis on mobility, security, and dynamism by implementing the D-Bus protocol. It provides an abstraction layer allowing it to run on multiple operating systems with the main goal of providing a software bus that offers distributed advertising and discovery of services in a secure mobile environment. In addition, the system supplies a Java-like location-transparent RMI. The object-oriented application programming interfaces (APIs) provided by AllJoyn represented the cornerstone of our implementation of the *Interest Spaces* framework [7] by providing the communication support for the *ONSIDE* [10] opportunistic dissemination algorithm.

The *HYCCUPS Tracer* application was employed in a 65-day-long tracing experiment [5] carried out in March–May 2012 at the Faculty of Automatic Control and Computers, University Politehnica of Bucharest, with a total of 66 volunteers varying in terms of year and specialization. We analyzed the collected HYCCUPS traces using the MobEmu emulator [11], an opportunistic network emulator that is able to replay a mobility trace and apply a desired algorithm when two nodes meet;[2] we concluded that Wi-Fi interactions using the AllJoyn framework were far more feasible than Bluetooth contacts as we observed that the long range coupled with higher bandwidth of Wi-Fi led to a staggering ≈ 21,000 interactions, while Bluetooth could only sum up to 34 percent of that value. Furthermore, after running a community detection algorithm, we discovered that Bluetooth has a tendency to isolate peers into micro-communities with infrequent inter-group message exchanges. However, as previously mentioned, infrastructure-based solutions are incomplete, as our opportunistic engine could not cover the cases where wireless APs were not present. As such, we turned our attention toward WFD, which, in modern mobile operating systems, enforces a requirement that the device's owner should personally accept pairing with another device, similar to Bluetooth.

In order to overcome this unacceptable need for human intervention, we have devised the following scheme over WFD:

1. Use the WFD legacy support (enforced by the WFD standard [12]) to enable the mobile device as a local Wi-Fi AP. The local WFD module generates a random session set identifier (SSID) and passphrase for it.
2. Start peer discovery:
   a) When peers are discovered, start advertising a *Bonjour* service over WFD with an instance name constructed as follows: *SSID:Passphrase:IP*. For security reasons the instance name is encrypted using a private symmetric key contained in the application.
   b) Start searching for other such services.

[1] http://hyccups.hpc.pub.ro, accessed on 2017-08-24.

[2] https://github.com/raduciobanu/mobemu, accessed on 2017-08-24.

c) If a service is discovered, decrypt its instance name, and connect to the node using the credentials from the instance name.

Given that AllJoyn is able to run over any stable network interface, it will be able to opportunistically connect to any other node connected to the WFD AP. Unfortunately, the node that is hosting the AP is not able to connect its own network interface; therefore, it will not be able to reach the nodes connecting to it. Furthermore, when a node is hosting a WFD AP, it does not disconnect from its main network interface and therefore shares it with the peers that access it. This proves to be extremely useful for data offloading and leads to the following cases:

1. Three or more nodes meet: One of them is assigned as an AP, and all other nodes interact among themselves as illustrated in Fig. 1, case (a). Unfortunately, two nodes in wireless range will not be able to interact with each other over AllJoyn without infrastructure support.
2. One node connected to a Wi-Fi AP establishes itself as a WFD AP: all nodes connecting to it will automatically connect to all peers from the Wi-Fi network as illustrated in Fig. 1, case (b).

However, the power consumption of this mechanism must be taken into account given that WFD is a notable battery consumer and, as such, cannot run continuously. Given that any decision to connect to a node is fully decentralized due to the nature of ONs, we need to synchronize the mobile devices to run the WFD connecting algorithm at specific times. Based on the fact that most modern mobile operating systems use Network Time Protocol (NTP) to keep the device's clock synchronized, we decided to run the algorithm 1 out of 10 min, every 10 minutes, starting from 12 a.m. This will guarantee that the connection schema will be able to find peers that are stationed nearby for at most 10 minutes. Furthermore, the algorithm contains an additional safe-guard against high power consumption, namely step 2a, which implies that the WFD AP will be created and advertised, if and only if peers are discovered at step 1.

Experimental Results

As previously mentioned, WFD requires a considerable amount of energy to function, which cannot be overlooked. In order to measure the impact it has over battery life, we have designed a battery drain experiment in which the device is left idle with no human interaction in order to observe the speed with which the battery is depleted. In this sense, case (a) in Fig. 2 illustrates a regular power drain without using WFD, while case (b) shows the impact of continuously using WFD. The results are straightforward: WFD drains approximately 80 percent from the battery life, which we consider to be unacceptable. However, in the approach proposed in the previous section, we show in case (c) that running WFD for 1 minute every other 10 minutes leads to 7 percent extra battery usage, which we consider to be more than acceptable. Furthermore, it proves that the WFD circuitry does not suffer from the cold-start effect, and its consumption is proportional to its usage, unlike GPS, for instance.

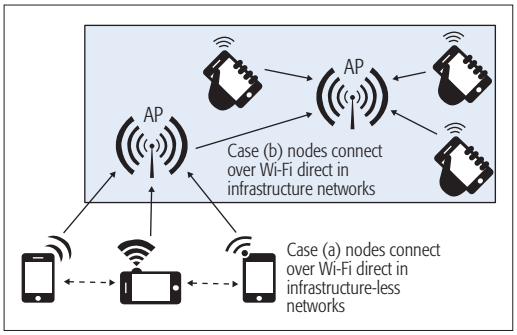


Figure 1. Wi-Fi Direct opportunistic networking case.

Next, we studied the influence of WFD over the ON performance by emulating the tracing datasets collected in [5] and replaying them through MobEmu. In order to emulate WFD interactions, we replaced the Bluetooth beacons used by the *HYCCUPS Tracer* and considered them to be contacts established using our previously proposed mechanism. The following metrics were taken into consideration:

- Hit rate: the percentage of messages that have successfully reached their intended destinations, computed as the ratio between the number of delivered messages and the total number of generated messages
- Delivery cost: the ratio between the total number of messages exchanged during the course of the test and the number of generated messages
- Latency: the time passed between the generation of a message and its eventual delivery to the destination
- Hop count: the number of nodes that carried a message until it reached the destination on the shortest path

While hit rate and latency are considered measures of network throughput, delivery cost and hop count are regarded as metrics of node and network congestion. Although the former two are consider to be of high importance, the latter two should be seriously taken into account and weighed against the former.

Furthermore, our analysis focuses on the two types of opportunistic communication:

1. Routing/forwarding: point-to-point communication between two peers. The data sent are not relevant for any of the other nodes on the path.
2. Dissemination: employs a modified form of publish/subscribe in which nodes can either act as publishers (which generate data marked with specific topics) or as subscribers (which subscribe to topics and expect to receive related data). Data dissemination in ONs is different from the classic publish/subscribe scheme due to the decentralized behavior of mobile networks.

To better understand the benefits of the proposed WFD connecting mechanism, we chose to run the simulations in three cases: Wi-Fi only — ignore any Bluetooth interactions and only focus on infrastructure-based contacts, WFD1 — adding the WFD mechanism on top of the Wi-Fi only case and replacing the Bluetooth contacts, and WFD2 — similar to WFD1, but choosing the

While hit rate and latency are considered measures of network throughput, delivery cost and hop count are regarded as metrics of node and network congestion. Although the former two are consider to be of high importance, the latter two should be seriously taken into account and weighed against the former.

ONSIDE is a dissemination strategy that leverages information about a node's social connections, interests and contact history, in order to improve hit rate and delivery latency. This is done by carefully selecting the nodes that act as forwarders, instead of simply flooding every node.

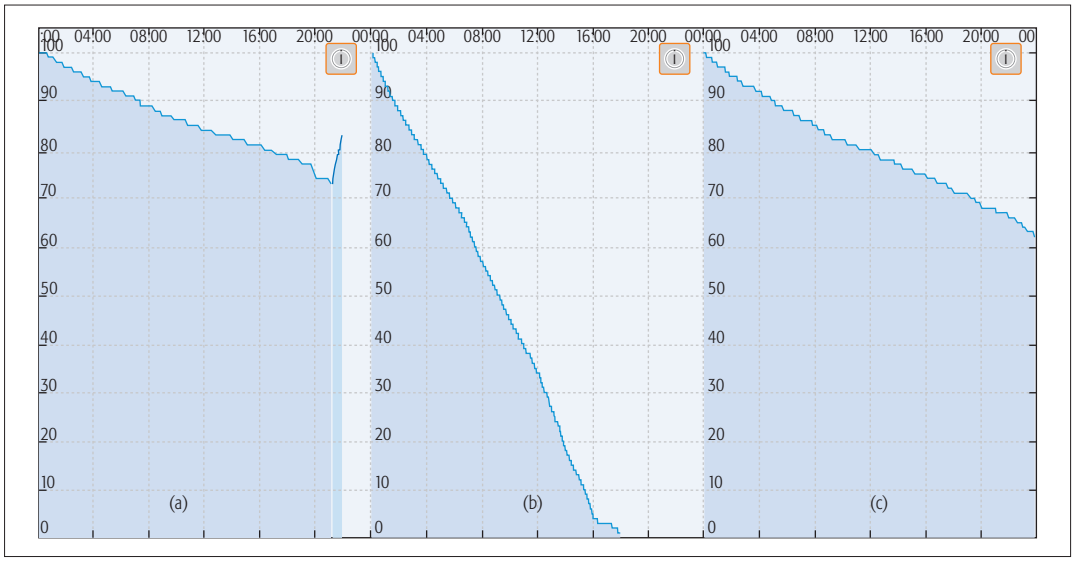


Figure 2. Battery drain experimental results (HYCCUPS Tracer screenshots). Axis X specifies the time of day, while axis Y is the device's battery percentage: a) regular power drain (no Wi-Fi Direct); b) continuously use Wi-Fi Direct; c) run Wi-Fi Direct for 1 minute every other 10 minutes.

Wi-Fi Only	500	5000	10,000	Unlimited
Hit rate	0.37	0.60	0.67	0.72
Delivery cost	84.14	86.80	103.07	55.92
Hop count	22.64	23.35	18.05	9.26
WFD1	**500**	**5000**	**10,000**	**Unlimited**
Hit rate	0.45	0.62	0.67	0.73
Delivery cost	144.54	196.85	264.70	86.67
Latency (s)	917.54	766.58	750.59	760.11
Hop count	45.70	46.30	42.83	9.69
WFD2	**500**	**5000**	**10,000**	**Unlimited**
Hit rate	0.47	0.65	0.71	0.77
Delivery cost	145.43	209.94	284.73	89.99
Latency (s)	840.69	681.36	653.99	670.00
Hop count	42.50	43.67	48.82	10.14

Table 2. Wi-Fi Direct influence over opportunistic routing.

nodes that become WFD APs in a round-robin fashion so that such hosts also get to receive the messages exchanged while they were altruistically acting as APs.

In order to analyze how the various tested algorithms behave in different conditions, we also vary a node's data memory size or the number of messages it can hold in memory. Thus, a node is able to store 500, 5000, 10,000 or an unlimited number of messages. We consider such values to map onto varied day-to-day situations and also to a diversity of mobile devices. As such, Table 2 shows a comparison over the three communication mechanisms when the Epidemic routing algorithm is employed. Epidemic [13] is one of the simplest forwarding strategies which simply floods the ON with all messages until the destination is reached.

As expected, the congestion of the network increased when using WFD, as more paths are generated between peers. However, not only does hit rate improve by ≈5–9 percent, but latency is decreased by up to 30 percent (summing up to 4 minutes).

Table 3 shows a comparison over the three communication mechanisms while running the ONSIDE [10] dissemination algorithm. ONSIDE is a dissemination strategy that leverages information about a node's social connections, interests, and contact history in order to improve hit rate and delivery latency. This is done by carefully selecting the nodes that act as forwarders, instead of simply flooding every node.

Similar to routing, dissemination is also affected by congestion due to more paths being generated in the network, but it gains a significant improvement in hit rate, up to 13 percent, with latency decreasing by ≈25 percent (summing up to almost 4 minutes as well).

Interestingly enough, there is a noticeable improvement in ON performance between WFD1 and WFD2, seeing as WFD altruistic nodes get a chance to receive their intended messages much faster in WFD2. Unfortunately, we have yet to find a way to balance AP hosts in real life as there is no centralized node that can coordinate and decide which node gets assigned as the next AP. However, clock skews in mobile devices might trigger a small degree of randomness when determining which node is first chosen by the rest of its peers to act as a WFD AP.

Conclusions

In this article, we have proposed a novel mechanism for connecting peers in opportunistic networks over Wi-Fi Direct in a secure and seamless fashion, without requiring any human intervention throughout the process of pairing. Furthermore, we have presented our opportunistic engine implementation over Wi-Fi and have shown how easily it is able to accommodate the addition of Wi-Fi

Direct without impacting the battery life of mobile devices. Furthermore, we have empirically proven through emulating real-life user traces that our mechanism improves the performance of ONs by increasing hit rate by up to 13 percent and reducing latency by 25 percent, while consuming only an additional 7–8 percent battery life.

Since we have only tested our proposal through simulations, for future work we would like to deploy an opportunistic engine enhanced with the Wi-Fi Direct connecting mechanism into a new tracing experiment of an even greater scale, in which we can properly observe human mobility and synergic patterns in both infrastructure-based and infrastructure-less opportunistic networks.

Acknowledgment

The research presented in this article is supported by project MobiWay, PN-II-PT-PCCA-2013-4-0321, and Traffic and Data Offloading in Mobile Networks — TTOff, H2020 as part of Measuring Mobile Broadband Networks in Europe.

References

[1] L. Pelusi, A. Passarella, and M. Conti, "Opportunistic Networking: Data Forwarding in Disconnected Mobile Ad Hoc Networks," *IEEE Commun. Mag.*, vol. 44, no. 11, Nov. 2006, pp. 134–41.
[2] J. M. Batalla *et al.*, "On Cohabitating Networking Technologies with Common Wireless Access for Home Automation System Purposes," *IEEE Wireless Commun.*, vol. 23, no. 5, Oct. 2016, pp. 76–83.
[3] Y. Nikoloudakis *et al.*, "A Fog-Based Emergency System for Smart Enhanced Living Environments," *IEEE Cloud Computing*, vol. 3, no. 6, Nov. 2016, pp. 54–62.
[4] M. Conti *et al.*, "From Opportunistic Networks to Opportunistic Computing," *IEEE Commun. Mag.*, vol. 48, no. 9, Sept. 2010, pp. 126–39.
[5] R.-C. Marin, C. Dobre, and F. Xhafa, "Exploring Predictability in Mobile Interaction," *Proc. 2012 3rd Int'l. Conf. Emerging Intelligent Data and Web Technologies*, 2012, pp. 133–39.
[6] R.-C. Marin, "Hybrid Contextual Cloud in Ubiquitous Platforms Comprising of Smartphones," *Int'l. J. Intelligent Systems Technologies and Applications*, vol. 12, no. 1, July 2013, pp. 4–17.
[7] R.-I. Ciobanu *et al.*, "Opportunistic Dissemination Using Context-Based Data Aggregation over Interest Spaces," *IEEE ICC 2015*, 2015, pp. 1219–25.
[8] A. Merlo, M. Migliardi, and L. Caviglione, "A Survey on Energy-Aware Security Mechanisms," *Pervasive and Mobile Computing*, vol. 24, no. C, Dec. 2015, pp. 77–90.
[9] AllJoyn; https://allseenalliance.org/framework, accessed Aug. 24, 2017.
[10] R.-I. Ciobanu *et al.*, "ONSIDE: Socially-Aware and Interest-Based Dissemination in Opportunistic Networks," *IEEE NOMS 2014*, May 2014, pp. 1–6.
[11] R. I. Ciobanu, C. Dobre, and V. Cristea, "Social Aspects to Support Opportunistic Networks in an Academic Environment," *Proc. 11th Int'l. Conf. Ad-hoc, Mobile, and Wireless Networks*, 2012, pp. 69–82.
[12] Wi-Fi Direct; https://www.wi-fi.org/discover-wi-fi/wi-fi-direct, accessed Aug. 24, 2017.
[13] X. Zhang *et al.*, "Performance Modeling of Epidemic Routing," *IEEE Computer Networks*, vol. 51, no. 10, 2007, pp. 2867–91.

Wi-Fi Only	500	5000	10,000
Hit rate	0.33	0.55	0.64
Delivery cost	5.67	7.62	6.67
Latency (s)	1012.55	880.37	866.15
Hop count	25.97	22.43	16.36
WFD1	**500**	**5000**	**10,000**
Hit rate	0.43	0.61	0.71
Delivery cost	10.38	16.99	19.33
Latency (s)	911.95	755.92	750.49
Hop count	45.39	44.23	38.69
WFD2	**500**	**5000**	**10,000**
Hit rate	0.46	0.63	0.72
Delivery cost	10.42	17.57	19.02
Latency (s)	834.24	670.62	666.01
Hop count	41.69	42.40	35.73

Table 3. Wi-Fi Direct influence over opportunistic dissemination.

Biographies

Radu-Corneliu Marin (radu.marin@cti.pub.ro) is a Ph.D. student within the Computer Science Department of University Politehnica of Bucharest, Romania. The topic of his Ph.D. is data and computation offloading through mobile cloud computing. His general interests regard distributed and collaborative systems, artificial intelligence and algorithms, embedded systems, Android application, and framework programming. His current research is focused on models and techniques for representing and reasoning on mobility for code execution in smartphone communities.

Radu-Ioan Ciobanu (radu.ciobanu@cs.pub.ro), Ph.D., is a lecturer and researcher at the Computer Science Department of the Faculty of Automatic Control and Computers at University Politehnica of Bucharest. His research interests include pervasive and mobile networks, DTNs, opportunistic networks, cloud computing, and so on. His research has led to the publishing of numerous papers and articles in important scientific journals and conferences. He is involved in several national and international research projects as a coordinator and technical manager.

Ciprian Dobre (ciprian.dobre@cs.pub.ro), Ph.D., has scientific contributions in the field of large-scale distributed systems concerning monitoring (MonALISA), data services (PRO, DataCloud@Work), high-speed networking (VINCI, FDT), large-scale application development (EGEE III, SEE-GRID-SCI), and evaluation using modeling and simulation (MONARC 2, VNSim). He was awarded a Ph.D. scholarship from Caltech and one from Oracle. His results received two CENIC Awards and three Best Paper Awards, and were published in numerous books, journal articles, and conference papers.

Since we have only tested our proposal through simulations, for future work we would like to deploy the opportunistic engine enhanced with the Wi-Fi Direct connecting mechanism into a new tracing experiment of an even greater scale, in which we can properly observe human mobility and synergic patterns in both infrastructure-based and infrastructure-less opportunistic networks.

Efficient Next Generation Emergency Communications over Multi-Access Edge Computing

Evangelos K. Markakis, Ilias Politis, Asimakis Lykourgiotis, Yacine Rebahi, George Mastorakis, Constandinos X. Mavromoustakis, and Evangelos Pallis

The authors study the challenges that next generation emergency services need to overcome in order to fulfill the requirements for rich-content, real-time, location-specific communications. The concept for next generation emergency communications as described in the project EMYNOS is presented, along with a vision of how this concept can fulfil the 5G requirements for ultra-reliable and ultra-low-latency emergency communications.

ABSTRACT

Traditionally, emergency communications between citizens and public authorities relied on legacy telecommunication technologies unable to cope with the agile, rich-media-content communications that mobile users are already using. This is due to the lack of harmonization and interoperable IP-based networking solutions. With the operators currently migrating to broadband IP infrastructures, emergency systems also need to follow this path and adapt their emergency communication platforms to fulfill next generation emergency services regulatory requirements. This becomes even more evident in light of the forthcoming 5G networks, which are envisioned to support an amalgam of diverse applications and services with heterogeneous performance requirements, including mission-critical IoT communication, massive machine-type communication, and gigabit mobile connectivity. Emergency service operators face an enormous challenge in order to synchronize their model of operation with the 5G paradigm. This article studies the challenges that next generation emergency services need to overcome in order to fulfill the requirements for rich-content, real-time, location-specific communications. The concept for next generation emergency communications as described in the project EMYNOS is presented, along with a vision of how this concept can fulfill the 5G requirements for ultra-reliable and ultra-low-latency emergency communications.

INTRODUCTION

Telecommunication networks are currently the dominant infrastructure for providing emergency services. The demand for emergency calls has been estimated at 320 million calls in the European Union annually [1]. and of these calls, 60 percent are originated by mobile devices. Evidently, mobile phones and particularly smartphone devices, having already proved to be the most convenient way for users to communicate and share multimodal information, are becoming a dominant factor in the emergency services sector as well. The boom in the mobile applications market, which provides a plethora of different text messaging, video, and picture sharing applications and social media networking, mainly fuels this. An emergency call placed by a smartphone with such capabilities can potentially increase the chances of a successful emergency response by providing, along with the multimedia call, precise location information. Almost fully deployed fourth generation (4G) broadband mobile networks are foreseen as the basis for supporting next generation emergency services. Although emergency systems today are based on outdated telecommunication technologies that cannot cope with IP-based services, eventually they will have to be upgraded in order to fulfil the next generation networks (NGN) regulatory requirements

This rapidly changing environment leaves the current emergency service operators lagging in the effort to identify a consistent roadmap for upgrading the current legacy systems (i.e., public safety answering points — PSAPs), while maintaining scalability and interoperability. Toward this end, several regulatory and standardization activities are already underway. Specifically, the National Emergency Number Association (NENA) and the European Emergency Number Association (EENA) are the key organizations promoting a universal emergency service number in the United States and in Europe, respectively. Moreover, the Internet Engineering Task Force (IETF) has chartered the Emergency Context Resolution with Internet Technologies (ECRIT) working group [2] to describe a framework based on existing protocols for emergency calling using Internet multimedia. Finally, the Third Generation Partnership Project (3GPP) enhanced the existing IP Multimedia Subsystem (IMS) with specialized tasks for emergency calls as well as location retrieval capabilities [3].

On the other hand, the technological progress in the mobile broadband networks is another key factor that emergency service operators need to consider. The future Internet is expected to be characterized by a higher number of services that in turn will satisfy the expectation of users and enterprises. The design of the fifth generation (5G) of mobile networks has at its heart the provision of an amalgam of services and applications across different industries including healthcare, transport, and industrial processes, which demand stringent network performance requirements. A

Digital Object Identifier:
10.1109/MCOM.2017.1700345

Evangelos K. Markakis, George Mastorakis, and Evangelos Pallis are with the Technological Educational Institute of Crete; Ilias Politis and Asimakis Lykourgiotis are with the Hellenic Open University; Yacine Rebahi is with Fraunhofer Fokus; Constandinos X. Mavromoustakis is with the University of Nicosia.

0163-6804/17/$25.00 © 2017 IEEE

roadmap has already been in place for the provisioning of these heterogeneous services by 5G designs, with ultra-low-latency and reliable service delivery at its core [4]. In this environment, next generation emergency services are expected to utilize all the emerging technical innovations to enhance both the user's experience when calling in an emergency and the first responder's situation awareness while in the area of the event. One of the potential enablers for these advances is the deployment of a massive number of sensors and haptic devices, expected to reach a number of 500 billion by the end of the decade [5] forming a global Internet of Things (IoT).

In this article, a number of new challenges that emergency service operators are going to face with the advent of 5G are described. Emphasis is given on the utilization of the recently described use cases such as remote healthcare monitoring and management and how current and future technological solutions — such as software defined networking (SDN), network functions virtualization (NFV), and mobile edge computing (MEC) — could be utilized for supporting enhanced emergency communications. Specifically, the article concentrates on proposed solutions developed in EMYNOS, an EU funded project that proposes a platform for managing next generation emergency services by supporting rich-media emergency calls (i.e., voice, text, and video), hence constituting a powerful tool for coordinating communication among citizens, call centers, and first responders. Extending the project' s outcomes, this study envisions EMYNOS as a 5G enabled emergency service platform and defines the main networking requirements that would enable it.

In the rest of the article, we discuss how emergency communications are perceived in the concept of 5G and how low-latency remote healthcare could be realized. We then dive deeper into the technological solutions of cloud networking and MEC that need to be adopted by the future emergency services operators. Next, the EMYNOS project and its related use cases are presented, and then we showcase the vision of EMYNOS as a 5G-based emergency platform for remote health care. Finally, we conclude the article.

5G Emergency Communications

In light of the forthcoming 5G networks, the legacy systems and protocols, and the services provided by the emergency communications organizations and the public safety sector in general, are going to be affected particularly by the increased performance of the next generation wireless and mobile networks, the enhanced security, and the improved device-to-device communications. Nevertheless, the envisioned 5G features will need to be tightly coupled with the stringent operational and management requirements of emergency services, which need to be maintained. Depending on the actual design of the 5G network, emergency communication is expected to support real-time, high-priority total conversation services (voice, video, real-time text) and improved immunity to data pollution and security threats that potentially could disrupt the reaction time of the first responders. Additionally, device-to-device communications will increase the availability of the communication channels and the uplink capacity, creating an "always connected" experience for the emergency workers and rendering them capable of utilizing high-quality multimedia content for improved awareness. Toward this end, the network slicing feature of the 5G network will enable network operators to dynamically adapt the transmission speed and latency of the network, ensuring high priority for first responders' communications [5].

Evidently, future emergency communications will be relying heavily on sensors either dispersed in buildings, road lights, or vehicles forming dense communication environments (e.g., smart cities) or carried as wearables by first responders or persons of interest (i.e., people in need of remote health monitoring), hence creating a massive IoT connected to the network on a massive scale. A major challenge for the next generation mobile networks is the management of such an unprecedented number of concurrent connections required by all these sensor devices and machines. Moreover, in order to satisfy the strict requirements imposed by emergency communications and public protection and disaster relief (PPDR) services in general, the network operators are expected to ensure ultra-low latency, ultra-high availability, and reliability for these services. Due to the low latency that will be provided by the underlying access network, next generation emergency services based on massive IoT [6] and device-to-device communications will be characterized by higher throughput, higher quality of service (QoS) and quality of experience (QoE), and low buffer requirements for the user devices.

Low-Latency Remote Healthcare

The recent advances in mobile communications give rise to paradigms and applications for the health sector, such as remote healthcare for providing access to medical services to rural areas, or healthcare emergency support in cases of critical life-threatening situations. To this end, IoT and wireless sensor networks based on mobile communication can provide remote monitoring for parameters including heart rate and blood pressure. In particular, current next generation mobile networks (i.e., 4G/LTE, 5G) are considered as the key enablers for the future applications and services of the healthcare industry [8]. Cloud computing and network virtualization offer the ability to healthcare providers to offload and process high volumes of medical data (e.g., high-definition medical images and real-time video), ensuring ultra-low latency during the date delivery to the appropriate node (e.g., MEC) and distributed data storage for optimum use of required processing resources. Moreover, ultra-dense device-to-device communication networks and cooperative multi-node/cell networks, as envisioned currently in the 5G architecture, can support reliable and low-latency uplink connectivity for very large numbers of connected devices, including wearable sensors for health monitoring and haptic devices [9].

5G use cases currently consider two main applications:

- Remote healthcare and precision medicine
- Remote robotic surgery and intervention

In this context, remote healthcare envisions the care provision to the home environment or on

Ultra-dense device-to-device communication networks and co-operative multi-node/cell networks, as envisioned currently in the 5G architecture, can support reliable and low-latency uplink connectivity for very large numbers of connected devices, including wearable sensor for health monitoring and haptic devices.

5G networking is no longer just about connectivity; 5G networks are envisaged to also accommodate, in-network services, virtualized and dynamically deployed at appropriate locations within the network, in line with the emerging paradigms of edge computing. Future network services will comprise both connectivity and network appliances in the form of "slicing" and edge computing.

the move (e.g., ambulance, emergency response units), leading eventually to the decentralization of hospitals. Additionally, sensitive medical data and records analysis will act as the basis for predictive diagnosis and proactive healthcare. In the context of remote surgery, geographical boundaries that prevent high-quality healthcare in complex medical interventions and surgeries can be overcome by highly available, reliable, low-latency access networks. Currently, the healthcare market is rapidly evolving by utilizing novel cloud computing services that allow remote care through broadband mobile networks, secure and timely medical data acquisition and analysis by medical service providers, use of assistive living technologies and e-health sensor wearables that enable chronic patient monitoring and management, as well as personalized treatment and coordinated healthcare.

Multi-Access Edge Computing Paradigm

5G networking is no longer just about connectivity; 5G networks are envisaged to also accommodate in-network services, virtualized and dynamically deployed at appropriate locations within the network, in line with the emerging paradigms of edge computing (i.e., fog, MEC, cloudlets). Future network services will comprise both connectivity and network appliances in the form of "slicing" and edge computing.

In this respect, edge computing technology has emerged in the last years with two main representative paradigms. The first is fog computing, which enables the provision of compute, storage, and networking services at the network edges. In this way, end users, IoT applications, vehicles, and so on can take advantage of conventional cloud services, but with the advantages of low communication latency and location awareness. Toward providing its services, a fog network relies on a number of heterogeneous devices, managed as clusters [11], named fog nodes, that can vary from low-end devices, such as set-top boxes, access points, switches, routers, and base stations, to high-end ones such as cloudlets and Cisco's IOx.

The second representative is MEC, led by the European Telecommunications Standards Institute (ETSI), targeting in phase 1 mobile/cellular environments. When fog computing started to be the main representative for IoT and end user computing, ETSI created the ISG MEC initiatives, moving MEC from mobile to multi-access edge computing [12]. ETSI MEC does not describe any business model for the ownership of the infrastructure at the edge; it usually assumes that the edge operator owns data centers, as this is the only entity that controls the access to the data plane and is able to inspect and offload certain traffic to the edge.

In this context, network "tenants" will not only need to request sliced capacity; they will also need to request information technology (IT) resources in order to be able to establish on demand a virtual edge infrastructure as part of their "virtual 5G network." This is directly analogous to the software-defined data center (SDDC) concept, in which entire data centers are virtualized, abstracted, and provided as a service through the enhanced slicing and isolation capabilities offered by certain emerging technologies [13].

5G envisions the slicing technology as a means to allow operators to split a single physical network to multiple logically isolated virtual networks, where each includes device, access, transport, and core networking functionalities for services with different types, characteristics, and requirements. The need for this type of slicing is because mobile networks have been optimized mainly for phones only. The use of these "slices" in case of emergency events can benefit from zero-time development of infrastructure that can be used for first responder deployment.

Next Generation Emergency Services Architecture

As already mentioned, EENA is leading the path toward the next generation emergency service architecture in Europe (known as NG112) [10], in collaboration with the IETF ECRIT and Geopriv working groups. According to the definition of this architecture, emergency calls are delivered through the Emergency Services IP network (ESInet), designed as an IP-based network. The border control function (BCF), acting as a firewall and a session border controller, is placed between the external networks and the ESInet and between the ESInet and the emergency service operators, performing network edge control and Session Initiation Protocol (SIP) message handling. The NG112 defines that the routing of the emergency calls to the appropriate PSAP will be based on the location information of the caller. Toward this end, the emergency call routing function (ECRF), which is a Location-to-Service Translation Protocol (LoST)-based entity, utilizes the location information and service uniform resource name (URN) received in a routing query and maps it to the destination uniform resource identifier (URI) for the call. The other key networking element is the emergency services routing proxy (ESRP), which is the basic routing function for emergency calls. According to the definition of the architecture, multiple "intermediate ESRPs" could coexist, forming different hierarchical levels of the ESInet. ESRP, which is SIP-enabled, is also responsible for:

- Evaluating a policy "rule set" for the queue on which the call arrives
- Querying the ECRF with the location included within the call to determine the "normal" next hop (smaller political or network subdivision, PSAP, or call taker group) URI
- Evaluating a policy rule set for that URI using other inputs available to it including headers in the SIP message, time of day, PSAP state, and so on

EMYNOS Case

Recently, the EU funded research and innovation project EMYNOS studied the integration of IP to emergency systems in Europe [10], in compliance with NG112. The project focuses on the transition of the legacy emergency operation systems to an operational model based on the IP protocol stack, GIS databases, and modern network functional elements. The proposed novel model of emergency communications is able to offer real-time voice, video, and real-time text services to PSAP operators, significantly enhancing their situation awareness. From the user's perspective, the proj-

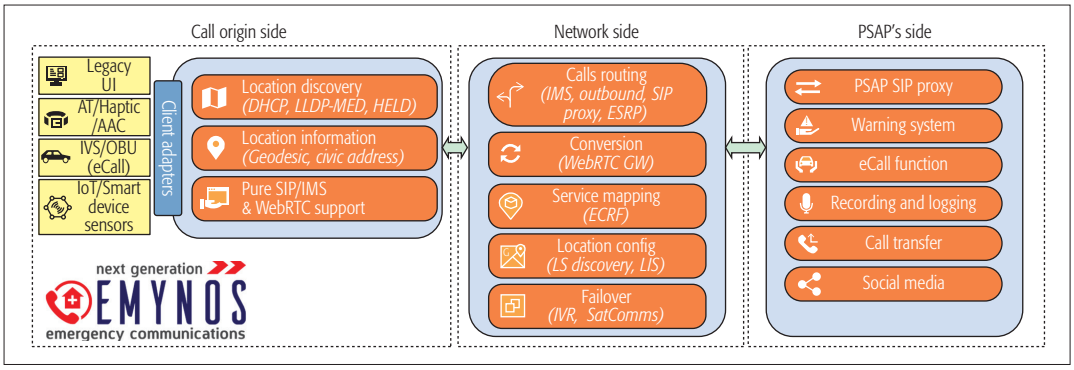


Figure 1. EMYNOS high-level architecture.

ect enables citizens to make IP-based emergency calls (i.e., to police, ambulance, and fire brigade) and supports remote e-health monitoring based on wearable sensor devices.

In this particular use case, the home environment of the caller is supported by a network of sensor devices that monitor different environmental parameters (temperature, humidity, air pressure, etc.) and a home gateway that aggregates these data or directly offloads them to the cloud. With the number of IoT devices increasing constantly and the volume of the data ever expanding, the resources required for processing such data without violating QoS and resource allocation requirements stretch to their limits the current network and cloud infrastructures. Fog computing has emerged as a novel solution for providing data computing, data storage, and local management of the sensors as well as mobility. Within the concept of fog computing, the IoT gateway needs to become the orchestrator of the device communication by adopting a service oriented approach. This way, it can cope with the heterogeneous nature of the different devices and their diverse behavior. A decision making engine processes these data and creates alerts as soon as certain threshold values are exceeded. Such alerts are sent over embedded messages in SIP signaling. Within the EMYNOS context, the Sensor Model Language (SensorML) was the standard of choice to convey the sensor data into the SIP signaling. SensorML is a standard of the Open Geospatial Consortium (OGC) that aims to provide interoperability, first at the syntactic level and later at the semantic level. An emergency call is initiated automatically by the alarm system, which triggers the PSAP.

Envisioning EMYNOS over 5G

IoT devices are resource constrained and cannot facilitate on-site data processing; thus, in the previously discussed scenario, the IoT gateway bore the responsibility to perform these tasks effectively. However, this approach lacks the elasticity and scalability inherent to cloud computing. IoT applications can run in cloud environments in order to acquire the desired computational resources, which are distributed throughout the network. As a result, IoT applications can utilize resource-demanding algorithms such as machine learning to infer knowledge and determine critical events regarding civilian disasters and medical complications. However, cloud resources, due to their distributed nature, are not competent enough for real-time scenarios where requirements for low latency and high bandwidth impose stringent constraints. Additionally, IoT big data will result in exponential traffic growth, and the current Internet does not have the ability to meet its demands.

On the other hand, MEC has recently been proposed to fill this niche not well served by existing cloud architectures. MEC is regarded as a key enabling technology, which provides the capability to have a high performance virtual environment residing at the network's edge. By being adjacent to the IoT environment, MEC can support applications and services with increased bandwidth, low latency, and improved QoS. At the same time, MEC architecture can drastically offload the core as IoT data will be processed and stored at the edge network, ready to be retrieved only when necessary. Finally, the IoT dynamic infrastructure with self-configuring capabilities encompasses various protocols such as Message Queue Telemetry Transport (MQTT), Constrained Application Protocol (COAP), and Extensible Messaging and Presence Protocol (XMPP). In general, these protocols define messages that are small in order to meet the resource constraint sensor environment based on the publish/subscribe interaction scheme. As IoT is still in its infancy without protocol interoperability, there is a need for a low-latency aggregation point to manage the various protocols and perform real-time processing and large-scale data analytics. MEC can fulfill this additional role, providing the ability to resolve all the aforementioned challenges.

Following the description of the EMYNOS concept, the migration of its proposed emergency service platform to the 5G era will have to incorporate the requirements for ultra-reliable, ultra-low-latency service delivery. Therefore, the vision of a 5G-based EMYNOS is based on a resilient, highly available, and dynamically managed system that brings the health monitoring and management service closer to the user, as depicted in Fig. 2. The proposed architecture relies on MEC. In this particular use case the application server operated by the healthcare provider is moved closer to the radio access network (RAN), thus reducing the end-to-end latency. Considering that each monitored person could wear a number of different sensors for monitoring heart rate, oxygen saturation, blood pressure, and so on, the number of such devices increases to very high levels in the case of rural areas. Each health sensor device could communicate directly with the RAN, minimizing the communication path to the healthcare

As IoT is still in its infancy without protocol interoperability, there is a need for a low-latency aggregation point to manage the various protocols and perform real-time processing and large-scale data analytics. MEC can fulfill this additional role, providing the ability to resolve all the aforementioned challenges.

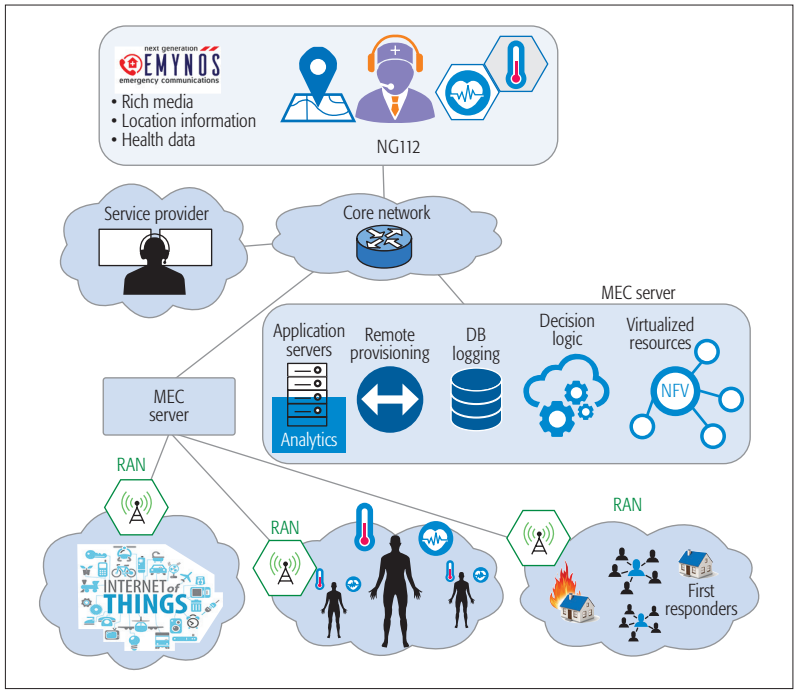


Figure 2. High-level vision of EMYNOS migration to 5G.

provider, reducing the latency, and ensuring continuity and seamless mobility when the monitored person is moving.

The described use case, the architecture of which is illustrated in Fig. 3, considers that the healthcare provider controls the MEC; therefore, all the monitored data are stored and processed to the cloud edge. Potential emergencies, which would require remote intervention (e.g., emergency callback to the user, remote calibration of pharmaceutical dose), are detected at the MEC level. As the current trend in decreasing number of caregivers and doctors moves non-urgent treatments to patients' homes, this scenario gains momentum. Specifically, for the case where the emergency requires the intervention of first responders to the area, the healthcare provider initiates the emergency call to the emergency service operator. Such a call will be augmented with the current health sensor data, along with the patient's location information, medical history, and insurance data. In the case of a fire detected by the building management system, the security service provider who owns the MEC platform can initially verify that it is not a false positive alert from video footage available from the site. Alternatively, an emergency call to the fire brigade is initiated, including a building floor plan for the first responders. Finally, the MEC's remote provisioning system can be utilized in order to decrease the sensors' reporting interval in order to provide near-real-time data delivery.

Opposite to the traditional case for emergency service networks based on GSM and LTE, where each one of the physical networks is reserved for one use case (e.g., GSM for voice, LTE for mobile data), the proposed network architecture for 5G emergency services is capable of creating and managing virtual instances of access networks, hence providing customized network resources to each emergency service agency (i.e., police, ambulance, fire brigade) to the area of the event. The network slicing will keep communication interference between the different networks, ensuring extremely high throughput and ultra-low latency. This flexible orchestration of network slices is realized by the use of software defined functions and programmable infrastructures. The RAN's backhaul is governed by the NFV infrastructure, the control of which relies on the MEC. In this case, the bandwidth allocated to each wearable for health monitoring, or the first responders' communication devices, the management of the traffic including delay, loss, active

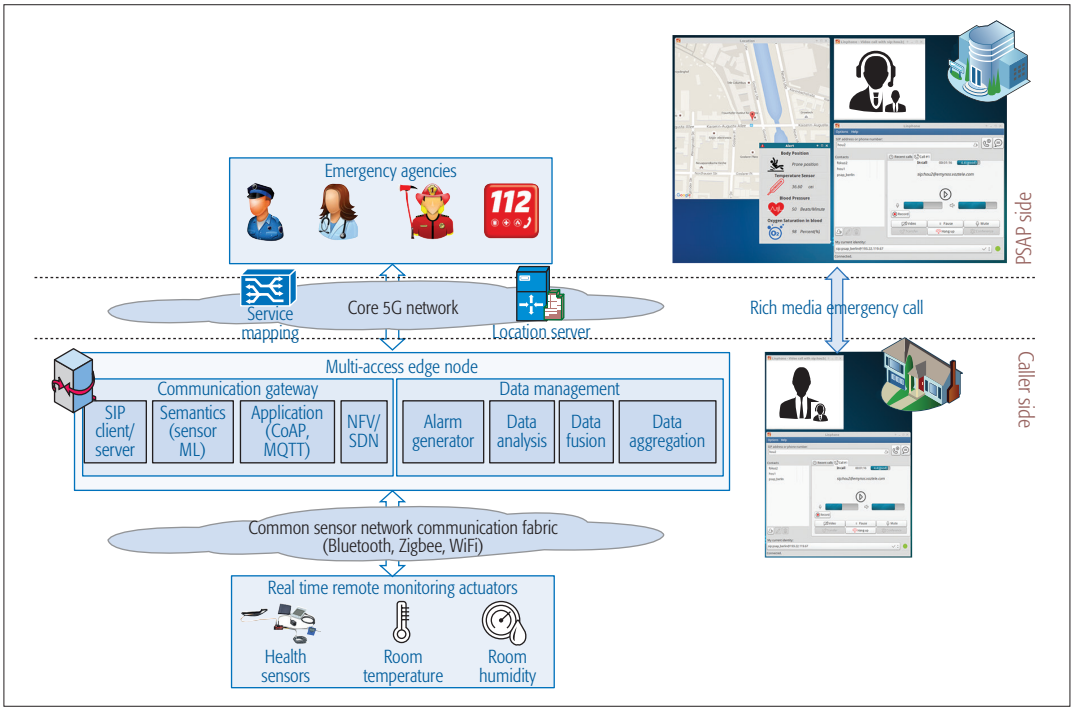


Figure 3. MEC-enabled emergency call component diagram.

bearers, and so on are synchronized by the NFV controller located in the MEC.

Conclusions

As most operators have already migrated to broadband IP infrastructures, emergency systems also need to follow this path and adapt their emergency communication platforms to fulfill regulatory requirements in terms of next generation emergency services. Emergency service operators face an enormous challenge in order to synchronize their model of operation with the 5G paradigm.

Thus, the transition to 5G for emergency service providers is a chance to migrate their systems to support an amalgam of diverse applications and services with heterogeneous performance requirements, including mission-critical IoT communication, massive machine-type communication, and gigabit mobile connectivity. This article presents the challenges that next generation emergency services need to overcome in order to fulfill the requirements for rich-content, real-time, location-specific communications of 5G networks, along with a vision of how this conception can fulfill the 5G requirements for ultra-reliable and ultra-low-latency emergency communications

Acknowledgment

This work has received funding from the European Union's Horizon 2020 research and innovation programme EMYNOS under grant agreement no. 653762.

References

[1] The European Emergency Number 112; http://ec.europa.eu/commfrontoffice/publicopinion/flash/fl_285_en.pdf, accessed Aug. 2017.
[2] H. Schulzrinne and R. Marshall, "Requirements for Emergency Context Resolution with Internet Technologies," IETF RFC 5012, 2008.
[3] Y. Rebahi *et al.*, "An IP Based Platform for Emergency Calls and Reporting." *Int'l. J. Critical Infrastructure Protection*, vol. 4, no. 3, 2011, pp. 137–53.
[4] K. Kusume, M. Fallgren, and O. Queseth, "Updated Scenarios, Requirements and KPIs for 5G Mobile and Wireless System with Recommendations for Future Investigations," METIS Deliverable ICT-317669-METIS D 1, 2015.
[5] V. Cristea, C. Dobre, and F. Pop, "Context-Aware Environments for the Internet of Things," *Internet of Things and Inter-Cooperative Computational Technologies for Collective Intelligence*, 460, 2013, pp. 25–49.
[6] A. Gorcin and H, Arslan, "Public Safety and Emergency Case Communications: Opportunities from the Aspect of Cognitive Radio," *IEEE DySPAN 2008*, 2008, pp. 1–10.
[7] M. Bornheim and M. Fletcher, 2016, "Public Safety Digital Transformation, The Internet of Things (IoT) and Emergency Services," EENA Technical Committee Doc., Brussels; http://www.eena.org/download.asp?item_id=170, accessed Aug. 2017.
[8] A. Osseiran *et al.,* "Scenarios for 5G Mobile and Wireless Communications: The Vision of the METIS Project," *IEEE Commun. Mag.*, vol. 52, no. 5, May 2014, pp. 26–35.
[9] J. Kvedar, M. J. Coye, and W. Everett, "Connected Health: A Review of Technologies and Strategies to Improve Patient care with Telemedicine and Telehealth," *Health Affairs*, vol. 33, no. 2, 2014, pp. 194–99.
[10] NG112, E. E. N. A., "Next Generation 112 Long Term Definition Standard for Emergency Services Document (Version 1.1)," 2013.
[11] E. Markakis *et al.*, "EMYNOS: Next Generation Emergency Communication," *IEEE Commun. Mag.*, vol. 55, no. 1, Jan. 2017, pp. 139–45.
[12] F. Bonomi *et al.*, "Fog Computing: A Platform for Internet of Things and Analytics," *Big Data and Internet of Things: A Roadmap for Smart Environments*, 2014, Springer, pp. 169–86.
[13] T. Taleb *et al.*, "On Multi-Access Edge Computing: A Survey of the Emerging 5G Network Edge Architecture & Orchestration," *IEEE Commun. Surveys & Tutorials*, 2017.
[14] C. Dobre and F. Xhafa, Eds., *Pervasive Computing: Next Generation Platforms for Intelligent Data Collection*, Morgan Kaufmann, 2016.
[15] D. Kakadia, *Apache Mesos Essentials*, Packt Publishing Ltd., 2015.

Biographies

Evangelos Markakis [M] (markakis@pasiphae.eu) holds a Ph.D. from the University of the Aegean. Currently he acts as a senior research associate for TEI of Crete and is the Technical Manager for HORIZON 2020 DRS-19-2014 EMYNOS. His research interest includes fog networking, P2P applications, and NGNs. He has more than 30 refereed publications in the above areas. He is a member of IEEE ComSoc and acts as Workshop Co-Chair for the IEEE SDN-NFV Conference.

Ilias Politis received his Ph.D. in multimedia networking from the University of Patras, Greece, in 2009. He is a postdoctoral research fellow at the School of Science and Technology at the Hellenic Open University and at the Wireless Telecommunications Laboratory of the Department of Electrical and Computer Engineering at the University of Patras. His research interests include multimedia networking, monitoring and management of multimedia QoE, and 3D video streaming.

Asimakis Lykourgiotis (asly@ece.upatras.gr) received his engineering diploma from the Electrical and Computer Engineering Department of the University of Patras in 2008 and a Ph.D. from the same institution. His main research interests include wireless local area networks, cellular networks, mobility management, and mobile multimedia. He has been involved in EU projects.

Yachine Rebahi [M] has a Ph.D. in mathematics and a Habilitation in computer science. He is currently a senior researcher at Fraunhofer FOKUS with 15 years' experience in the context of NGN. He has over 60 publications. His research activities span, in particular, next generation emergency services and next generation networks security, namely, DoS attack detection, SPAM mitigation, and fraud and service misuse detection. He is the EMYNOS project coordinator.

George Mastorakis received his B.Eng. in electronic engineering from the University of Manchester Institute of Science and Technology in 2000, his M.Sc. degree in telecommunications from University College London in 2001, and his Ph.D. degree in telecommunications from the University of the Aegean in 2008. He is currently serving as an associate professor at the Technological Educational Institute of Crete, Greece. He has published more than 150 research articles.

Constandinos X. Mavromoustakis [M] is currently a professor in the Department of Computer Science at the University of Nicosia, Cyprus, where he is leading the Mobile Systems Lab. He has been an active member (Vice-Chair) of the IEEE/ R8 regional Cyprus section since January 2016, and since May 2009 he has served as the Chair of the C16 Computer Society Chapter of the Cyprus IEEE section.

Evangelos Pallis [M] holds M.Sc. and Ph.D. degrees in telecommunications from the University of East London, United Kingdom. He currently serves as an associate professor at TEI of Crete in the Department of Informatics Engineering and director of PASIPHAE Lab. His research interests are in the fields of wireless and mobile networking. He has more than 200 refereed publications. He is member of IEE/IET and IEEE ComSoc, and a Distinguished Member of the Union of Regional Televisions in Greece.

The transition to 5G for emergency service providers is a chance to migrate their systems to support an amalgam of diverse applications and services with heterogeneous performance requirements, including mission critical IoT communication, massive machine-type communication and Gigabit mobile connectivity.

Pseudo-Dynamic Testing of Realistic Edge-Fog Cloud Ecosystems

Massimo Ficco, Christian Esposito, Yang Xiang, and Francesco Palmieri

Testing a software artifact to be deployed in the nodes composing an edge-fog ecosystem could be extremely challenging. Pure simulated environments and real testbeds could be not representative enough of realistic scenarios or unacceptably expensive. The authors explore such issues and present a pseudo-dynamic testing approach.

ABSTRACT

Currently, our society is undergoing a radical change due to the increasing pervasive use of ICT within all of its processes. Such an evolution has been triggered by the advent of the Internet of Things vision, where smart sensing devices can be integrated in the daily objects surrounding any human being. The increasing demand of dealing with the big data generated, managed, and stored by the applications built on top of such novel sensory networks has called for innovative architectures, such as edge and fog computing, which have not yet been fully standardized. Considering the large scale and complexity of these architectures, testing a software artifact to be deployed in the nodes composing an edge-fog ecosystem could be extremely challenging. Pure simulated environments and real testbeds could be not representative enough of realistic scenarios or unacceptably expensive. We explore such issues and present a pseudo-dynamic testing approach, where a portion of the experimental scenario is simulated, while the edge and fog nodes under test are emulated or executed in a real environment.

THE ERA OF EDGE-FOG COMPUTING

The success of the Internet of Things (IoT) is essentially due to the diffusion of a huge amount of very cheap embedded devices equipped with smart sensors and flexible actuators, which can easily be used to monitor and control a significant variety of elements in almost any daily life scenario. This paradigm promises to make anything part of the Internet in order to share sensory data and execute advanced control logic, opening up new horizons to innovative services and applications, which will see massive interaction among things and humans. IoT applications are currently expected to grow in scale, due to the phenomenon of federating different IoT solutions in order to realize more complex ones. This demands increasing memory and processing resources that easily exceed the possibility of current single-server architectures, as traditionally adopted by wireless sensor networks (WSNs). On the contrary, it calls for more elastic resource provisioning, by using a collection of servers and, in a later evolution, a cloud infrastructure providing enough capacity for quickly processing all the data generated by IoT objects, and making sense of them through proper analytics applications and practices. Cloud computing, thanks to its elasticity, flexibility, and scalability, offers the right infrastructure-level facilities to support the IoT runtime and storage requirements, and easily adapt to any change in the associated demand and operating scenario. However, the network traffic generated from hundreds of thousands of IoT devices toward the cloud may also become a major challenge in the presence of significant bandwidth resources, resulting in transmission and hence processing-response delays that may become crucial for many applications, such as power control, assisted driving, and health monitoring.

On the other hand, in order to extend the cloud computing paradigm to the edge of the network, where delays are critical and bandwidth is limited, the recent trend is to allow processing tasks, analytics, and knowledge generation to be handled closer to the data sources (i.e., sensors and actuators; edge computing), or between the edge and the cloud (fog computing), managing massive temporary storage and heavier analytics tasks. Both evolutions push the frontier of modern data processing applications away from centralized nodes, making the cloud as close as possible to the entities that produce and act on IoT data, in order to provide additional distributed intelligence and to deliver a faster response in analytics [1]. Edge and fog-level facilities will also participate in the management of network switching, routing, load balancing and security tasks, becoming the ingress points for the data coming from multiple heterogeneous sources and deciding if it has to be analyzed locally or conveyed through a specific path to the cloud for further processing. Such a complex ecosystem is referred to as edge-fog cloud computing. Its scalability, flexibility, and performance characteristics represent a driving force for a new breed of critical IoT services and applications that involve effective and efficient data management and analytics, such as latency-sensitive applications for smart grids, traffic management of vehicles, connected vehicles, smart cities, and services for enhancing the quality of life.

EDGE-FOG CLOUD AS A LARGE-SCALE COMPLEX SYSTEM

Applications and services tagged by the edge-fog cloud paradigm are characterized at the lower level by thousands or millions of networked objects generating and consuming data, deployed

Digital Object Identifier:
10.1109/MCOM.2017.1700328

Massimo Ficco is with the Università degli Studi della Campania Luigi Vanvitelli; Yang Xiang is with Deakin University; Christian Esposito and Francesco Palmieri are with the Università di Salerno.

0163-6804/17/$25.00 © 2017 IEEE

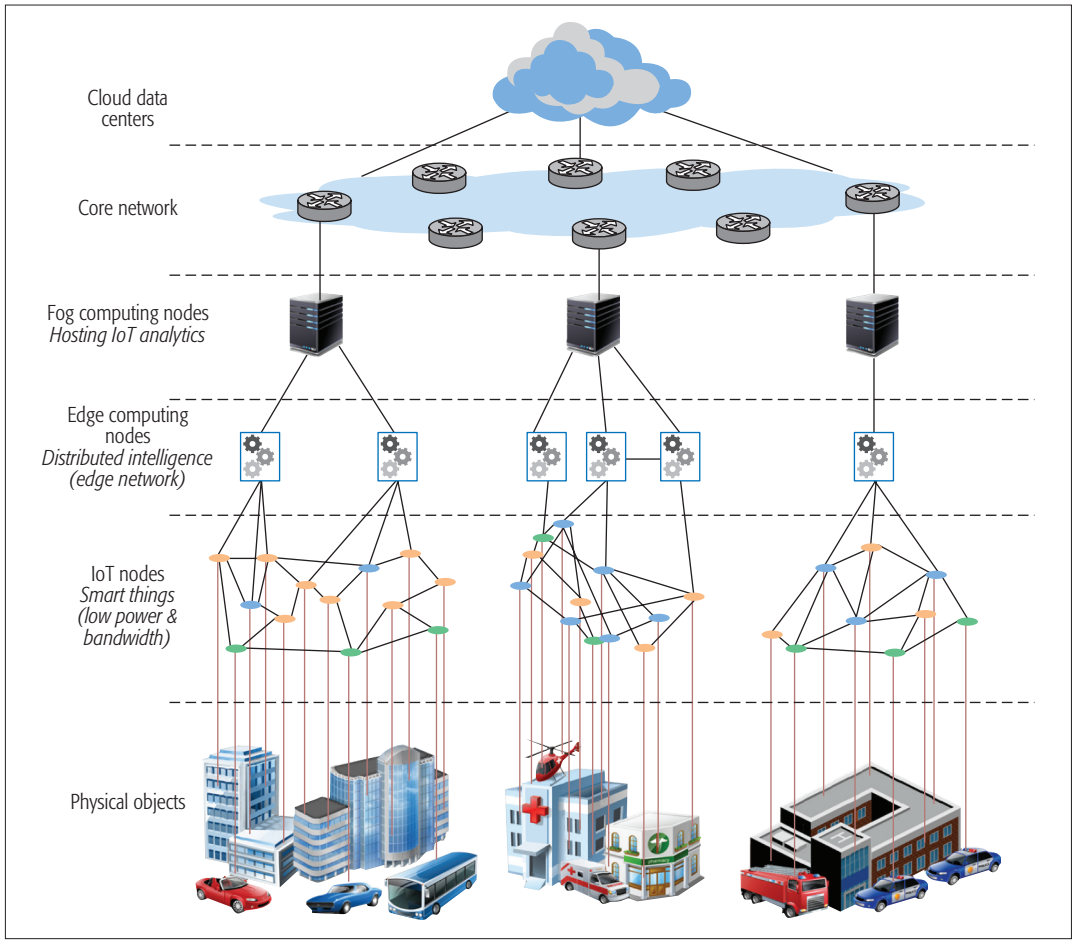


Figure 1. Edge-fog cloud computing architecture as an ecosystem of different cyber-physiscal systems.

In order to extend the services provided by the Cloud to the edge of the network, the new Fog paradigm involves application components, as well as virtualized network entities running both in the Cloud and on devices between the Cloud and the network edge.

across a large geographic area. Figure 1 presents a hierarchical distributed architecture supporting the future IoT applications, which shows the role of edge-fog cloud computing facilities. Specifically, the figure illustrates a typical setting for a smart city, which includes three interleaved and promiscuous applications: one (with orange IoT nodes) for the structural and environmental monitoring of the buildings, one (with blue IoT nodes) for patient management and healthcare, and the last one (with green IoT nodes) for traffic management. All the physical objects in the city, such as vehicles, buildings, and even humans, are equipped with proper devices, which can be abstracted as IoT nodes. For example, the orange IoT nodes can be sensors used to measure oscillations, temperature, or humidity in a room; the blue IoT nodes can be devices monitoring patient movements, vital signs, or drug consumption; and the green IoT nodes can be devices monitoring driving activities, checking traffic conditions, or alerting in case of emergencies.

Close to the IoT nodes (i.e., within a one- or two-hop distance from the IoT devices, and hence on the *network edge*), we can find a collection of loosely coupled and often human-operated nodes, such as tablets, laptops, workstations, wireless access points, and nano data centers. These edge nodes are usually provided with wireless device-to-device connectivity among them, and reliable connectivity to fog nodes. Their main task is to collect the sensor data and perform preliminary processing activities on them. In the example illustrated in Fig. 1, the edge nodes can be entitled to summarize the environmental data of the orange IoT nodes and to rapidly detect possible events (e.g., fire or structures that are about to collapse), or in the case of the green IoT nodes, an edge node can be used to detect a sudden accident on a road and to alert other reachable cars that are on the same route or planning to take it.

The fog nodes are more distant from the IoT objects and consist of more powerful machines, whose main task is supporting massive big data analytics and temporary storage activities. However, even if such nodes can provide significant computational capabilities, they are not able to cope with the increasing complexity of the envisioned applications, so most of the analytics and runtime capabilities have to still be allocated to the central cloud. Thus, in order to extend the services provided by the cloud to the edge of the network, the new fog paradigm involves application components, as well as virtualized network entities running both in the cloud and on devices between the cloud and the network edge. Specifically, they may consist of virtualized appliances or networking devices, such as smart gateways, next generation firewalls, and routers with high computing capabilities, interconnected with high speed and reliable links to the core network. Fog nodes are enabled to run cloud application logic on their native architecture, and can be managed

A real testbed, although desirable, in many cases is too expensive and does not provide a repeatable and controllable testing environment. An alternative approach consists of the emulation of a portion of the IoT solution by using a virtual network and virtual devices, typically for the sensing and embedded devices, where the real software is executed and tested.

and deployed by cloud providers. They are used to perform time-sensitive data analysis, as well as computationally intensive tasks offloaded by edge nodes, in order to reduce latency and traffic in the core network. Data that is less time-sensitive is sent to the nodes in the cloud for historical analysis of large sets of data collected from multiple fog nodes, big data analytics, and long-term storage. In the scenario shown in Fig. 1, the fog nodes can be used in order to make analysis on the traffic in the given portion of the city and to disseminate this information to all the cars placed in this portion, or to detect a critical situation for a patient and to call an ambulance. The fog nodes may be shared among the different applications running within the IoT infrastructure, so in case of a call for an ambulance, its driver can be informed of the best path to reach the patient, and all the cars along the path are instructed to yield to the ambulance.

On the other hand, the structural and dynamic complexity of this kind of information and communications technology (ICT) infrastructure makes the testing and vulnerability analysis of applications running on top of an edge-fog cloud ecosystem extremely challenging, due to the lack of a proper testing environment able to provide a realistic representation of these infrastructures. Large-scale and structural complexity derives from a wide and dense geographically distributed deployment of a huge number of IoT, edge, and fog nodes (in general not belonging to the same provider or organization), running in a wide variety of heterogeneity environments. They should be in charge of interacting and interoperating across different communications technologies, different providers, and federated domains. Their dynamic complexity can introduce behaviors that usually become evident only during on-site system operations, as well as during unpredictable fluctuations in both workloads and computational and communication resources needed to process these workloads. In addition, a proper testing environment should be not only limited to the hardware and software perspectives of an edge-fog cloud ecosystem, but it should also consider the human counterpart in order to have human-driven testing by mimicking how humans behave and interact with the cyber-physical components at the edge of the network. Identifying, understanding, and representing such complexity represent a challenge to support effective test and vulnerability analysis activities [2]. Testing and simulation environments should be provided to reproduce such complex and distributed systems locally in order to gain knowledge about their real behavior as it would be on site.

Testing Edge-Fog Cloud Ecosystems

In order to support development and test of software artifacts and services located at different levels of edge-fog cloud architecture, and assess the offered functionalities and perceived quality of service properties, several commercial and open source solutions have been proposed.

FIT IoT-LAB is a testbed equipped with thousands of wireless nodes located in six different sites across France, which can be used to test protocols and applications in a large-scale wireless IoT environment [3]. SmartSantander is a European project that offers a city-scale experimental facility for testing smart city applications [4]. The testbed comprises a large number of devices deployed in several urban locations. The logical architecture is organized into two tiers: *IoT nodes*, responsible for sensing the environmental parameters; and *gateways*, linking the IoT nodes at the network edge to a core infrastructure, providing more powerful server devices. Gateways enable interworking and integration testing between different IoT and server solutions. Cisco proposed an end-to-end fog application framework called IOx [5], providing facilities for orchestrating and managing applications on thousands of fog nodes in order to enable operations at multiple scales, as well as monitoring the application performance. Specific IOx middleware services and application programming interfaces (APIs) available on fog nodes make runtime and storage capabilities available to applications hosted on the network edge. By using device abstractions provided by Cisco IOx APIs, applications running on the fog nodes can communicate with IoT devices and applications hosted in the cloud.

However, a real testbed, although desirable, in many cases is too expensive, and does not provide a repeatable and controllable testing environment. An alternative approach consists of the emulation of a portion of the IoT solution by using a virtual network and virtual devices, typically for the sensing and embedded devices, where the real software is executed and tested. For example, the solution presented in [6] emulates an IoT environment by using the OpenStack cloud infrastructure, and in [7] a solution called MAMMotH is illustrated for the emulation of IoT nodes. Although this solution allows the implementation of experimental scenarios that are almost identical to real deployment settings, both at the hardware and software level, it can be used to emulate only a part of these complex infrastructures, and is not able to mimic and reproduce the effects of humans in the control loop, mainly referred to as human interactions with IoT devices.

Thus, simulation represents a valuable and cost-effective support for reproducing the more dense architectures that can be found at the edge of the network, as well as their iteration with the fog nodes and the cloud. It could be exploited to perform scheduling, migration, and resource management analysis at the edge-fog level by accurately modeling the involved reality and by carefully quantifying the defined performance metrics. In this direction, SimpleIoTSimulator is a commercial tool for simulating IoT-edge experimental scenarios, consisting of sensors and gateways interacting though common publish/subscribe protocols, including Constrained Application Protocol (CoAP) and MQTT [8]. Brambilla *et al.* [9] proposed a methodology for modeling and simulating large-scale IoT system deployments. It was designed to study low-level networking aspects by analyzing different mobility, network, and energy consumption models. Other concrete examples are [10, 11], which use the Omnet++ and NS3 simulators to reproduce IoT solutions and test their correctness. These simulators provide ad hoc software modules dealing with all the issues related to movement trails, network discovery, ongoing transmissions and

receptions, radio signals, battery consumption, and much more in order to build realistic simulated sensor networks. The application logic can be coded as it would be in a real deployment by means of the programming language supported by the simulator.

However, the presented simulators can be mainly used to model IoT environments. A specific framework designed to model fog environments along with IoT and cloud is iFogSim [12]. It enables performance evaluation of resource management policies applicable to fog environments with respect to their impact on latency, energy consumption, network congestion, and operational costs. It simulates edge devices, cloud data centers, and network links to measure performance metrics. This enables performance evaluation of resource management and scheduling policies across edge and cloud resources under multiple scenarios, such as real-time stream processing in a comprehensive end-to-end environment.

However, the representation and modeling capabilities of all the above simulation-based solutions and approaches are not able to fully capture the characteristics of the most sophisticated and articulated architectures, and in such scenarios a significant degree of uncertainty remains always present and impossible to represent and quantify in a reliable way, due to the heterogeneity as well as the structural and dynamic complexity of the involved systems. Moreover, considering the complexity of such systems and the huge number of involved entities at the different architectural levels, the processing of a single monolithic simulation system can lead to very high simulation cost and time. Thus, to be effective, simulations should manage these complexity aspects, and at the same time be realistic, time-optimal, and cost-effective, which, of course, are objectives that are in contrast to each other. Therefore, specific hybrid and distributed modeling strategies represent a viable alternative to design the simulation environments needed to support the evaluation of such complex systems [13]. In particular, pseudo-dynamic testing, which integrates simulation, emulation, and real components, represents the most effective solution for testing an extremely complex and large ecosystem such as the one composing the upcoming edge-fog cloud computing. This means that all or some of the above-mentioned approaches have to be jointly employed in the testing environment, where each approach will focus on a particular aspect or portion of the architecture. Moreover, such a solution is able to put humans in the loop by encompassing human activities and interactions in the testing activities of the ecosystem.

Pseudo-Dynamic Testing Approach

In order to enable testers to assemble complex, distributed, and cost-efficient edge-fog cloud experimental environments, pseudo-dynamic testing can be exploited. Such an approach leverages multiple testing methodologies and architectures, where traditional modeling practices are hybridized with experimental testing of systems of systems whose dynamic behavior can be condensed into a reduced number of degrees of freedom. In particular, it combines simulation and emulation, and also supports interaction with real systems. The emulated parts are the components under test, such as fog routers and edge devices, which can easily be assessed as prototypes that can be refined and strengthened during the testing activities. Simulation can be exploited to reproduce the behavior of the external systems, such as IoT objects used to generate the experimental workload and the testing stimulation. Unfortunately, some specific devices and phenomena cannot be simulated in a reliable way, so that real IoT devices, or specific activities associated with real human users, can be linked to the emulated fog/edge computing environment, supporting the testing and analysis of extremely realistic interactions between existing sensors and novel system architectures to be integrated into edge-layer services. Moreover, infrastructural components, such as the underlying communication network and cloud data services can also be simulated in order to abstract the testing scenario from unnecessary details and focus only on the elements that are really meaningful for a specific evaluation task, such as evaluating the effect of specific modifications to communication protocols and/or architectures in a fully controllable way. This results in cost-effective, more scalable, stable, and reliable testing frameworks, where the network architecture/layout is not constrained by economic or technological availability factors, and the networking behavior, including the link error rate, delay, and so on, is always verifiable and reproducible during the whole testing process.

Summing up, all the components involved in the experimental scenario, which are not the target of the test analysis, can be simulated, whereas external sources from which it is possible to obtain, in real time, the data streams needed to reproduce reality with a high degree of verisimilitude (e.g., meteorological services) can be associated with real systems connected to the hybrid testing scenarios. Furthermore, emulation of the edge nodes favors the interaction of humans with the systems involved in the scenario under evaluation, by providing a more realistic user experience to the whole testbed through the support of new experiments involving human in the loop without the necessity of a real-world infrastructure based on a complex communication environment. For example, the effect of human operators on edge nodes can be assessed by relying on the real contribution of human perception and decision making capabilities.

The Reference Architecture

Figure 2 illustrates the proposed solution for pseudo-dynamic testing, which encompasses the integration of simulated, emulated, and real components. Specifically, the sensory part made of IoT nodes and network interconnections among them can be simulated by means of several event-based simulators, such as TOSSIM, OMNET++, and NS3. Such tools have been extensively used in the academic literature in the context of WSNs, ad hoc wireless networks, and IoT. Despite being affected by the problem of accurately simulating wireless channels and sensors, they are a suitable solution for reproducing the behavior of a massive number of sensors characterizing an IoT deployment. Moreover, the simulators can be populated by statistics taken from measurement campaigns on

Pseudo-dynamic testing, which integrates simulation, emulation, and real components, represents the most effective solution for testing an extremely complex and large ecosystem such as the one composing the upcoming edge-fog cloud computing.

In order to exploit such a hybrid simulation approach, specific features should be offered to support interoperability and synchronization among the different simulators, as well as communication among simulated and emulated parts for correct evaluation of the scenario.

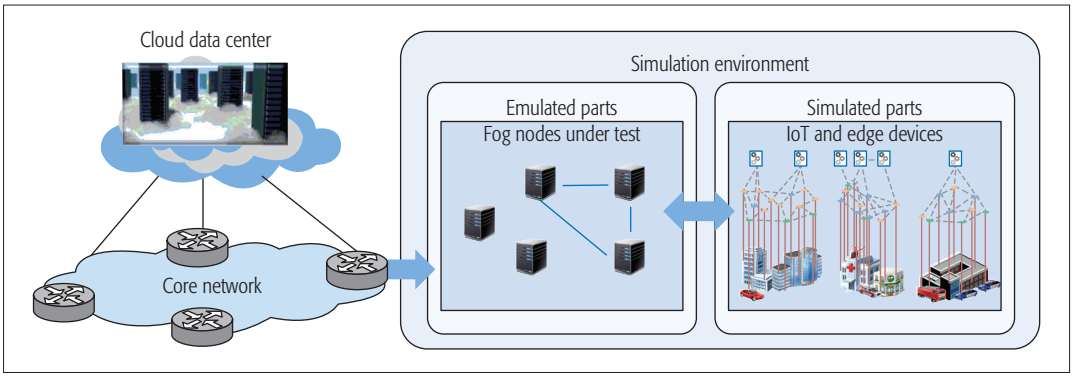


Figure 2. Simulation environment.

real sensor testbeds in order to improve the representativeness of the simulation. Each simulator is characterized by a set of key features and is specialized for selected scenarios (e.g., TOSSIM is dedicated to model sensor networks made of devices running the TinyOS operative system), while others are more general-purpose solutions (e.g., OMNET++ covers the broad types of networks from wired to wireless ones). Moreover, TOSSIM is more efficient in dealing with energy-related issues and obtains more realistic results than OMNET++. Finally, for simulating a realistic experimental scenario, specific sensor behaviors should be simulated, and human activity should also be reproduced to provide inputs to the IoT objects (e.g., by using agent-based models).

However, considering the complexity of IoT models to be simulated, the implementation of realistic large monolithic simulation systems from scratch, by using a single tool, could be excessively complex and time consuming. The simulation approach should exploit facilities for reuse that already exist: system components simulated by dedicated tools. For example, as for disaster management due to earthquakes, several simulators for disaster prediction, intensity analysis, damage estimation, and response have been implemented to provide meaningful outcomes. Therefore, it is more time- and cost-efficient to integrate multiple independent and heterogeneous simulation environments, each with its own features, languages, and operating systems, within a more complex federated simulation system, by enabling the reuse of already existing solutions for IoT simulation. By exploiting such federation dynamics and considering the elastic nature of modern federated computing environments, the resulting testbeds can reach a degree of scalability that is practically impossible in traditional architectures. Indeed, by relying on modern cloud-based runtime capabilities, we can use a virtually unlimited number of systems, located everywhere on the Internet, to host the virtual machines running the emulated or simulated entities, resulting in fully distributed testing architectures.

Moreover, according to the presented solution, the edge and fog parts of the envisioned ecosystem are the components under test. Therefore, they are reproduced using an emulation approach, by running the real software to be deployed on virtual nodes for representative machinery of the operational environment. In the same way, the networks connecting emulated nodes, the IoT objects and the external systems (i.e., core network and the cloud) can be emulated or simulated depending on the testing objectives.

On the other hand, in order to exploit such a hybrid simulation approach, specific features should be offered to support interoperability and synchronization among the different simulators, as well as communication among simulated and emulated parts for correct evaluation of the scenario. In this direction, the high-level architecture (HLA) represents the IEEE standard for distributed simulation [14]. It is an architecture developed to facilitate the reuse, interoperability, and synchronization of different simulation tools and assets, implementing a federation of interoperable simulation members called federates. HLA defines the runtime infrastructure (RTI) specification, which describes how to communicate within the federation. Each federate is represented by a simulation object model (SOM), which specifies the types of information that the federate can provide to or receive from the federation. The interactions among the federates are implemented by exploiting the publish/subscribe paradigm, and are described by the federation object model (FOM). Moreover, RTI features can be exploited in order to coordinate the synchronization among federates, which evolves according to a different emulated and simulated temporal model, as well as manages how fast the simulators advance. As Fig. 3 shows, in order to enable the interoperability among the federates, an *Ambassador* must be implemented for integrating each simulated and emulated component in the federation.

In particular, the interoperability among the simulation components with the emulated nodes is enabled by using a specific gateway (named *Emu-Ambassador*). The emulated fog and edge nodes run on virtual nodes (i.e., virtual machines or containers) with specific IP addresses assigned. When the simulation starts, each Ambassador registers the associated nodes to the federation, and a symbolic name is associated with each IP address, which is used to identify the involved component in the federation. The Emu-Ambassador operates as a bridge, which publishes and subscribes messages in the federation and interacts with the fog nodes by using the traditional socket interface.

Finally, in order to support the setup of such large-scale complex testing scenarios, massive virtualization technologies are exploited. Each scenario could involve several simulation tools

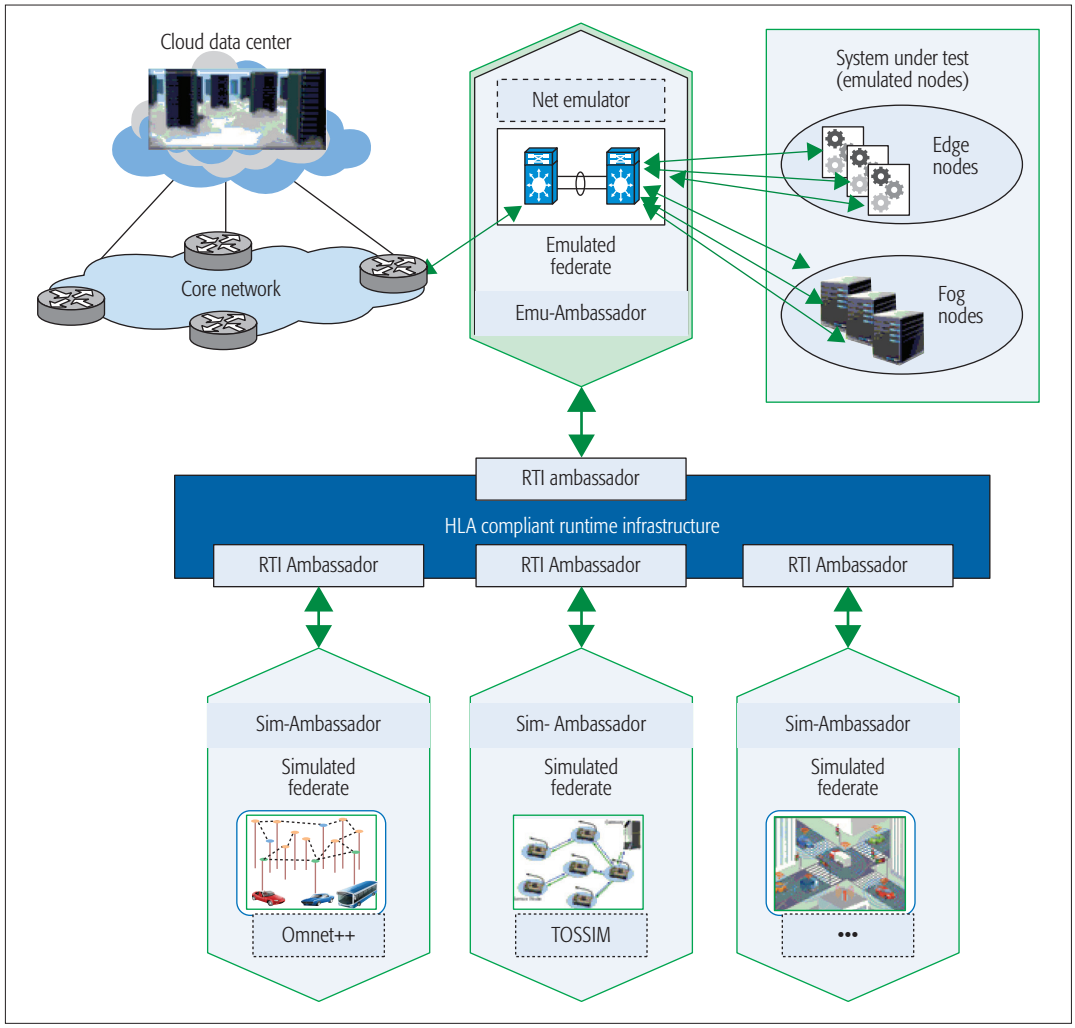


Figure 3. Hybrid simulation environment.

A distributed runtime platform, such as a cloud, can be adopted to provide the set of virtual nodes hosting the simulation tasks and the emulated nodes needed to reproduce a realistic edge-fog ecosystem. Moreover, specific mechanisms can be adopted to dynamically add or remove virtual nodes during the the experiments in order to simulate specific scaling behaviors within the edge-fog cloud ecosystem.

(which simulate dozens or hundreds of heterogeneous IoT objects), and edge and fog nodes to reproduce the experimental environment. A distributed runtime platform, such as a cloud, can be adopted to provide the set of virtual nodes hosting the simulation tasks and the emulated nodes needed to reproduce a realistic edge-fog ecosystem. Moreover, specific mechanisms can be adopted to dynamically add or remove virtual nodes during the the experiments in order to simulate specific scaling behaviors within the edge-fog cloud ecosystem.

Prototype Implementation Details

The fundamental architectural choices for a proof-of-concept implementation, entirely based on open-source components, of the proposed pseudo-dynamic testing framework, are described in the following. Specifically, in order to reduce both the number of virtual nodes to be scheduled (i.e., the required computational resources) and the boot-up time needed for the setup of the whole test scenario, lightweight Linux containers can be adopted for the simulation part, whereas for the emulated components that must be "physically" tested, KVM-based full virtualization technology is a better choice. A container is a packaged, self-contained, ready-to-deploy set of parts of an application, represented by a lightweight virtual image that can include both the middleware and business logic (binaries and libraries) needed to run it. Instead, a virtual machine (VM) is a full monolithic image, which requires guest OS images in addition to the binaries and libraries necessary for the applications. The life cycle of containers is managed by the Docker virtualization technology, supported by the OpenShift cloud platform as a service (PaaS) framework, whereas Kubernetes is used to orchestrate Docker containers on cluster nodes. Finally, OpenStack provides the needed cloud infrastructure as a service (IaaS). This solution acts as a container manager, which enables a registry for the images of the simulated components to be deployed on the virtual nodes. It can be exploited to keep track of the images executed on each node, and identify the virtual nodes on which the images are deployed and downloaded from the registry, needed for instantiating the test scenario. Moreover, the IaaS management layer is exploited in order to access and control the virtualization infrastructure, for supporting on-demand resources provisioning, running simulations on the cloud efficiently, and improving load-balancing capability. Specifically, the Chef technology has been used to simplify the configuration and deployment of the OpenShift nodes and emulated components under test on cloud resources.

In order to cope with the complexity and scale of the edge-fog cloud ecosystem, hybrid and distributed simulation practices supported by the virtualization technologies should be exploited to reproduce reality with a high degree of verisimilitude in order to set up realistic test environments.

The hybrid simulation platform proposed in Fig. 3 has been adopted to reproduce the edge-fog cloud computing scenario presented in Fig. 1, which includes about 1000 IoT nodes simulated through two different tools (i.e., NS3 and OMNET++), interoperating via HLA-RTI infrastructure, and nine edge-fog computing nodes, emulated by the Common Open Research Emulator (CORE) [15] and the OpenvSwitch network emulator. The whole solution has been hosted on the top of a cluster consisting of 8 Dell PowerEdge M610 Blade servers, each equipped with two Quad-Core Intel Xeon E5420 2.50 GHz processors, 16 GB of RAM memory, and 4 Gigabit Ethernet adapters, and running Linux CentOS 6.4. L3 Switching module (Dell M6220), which provides network connectivity to the nodes.

Conclusions

In order to cope with the complexity and scale of the edge-fog cloud ecosystem, hybrid and distributed simulation practices supported by virtualization technologies should be exploited to reproduce reality with a high degree of verisimilitude in order to set up realistic test environments.

Pseudo-dynamic simulation represents one of the most promising ways to support the key players in the edge-fog services market in dealing with the new challenges posed by IoT applications, including:

- The development of new scheduling algorithms for balancing load distribution between edge and cloud resources in order to minimize latency and maximize throughput
- The definition of effective resource management strategies for managing tenant environments, in which multiple application instances with different quality of service requirements share the same edge and fog nodes, networks, and sensing resources
- The definition of new policies for dynamic migration of processing tasks among edge, fog, and cloud nodes (based on the battery life of devices, the kind of operators, etc.) — the question is when and where to migrate what
- The introduction of authentication and authorization techniques that can work with multiple fog nodes belonging to different operators
- The simulation and analysis of the effects of advanced cyber attacks and catastrophic events that could compromise the edge-fog cloud infrastructure, as well as the definition and evaluation of possible recovery policies

References

[1] H. Gupta *et al.*, "iFogSim: A Toolkit for Modeling and Simulation of Resource Management Techniques in Internet of Things, Edge and Fog Computing Environments," June 2016, pp. 1–22; https://arxiv.org/pdf/1606.02007.pdf, accessed Feb. 2017.

[2] G. Kecskemeti *et al.*, "Modelling and Simulation Challenges in Internet of Things," *IEEE Cloud Computing*, vol. 4, no. 1, Jan.–Feb. 2017, pp. 62–69.

[3] C. Adjih *et al.*, "Fit IoT-Lab: A Large Scale Open Experimental IoT Testbed," *Proc. 2nd IEEE World Forum on Internet of Things*, 2015.

[4] L. Sanchez *et al.*, "Smartsantander: IoT Experimentation over a Smart City Testbed," *Computer Networks*, vol. 61, 2014, pp. 217–38.

[5] IOx — Cisco Framework; https://developer.cisco.com/site/iox/documents/developer-guide/?ref=overview, accessed Mar. 2, 2017).

[6] Q. Le-Trung, "Towards an IoT Network Testbed Emulated over OpenStack Cloud Infrastructure," *Proc. Int'l. Conf. Recent Advances in Signal Processing, Telecommu. & Computing*, 2017, pp. 246–51.

[7] V. Looga *et al.*," MAMMOTH: A Massive-Scale Emulation Platform for Internet of Things," *Proc. IEEE 2nd Int'l. Conf. Cloud Computing and Intelligence Systems*, 2012, pp. 1235–39.

[8] Simpleiotsimulator: The internetofthings simulator, http://www.smplsft.com/SimpleIoTSimulator.html, accessed Mar. 2, 2017.

[9] G. Brambilla *et al.*, "A Simulation Platform for Large-Scale Internet of Things Scenarios in Urban Environments," *Proc. 1st Int'l. Conf. IoT in Urban Space*, 2014, pp. 50–55.

[10] P Wehner, and D. Göhringer, "Internet of Things Simulation Using OMNeT++ and Hardware in the Loop," *Components and Services for IoT Platforms*, Sept. 2016, pp. 77–87.

[11] S. Tozlu *et al.*, "Wi-Fi Enabled Sensors for Internet of Things: A Ppractical Approach," *IEEE Commun. Mag.*, vol. 50, no. 6, June 2012, pp. 134–43.

[12] H. Gupta *et al.*, "iFogSim: A Toolkit for Modeling and Simulation of Resource Management Techniques in Internet of Things, Edge and Fog Computing Environments," *Report no. CLOUDS-TR-2016-2*,June 2016.

[13] M. Ficco *et al.*, "An HLA-Based Framework for Simulation of Large-Scale Critical Systems," *Concurrency Computation*, vol. 28, no. 2, 2016, pp. 400–19.

[14] IEEE Std. 1516-2000, "1516–2010 — IEEE Standard for Modeling and Simulation (M&S) High Level Architecture (HLA) — Framework and Rules," Aug. 2010; http://ieeexplore.ieee.org/document/5553440/, accessed Nov. 2016.

[15] CORE — Common Open Research Emulato;: http://www.nrl.navy.mil/itd/ncs/products/core, accessed Jan. 2017.

Biographies

Massimo Ficco (massimo.ficco@unicampania.it) is an assistant professor at the Universita degli Studi della Campania Luigi Vanvitelli. He has a Ph.D. in computer engineering from the University of Napoli Parthenope, Italy. His research interests include security, cloud computing, and pervasive systems.

Cristian Esposito (christian.esposito@dia.unisa.it) is an adjunct professor at the University of Napoli Federico II and a research fellow at the University of Salerno. His research interests include information security and reliability, middleware, and distributed systems. He has a Ph.D. in computer engineering from the University of Napoli Federico II, Italy.

Yang Xiang (yang.xiang@swin.edu.au) received his Ph.D. in computer science from Deakin University, Australia. He is the Dean of Digital Research & Innovation Capability Platform, Swinburne University of Technology, Australia. His research interests include cyber security, which covers network and system security, data analytics, distributed systems, and networking.

Francesco Palmieri (fpalmieri@unisa.it) received his M.S. and Ph.D. degrees in computer science from he University of Salerno. He is an associate professor at the same university. His research interests concern networking protocols, architectures, and security. He directed the Networking Division of the University of Napoli Federico II and is a Senior Member of the Technical-Scientific Advisory Committee e of the Italian NREN.

Vehicular Fog Computing: Architecture, Use Case, and Security and Forensic Challenges

Cheng Huang, Rongxing Lu, and Kim-Kwang Raymond Choo

ABSTRACT

Vehicular fog computing extends the fog computing paradigm to conventional vehicular networks. This allows us to support more ubiquitous vehicles, achieve better communication efficiency, and address limitations in conventional vehicular networks in terms of latency, location awareness, and real-time response (typically required in smart traffic control, driving safety applications, entertainment services, and other applications). Such requirements are particularly important in adversarial environments (e.g., urban warfare and battlefields in the Internet of Battlefield Things involving military vehicles). However, there is no one widely accepted definition for vehicular fog computing and use cases. Thus, in this article, we formalize the vehicular fog computing architecture and present a typical use case in vehicular fog computing. Then we discuss several key security and forensic challenges and potential solutions.

INTRODUCTION

An observation on the Internet of Things (IoT) trend in a 2017 Gartner report [1] is the movement away from cloud-, Thing-, and gateway-centric IoT implementations to the edge (also referred to as fog computing or edge computing in the literature). In such an implementation, the bulk of the application logic, data storage, and analytics are placed on the actual device instead of a cloud or gateway server. This can result in significant bandwidth saving for the heterogeneous communication network.

One potential application of fog computing is in vehicle-based settings, such as the integration of fog computing with conventional vehicle ad hoc networks (VANET) to form the Internet of Vehicles (IoV) or vehicular fog computing. In the latter architecture, vehicles are regarded as intelligent devices that are mobile and equipped with multiple sensors, and have the computational/communication capability to gather useful traffic information. The information is gathered not only from the intra-vehicle sensors but also from the environment external to the vehicle(s). Fog nodes can be deployed at the edge of vehicular networks to efficiently and effectively collect, process, organize, and store traffic data in real time. When acquiring and processing a large amount of data from urban/highway areas via smart vehicles, vehicular fog computing architecture can facilitate or provide a wide range of vehicle-based services to the driver and passengers, such as smart traffic control, road safety improvement, and entertainment services. Similarly, there are potential applications in Internet of Battlefield Things deployment.

Fog computing, especially vehicular fog computing, is still in its early stage, with many unresolved and under-explored technical and operational challenges, ranging from architecture to clear use cases to security issues and so on. There has been interest in fog computing not only from academia, but also from the industry such as the establishment of the OpenFog Consortium [2]. Vehicular fog computing is one area that is relatively under-studied, despite the increasing trend in smart vehicles in practice. For example, in a recent report on connected vehicles by the IHS automotive company [3], it is estimated that there will be 152 million actively connected cars on the road by 2020, and an average car will produce up to 30 TB of data each day. This will result in a significant increase in bandwidth consumption and competition, in the sense that a connected vehicle would need to compete against other devices for finite bandwidth.

One potential solution to reduce the communication overhead is to have the server be geographically closer to the vehicle to serve the vehicle-based applications' demands in real time. This will require a significant investment in the underpinning infrastructure. However, to ensure optimal quality of protection (QoP) and quality of service (QoS), we need to strike a balance between performance, security, and privacy requirements.

In this article, we discuss the architecture, use cases, and security issues in this emerging vehicular fog computing paradigm. In the next section, we present a high-level overview of vehicular fog computing architecture and describe its benefits.

VEHICULAR FOG COMPUTING ARCHITECTURE: OVERVIEW

SYSTEM ARCHITECTURE

A high-level architecture of vehicular fog computing is presented in Fig. 1, which comprises three types of entities, namely smart vehicles as the data generation layer, roadside units/fog nodes as the fog layer, and cloud servers as the cloud layer.

The authors formalize the vehicular fog computing architecture and present a typical use case in vehicular fog computing. They discuss several key security and forensic challenges and potential solutions.

Cheng Huang is with the University of Waterloo; Rongxing Lu is with the University of New Brunswick; Kim-Kwang Raymond Choo is with the University of Texas at San Antonio as well as the University of South Australia.

Digital Object Identifier: 10.1109/MCOM.2017.1700322

0163-6804/17/$25.00 © 2017 IEEE

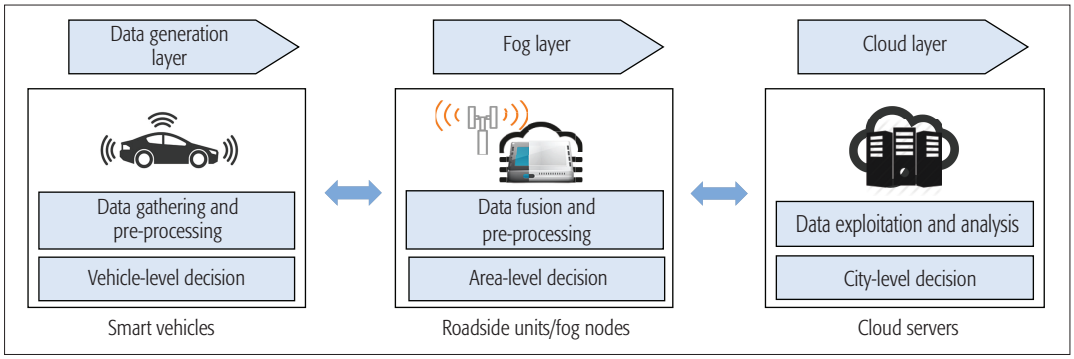


Figure 1. Architecture of vehicular fog computing.

Application type	Service	Description
Traffic control	Smart navigation	Plan optimal routes for smart vehicles
	Smart traffic lights	Schedule traffic lights of each intersection in the city to control traffic flows
Driving safety	Road condition detection	Detect environment information of smart vehicles and make adjustments accordingly
	Emergency warning	Broadcast emergency warning information to nearby smart vehicles, such as car accidents and work zones
Entertainment	Commercial advertisement	Publish advertisements of public interest (e.g., Amber alerts) to nearby smart vehicles
	Multimedia	Provide multimedia services for smart vehicles, such as music and video

Table 1. Application examples of vehicular fog computing.

Smart Vehicles: Smart vehicles play an important role as the key data generator in a vehicular fog computing system, due to their real-time computing, sensing (e.g., cameras, radars and GPS), communication, and storage capabilities. The amount of data collected by the various sensors in a smart vehicle has been estimated to be around 25 GB/h in a single day (e.g., 20–60 MB/s for cameras, 10 kB/s for radar, and 50 kB/s for GPS). Some of these data can be processed by the smart vehicle itself, in order to inform real-time decision making (i.e., vehicle-level decision), while other data will be shared and uploaded to the fog nodes for analysis and used for other purposes (e.g., traffic and infrastructure planning, as well as surveillance).

Roadside Units/Fog Nodes: Roadside units, generally deployed in different areas of a city, can easily be upgraded to act as fog nodes. This will allow the collection of data sent by smart vehicles, processing of the collected data, and reporting of the (processed) data to the cloud servers. These units/nodes also act as the middleware/intermediate devices on the function of a connecting link between the cloud servers and the smart vehicles in a vehicular fog computing system. Unlike existing vehicular networks, these units/nodes will have more functions and provide more diverse services for smart vehicles, such as navigation, video streaming, and smart traffic lights. In other words, these units/nodes are not just relays or broadcasters; they also process data, store data, and make decisions as a fog layer. (i.e., area-level decisions).

Cloud Servers: Cloud servers provide city-level monitoring and centralized control from a remote location. These servers will obtain the data uploaded by the fog nodes while performing computationally intensive analytics to make optimal decisions from a holistic perspective (e.g., city-level decision). For instance, they will monitor, manage, and control the city's road traffic infrastructures to achieve optimal city-level traffic control.

Potential Benefits

The vehicular fog computing architecture, if implemented correctly, can deliver wide-ranging benefits such as those shown in Table 1. Although many current and fast-developing vehicular fog computing systems may have unique features, most vehicular fog computing systems are generally organized in an architecture similar to that shown in Fig. 1, with the following common characteristics: The fog nodes are extensions of the cloud servers from remote areas to the edge in order to offer more efficient and effective services. This will allow vehicle-based applications to benefit in terms of response time, communication, and storage. These properties are particularly important in an adversarial setting (e.g., Internet of Battlefield Things involving military vehicles).

Response time: Most vehicular applications require real-time response, especially for traffic control and safety enhancement applications. However, conventional vehicular cloud computing architecture is not designed to meet this low-latency requirement, since data collected from smart vehicles will be processed remotely instead of locally. Due to the transmission delay and any potential connectivity issues (e.g., out of range), the average response time for cloud-based applications and locally processed applications will likely be more than a second and under 10 ms, respectively. Hence, fog nodes in a vehicular fog computing system, located in proximity to smart vehicles, can significantly reduce the response time for vehicular applications.

Communication: In the foreseeable future, the number of smart vehicles (including smart military vehicles) is likely to increase and perhaps become the norm. Thus, it is likely that the amount of data generated and transmitted by such vehicles will increase exponentially at a high frequency (similar to the current big data trend). In conventional vehicular cloud computing scenarios, raw data

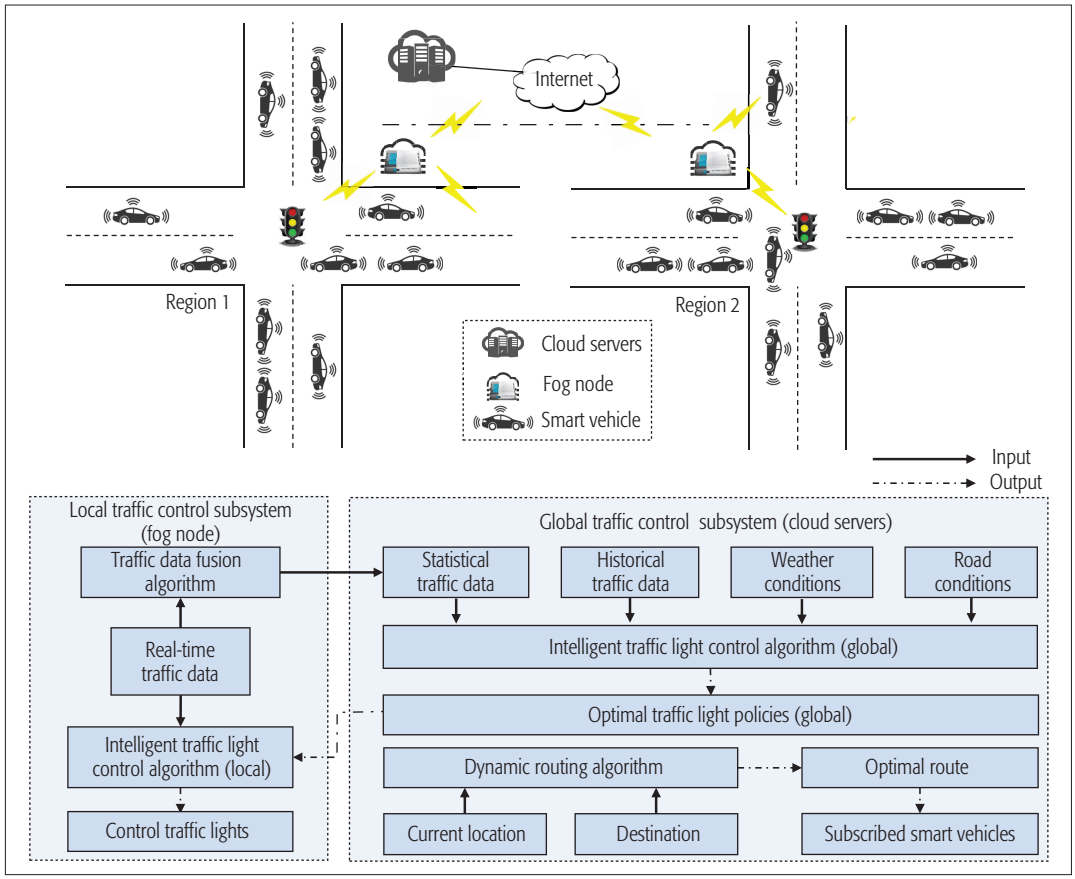


Figure 2. A fog-assisted traffic control system: an overview.

is directly uploaded to the cloud servers for subsequent processing. Despite potential advances in communication technologies, the bandwidth required for efficient transmission of such a big volume of data is not guaranteed due to a wide range of logistical, political, and geographical factors, particularly in a conflict zone. If the data is too large and frequent, communications will be a bottleneck for most vehicular applications. Therefore, the fog nodes in a vehicular fog computing system can alleviate such limitations by pre-processing the collected data so that the data can be aggregated/filtered prior to uploading. This allows data volume and frequency to be reduced.

Storage: For conventional vehicular cloud computing architecture, almost all application data will be stored in the remote cloud servers. This may not be practical due to the changing nature of vehicular applications and collected data. For example, data and vehicular applications are increasingly becoming location-aware. Thus, the ability to access stored data in real time (e.g., data stored in decentralized location-aware fog nodes) will reduce the storage burden on the remote cloud servers.

A Potential Use Case: A Fog-Assisted Traffic Control System

We use a fog-assisted traffic control system as a use case to explain the vehicular fog computing architecture. A fog-assisted traffic management system is designed to deliver benefits such as reducing road traffic congestion and car accidents. A typical implementation will consist of two subsystems: one responsible for the local area and one responsible for the global area (a.k.a city-wide area).

Local Traffic Control Subsystem

The local traffic control subsystem is tasked with monitoring and managing traffic flow in a local area. A fog node's communication range covers a region of the city and can involve several intersections, as shown in Fig. 2. If a smart vehicle is physically located within the communication range of a fog node, it can send and receive messages to and from the fog nodes. Specifically, when a smart vehicle drives into a region within the coverage of a fog node, it will frequently report its current location, speed, weather conditions, and road conditions to the specific node until it leaves this region.

Based on the data received from the smart vehicles, the local traffic control subsystem can perform the following. For example, in the first phase, the fog node will monitor and control the local traffic flow by scheduling the traffic light at each intersection for the smart vehicles in its region. An intelligent traffic light control algorithm (local) is implemented at the fog node. By using each vehicle's reported data as the input, the fog node calculates the traffic information such as road segment occupancy, and then runs the intelligent traffic light control algorithm to avoid traffic build-up by managing the red and green phase proportion of each traffic light. This phase should be operated in real time and have low latency. When smart vehicles are being operated at high speed and are constantly on the move, the

If the data is too large and frequent, communications will be a bottleneck for most vehicular applications. Therefore, the fog nodes in a vehicular fog computing system can alleviate such limitations by pre-processing the collected data so that the data can be aggregated/filtered prior to uploading. This allows data volume and frequency to be reduced.

A number of traffic management algorithms, which may be deployed in a fog-assisted traffic control system, have been proposed in recent years. These include the traffic scheduling algorithms, ITLC and ATL, for controlling the traffic lights of an isolated traffic intersection and the entire road network [4], and a distributed real-time routing algorithm to avoid traffic congestion.

response time for controlling the traffic flow will be significantly impacted by the traveling speed of these vehicles. It would not be useful to update traffic lights for vehicles leaving this intersection. During the second phase, the fog node will pre-process and aggregate these received data as the statistical traffic information, and report such information to the cloud servers. Concretely, the data reported to the cloud servers will not be the same as each vehicle's data has a different data format (e.g., location and speed). Using a traffic data fusion algorithm, the fog node will integrate the received data as traffic volumes (e.g., number of vehicles, average speed of the vehicles, and average waiting time of the vehicles at each intersection) and then report the output of this algorithm to the cloud servers.

Global Traffic Management Subsystem

The global traffic control subsystem is responsible for controlling and managing the traffic flow from a city-wide perspective. As shown in Fig. 2, the cloud servers are remote, gathering data sent by the fog nodes and performing (big) data analytics to mine the traffic information. Traffic controlling algorithms used in the cloud servers include an intelligent traffic light control algorithm (global) and a dynamic routing algorithm.

The intelligent traffic light control algorithm in the cloud servers is more complicated compared to that at the fog nodes. The algorithm in the cloud servers is intended to predict and adjust traffic control systems (e.g., traffic lights) by considering not only real-time traffic volume but also other relative information (e.g., historical traffic records, weather conditions, and road conditions). Thus, this is a more time-consuming exercise at the cloud server. However, the response time will not be a critical metric for this algorithm since the traffic volume of the whole city and weather conditions are unlikely to vary significantly over a short period. The mining results of this algorithm will be the optimal traffic light policies, and the cloud servers will distribute these policies as the feedback to all fog nodes in the city. Additionally, the cloud servers can provide a navigation service for some smart vehicles to help control the traffic flow. A dynamic routing algorithm is also required. Using a participating smart vehicle's current location and destination as the input, the algorithm will output the smart vehicle's optimal route by predicting and simulating the traffic conditions.

A number of traffic management algorithms, which may be deployed in a fog-assisted traffic control system, have been proposed in recent years. These include the traffic scheduling algorithms ITLC and ATL for controlling the traffic lights of an isolated traffic intersection and the entire road network [4], and a distributed real-time routing algorithm to avoid traffic congestion [5].

Security and Forensic Challenges in Vehicular Fog Computing

Research in understanding and mitigating security risks in vehicular fog computing is still in its infancy. Existing security research mainly focuses on the identification of potential attacks, threats, and vulnerabilities of fog-assisted vehicular applications. Generally speaking, attacks in vehicular fog computing can be categorized into passive and active attacks, and there are two kinds of attackers: an external attacker and an insider. An external attacker is not equipped with key materials in a vehicular fog computing system, while an insider attack is one originating from compromised smart vehicles, fog nodes, or cloud nodes that hold the key materials.

A passive attack does not destroy the functionality of a vehicular fog computing system but attempts to disclose private information (e.g., eavesdropping). Passive attacks by an internal attacker are generally more damaging than those conducted by an external attacker, particularly in an adversarial setting, since the insider is more likely to be able to circumvent existing security controls. An active attack is an attempt to deliberately disrupt the operations of a vehicular fog computing system (e.g., distributed denial of service [DDoS] attacks, modifying data of smart vehicles or the decision data of fog nodes and cloud servers, and data exfiltration). An active attack is easy to detect as long it has an enormous impact on the system. However, sometimes the attacker is prone to perform an inconspicuous attack (insider or outsider) during an ultra short period, which is hard to find.

Security and Forensic Requirements

A secure vehicular fog computing implementation should provide the following baseline security and forensic properties.

Confidentiality: Confidentiality ensures that any unauthorized access attempts to either data-at-rest and data-in-transit in a vehicular fog computing system will be detected and prevented.

Integrity: Integrity ensures that any unauthorized attempts to modify data being transmitted or stored will be detected. In a vehicular fog computing system, it is critical to meet the integrity requirement since unauthorized modification may result in serious and/or fatal consequences, especially in life-critical vehicular application contexts such as a traffic control system.

Authentication: Authentication ensures that any two communication entities are able to corroborate the data in transmission.

Access control: Access control is designed to limit fog node access only to authorized entities (e.g., participating and non-compromised smart vehicles to gain access to the fog nodes for some subscribed services like navigation and entertainment).

Non-repudiation: Non-repudiation ensures that any entity in the system is not able to deny a previous action (e.g., sending data).

Availability: Availability ensures that whenever a vehicular application attempts to access the fog nodes or cloud servers, they are always available.

Reliability: Reliability ensures that the data collected from smart vehicles and fog nodes has not been modified or fabricated.

Forensics: Forensics ensures that the capability to identify, collect, and analyze data from smart vehicles, fog nodes, and the underlying infrastructure for tracing and identifying the malicious sources.

In general, most of the above-mentioned security requirements can be achieved partly using cryptographic techniques. For example, fully homomorphic encryption primitives can be employed to achieve confidentiality and functionality at the same time. However, most security mechanisms only effectively defend against passive attacks, and there is no foolproof security solution. Once one or more fog nodes have been compromised [6], for example, to launch attacks within a fog-assisted traffic control system, more sophisticated security mechanisms will be necessary to detect and deter such attacks.

An Example: A Compromise Attack on Fog-Assisted Traffic Control System

In a fog-assisted traffic control, fog nodes are regularly deployed at public roadsides without any physical isolation due to their location-aware nature. Hence, such nodes are more vulnerable to physical compromise attacks compared to cloud servers that are generally protected physically. With public access to fog nodes, attackers can attempt a variety of attacks, such as false data injection, black/gray hole attack, and on-off attack. A node compromise attack can be broadly generalized into the following three stages:

- The attacker gains administrative access to a fog node by physically capturing and compromising the particular node.
- The attacker alters the functions of the compromised fog node and redeploys it back to the system.
- The attacker manages the compromised fog nodes and launches different attacks to disrupt the process of traffic control.

Two types of attackers are considered after the fog nodes are compromised. Specifically, an attacker who seeks to degrade the performance of the system (hereafter referred to as an evil attacker) and an attacker who seeks to benefit himself/herself in the system (i.e., a selfish attacker). Specifically, the evil attacker will minimize the road traffic network's utility to maximize the total travel time in this network (i.e., average travel time increases). Suppose that T' denotes the total travel time computed from a traffic model for the attacked network, while T is the total travel time computed from the same traffic model for a normal network. The goal of this attacker is to find the maximum $T' - T$ using the compromised fog nodes. In contrast, the selfish attacker will maximize his/her own interests by changing the traffic flow of the network. That is, the attacker will minimize travel time between locations A and B by amending the strategies of compromised fog nodes since he/she would likely travel from A to B. Let $T'_{A,B}$ denote the travel time computed from a traffic model for the target network, while $T_{A,B}$ is the travel time computed from the same traffic model for the normal network. The purpose of this attacker is to locate the maximum $T_{A,B} - T'_{A,B}$.

However, in comparison to the evil attackers, selfish attackers are more difficult to detect as attack time may be extremely short, and the compromised fog node is most likely to behave normally outside the attack. In addition, in an attack performed by a selfish attacker, modifications to the system are likely to be minimal (e.g., only sufficient to reduce the waiting time at an intersection the attacker is approaching). For fog-assisted traffic control, traffic light control in each intersection is determined by the fog node itself, and a minor modification in the traffic light's strategy will not result in a significant influence or impact on the entire system. Thus, we introduce two security mechanisms as potential countermeasures for selfish attackers.

Potential Countermeasures for Selfish Attacks

To deal with selfish attackers, it is important to identify and detect compromised fog nodes in the system. It is not practical to physically check all the fog nodes deployed in the system (e.g., the state of Texas) for compromise, and real-time detection is not realistic. We also need to ensure that the security solutions do not significantly impact functionality and performance. Hence, we posit the use of an evidence-based digital forensic approach and a traffic-based analysis approach based on real-time and historical traffic data.

Evidence-based digital forensic approach: A potentially effective way to locate compromised fog nodes is to forensically analyze artifacts from smart vehicles and fog nodes that have been or are believed to be compromised. For example, smart vehicles can directly communicate with the fog node and make judgments based on the behavior of each fog node. If any smart vehicle or fog node flags another node or vehicle as suspicious or exhibits abnormal behavior (e.g., an usually short/long waiting time at a particular intersection), a forensic investigation into these vehicles or nodes can be conducted. Based on the findings of the forensic investigation, the vehicles may be restricted from further interacting with the system or the nodes replaced. Depending on the findings and the context, we could monitor the behavior of the vehicle or node before making the final determination (e.g., compromised, malicious, or false alarm). To generate evidence, an authentication mechanism is needed between the smart vehicles and the fog nodes, and each fog node needs to periodically broadcast its digitally signed identity/signature that can be verified by the smart vehicles.

To investigate the utility of this approach, we simulate the traffic condition based on SUMO [7] and OpenStreetMap [8]. As shown in Fig. 3a, we choose two random junctions (one is compromised and the other is normal) in the city of Waterloo, Canada. We then generate different traffic rates (i.e., 1 vehicle/s, 1 vehicle/5 s, and 1 vehicle/30 s). The routes for these smart vehicles are created randomly during one hour. When smart vehicles pass through a junction, they have a probability p_d to accurately identify a compromised fog node and a probability p_e to mistakenly regard a normal fog node as a compromised one. Formally, the probability $1 - p_d$ indicates the false negative rate, while the probability p_e indicates the false positive rate. We define p_d as 0.8, 0.4, and 0.2 and p_e as 0.1, 0.2, and 0.4, and the numerical results of the simulation are shown in Fig. 3b. The number of reports from the compromised fog node is slightly more than that of those of the normal fog node. Since the smart vehicles' diverse abilities (low p_d or high p_e) make a lot of "noises," it is a challenge to identify the compromised fog node from the received reports.

In a fog-assisted traffic control, fog nodes are regularly deployed on public roadside without any physical isolation due to its location-awareness nature. Hence, such nodes are more vulnerable to physical compromise attacks, as compared to cloud servers that are generally protected physically.

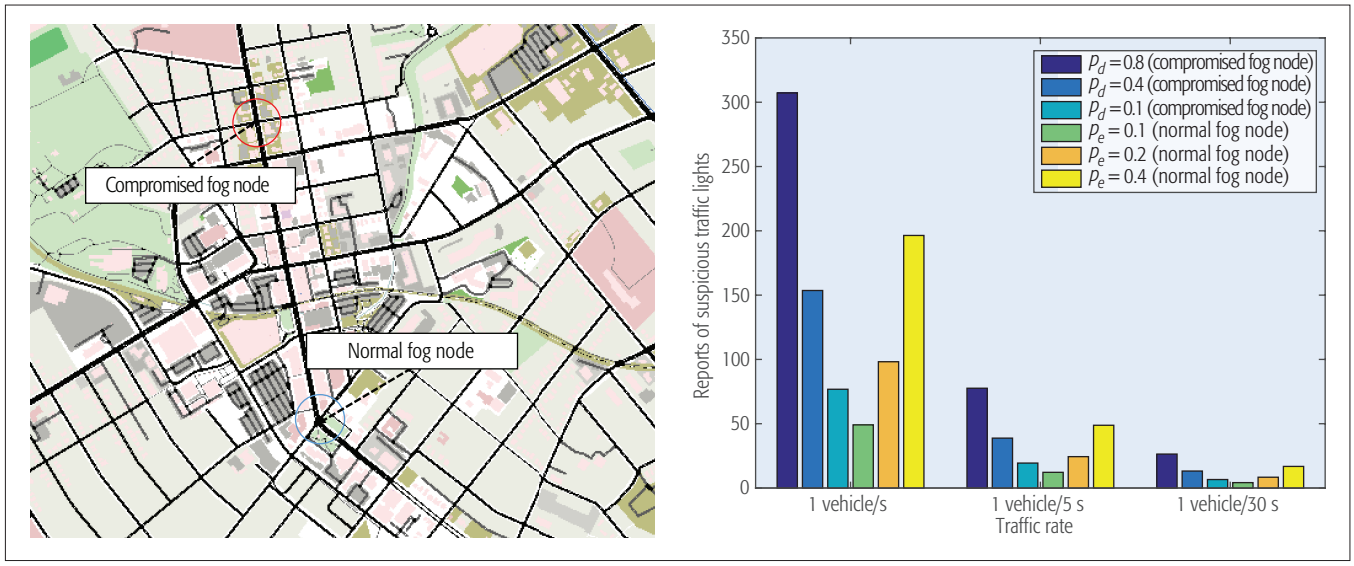


Figure 3. Simulation of evidence-based digital forensic approach: a) Waterloo map downloaded from OpenStreetMap; b) numerical results of simulation based on different settings.

Therefore, a reputation system needs to be in place to prevent "badmouthing" attacks from colluding smart vehicles. Finally, the suspicious vehicles or fog nodes should be manually examined to confirm the abnormal behaviors from the smart vehicles, and the reputation of smart vehicles should be updated based on the feedback.

Traffic-based analysis approach: Another possible practical approach is based on big data analytics and deep learning algorithms. The cloud servers have access to historical/archival traffic data reported by each fog node. No matter how sophisticated an attacker is, data from compromised fog nodes will display/have some (minor) different behavior and characteristics than data from normal nodes. For example, as long as normal fog nodes are deployed around the compromised ones, which modify the traffic lights without considering logical traffic flow, the traffic flow at the normal fog nodes will be irregular to some extent. The association between the fog nodes can also be mined to identify the compromised fog nodes based on the real-time traffic flow changes. A simple example is the traffic prediction model. The current traffic data is utilized to forecast the future traffic flow, which is compared to the real traffic flow at that time to locate fog nodes that may be compromised. In addition, the historical traffic data can also be a baseline reference to identify compromised fog nodes. The relationship between the present traffic flow and the historical traffic flow can also help to distinguish abnormal traffic data and then can be deeply mined to identify compromised fog nodes. That is, the traffic features extracted from the statistical traffic information, such as the average velocity and number of vehicles in a junction, can be analyzed using data clustering (e.g., *K*-means and outlier detection) and classification methods (e.g., naive Bayes and support vector machine).

In summary, a node compromise attack is likely to be mitigated by an evidence-based digital forensic approach and/or a traffic-based analysis approach based on real-time and historical traffic data. Specifically, in the case of established authentication schemes, certificate-based and identity-based encryption/signatures can easily be applied and implemented, despite having limitations such as authentication efficiency and revocation costs [9]. In terms of efficiency, the group authentication scheme in [10] is a potential solution, but the use of such a scheme may complicate forensic investigations. A recently proposed reputation-based system [11] for vehicular networks, based on the Dirichlet distribution, can potentially be extended to help detect malicious vehicles and compromised fog nodes. One could also combine reputation-based systems and truth discovery approaches, such as majority voting and weighted averaging, for enhanced accuracy. Although there are a number of traffic monitoring and prediction systems in the literature (e.g., [12]), how to detect the abnormal traffic flow and finally detect the compromised fog nodes based on a large volume of traffic data still remains a research and an operational challenge.

Conclusion

In this article, we present an architecture for vehicular fog computing, and discuss the potential benefits, security, and forensic challenges and mitigation strategies using the fog-assisted traffic control system as a use case.

To keep pace with technological advances and the changing nature and needs of our society, there are a number of research opportunities in this space. One such challenge is to effectively strike a balance between functionality, security, and privacy in specific vehicular application contexts (e.g., data privacy [13] and location privacy [14]). Extending the work of Ab Rahman, Glisson, Yang, and Choo [15]}, how to best integrate forensics techniques and best practices into the design and development of a vehicular fog computing system so that it is forensically ready/friendly is another potential research topic. Having a forensically ready/friendly vehicular fog computing system will allow the real-time identification, collection, and analysis of data that can be used to inform mitigation strategies.

References

[1] B. Menezes and S. B. Alaybeyi, "Iot Components Will Require Changes to Enterprise Networks," Gartner Report, vol. G00316844, 2017, pp. 1–14.

[2] O. Consortium, "Openfog Architecture Overview"; https://www.openfogconsortium.org/wp-content/uploads/OpenFog-Architecture-Overview-WP-2-2016.pdf, 2016, accessed 3 Mar. 2017.

[3] S. Institute, "The Connected Vehicle: Big Data, Big Opportunities"; https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/connected-vehicle-107832.pdf, 2016, accessed 11 Mar. 2017.

[4] M. B. Younes and A. Boukerche, "Intelligent Traffic Light Controlling Algorithms Using Vehicular Networks," *IEEE Trans. Vehic. Tech.*, vol. 65, no. 8, 2016, pp. 5887–99.

[5] S. J. Pan, I. S. Popa, and C. Borcea, "DIVERT: A Distributed Vehicular Traffic Re-Routing System for Congestion Avoidance," *IEEE Trans. Mob. Comp.*, vol. 16, no. 1, 2017, pp. 58–72.

[6] A. Laszka *et al.*, "Vulnerability of Transportation Networks to Traffic-Signal Tampering," *7th ACM/IEEE Int'l. Conf. Cyber-Physical Systems 2016*, Vienna, Austria, Apr. 11–14, 2016, pp. 16:1–16:10.

[7] D. Krajzewicz *et al.*, "Recent Development and Applications of SUMO — Simulation of Urban MObility," *Int'l. J. Advances in Systems and Measurements*, vol. 5, no. 3&4, Dec. 2012, pp. 128–38.

[8] OpenStreetMap, "Waterloo Map Downloaded from Openstreetmap Dataset"; https://www.openstreetmap.org/#map=13/43.4572/-80.5035, 2017, accessed 4 Apr. 2017.

[9] F. Qu *et al.*, "A Security and Privacy Review of VANETS," *IEEE Trans. Intelligent Transportation Systems*, vol. 16, no. 6, 2015, pp. 2985–96.

[10] C. Lai *et al.*, "GLARM: Group-Based Lightweight Authentication Scheme for Resource-Constrained Machine to Machine Communications," *Computer Networks*, vol. 99, 201, pp. 66–816.

[11] H. Hu *et al.*, "Tripsense: A Trust-Based Vehicular Platoon Crowdsensing Scheme with Privacy Preservation in VANETS," *Sensors*, vol. 16, no. 6, 2016, p. 803.

[12] R. Lu *et al.*, "A Lightweight Conditional Privacy-Preservation Protocol for Vehicular Trafficmonitoring Systems," *IEEE Intelligent Systems*, vol. 28, no. 3, 2013, pp. 62–65.

[13] R. Lu *et al.*, "A lightweight Privacy-Preserving Data Aggregation Scheme for Fog Computing-Enhanced IoT," *IEEE Access*, vol. 5, 2017, pp. 3302–12.

[14] H. Zhu *et al.*, "An Efficient Privacy-Preserving Location-Based Services Query Scheme in Outsourced Cloud," *IEEE Trans. Vehic. Tech.*, vol. 65, no. 9, 2016, pp. 7729–39.

[15] N. H. Ab Rahman *et al.*, "Forensic-by-Design Framework for Cyber-Physical Cloud Systems," *IEEE Cloud Computing*, vol. 3, no. 1, 2016, pp. 50–59.

Biographies

Cheng Huang [S'15] (c225huan@uwaterloo.ca) received his B.Eng. and M.Eng. from Xidian University, China, in 2013 and 2016, respectively, and was a project officer with the INFINITUS laboratory at the School of Electrical and Electronic Engineering, Nanyang Technological University until July 2016. Since September 2016, he has been a Ph.D. student with the Department of Electrical and Computer Engineering, University of Waterloo, Ontario, Canada. His research interests are in the areas of applied cryptography, cyber security, and privacy.

Rongxing Lu [S'09-M'11-SM'15] (rlu1@unb.ca) is an assistant professor at the Faculty of Computer Science, University of New Brunswick. He was awarded the Governor General's Gold Medal, Canada, in 2012, and the IEEE ComSoc Asia Pacific Outstanding Young Researcher Award in 2013. He is Secretary of IEEE ComSoc CIS-TC. His research interests include applied cryptography, privacy enhancing technologies, and IoT-big data security and privacy.

Kim-Kwang Raymond Choo [SM'15] (raymond.choo@fulbrightmail.org) holds the cloud technology endowed professorship at the University of Texas at San Antonio. His awards include the ESORICS 2015 Best Research Paper Award, the 2015 Winning Team of Germany's University of Erlangen-Nuremberg Digital Forensics Research Challenge, the 2014 Australia New Zealand Policing Advisory Agency's Highly Commended Award, the 2010 Australian Capital Territory Pearcey Award, a Fulbright Scholarship, a 2008 Australia Day Achievement Medallion, and the British Computer Society's Wilkes Award. He is an Australian Computer Society fellow.

To keep pace with technological advances and the changing nature and needs of our society, there are a number of research opportunities in this space. One such challenge is to effectively strike a balance between functionality, security and privacy in specific vehicular application contexts.

NETWORK SERVICES CHAINING IN THE 5G VISION

Jordi Mongay Batalla
George Mastorakis
Constandinos X. Mavromoustakis
Ciprian Dobre
Naveen Chilamkurti
Stefan Schaeckeler

Among all the propositions for the future Internet, the fifth generation (5G) seems to be very well positioned for becoming a reality in the near future. The 5G network aims to converge mobile and fixed networks for supporting end-to-end applications and services. 5G will be resilient, secure, and available at all times. It is unthinkable to build such a convergent network without slicing at all levels: from the physical (slicing is enforced by the integration of several radio access technologies) to application levels. Slicing will be achieved thanks to service chaining: data processing will become a sequence of services, potentially managed by different stakeholders. This will require more collaboration between stakeholders and/or greater openness of the offered services.

There are a number of required 5G features that pose real challenges to network service chaining:

- 5G will count massive concurrent sessions, even in small cells, since mobile edge computing (MEC) is needed to be the running space for many applications with ultra-low latency requirements. Thus, services in 5G should be modular (which will increase the importance of approaches such as micro-services) and distributed in the heterogeneous network.
- 5G will build multi-operator heterogeneous scenarios, where management will be distributed. Therefore, management will be based on service chaining, including virtual network functions and external management applications. Two main scopes of management will be the separation between data storage and processing, and distributed management (e.g., operators' dashboards should be visible to other operators). Multi-technology and multi-operator scenarios bring important challenges to service chaining since concatenated services may be of different natures and have different scopes. An example of the latter is how to localize an error in one slice that spans different domains.
- Agile network operations of the radio are necessary in 5G, so radio access is mainly based on virtualizing and chaining the services (cloud radio access network [C-RAN], virtualized RAN, etc.). The slicing at the physical layer and at the service layer for RAN functions makes the integration more difficult. Moreover, automation in radio access services is necessary so that they may be self-managed and more dynamic (real-time management and control).
- 5G requires dynamicity, understood as the capacity of offering ad hoc solutions, some of them managed by the customer. This requires flexibility in service chaining, so new solutions must be provided. Fault, configuration, accounting, performance, and security management should be modular and open to core and/or access ad hoc services defined by customers (enterprises).

All these features will require the use of non-monolithic developments of service chaining, based more on reusable (micro-) services. Services will be much more abstract (developed in an out-of-context way) and service chains much more modular. In an all-layers slicing architecture such as 5G, service chain orchestration takes on significant importance. Orchestration must consider the service chain reliability, which is a challenging requirement due to multi-nature (different procedures and scopes) services.

Applicability of advanced service chains in 5G networks is presented in the articles of this Feature Topic. The authors' research shows foresight of networking into the future Internet for the readers.

The article "Network Service Chaining in the Fog and Cloud Computing for the 5G Environment: Data Management and Security Challenges" presents a novel architecture for providing cloud and fog facilities in 5G networks. The presented architecture makes use of advances in network functions virtualization and software defined networking (SDN) in order to provide access to data analytics and processing in case the end user is on the move. A similar approach is defended by Datsika *et al.* in the article "Software Defined Network Service Chaining for OTT Service Providers in 5G Networks," where the authors analyze the position of over-the-top (OTT) service providers in heterogeneous environments and propose prioritization of network service chaining for OTT applications based on SDN.

For their part, Qiu *et al.* present the problem of massive data collection with ultra-low latency and low energy consumption requirements in "A Lifetime-Enhanced Data Collecting Scheme for the Internet of Things," and propose a solution for improving routing decisions to data storage.

Edge computing is discussed in "Computing, Caching, and Communication at the Edge: The Cornerstone for Building a Versatile 5G Ecosystem," which deals with user-driven ad hoc solutions in the 5G edge and proposes that end users build virtual fogs for providing QoS/QoE requirements (e.g., latency) into the edge. Ultra-low latency is also a crucial aspect of providing edge computing, as shown in "Bringing Computation Closer toward the User Network: Is Edge Computing the Solution?," where research efforts and challenges of edge computing are analyzed together with the requirements of the network in order to provide the edge computing principles.

At a lower level, a solution for radio access in cognitive radio networks by using service chaining is demonstrated in "Cognitive Radio Network and Network Service Chaining toward 5G: Challenges and Requirements." Kakalou *et al.* discuss the deployment of cognitive radios into the 5G horizon.

In conclusion, the six articles in this Feature Topic give a spherical view of the challenges of service chaining in 5G networks for providing sliced networking at all levels.

BIOGRAPHIES

JORDI MONGAY BATALLA (jordim@interfree.it) is currently the head of the Internet Technologies and Applications Department at the National Institute of Telecommunications. He is also with Warsaw University of Technology. He is an editor of several books and an author or co-author of more than 150 papers published in international journals and conferences in the fields of technologies (radio: 4G and 5G; wired: network services chain, SDN; and applications (the Internet of Things, smart cities, multimedia) for the future Internet.

GEORGE MASTORAKIS (gmastorakis@ieee.org) obtained his M.Sc. in telecommunications from University College London, United Kingdom, in 2001 and his Ph.D. degree from the University of the Aegean, Greece, in 2008. He currently serves as an associate professor at the Technological Educational Institute of Crete, Greece. His research interests include mobile networks, multimedia applications and services, cognitive radio networks, radio resource management, network management, quality of service, the Internet of Things, and energy-efficient networks.

CONSTANDINOS X. MAVROMOUSTAKIS [SM] (mavromoustakis.c@unic.ac.cy) is currently a professor with the Department of Computer Science at the University of Nicosia, Cyprus, where he leads the Mobile Systems Lab (MOSys Lab., http://www.mosys.unic.ac.cy/) in the Department of Computer Science at the University of Nicosia. He has been an active member (Vice-Chair) of IEEE/Region 8 Cyprus Section since January 2016, and since May 2009 he has served as the Chair of the C16 Computer Society Chapter of the Cyprus IEEE Section.

CIPRIAN DOBRE (ciprian.dobre@cs.pub.ro) is a professor at University Politehnica of Bucharest. He leads the activities within the Laboratory on Pervasive Products and Services, and MobyLab. His research interests involve mobile wireless networks and computing applications, pervasive services, context awareness, and people-centric sensing. He is Director or Principal Investigator for national and international research projects, and has received the IBM Faculty Award, CENIC Awards, and Best Paper Awards. He serves on the Steering and Organization Committees of major conferences.

NAVEEN CHILAMKURTI (n.chilamkurti@latrobe.edu.au) is currently cybersecurity program coordinator in the Department of Computer Science and Information Technology, La Trobe University, Melbourne, Australia. He is also the inaugural Editor-in-Chief of the *International JWNBT*. He has published about 200 journal and conference papers. Some of his publications are in IEEE journals. His current research areas include intelligent transport systems, vehicular security, the Internet of Things, SDN, smart grid, and security in wireless networks.

STEFAN SCHAECKELER (sschaeck@cisco.com) works for Cisco Systems, Inc, San Jose, California. He received his Ph.D. degree from Santa Clara University, California, in 2010. He has published 12 papers in refereed journals and conference proceedings. His works for Cisco on edge, core, and data center routers, helping to pave the road for next generation Internet technologies.

Network Service Chaining in Fog and Cloud Computing for the 5G Environment: Data Management and Security Challenges

Rajat Chaudhary, Neeraj Kumar, and Sherali Zeadally

The authors present an architecture that integrates cloud and fog computing in the 5G environment that works in collaboration with advanced technologies such as SDN and NFV with the NSC model. The NSC service model helps to automate the virtual resources by chaining in a series for fast computing in both computing technologies. The proposed architecture also supports data analytics and management with respect to device mobility.

ABSTRACT

In the last few years, we have seen an exponential increase in the number of Internet-enabled devices, which has resulted in popularity of fog and cloud computing among end users. End users expect high data rates coupled with secure data access for various applications executed either at the edge (fog computing) or in the core network (cloud computing). However, the bidirectional data flow between the end users and the devices located at either the edge or core may cause congestion at the cloud data centers, which are used mainly for data storage and data analytics. The high mobility of devices (e.g., vehicles) may also pose additional challenges with respect to data availability and processing at the core data centers. Hence, there is a need to have most of the resources available at the edge of the network to ensure the smooth execution of end-user applications. Considering the challenges of future user demands, we present an architecture that integrates cloud and fog computing in the 5G environment that works in collaboration with the advanced technologies such as SDN and NFV with the NSC model. The NSC service model helps to automate the virtual resources by chaining in a series for fast computing in both computing technologies. The proposed architecture also supports data analytics and management with respect to device mobility. Moreover, we also compare the core and edge computing with respect to the type of hypervisors, virtualization, security, and node heterogeneity. By focusing on nodes' heterogeneity at the edge or core in the 5G environment, we also present security challenges and possible types of attacks on the data shared between different devices in the 5G environment.

INTRODUCTION

Cloud computing and the Internet of Things (IoT) have become popular with the exponential usage of smart devices in recent years. Cloud computing is a platform for data storage, analytics, visualization, and shared pools of resources located across the globe through which various services can be accessed from anywhere using the Internet. On the other hand, IoT provides connectivity to various smart devices that can be used for computation and storage. Each device (smart object) has its own unique IP address for communicating with the other devices. Compared to IoT, cloud computing has a centralized architecture in which various data service providers are used to reduce data warehousing costs while providing virtually unlimited storage space. In contrast, IoT is a distributed architecture and acts as a data receiver with limited storage capacity.

As per the CISCO report [1], by 2020, the IoT will be made up of nearly 50 billion devices connected to the Internet (from 500 million in 2003, 12.5 billion in 2010, and 25 billion in 2015). The massive amount of data generated from the IoT devices is stored at the cloud data centers (DCs), which exponentially increases the load on the network. Network congestion is a major challenge for processing large amounts of data from different geo-distributed database repositories. The other issues related to data processing at the DCs include slow data rates, low bandwidth, high end-to-end latency, high cost, fault tolerance, and security. Due to these challenges, real-time data analytics on large amounts of data becomes a challenging task. To meet the requirements of higher capacity and higher data rates for most real-time business applications, fifth generation (5G) technology has emerged.

To improve the performance of cloud DCs, a new infrastructure model called a cloudlet (cloud servers) has become popular. The cloudlet model makes the cloud DCs' capabilities accessible at the edge of the mobile network, also known as mobile edge computing (MEC) or fog computing (FC), which is considered as the future evolution of cloud computing. To address the above challenges, fifth generation (5G) technology works in collaboration with other promising technologies such as- software-defined networking (SDN), network functions virtualization (NFV), network service chaining (NSC), and massive mutiple-input multiple output (MIMO) technology [2] in order to provide high quality of service (QoS) to smart objects.

The NSC service model integrates SDN and NFV services in order to perform fast computation of services in the 5G network with the help of different types of network and communicating protocols. The types of network connectivity used by the NSC model includes cellular tech-

Digital Object Identifier:
10.1109/MCOM.2017.1700102

Rajat Chaudhary and Neeraj Kumar are with Thapar University; Sherali Zeadally is with the University of Kentucky.

0163-6804/17/$25.00 © 2017 IEEE

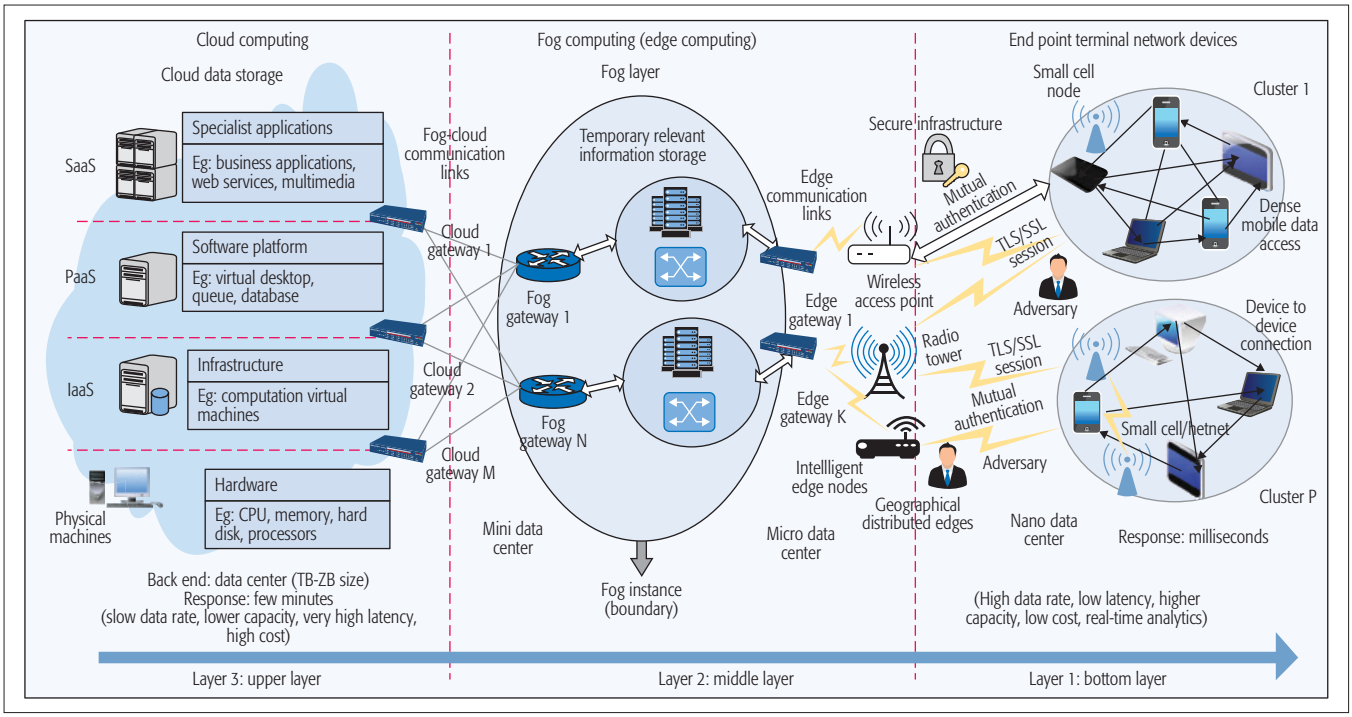


Figure 1. Detailed system architecture for integration of cloud, fog, and end point terminals.

nologies, Wi-Fi networks, and small cell base stations (e.g., femtocells, picocells, macrocells) [2] deployed to provide services to the smart objects. Moreover, the NSC model uses communication protocol such as: OpenFlow, routing protocol for low-power lossy-networks (RPL) for routing, Constrained Application Protocol (CoAP) for messaging, and transport layer security (TLS) for providing security [1] to enable the operations of smart objects.

The integration of the cloud computing model and the FC model with IoT in the 5G network environment opens up various security challenges. Risks and possible attacks, such as distributed denial of service (DDoS, at the nearby cloudlet instead of the remote DCs exist. The perpetrator (attacker) using mobile devices can make services unavailable by disrupting the cloudlet services in order to compromise the QoS of the edge devices. In a cloudlet mesh architecture, the authentication of mobile devices is mandatory in order to prevent DDoS attacks. In this context, Kerberos, a secure reliable authentication server [3], checks the authenticity of every mobile device and generates the ticket to access services from the cloudlet mesh. For every cloudlet at a distinct location, Kerberos is implemented in our proposed architecture to protect the cloudlet servers from possible DDoS attacks.

Figure 1 shows the proposed architecture of cloud, fog, and end terminal devices. We have divided the proposed system architecture into the following three layers, discussed below:

- **Bottom Layer (Layer 1)**. Endpoint terminal devices
- **Middle Layer (Layer 2)**. Fog/edge computing model
- **Upper Layer (Layer 3)**. Cloud computing model

The detailed description of these layers is provided in the coming sections.

Contributions of This Work

We summarize the main research contributions of this work as follows:

- We propose a system architecture that enables the integration of cloud and fog computing in context with NSC.
- In the context of the 5G environment, we present a unified NSC model in SDN and NFV architecture for increasing the response time on the cloud computing and FC model.
- We explore a security model for protecting the cloudlet mesh from DDoS attack by using a Kerberos authentication server.

The rest of the article is organized as follows. We describe cloud computing in the 5G environment. We describe virtualization and its needs. We discuss fog computing and its integration with cloud computing and IoT in the 5G environment. DDoS attacks and their impact on the proposed architecture is evaluated. Finally, the article is concluded.

Cloud Computing in the 5G Environment

The traditional cellular radio access network (RAN) architecture has limited spectral efficiency and causes frequency reuse interference issues among adjacent cells. The modern cloud RAN (CRAN) architecture used in 3G/4G uses technologies like dense wavelength-division multiplexing (DWDM), and millimeter-wave (mmWave) [4] for delivering high performance. Although CRAN uses the cloud computing capabilities for virtualizing the operations of base stations, it still has limited capacity and incurs long delay. In order to upgrade the overall performance, 5G technology uses heterogeneous networks (HetNets) that combine different RANs and distinct small cells to address the issues of capacity, coverage, and delay. Different CRANs use networks such

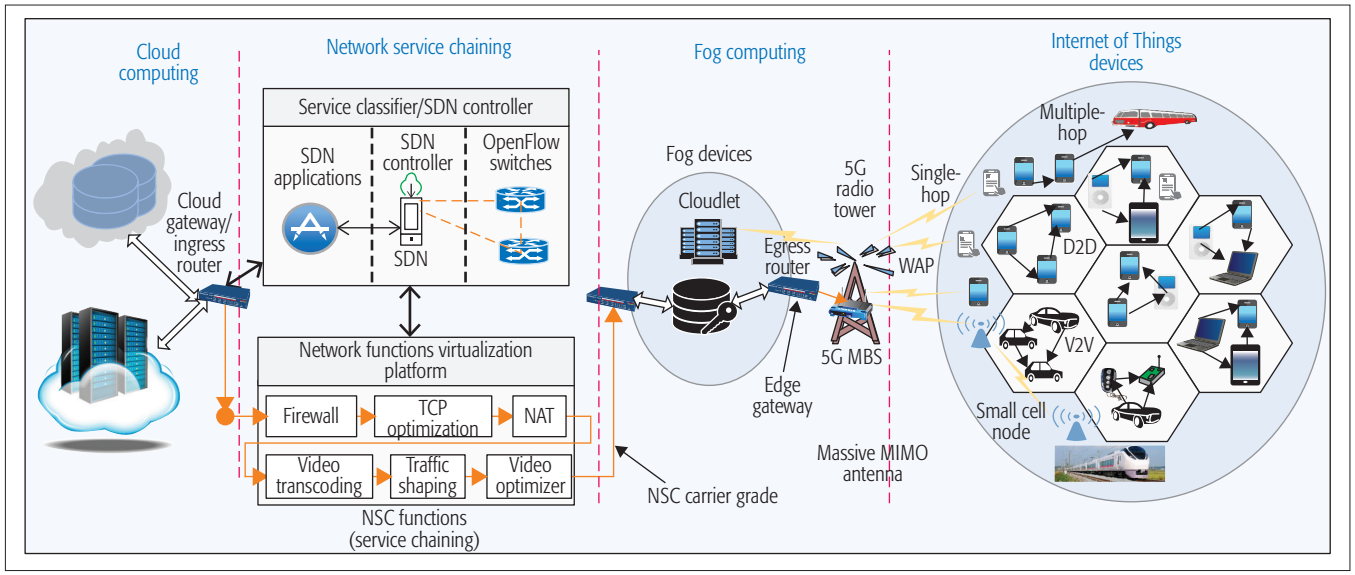


Figure 2. Network service chaining model in 5G wireless architecture.

as: High-Speed Packet Access (HSPA), Global System for Mobile Communication (GSM), enhanced data rates for GSM evolution (EDGE), code-division multiple access (CDMA), and Wi-Fi [4]. Small cells are low-power nodes (LPNs) with low cost and have a coverage range of femtocells for buildings, picocells and microcells for dense areas such as shopping malls, railway stations, and hospitals, and macrocells for large coverage areas used in vehicular networking and smart cities [2].

The small cell, CRAN, and HetNet architectures have challenges (related to inter-cell interference mitigation, high energy consumption, and low spectral efficiency) that affect data management in cloud computing. The concept of heterogeneous CRAN (H-CRAN) services with 5G architecture was introduced to overcome the challenges of CRAN and HetNets. H-CRAN uses orthogonal frequency-division multiple access (OFDMA) to improve the frequency and time domain variations. In [4], Peng *et al.* proposed a scheme known as soft fractional frequency reuse (S-FFR) for reusing frequency that is free from interference, which improves the energy and spectral efficiency in H-CRAN.

Cloud-Based Architecture

Cloud computing relies on DCs (groups of servers) to ensure control and management of resources by shifting data at a centralized location. The DCs troubleshoot servers locally and remotely with physical security authentication protocols such as Lightweight Directory Access Protocol (LDAP). Figure 1 shows the cloud computing model comprising services such as software as a service (SaaS), platform as a service (PaaS), and infrastructure as a service (IaaS). SaaS is used in software licensing, delivering web services, and business applications. SaaS examples include Google Docs, Google apps, Microsoft Office 365, Salesforce CRM [5], and so on. PaaS allows the application developers to develop their own applications online, without being required to set up and manage individual hardware and software layers. Examples of PaaS providers include Google appengine, Windows azure, and salesforce. IaaS incorporates the provision of physical assets in various forms such as physical machines (PMs), network devices, virtual machines (VMs), storage disks, and load balancers.

The underlying layer of IaaS is the physical hardware on which end users can install their application software and OS. Examples of IaaS are Google compute engine, HP Cloud, SQL Azure, and Amazon S3 [5]. In [6], John *et al.* proposed a solution to improve the network performance at the cloud to the edge devices by focusing on the NSC model in terms of carrier-grade infrastructure networks. The detailed description of the NSC model is discussed in the next section.

Network Service Chaining Model in 5G

NSC is a flexible service model used in SDN and NFV to automate virtual network devices instead of manual connections. NFV is the ability to perform network slicing that allows the virtualization of a physical network to create a logical network that consists of device instances. Some examples of middlebox NFV functions are intrusion detection and prevention system (IDS/IPS), firewalls, network address translation (NAT), video transcoding, TCP optimization, TCP proxy, traffic shaping, and load balancing [6].

SDN is a promising technology that helps to simplify network design and management. It relies on software-based programming rather than hardware-based, thereby enabling software reconfiguration for upgrading network policies. SDN consists of three planes: the data plane (OpenFlow switches), the control plane (SDN controller), and the application plane. All the network policies are programmed at the SDN controller, and these policies are reflected at both the application and data planes [2].

Figure 2 shows the NSC model in the 5G network where the cloud and FC are integrated to serve IoT devices. NSC performs data flow processing using the service chaining model connected by using SDN-based approaches. Data is transmitted from the core DCs with the help of the ingress router and gateway to the edge devices. For monitoring the global traffic, an SDN

controller works with the NFV platform. In Fig. 2, the dark orange line represents a continuous chain of network functions that combine a series of carrier-grade networks to automate the virtual functions of the network resources and to provide high QoS to the IoT devices using an egress router. The aim of the NSC model is to reduce the capital expenditure (CAPEX) and operational expenditure (OPEX), enable quick failure recovery, and simplify the installation/modification of new services at the SDN controller.

Moreover, by using the NSC model in 5G, data offloading can be done to achieve the best QoS for the end users. Data offloading is needed when there is a shortage of bandwidth, thereby distributing the load on nearby networks. One possible solution for offloading is switching from the cellular bands to either the Wi-Fi network, small cells, or delay-tolerant network (DTN) [2]. In addition to SDN and NFV technology along with the NSC model, the 5G network also works in collaboration with different wireless communication technologies namely, MIMO, and small cells, to accommodate high data rates. Massive MIMO involves multiple arrays of antennas merged into each base station in order to transmit high data streams simultaneously [2]. 5G technology provides the best services to IoT devices by direct vehicle-to-vehicle (V2V) and device-to-device (D2D) communications through single-hop and multihop paths. We discuss the virtualization concept in the NSC model below.

Virtualization

Virtualization is a technique that allows the abstraction of physical resources by creating their specimen and represents them as logical resources. It is put into effect at the computing, storage, network, and application levels. A physical switch consists of multiple PMs linked by a physical network interface card (pNIC). With the help of hypervisor software or the virtual machine monitor (VMM), multiple VMs (guests) reside on the same PM. At the front-end, a virtual switch (vswitch) is connected to the physical switch, whereas at the back-end, a vswitch is connected to all the vNICs. To provide VM connectivity, each VM is assigned an IP address over the virtual NIC (vNIC). The virtual Ethernet port aggregator (VEPA) [7] allows the switching among the VMs. Once a virtualized system is initiated, the hypervisor states are then transferred to the memory. The underlying hardware (e.g., AMD Processor and Intel VMX) assists the transition between the VMs and hypervisor. After a VM exits, the CPU is loaded with the execution context of the hypervisor, which in turn operates on the data residing in memory.

Kim *et al.* [8], explored different types of existing hypervisors, such as Citrix Xenserver, VMware ESXi, KVM, and Hper-V hypervisor. The authors proposed a scheme known as VM placement for addressing the issues of access latency recommended for each VM running on the non-uniform memory access (NUMA) system. Hypervisors are broadly classified into various categories [7, 8]:

- **Para-Virtualization Hypervisor:** It helps to modify the guest OS and is used by the Citrix XenServer hypervisor.
- **Fully Virtualized Hypervisor:** It installs all the hardware drivers and software for detecting malicious instructions that attempt to update the hardware. VMware ESXi is a variant of the fully virtualized hypervisor.
- **Hybrid Model Hypervisor:** One of the popular variants of this category is the kernel VM (KVM), which enables the virtualization capabilities for the guest processing and input/output (I/O) scheduling.
- **Micro-Kernelized Design Hypervisor:** It executes on micro-kernel design architecture and is independent of the device drivers; for example, the Hyper-V hypervisor.

The hypervisor models can be broadly classified into two categories, type 1 and type 2. Type 1 operates on the system hardware, whereas type 2 operates on the host OS to provide virtualization services such as: memory management, CPU scheduling, and I/O device support. Table 1 summarizes the features of all the types of hypervisor software.

Fog Computing

To get fast response time with respect to resource depletion, offloading of computation services to the nearby cloud resources is required. In this context, the cloudlet/MEC is a prototype model for offloading the remote cloud services to the edge of the network to serve the nearby mobile devices (e.g., smartphones, tablets, wearable smart devices). The cloudlet model operates at the LAN level via a Wi-Fi or mobile network, and it addresses issues such as: battery consumption, latency, and cost incurred when executing applications on mobile devices. The mobile users are served by the FC model by offloading the computation services in a single-hop wireless access, thereby providing a fast response time as depicted in Fig. 2. The cloudlet services work in smartphone applications, such as facial and speech recognition, GPS navigation, and healthcare. Chen *et al.* [10] presented a trust model to protect data privacy and content sharing, and proposed an IDS based on secure cloudlet mesh for remote healthcare applications.

In [11], Sarkar *et al.* addressed the issues concerned with the integration of cloud computing and the FC model with respect to IoT devices. The author focused on the performance of the DCs that depletes high power and returns a massive amount of carbon dioxide (CO_2). The author showed that the FC helps to reduce CAPEX and OPEX up to 50.09 percent compared to cloud DCs. The FC architecture is a distributed approach in which devices have peer-to-peer (P2P) connectivity for data computation and storage. Figure 1 shows the FC model in a 5G network and how it operates when integrated with the cloud computing scenario.

Architecture of the Fog Computing Layer

FC is characterized by benefits such as- interoperability, mobility support, open communication, robust performance, autonomous security, agility, and relatively low latency of a few milliseconds [11]. With FC, the distributed computing infrastructure where the user services are hosted by the network edge devices such as gateways, access points, routers, and intelligent switches

To get fast response time, with respect to resource depletion, offloading of computation services to the nearby cloud resources is required. In this context, the Cloudlet/MEC is a prototype model for offloading the remote cloud services to the edge of the network to serve the nearby mobile devices (such as smartphones, tablets, wearable smart devices).

Hypervisor Models	XenServer	VMWare ESXi	KVM	Hyper-V
Developer	Linux	VMWare	OVA	Microsoft
Hypervisor type	1	1	2	1
Physical memory	1 TB	2 TB	2 TB	4 TB
Logical processors	64	160	160	320
Nodes per cluster	16	32	No cluster support	64
Maximum VMs	800-960	3000-4000	–	8000
Monitoring protocol	HTTPS	SOAP	HTTPS	WMA
IPV6 support	No	No	Yes	No
Host OS	Linux, Solaris	bare Metal	Linux	Windows Server
Guest OS	Windows, Linux	Linux Kernel	Windows, Linux	Windows, Linux
N/W virtualization	No	Third party	Yes	Yes
Cold migration	Yes	Yes	Yes	Yes
Risk of vulnerabilities	High	High	Low	Medium
Packet sniffing testing tools	Wireshark, dsniff	Wireshark, NetworkMiner	Wireshark, tcpdump	Wireshark, mysql-sniffer
Intrusion detection tools	Snort, WirelessIDS	Snort	Snort	Snort
Man-in-the-middle attack tools	Ettercap, Cain e Abel	Ettercap, PacketCreator	Dsniff, Cain e Abel	Armitage, Ettercap

Table 1. Characteristics of hypervisor software models.

using which the data is transmitted, instead of accessing it from the remote DCs. The fog devices are intelligent network devices making smart decisions and store mini DCs to provide computation and routing functions.

Figure 1 shows the fog layer, consisting of fog instance (FI) and fog nodes (FNs). It is the fog boundary in which only the relevant information is mined and stored temporarily. With the help of a communication link, packets are transferred from the cloud end gateways to fog gateways and vice versa. FNs enforce local processing, computation, and storage, and perform analytics on the data. FC architecture is subdivided into two different components,

- *Fog abstraction nodes* are closer to the cloud gateway for providing analytics, visualization, and privacy.
- The *fog orchestration layer* consists of a "foglet," which is closest to IoT devices to take input requests from end users. This layer takes care of the fault tolerance, resource management, and security with respect to the service deployment model at the edge [11].

Fog Computing Based on 5G Wireless Network

According to [9], FC-RAN is the most popular paradigm in 5G wireless networks for channel assignments, energy, and spectrum efficiency. The advantages of a 5G network over edge devices are described as follows:

- 5G supports IoT devices 100 times faster compared to a 4G/LTE network.
- In 5G, the user data rate is approximately around 10–32 Gb/s compared to 4G having 100–150 Mb/s, reaching a peak terminal data rate increase of almost 30 times.
- In 4G, high latency is incurred in cloud computing, but using FC in association with 5G network, the end-to-end (E2E) latency is reduced by almost 10 times.
- In 5G, the battery power life is increased by 10 times in D2D communication.
- The bandwidth is higher (around 60 GHz) in 5G compared to 4G with low bandwidth (10 MHz).

Table 2 shows the comparative analysis of the cloud computing and FC model along with the existing proposals. The integration of cloud computing and IoT devices is called the CloudIoT model, and security remains a significant challenge in this model.

Security Model at the Cloudlet Mesh

In 5G networks, an E2E security framework is necessary in the business model. Although the response time in 5G wireless networks is increased by locating the cloud nearby the mobile devices, the cloudlet mesh architecture is vulnerable to attacks. In [12], Modi *et al.* discussed various possible attacks at various layers for cloud computing. An adversary can flood a single server with multiple requests (i.e., DDoS attack), making it unable to handle valid requests, as shown in Fig. 3. Here, the two attackers issue multiple service requests simultaneously via a wireless network in a single hop. The attackers disrupt the cloudlet services in order to compromise the QoS available to the mobile devices. The remote cloud DCs and SDN controller are connected to each other via the Internet. The SDN controller is used for monitoring and controlling the global network traffic. Data flows through the OpenFlow switches to the edge gateways, which forward the data to

the cloudlet. Cloudlets are interconnected with each other in a P2P mode through the OF-switches and make up the cloudlet mesh architecture. The cloudlet is located in distinct geographical locations (say X, Y, and Z). Cloudlet mesh locations Y and Z are unsecured because in the communication session, an attacker has launched a DDoS attack.

To defend the cloudlet mesh, various security protocols are used, such as- multi-party authentication protocol (MAP), inter-cloudlet protocol (ICP), trusted cloud transfer protocol (TCTP), and secure socket layer (SSL) [13]. The MAP supports mobility management, content filtering and multiway authentication between the cloudlet and the mobile device. The ICP supports IDS and load balancing operations within the cloudlet mesh. The TCTP performs the encryption of the file before data gets uploaded to the public cloud. Moreover, the authentication of services at the cloudlet mesh is necessary in order to verify the authenticity of the genuine mobile user served by the cloudlet. The solution for providing the authentication of services at the cloudlet server is the Kerberos mechanism, which is discussed below.

Kerberos Authentication Server

Kerberos is a secure reliable authentication server [3] for providing authentication service quickly at the cloudlet mesh for every incoming request from mobile devices. For mutual authentication between the mobile device and the cloudlet server (CS), the mobile device performs authentication first by the Kerberos servers, such as the authentication server (AS) and service granting server (SG). It uses SSL/TLS protocol for secure communication between the communication parties. Figure 3 shows the communications between the Kerberos server and the mobile device user '*k*'.

The steps involved in the communication session are as follows:

- The mobile device (MD_k) first initializes the login session and sends the request to the AS. The (MD_k) sends his/her attributes such as (ID_{MDk}, OTP_{MDk}, IP_{MDk}, ID_{SG}, $tstamp_1$) all appended in (P_{MDk}). These attributes are the identity, one-time password (OTP), network IP address, identity of the SG, and timestamp to synchronize his/her clock with the AS.
- The AS verifies the (MD_k) information in the database. If the information matches successfully, the AS computes ($ticket_{SG}$) as a ticket for the SG. The ticket contains the (ID_{MDk}) along with the timestamp and the ticket lifetime (lt_1), which are all encrypted with the key (K_{SG}) which is known only to the AS and SG.
- After receiving the ticket from the AS, (MD_k) sends the ($ticket_{SG}$) along with ($tstamp_3$) and (ID_{CS}) as an identity of the cloudlet server (CS) to the SG. The third timestamp is used to inform the SG of the time at which the authentication was performed by the AS.
- The SG authenticates the ticket, and if the authentication is successful, it then creates ($ticket_{CS}$) as a ticket for the CS. The ticket contains the (MD_k) attributes along with the ($tstamp_4$), (lt_2), (ID_{CS}) all encrypted with the (K_{CS}) key to prevent tampering; the key is known only to the SG and CS.

(a) Comparative analysis of cloud and edge computing		
Requirements	**Cloud Computing**	**Fog Computing**
Latency	High	Low
Delay jitter	High	Very low
Distance between client and server	Multiple hops	Single hop
Location awareness	No	Yes
Location of servers	Within Internet	Edge nodes
Type of connectivity	Leased line	Wireless
Geographical distribution	Centralized	Distributed
Response time	Minutes	Milliseconds, sub-seconds
Number of server nodes	Few	Very large
Security	More secure	Less secure
N/W bandwidth	High	Less
Risks of man-in-the-middle attack	Less	High
Interference mitigation	No	Yes
Service type	Global information	Localized information services
Target user	Internet users	Mobile users

(b) Comparative analysis of existing proposals					
Proposals	**Peng *et al.* [9]**	**Aujla *et al.* [2]**	**Kim *et al.* [8]**	**Chen *et al.* [10]**	**Sarkar *et al.* [11]**
Technique	S-FRR	Stackelberg Game	VM Placement	Number Theory	Mathematical Model
Latency	Yes	Yes	Yes	Yes	Yes
Energy	Yes	Yes	No	No	Yes
Spectrum	Yes	Yes	No	No	No
Data offloading	No	Yes	Yes	No	No
Security	No	No	No	Yes	No

Table 2. Comparative analysis.

- Finally, the (MD_k) sends his/her attributes along with the ticket ($ticket_{CS}$) to the CS.
- The CS checks the ticket ($ticket_{CS}$) with the authenticator, and if the ticket matches, the mobile device is granted access to the CS services.

Hence, instead of implementing as a centralized Kerberos for linking all the cloudlet web servers, a distributed Kerberos architecture is designed for providing mutual authentication at every cloudlet web server location. The DDoS is the most common attack that occurs at the cloudlet web servers. The attacker floods the web server with multiple requests like http or https over TLS to deplete server resources. The high risks of various network layer DDoS attack types such as: volume-based attacks (UDP flooding, ICMP flooding) and protocol attacks (SYN flooding, ping of death) may be launched by an attacker. Flooding consumes both incoming and

5G technology works in both of the computing models by integrating advanced technologies such as SDN and NFV. Moreover, the NSC model is integrated into the SDN architecture for fast chaining of the network virtualized services in order to deliver high QoS performance to the IoT devices.

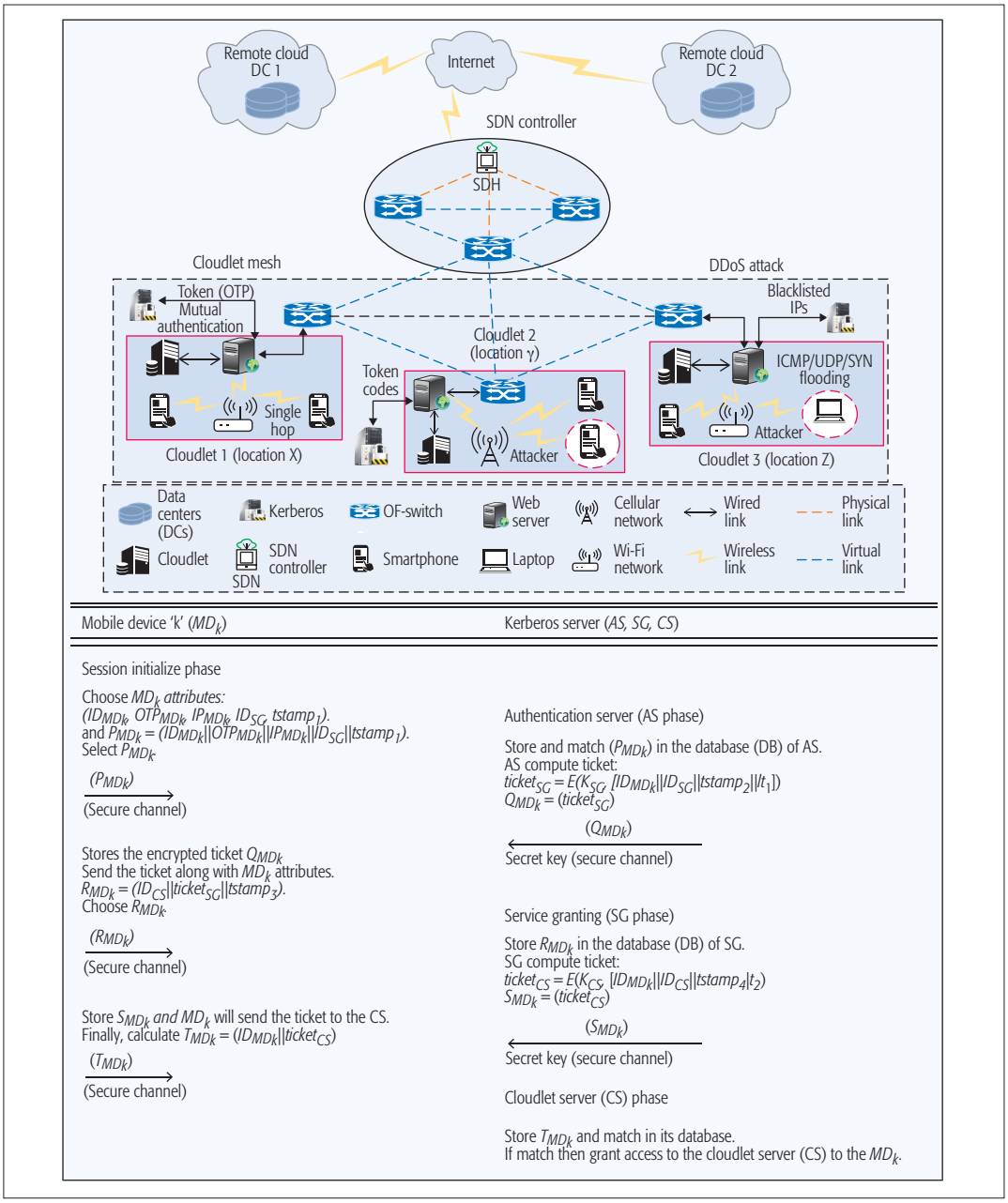


Figure 3. Mobile devices authentication at cloudlet mesh using Kerberos as a countermeasure for DDoS attack.

outgoing bandwidth. To protect from unauthorized and malicious traffic, a Kerberos server is placed as an external entity that uses hardware tokens authentication. It generates token codes or one-time passwords (OTPs) over a fixed time interval with a timestamp. Figure 3 shows how Kerberos authenticates the client's credentials based on the token codes. In the case of flooding using IP spoofing, Kerberos filters out those IP address and considers them as blacklisted IP addresses that belong to spam resources. These blacklisted IP addresses remain blocked to prevent sending requests and consuming network bandwidth.

Therefore, the Kerberos authenticate mechanism helps in mitigating the DDoS attacks against cloudlet web servers, but still Kerberos has some limitations. In Kerberos version 5, two technical deficiencies are noted in performing the fast authentication service. First, double encryption is not required after the ticket for the targeted server gets encrypted. Hence, single encryption is sufficient for performing fast processing. Second, symmetric key encryption is performed for fast execution times. Keeping these constraints in mind, we investigate the performance analysis of the Kerberos-based system.

Performance Analysis

We compute the running time of various operations during the communication among the various parties. The list of various operations performed are listed below:

- T_h: Execution time of a general hash operation
- T_{ap}: Execution time of an append operation
- T_{BF}: Execution time of the blowfish algorithm
- T_{DES}: Execution time of the DES algorithm

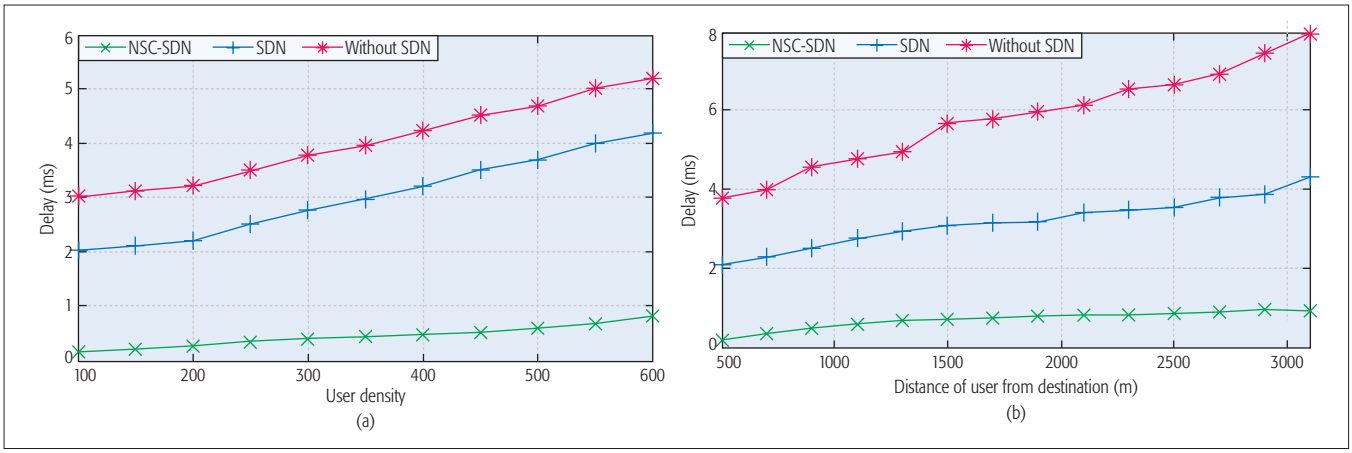


Figure 4. Results obtained with/without SDN and NSC-SDN approaches: a) delay vs. user density; b) delay vs. distance between user and destination.

Three main operations are used in the proposed Kerberos-based system. First, the double hashing operation is used by the AS for matching the client information in the database; hence, the two hash operations are used, and each T_h takes 0.007 ms [14] standard running time at the server. Second, the append operation is used to merge the attributes. In the proposal, 13 append operations are executed, and each T_{ap} takes 0.30 ms running time to execute on Python. Third, the symmetric key encryption is performed by using the blowfish or DES algorithm. The blowfish 448-bit key size algorithm incurs an execution time of 3976 ms, and the DES 56-bit key size algorithm takes 5998 ms for processing 256 MB of data [15]. The ticket for the targeted server is generated twice by using a symmetric encryption algorithm. Therefore, the computation cost of Kerberos using the blowfish algorithm is $(2 \times T_h + 13 \times T_{ap} + 2 \times T_{BF}) \approx 3979.914$ ms or 3.979 s. In the case of the DES algorithm, the computation cost is $(2 \times T_h + 13 \times \{T_{ap} + 2 \times T_{DES}) \approx 6001.91$ ms or 6.0 s. The computation cost shows that the blowfish algorithm is better for fast authentication.

Finally, the performance of the proposed NSC model with SDN is evaluated using lightweight simulations. The results obtained are shown in Fig. 4. We note that the proposed model incurs lower delay compared to SDN or without SDN with respect to user density and the distance of the user from the destination.

Conclusion

In this article, we have presented a relative comparison and analysis of fog and cloud computing with respect to network service chaining in the 5G environment. We have discussed the major issues of latency, cost, security, and data offloading at the core DCs or located at the edge in the virtualized environment. 5G technology works in both of the computing models by integrating advanced technologies such as SDN and NFV. Moreover, the NSC model is integrated into the SDN architecture for fast chaining of the network virtualized services in order to deliver high QoS performance to the IoT devices. We have also presented different types of hypervisors and their properties with respect to storage, operating system, and number of nodes. Finally, we have focused on the security attack at the cloudlet mesh architecture connected to remote cloud DCs. At the cloudlet mesh, the high risk of DDoS attacks is a major challenge. Hence, Kerberos is designed for authentication of services to protect secure communication with authorized mobile devices only.

In the future, we will explore how the SDN can be used in the 5G environment for accessing various resources. We will also explore more security features of SDN-based cloud infrastructures.

Acknowledgments

We are thankful to all the anonymous reviewers for their valuable suggestions, which improved the overall quality and presentation of the article. The work presented in this article is supported by the Council of Scientific and Industrial Research, New Delhi (no. 22/717/16/EMR-II).

References

[1] J. Frahim *et al.*, "Securing the Internet of Things: A Proposed Framework," Cisco White Paper, 2015.
[2] G. S. Aujla *et al*, "Data Offloading in 5G-Enabled Software-Defined Vehicular Networks: A Stackelberg Game-Based Approach," *IEEE Commun. Mag.*, vol. 55, no. 7, July 2017.
[3] W. Stallings, *Cryptography and Network Security: Principles and Practices*, Pearson Education India, 2006.
[4] M. Peng *et al*, "Energy-Efficient Resource Assignment and Power Allocation in Heterogeneous Cloud Radio Access Networks," *IEEE Trans. Vehic. Tech.*, vol. 64, no. 11, Nov. 2015, pp. 5275–87.
[5] D. Gonzales *et al*, "Cloud-trust — A Security Assessment Model for Infrastructure as a Service (Iaas) Clouds," *IEEE Trans. Cloud Computing*, vol. PP, no. 99, 2015, pp. 1–1.
[6] W. John *et al*, "Research Directions in Network Service Chaining," *2013 IEEE SDN for Future Networks and Services*, Nov. 2013, pp. 1–7.
[7] S. Varrette *et al*, "HPC Performance and Energy-Efficiency of xen, kvm and Vmware Hypervisors," *2013 25th Int'l. Symp. Computer Architecture and High Performance Computing*, Oct. 2013, pp. 89–96.
[8] C. Kim and K. H. Park, "Credit-Based Runtime Placement of Virtual Machines on a Single Numa System for QoS of Data Access Performance," *IEEE Trans. Computers*, vol. 64, no. 6, June 2015, pp. 1633–46.
[9] M. Peng *et al*, "Fog-Computing Based Radio Access Networks: Issues and Challenges," *IEEE Network*, vol. 30, no. 4, July 2016, pp. 46–53.
[10] M. Chen *et al*, "Privacy Protection and Intrusion Avoidance for Cloudletbased Medical Data Sharing," *IEEE Trans. Cloud Computing*, vol. PP, no. 99, 2016, pp. 1–1.
[11] S. Sarkar, S. Chatterjee, and S. Misra, "Assessment of the Suitability of Fog Computing in the Context of Internet of Things," *IEEE Trans. Cloud Computing*, vol. PP, no. 99, 201, pp. 1–15.
[12] C. Modi *et al*, "A Survey on Security Issues and Solutions at Different Layers of Cloud Computing," *J. Supercomputing*, vol. 63, no. 2, 2013, pp. 561–92.

[13] Y. Shi, S. Abhilash, and K. Hwang, "Cloudlet Mesh for Securing Mobile Clouds from Intrusions and Network Attacks," *2015 3rd IEEE Int'l. Conf. Mobile Cloud Computing, Services, and Engineering*, Mar. 2015, pp. 109–18.
[14] D. He *et al*, "Efficient and Anonymous Mobile User Authentication Protocol Using Self-Certified Public Key Cryptography for Multi-Server Architectures," *IEEE Trans. Info. Forensics and Security*, vol. 11, no. 9, Sept. 2016, pp. 2052–64.
[15] O. P. Verma *et al*, "Notice of Violation of IEEE Publication Principles Peformance Analysis of Data Encryption Algorithms," *2011 3rd Int'l. Conf. Electronics Computer Technology*, vol. 5, Apr. 2011, pp. 399–403.

Biographies

Rajat Chaudhary [S'17] (rajatlibran@gmail.com) is pursuing a Ph.D. from Thapar University, Patiala, Punjab, India. He received his B.Tech degree in computer science and engineering from UPTU, Lucknow, India, in 2010, and his M.Tech degree from UTU, Dehradun, India, in 2012. His research interests focus on software-defined networking, network functions virtualization, IIoTs, cloud computing, fog computing, and security.

Neeraj Kumar [M'16, SM'17] (neeraj.kumar@thapar.edu) is working as an associate professor in the Department of CSED, Thapar University. He received his Ph.D. from SMVD University, Katra (J&K) in computer science and engineering. He was a postdoctoral research fellow at Coventry University, United Kingdom. He has more than 150 research papers in leading journals and conferences. His research is supported by UGC, DST, CSIR, and TCS. He is an Associate Editor of *IJCS*, Wiley and *JNCA*, Elsevier.

Sherali Zeadally [SM'08] (szeadally@uky.edu) is an associate professor in the College of Communication and Information at the University of Kentucky. He received his doctoral degree in computer science from the University of Buckingham, England. His research interests include cybersecurity, privacy, the Internet of Things, and energy-efficient networking. He is a Fellow of the British Computer Society and the Institution of Engineering Technology, England.

CALL FOR PAPERS

IEEE COMMUNICATIONS MAGAZINE

EXPLORING CACHING, COMMUNICATIONS, COMPUTING AND SECURITY FOR THE EMERGING SMART INTERNET OF THINGS

BACKGROUND

It is estimated that more than 50 billion devices will be interconnected by 2020, and the generated data traffic will grow by another 1000 times. The evolution of Internet-of-things (IoT) paradigm not only comes with challenges to the research and industrial communities on finding efficient solutions for massive connectivity, but also poses a profound opportunity to reshape the current IoT architectures by exploring the inherent nature of the huge number of smart devices with caching, communications, computing and security capabilities.

Smart IoT cannot be simply regarded as an upgrade of current IoT by just adding to or replacing sensors/actuators/RFID tags in smart devices. It should be redesigned from the physical layer to the application layer in a bottoms-up way. In the context of smart IoT, communication is essential to guarantee end-to-end connectivity. Together with communications technology, other features, such as caching, computing and security should be fully utilized to complement the current study of IoT. The resulting new structure, by utilizing these inherent features, may have a wider application over current infrastructure-based cellular networks and traditional sensor-based networks by adopting all these features. However, new features also create unexpected problems which can not be directly addressed through the traditional approaches designed for low-rate IoT systems. Thus, how to effectively utilize existing capabilities to address the fundamental challenges in smart IoT remains challenging. Despite the evolution of IoT, some fundamental problems are still open and require immediate investigation.

This Feature Topic (FT) aims at providing timely and comprehensive overviews of the current state-of-the art in terms of fundamental theoretical innovations and technological advances towards exploiting smart caching, communications, computing and cybersecurity for smart IoT networks. Topics include, but are not limited to:

- Smart mobile computing, edge computing, cloud computing for IoT
- Smart content caching, pushing, and distribution for IoT
- Advances in technology and architecture for smart IoT
- Smart data collection, sensing, caching, processing, aggregation, communication, and analysis for IoT
- Software-defined solutions and network functions virtualization for IoT
- Efficient resource allocation, QoS, and QoE of IoT
- Novel authentication, access control and intrusion detection for smart IoT
- Novel cryptography algorithms for content privacy for IoT
- Experimental results, prototypes and testbed of IoT
- Energy efficiency and energy harvesting in IoT
- Emerging IoT applications in 5G networks such as ITS, CPS, smart grid, smart city, smart health, etc.
- Standardization of caching, communications, computing and security of IoT.

SUBMISSION GUIDELINES

Articles should be tutorial in nature, with the intended audience being all members of the communications technology community. They should be written in a tutorial style comprehensible to readers outside the specialty of the article. Mathematical equations should not be used (in justified cases up to three simple equations are allowed). Articles should not exceed 4500 words (from introduction through conclusions, excluding figures, tables and captions). Figures and tables should be limited to a combined total of six. The number of archival references is recommended not to exceed 15. In some rare cases, more mathematical equations, figures, and tables may be allowed, if well-justified. In general, however, mathematics should be avoided; instead, references to papers containing the relevant mathematics should be provided. Complete guidelines for preparation of the manuscripts are posted at http://www.comsoc.org/commag/paper-submission-guidelines. Please submit a PDF (preferred) or MSWORD formatted paper via Manuscript Central (http://mc.manuscriptcentral.com/commag-ieee). Register or log in, and go to Author Center. Follow the instructions there. Select "August 2018/Exploring Caching, Communications, Computing and Security for the Emerging Smart Internet of Things" as the Feature Topic category for your submission.

IMPORTANT DATES

- Manuscript Submissions Deadline: December 1, 2017
- Decision Notification: April 1, 2018
- Final Manuscripts Due: May 15, 2018
- Publication Date: August 2018

GUEST EDITORS

Jun Huang
Chongqing Univ, of Post and Telecommunications, China
xiaoniuadmin@gmail.com

Zheng Chang
University of Jyvaskyla, Finland
zheng.chang@jyu.fi

Chonggang Wang
InterDigital Communications, USA
gwang@ieee.org

Yi Qian
University of Nebraska-Lincoln, USA
yqian2@unl.edu h

Hamid Gharavi
NIST, USA
amid.gharavi@nist.gov

Zexian Li
Nokia, Finland
Zexian.li@nokia-bell-labs.com

Software Defined Network Service Chaining for OTT Service Providers in 5G Networks

Eftychia Datsika, Angelos Antonopoulos, Nizar Zorba, and Christos Verikoukis

The authors describe 5G network management architectures and propose virtualization components that enable OSP-oriented NSC. They also outline the arising issues for OSPs in NSC and introduce a distributed prioritization NSC management scheme for OTT application flows, based on matching theory.

ABSTRACT

The fifth generation wireless networks are expected to offer high capacity and accommodate numerous over-the-top applications, relying on users' Internet connectivity, thus involving different stakeholders, that is, network service providers and over-the-top service providers. For the efficient management of over-the-top application flows, the implementation of service functions and their interconnection in service chains, namely network service chaining, should consider the over-the-top service providers' performance goals and user management strategies. However, in current wireless network deployments, the network service providers have full control of network service chaining. Considering that user satisfaction from the offered services is a common interest for both types of stakeholders, the over-the-top service providers need to participate in network service chaining, and apply QoS and user prioritization policies to network service chain resource management, which involves users connected at different network points, in a distributed manner. In this article, we describe 5G network management architectures and propose virtualization components that enable over-the-top service-provider-oriented network service. We also outline the arising issues for over-the-top service providers in network service chaining and introduce a distributed prioritization network service chain management scheme for over-the-top application flows, based on matching theory. The evaluation results indicate the performance gains in over-the-top service providers' service levels that stem from the proposed scheme, demonstrating the benefits of introducing prioritization in network service chain deployment.

INTRODUCTION

Current research on wireless systems revolves around the accommodation of future networking demands of both end users and business enablers in the wireless market. Wireless traffic is expected to increase almost 10,000 times by 2030 compared to 2010.[1] Along with the growth of circulating mobile data, the development of fifth generation (5G) wireless networks is triggered by the appearance of novel Internet-based applications and business models, introducing multifaceted challenges in network operation.

Recent advances in wireless technology have created a plethora of over-the-top (OTT) services relying on broadband Internet service technologies (e.g., live video streaming, gaming) offered by OTT service providers (OSPs). OTT services have different quality of service (QoS) requirements, depending on their data traffic type; for example, video delivery requires high bandwidth, while VoIP needs low latency. These services encompass different user categories (e.g., free users or premium users paying for advanced QoS).

The users access the OTT content through various devices (e.g., smartphones or tablets), and their Internet access is often based on cellular connectivity. In this case, the users are also customers of mobile network operators, which own the cellular network infrastructure and spectrum, acting as network service providers (NSPs). Therefore, a user may access different OTT applications, whereas the users of a specific OTT application may belong to different NSPs.

The coexistence of multiple network services and OTT applications' QoS demands stresses the need for dynamic network configuration at the software level without modifying the network equipment. For this purpose, network service chaining (NSC) can be employed, which allows on-demand network services adjustment using service chains (SCs), that is, sets of interconnected software-based service functions (SFs) [1]. The SFs determine the way packets of flows are treated while circulating through network elements. An SC defines the set and sequence of SFs related to a flow, namely the actions applied to a flow. For example, it may refer to a policy with two SFs, one that enforces all HTTP traffic passing through a firewall and another that applies content filtering. In brief, NSC is the process of flow classification, flow forwarding to appropriate SFs, and instantiation of SCs that implement network services.

The development of modern OTT applications emphasizes the need for efficient NSC management in radio access networks (RANs). SCs are configured specifically per data connection type (e.g., Internet connection or connection for multimedia messaging services in cellular networks). Nonetheless, as the variety of network services increases, SCs should be deployed in a fine-grained manner (e.g., per user type), creating sophisticated SF combinations. A flexible and cost-effective solution is the networking paradigm of *cloud computing*, which allows the manage-

This work has been funded by the Research Projects AGAUR (2014-SGR-1551) and CellFive (TEC2014-60130-P).

[1] ITU-R, "Report M.2370-0, IMT traffic estimates for the years 2020 to 2030"; http://www.itu.int/pub/R-REP-M.2370-2015, July 2015, Accessed on: 2017-06-21.

Digital Object Identifier: 10.1109/MCOM.2017.1700108

Eftychia Datsika is with IQUADRAT Informatica S. L.; Angelos Antonopoulos and Christos Verikoukis are with the Telecommunications Technological Center of Catalonia (CTTC/CERCA); Nizar Zorba is with Qatar University.

0163-6804/17/$25.00 © 2017 IEEE

ment of full-fledged services in centralized data centers, forming cloud-based RANs (C-RANs) [2]. Offering a distributed alternative, *fog computing* is a variation of cloud computing that can improve the experienced QoS. It brings services closer to end users, allowing them to be hosted away from cloud data centers, at edge nodes. These nodes form a distributed networking structure that acts as an intermediate management unit between cloud and users, implementing fog-based RANs (F-RANs).

The resource and NSC management is feasible through direct programmability of network services using the software defined networking (SDN) and network functions virtualization (NFV), which allow the softwarization of network functions and virtualize the network infrastructure [3]. The SDN architecture offers to the stakeholders access to RAN software-defined controllers that implement functionalities of the control, data, and application planes [4]. NFV implements SFs as software programs running on servers [5]. Exploiting the SDN/NFV assets, the NSPs can deploy SCs over the RAN.

In 5G networks, stakeholders like NSPs and OSPs may coexist, with common interests in the provision of high-quality services and users' satisfaction. As the OTT applications' performance is intertwined with network connectivity service levels that affect the overall user experience, the OTT QoS is both NSPs' and OSPs' concern. However, with the current network architectures, the NSPs have total control of NSC; thus, the OSPs cannot supervise the OTT applications' key performance indicators (KPIs), such as grade of service, or fully manage their users, as the Internet connections are controlled by the NSPs. Even in cases where users with high priority should be accommodated first by the SFs, the OSPs are not able to apply their user prioritization policies. Therefore, 5G network architectures should allow the OSPs to intervene in the NSC customization in two ways:

1. Select the SFs that should be implemented.
2. Indicate the resources required for the SCs' implementation.

As multiple OTT applications might access the same network concurrently, centralized optimization methods that require the aggregation of all flow information become impractical, due to the high overhead of control data transmissions and poor adaptability to dynamic network conditions. Hence, distributed self-organizing approaches should be employed for the OSP-oriented NSC deployment.

Even though wireless network virtualization facilitates OTT applications' management through the exposure of network resources, several issues may arise for the OSPs regarding their participation in NSC.

Motivated by the lack of literature that studies the NSC deployment from the OSPs' viewpoint, in this article, our aim is threefold:

- We describe network management architectures and propose virtualization components that enable dynamic SC configuration by OSPs, exploiting SDN/NFV.
- Considering the characteristics of the OTT applications and the needs of the OSPs as industry verticals in 5G wireless networks, we investigate the challenges that arise in the NSC management process.
- We study the realization of flexible OTT-oriented NSC and propose a matching-theoretic NSC management algorithm for OTT application flows that enables the OSPs to make decisions regarding the user prioritization policy and dynamically select suitable resources. The performance evaluation demonstrates that the OTT applications' service levels are improved when the OSPs declare their preferences over the resource assignment.

OSPs' Requirements for the Deployment of Network Service Chains

Numerous OTT applications already exist (Skype, WhatsApp, etc.), which rely on Internet connectivity, often provided by the users' NSPs. Hence, the OTT flows circulate over different NSPs' network infrastructure. Furthermore, the OSPs face the dynamic nature of their services, as business decisions may require new SFs in order to capture the users' demands for new features. In this context, we next describe the OSPs' requirements regarding the NSC deployment.

OSPs' Access to NSPs' Networks

The OSPs should interact with the NSPs for the orchestration of NSC, monitoring their users' status (connection quality, location, subscription details, etc.) and combining this information for the construction of OTT flow profiles. Allowing the OSPs' intervention in NSC implies their access to NSPs' network resources, which can be financially advantageous for both parties. As their revenues seem to be correlated, achieving high OTT QoS can also be in the NSPs' best interests if the OSP-NSP cooperation is balanced through negotiation of agreements that regulate the degree of OSPs' intervention and sharing of gains [6].

The OSP-NSP interaction requires that the NSPs expose their service capabilities through properly designed application programming interfaces (APIs), as described in the service capability exposure function concept of the Third Generation Partnership Project.[2] As multiple NSPs coexist, the OSPs may be associated with multiple network slices, with different characteristics and services. Therefore, the OSPs should modify the SCs according to network service capabilities and available resources of the involved network slices.

Enabling the OSPs to develop SCs might entail preferential management of certain flows over the Internet. If OSPs apply flow prioritization in NSC, the NSPs' resources may not be shared fairly among OTT flows, creating concerns about network neutrality. Although prioritization policies are necessary in certain cases (e.g., for gaming applications with low latency requirements), NSPs' resources should be accessed in an impartial manner, without monopolizing their utilization by some OSPs only. Thus, OSP-oriented NSC should balance flow prioritization and fair access to NSPs' networks.

Adaptation to OTT Service Market Dynamics

The OSPs need to deploy services dynamically over static networks and compose business strategies, customizing the SCs. As the OSPs' revenue

Numerous OTT applications already exist (Skype, WhatsApp, etc.), which rely on Internet connectivity, often provided by the users' NSPs. Hence, the OTT flows circulate over different NSPs' network infrastructures. Furthermore, the OSPs face the dynamic nature of their services, as business decisions may require new SFs in order to capture the users' demands for new features.

[2] Third Generation Partnership Project, "Technical Specification Group Services and System Aspects; Architecture Enhancements for Service Capability Exposure (3GPP TR 23.708 version 13.0.0 Release 13)"; https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=869), June 2015, accessed June 21, 2017.

The NFV is a key technology for the customization of SCs and provides the essentials for virtualized management and organization. It enables the instantiation of virtual network functions, manages the NFV infrastructure resource requests, and offers complete services by combining the VNFs.

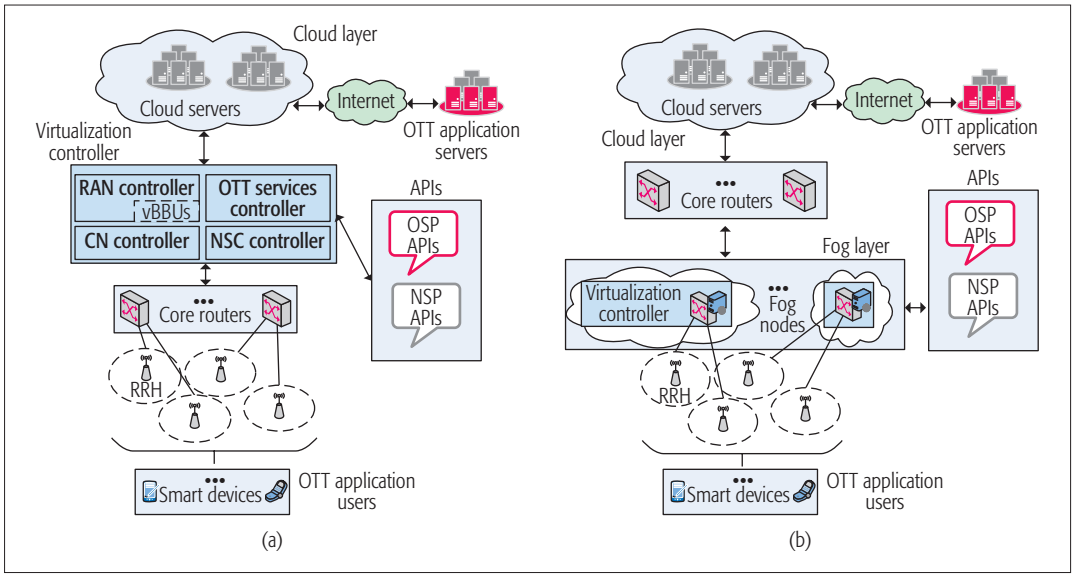


Figure 1. Network management architectures for OSPs: a) cloud-based architecture (C-RAN); b) fog-based architecture (F-RAN).

is highly dependent on the timely development of high-quality services, the flexibility in constructing SCs in real time is of crucial importance for the acceleration of OTT services' time to market.

The analysis of OTT application traffic is useful for the customization of SCs as a response to OTT service market dynamics, which may require the development of novel services or the addition of new features in OTT services. This update process might induce the addition of SFs in OSPs' SCs; for example, for the enrichment of an online gaming application, an SF that adapts graphics rendering to the capabilities of users' devices might be added. Essentially, dynamic NSC implies end-to-end network intelligence and adaptive network service composition [1]. These features can be achieved by examining the OTT users' behavior and the market trends using the flow information of properly designed SFs. A typical example of such SFs is the deep packet inspection (DPI) that provides the OSPs with network analytics needed to define user policies.

Network Management for OTT Service Delivery

The OTT application users are attached to RANs composed of heterogeneous network nodes (small cells, Wi-Fi access points, etc.) owned and managed by different NSPs, and are served by data centers with different capabilities and architectural designs. Hence, the 5G network design should facilitate NSC for OSPs, offering two fundamental functionalities [7]:

- Holistic network view: OSPs should be aware of NSPs' resources and networking capabilities in order to create the appropriate SCs using virtualization techniques.
- Support for network slicing for different OSPs: NSPs should be able to allocate resources to multiple OSPs, creating proper network slices.

OSP-Friendly Network Management Architectures

For SC development, information of OTT users, often scattered at different RAN connection points, should be aggregated. A well-known RAN-wide management technology is cloud computing networking [2]. A C-RAN consists of three main structural elements:

- Several access points (APs) with remote radio heads (RRHs)
- A virtual baseband unit (vBBU) pool, connected to the APs, that performs baseband operations
- Core routers that connect the RRHs with the cloud (Fig. 1a)

Even though the centralized approach is useful for NSC, locating the RAN management unit away from users leads to high latency and overhead. Alternatively, fog computing places cloud services close to the network edge [2]. An F-RAN is a distributed system that controls a set of RRHs through fog nodes (FNs), network devices as local servers with storage and computing capabilities (Fig. 1b). FNs bring the network management operations closer to end users, are connected with switches or routers at the RAN edge, and communicate with the vBBU pool. Each FN is a small data center implementing SFs for users connected to the APs it manages. The cloud data center acts as a global administration point.

NSC Virtualization Frameworks for OSPs

In the aforementioned architectures, SFs are applied in different network nodes with different order and features. This procedure requires the programmability of network functions and the abstraction of RAN resources for network slicing. For this purpose, the NFV and SDN frameworks can be incorporated in the network management architectures and provide virtualization components [4] (Fig. 2).

NFV is a key technology for the customization of SCs and provides the essentials for virtualized management and organization (MANO). It enables the instantiation of virtual network functions (VNFs), manages the NFV infrastructure (NFVI) resource requests, and offers complete services by combining the VNFs. The VNF man-

agers initialize and configure VNF instances and their interconnections with the NFVI that contains virtual computing resources (CPU, memory, etc.), storage resources, and virtual machines. The virtualized infrastructure manager (VIM) controls the underlying resources of NFVI, allocating them appropriately to VNF instances.

For SFs' management, SDN offers the capability of VNF orchestration using various components. In the considered architectures, a virtualization controller consists of four types of controllers:

- The RAN controller, which orchestrates the RRHs, allocates the spectrum resource blocks (RBs), and performs flow scheduling at each RRH
- The core network (CN) controller, which manages the gateways
- The NSC controller, which stores the information for NSC deployment and coordinates the VNFs
- The OTT services controller, which is used by the OSPs for OTT service surveillance and submission of NSC requests

The OTT services controller communicates with the VNFs and provides an overview of the implemented SFs (VNF instances), allowing the OSPs to assess the NSC performance and decide on the SCs by submitting NSC requests using the OSP APIs.

Open Issues in NSC Deployment for OSPs

The SCs are ordered sequences of SFs combined in order to process application flows, which are forwarded to the APs. For the OSPs, the NSC should be performed according to the flows' QoS requirements, the user information, and the OSPs' preferences regarding the KPIs. Although the current network management architectures and virtualization frameworks are useful for NSC, several issues arise for the OSPs, which are outlined in this section (Table 1).

Assessing the OTT Users' Requirements

The OTT applications access RANs via the users' Internet connections. OTT flows have different network and application related user characteristics, which influence the experienced QoS. Thus, it is important for OSPs to evaluate the NSC requirements and appropriately coordinate the SCs related to a heterogeneous set of users over multiple network slices.

As flows are related to different users, the OSPs should obtain information regarding users' location and downlink channel conditions. Even if the OSPs' KPIs characterize the OTT applications, the users' specific context may affect the SC construction. For example, if the users of a video streaming application experience poor downlink channel conditions, it might be unreasonable to use an SF for video optimization. Moreover, the OSPs might serve users associated with different network slices, APs, and data centers with different capabilities possibly owned by different NSPs [1]. The NSPs' resources made available to users may impose bounds on NSC efficiency, for example, a small FN at the network edge may not support advanced flow processing for all users.

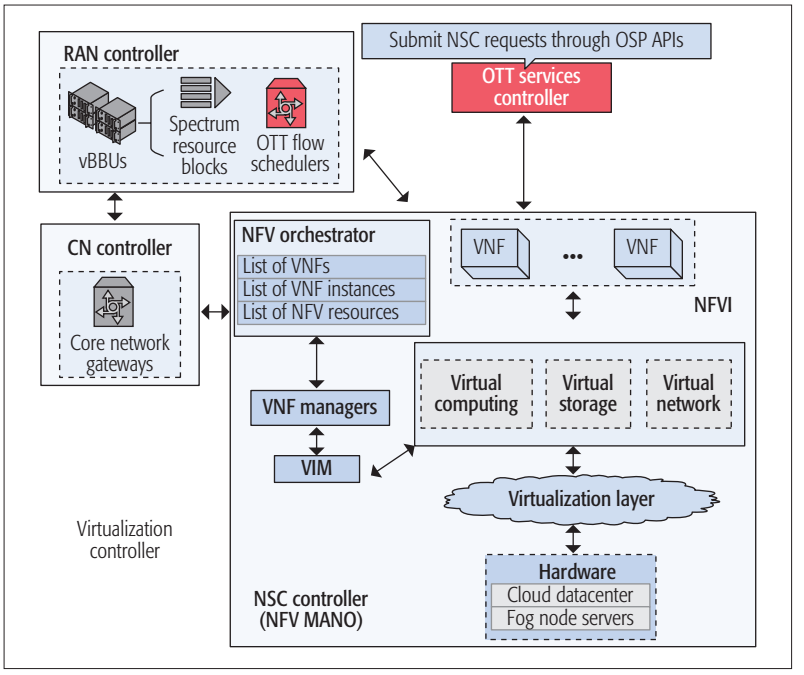


Figure 2. SDN/NFV framework for OSPs' NSC.

Mapping the OSPs' SCs to the NSPs' Resources

In SDN/NFV enabled networks, the composition of SCs implies the selection of network services that will be implemented as VNF instances, the orchestration of VNFs on a server or cluster of servers, the establishment of proper traffic routing paths among the VNF instances, and the allocation of resources to VNF instances (SC embedding problem). The resources are computational and storage resources of hardware and virtual machines or infrastructure and spectrum elements corresponding to different virtual networks, managed by NSPs' data centers.

Mapping the SCs to resources is a complex process that matches various types of physical resources with the VNF instances related to the SFs. For instance, virtualized C-RAN resources (e.g., RRHs and fiber links) are assigned to different tenants using auction mechanisms [8]. Hybrid NFV-based networks that incorporate network functions provided by dedicated physical hardware and virtualized instances may also exist [9]. Spectrum resource blocks can be allocated to APs associated with the VNF instances that serve the users connected to them (e.g., using VNFs that schedule the wireless resources of network slices in RAN nodes) [10].

The SC embedding problem is multi-fold, since decisions for NFV placement affect resource allocation and vice versa [11]. As NSPs manage the network resources and are concerned about overall VNF operational cost, VNF instances can be deployed in such a way that VNF host selection cost, traffic forwarding cost, and energy consumption are minimized [5]. The VNF orchestration might imply a trade-off between the latency induced by the VNF placement and chaining in servers and switches and the efficiency of resource utilization [12]. Suitable VNF locations and flow routing paths can also be defined according to capacity constraints of virtual machines that host the VNFs and of the links among them, optimizing

Reference	NSC issue	Methodology	Provision for OSPs
[5]	Optimization of network operational cost and resources (servers, links) utilization (optimal number and location of VNFs)	Viterbi algorithm applied in multi-staged directed graph that models sequence of VNFs	No, but considers service level agreements
[9]	Resource allocation to SCs	Selection of SFs, their network location, and their interconnections using integer linear programming (ILP)	No, but supports multi-tenancy
[10]	Selection of optimal VNF placement considering availability of radio resources	ILP-based algorithm and heuristics that maps nodes and links of SC requests to substrate network	No, but supports multi-tenancy
[11]	Joint optimization of VNF forwarding graph embedding and VNF scheduling	Heuristic-based algorithm for traffic scheduling between adjacent VNF instances and mixed ILP algorithm for selection of paths between NFV nodes	No
[12]	Optimal deployment of VNF forwarding graphs	Eigen decomposition and Hungarian method used to derive optimal matching of VNF graphs to infrastructure or NFVI	No, but supports multi-tenancy
[13]	Allocation of link capacity and virtual machines to SCs in order to maximize the number of served requests	SC deployment algorithm that selects routing path length and decides on use of additional or existing servers' resources	No
[14]	NFV placement and routing path selection considering network security defense patterns	Heuristic-based algorithm for security function placement in network partitions	No, but supports multi-tenancy

Table 1. NSC issues and existing solutions.

the amount of these resources that are allocated to SFs [13].

Existing works arrange the VNFs aiming to optimize aspects. Nonetheless, OSPs should participate in NFV orchestration by expressing their preferences over SFs. Users of the same OSP may be connected to different APs and served by different data centers (e.g., in F-RANs). Different FNs may serve users; thus, different VNFs are required in each FN, considering the OSPS' flow management policies.

Moreover, the OSPs usually have their own policies regarding service differentiation, which provide the rules that deem the flows to be of higher or lower importance. The flows have different characteristics, and different user priorities exist. This prioritization should be depicted in NSC, not only during NFV placement, but also in resources allocated to SCs, as VNF instances related to higher-priority flows should be arranged first. For example, in an F-RAN accessed by flows with multiple priorities, the SC embedding involves not only the prioritized arrangement of VNF instances in FNs, but also the prioritization of spectrum allocation to users connected to APs.

Constructing Secure SCs

An important aspect of the NSC procedure is the deployment of safe SFs considering different security standards. Selecting the location and order of SFs based solely on the SC performance estimation does not always lead to secure NSC. Particularly in RANs accessed by various OTT applications, the instantiation of SCs in a secure manner is not trivial, as security constraints of different OSPs have to be imposed on NSC. A recent work proposes the use of network security patterns in order to capture the network security constraints in a C-RAN [14]. Still, the OSPs need to devise their own security policies that may change according to their users' demands or OTT application characteristics. These policies should be "translated" to SFs organized jointly for all OSPs accessing a RAN, in a way that all OTT services' security needs are met.

Flexible NSC for OSPs

The implementation of SCs should match the OTT applications' particularities and OSPs' policies. Considering the need for distributed control of prioritization in NSC over networks accessed by multiple OSPs, we propose an NSC management algorithm that allows the OSPs to define their policies and declare their preferences over SFs and resources in a distributed manner, based on matching theory. Moreover, we examine the effects of flow prioritization in NSC by assessing the performance of the proposed algorithm.

OTT Flow Prioritization Using Matching Theory

OTT users may be connected to different network points, and prioritization has to be applied in various abstraction levels (i.e., VNF instantiation and RAN resource allocation). For the composition of SCs, the OSPs need information related to:

- The availability of network resources (e.g., spectrum RBs) and SFs provided by the NSPs (e.g., DPI)
- The OTT application flows' characteristics

These characteristics include required data rates, user subscription status, content type, and so on, which are known to the OSPs. Flows' characteristics referring to the OTT users' cellular connections, that is, parameters related to downlink channel conditions (e.g., supported modulation and coding schemes), are provided by the NSPs, along with the information about network resources and functions, and can be accessed through the OSP APIs.

Introducing the concept of matching theory in NSC enforces the role of OSPs in OTT flow management. NSC management employs the notion of a matching game, which models the interactions between NSPs and OSPs [15]. The OSPs create ranked lists of preferences over virtual resources according to flows' requirements. Each list item is a virtual resource request (VR), that is, a combination of parameters related to each flow i (id_i, AP_i, RB_i, $priority_i$, $chain_i$), where AP_i is the AP

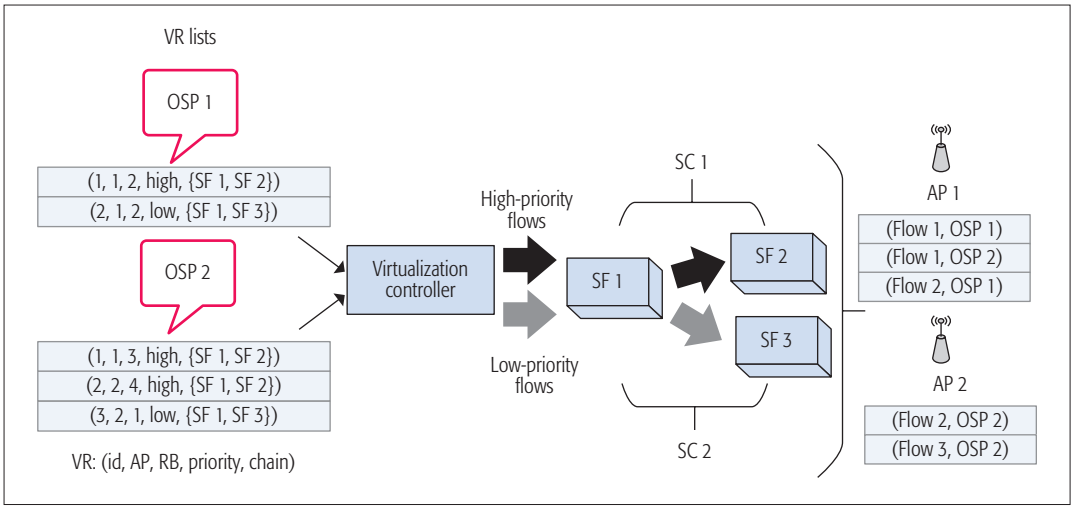


Figure 3. NSC example using the matching-theoretic OTT flow prioritization algorithm.

Considering the need for distributed control of prioritization in NSC over networks accessed by multiple OSPs, we propose an NSC management algorithm that allows the OSPs to define their policies and declare their preferences over SFs and resources in a distributed manner, based on matching theory.

of the user with flow i, RB_i is the number of spectrum RBs needed for the flow's QoS demand, in terms of data rate, latency or other metric, when the user is connected in AP_i, $priority_i$ is the flow's priority level, and $chain_i$ is an ordered set of SFs that declares the chain of VNFs required for the specific flow's processing. Multiple VRs may be related to the same flow, but at most one will be finally accommodated.

The VRs are added to the OSPs' preference lists according to flows' priority levels. The priorities are determined by each OSP according to the performance goals (e.g., maximize the number of flows that achieve the QoS demands). Therefore, the QoS metrics of each flow and KPIs of each OSP that affect the construction of the preference list might be different. Each flow may be associated with different SFs; thus, different SCs may be added in a VR; for example, an OSP can apply an SF, such as DPI only for premium users.

When new flows arrive, the OSPs decide about the necessary SFs and RBs and create preference lists, placing VRs for high-priority flows first (Fig. 3). The VRs of flows with the same priority are ordered by ascending number of RBs; for example, for OSP 2, flow i = 1 needs three RBs and is placed before flow i = 2. Subsequently, the matching process begins using the various components of the virtualization controller (Fig. 2). The VRs are submitted through the two OSPs' APIs to the OTT services controller, and the NSC controller initiates the matching process, handling the requests by priority. The RAN controller allocates the RBs in such a way that in both APs, the high-priority flows receive RBs first. The VNF manager organizes the VNF instances and their interconnections in the data center connected to the two APs. The requests for each VNF are aggregated, and suitable resources (CPU, memory) in NFVI are assigned by the VIM according to flows' priorities. For all flows, one VNF instance for each of the SFs (SF 1, SF 2, SF 3) is required. The high-priority flows access the SF 1 and SF 2 instances, whereas the low-priority flows pass through the SF 1 and SF 3 instances. Thus, two SCs are created, namely SC 1: {SF 1, SF 2} and SC 2: {SF 1, SF 3}. Once RBs and VNF instances are organized, SCs can be implemented.

Finally, a stable matching between flows and resources is reached, including allocations acceptable by both OSPs and NSPs. Each acceptable matching is individually rational and is not blocked by a flow-resource combination, making it the most preferable match for the flow [15].

Matching-Theoretic NSC Performance

We assess the performance of the OTT application flow prioritization matching algorithm (MAFP) against a best effort approach without prioritization (BE) and a fair allocation (FA) scheme that splits radio resources evenly among all OSPs in each AP. Considering the importance of user satisfaction, the OTT service levels are evaluated in terms of grade of flow accommodation (GFA), namely the percentage of flows that are not served with the requested QoS out of the flows of all OSPs. We also study the resource utilization levels, estimating the number of NFV instances required to serve the users.

Two scenarios with different users' minimum data rate demands are tested:

- A file downloading (FD) scenario with minimum acceptable data rate equal to 64 kb/s
- A video streaming (VS) scenario with data rate equal to 128 kb/s.

In the presented results, a 95 percent confidence interval is considered.

The FN implements two SC types, one that offers a full DPI service where all flows access the DPI SF, and one for a sampled DPI service, where a portion of flows access the DPI SF, as regulated by OSPs (Table 2). In our simulations, two OSPs serve users of either high or low priority. Half of the users belong to one OSP, and 60 percent of each OSP's users have high priority. High-priority flows access the full DPI service, and low-priority flows use the sampled DPI service.

In Fig. 4a, the GFA performance of all schemes is depicted. The increase of flows degrades the performance of both schemes, as fewer flows are served using the available RBs at each AP. Nonetheless, the MAFP achieves lower GFA than BE and FA, and enables the accommodation of more flows, reaching a reduction of 54–86 percent (FD scenario), and 50–77

Setting	Value
Network	F-RAN
APs	8
RBs per AP	50
Minimum required data rate	64 (FD), 128 (VS) kb/s
AP range	200 m
AP transmission power	33 dBm*
FN capabilities	10 CPU cores, 500 GB memory
Resources per VNF type (for 100 flow requests/s) [9]	Routing, firewall: 1 CPU core, 10 MB memory (each); DPI: 1 CPU core, 500 MB memory
OTT application flows' number	{100, 200, 300, 400, 500}
Full DPI service SC	{DPI SF}
Sampled DPI service SC	90% of flows: {routing SF, firewall SF} 10% of flows: {routing SF, DPI SF}

* Miao *et al.*, "Energy Efficient Design in Wireless OFDMA," *IEEE ICC*, 2008, pp. 3307–12.

Table 2. Simulation settings.

percent (VS scenario) for 300 and 200 users, respectively. As flows are sorted by number of required RBs and priority, more flows are served, and the high-priority flows are guaranteed to receive resources first. The performance gain is lower when higher data rate is required, as more RBs are needed.

Focusing on the VS scenario, Fig. 4b shows the VNF instances for the implementation of full and sampled DPI services. The GFA levels influence the NSP's resource utilization, as different numbers of VFNs must be instantiated. Three VNF instances, including one DPI VNF instance for 100 flows are needed for all schemes. In contrast, for 400 or more flows, five or more VNF instances are used (three DPI VNF instances). As the number of flows increases, more VNF instances are required. With BE and FA, more VNF instances are needed for more than 400 flows, as more flows of low priority are served and require routing and firewall VNFs.

Overall, prioritization in NSC affects the resource allocation in the APs and the VFN instantiation. It improves the OTT service levels, as the OSPs declare their preferences. In reality, though, the NSPs may supervise the resource utilization, ensuring the application of cooperation terms and fairness among OSPs. Still, using matching theory, the OSPs can express their preferences over resources and SFs. Therefore, the available resources are no longer allocated in a best effort manner, but the OTT application flows' requirements are considered in the NSC. Last, the OSPs can request the prioritization of certain flows according to the required QoS, and their KPIs and flows with higher priority are guaranteed to receive resources first.

Conclusions

In this article, we have described virtualization components for 5G network architectures that serve OTT application users and the challenges that arise in NSC deployment for OSPs. The network elements and their resource availability are different from one NSP to another. Furthermore, content and user types, QoS levels, or KPIs change with the evolution of OTT applications and OSPs' business decisions. Thus, the successful deployment of OTT applications requires flexible adaptation of network services, according to OSPs' flow management and prioritization strategies. Considering this context, we have presented a matching-theoretic OTT flow prioritization algorithm for NSC, which improves OTT applications' service levels, achieving more efficient resource management. The performance evaluation results can provide valuable insights for OSPs in the 5G wireless market.

We should note that the NSC deployment creates various practical challenges for both NSPs and OSPs. First of all, the NSC deployment implies the exposure of the NSPs' network resources and SFs. As the resources provided to OSPs are affected by both the network capabilities (feasibility of exposure) and the NSPs' business goals (expected profit from exposure), the decision about the network exposure levels requires joint consideration of financial parameters, which is an open research issue of NSC resource management. Moreover, the increasingly complex OSPs' requirements should be

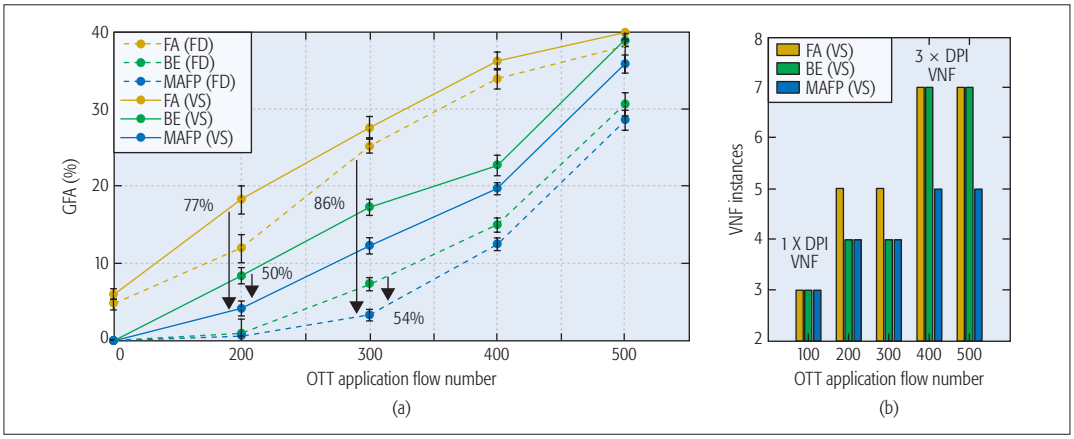


Figure 4. Performance with and without flow prioritization: a) grade of flow accommodation (%); b) VNF instances for VS scenario.

efficiently mapped in SFs, stressing the need for sophisticated and self-organizing solutions that customize the SCs properly. Still, this mapping process should not compromise the security of SCs. Ensuring that the security levels requested by OSPs and NSPs are maintained in NSC deployment can be an issue with high technical complexity. To this end, we believe that our study has shed some light on the OSPs' requirements and can motivate further investigation of NSC for OTT applications.

References

[1] F. Paganelli, M. Ulema, and B. Martini, "Context-Aware Service Composition and Delivery in NGSONs over SDN," *IEEE Commun. Mag.*, vol. 52, no. 8, Aug. 2014, pp. 97–105.

[2] C. X Mavromoustakis, G. Mastorakis, and C. Dobre, "Advances in Mobile Cloud Computing and Big Data in the 5G Era," *Studies in Big Data*, vol. 22, 2016.

[3] A. M. Medhat *et al.*, "Service Function Chaining in Next Generation Networks: State of the Art and Research Challenges," *IEEE Commun. Mag.*, vol. 55, no. 2, Feb. 2017, pp. 216–23.

[4] V.-G. Nguyen, T.-X. Do, and Y. Kim, "SDN and Virtualization-Based LTE Mobile Network Architectures: A Comprehensive Survey," *Wireless Personal Commun.*, Springer, vol. 86, no. 3, Feb. 2016, pp. 1401–38.

[5] F. Bari *et al.*, "Orchestrating Virtualized Network Functions," *IEEE Trans. Network and Service Management*, vol. 13, no. 4, Dec. 2016, pp. 725–39.

[6] A. Antonopoulos *et al.*, "Shedding Light on the Internet: Stakeholders and Network Neutrality," *IEEE Commun. Mag.*, vol. PP, no. 99, 2017, pp. 2–9.

[7] G. Tseliou *et al.*, "A Capacity Broker Architecture and Framework for Multi-Tenant Support in LTE-A Networks," *IEEE ICC*, 2016, pp. 1–6.

[8] S. Gu *et al.*, "Virtualized Resource Sharing in Cloud Radio Access Networks through Truthful Mechanisms," *IEEE Trans. Commun.*, vol. 65, no. 3, Mar. 2017, pp. 1105–18.

[9] H. Moens and F. De Turck, "Customizable Function Chains: Managing Service Chain Variability in Hybrid NFV Networks," *IEEE Trans. Network and Service Management*, vol. 13, no. 4, Dec. 2016, pp. 711–24.

[10] R. Riggio *et al.*, "Scheduling Wireless Virtual Networks Functions," *IEEE Trans. Network and Service Management*, vol. 13, no. 2, Apr. 2016, pp. 240–52.

[11] L. Wang *et al.*, "Joint Optimization of Service Function Chaining and Resource Allocation in Network Function Virtualization," *IEEE Access*, vol. 4, Nov. 2016, pp. 8084–94.

[12] M. Mechtri, C. Ghribi, and D. Zeghlache, "A Scalable Algorithm for the Placement of Service Function Chains," *IEEE Trans. Network and Service Management*, vol. 13, no. 3, Sept. 2016, pp. 533–46.

[13] T. W. Kuo *et al.*, "Deploying Chains of Virtual Network Functions: On the Relation Between Link and Server Usage," *IEEE INFOCOM 2016*, 2016, pp. 1–9.

[14] A. S. Sendi *et al.*, "Efficient Provisioning of Security Service Function Chaining Using Network Security Defense Patterns," *IEEE Trans. Services Computing*, vol. PP, no. 99, 2017, pp. 1–1.

[15] Y. Gu *et al.*, "Matching Theory for Future Wireless Networks: Fundamentals and Applications," *IEEE Commun. Mag.*, vol. 53, no. 5, May 2015, pp. 52–59.

Biographies

EFTYCHIA DATSIKA (edatsika@iquadrat.com) received her B.Sc. and M.Sc. degrees in computer science from the Computer Science Department, University of Ioannina, Greece, in 2010 and 2012, respectively. She is currently a researcher in IQUADRAT Informatica S.L., Barcelona, Spain. Her research interests lie in the areas of resource management in Long Term Evolution Advanced networks, software defined networking, network service chaining, and matching theory.

ANGELOS ANTONOPOULOS (aantonopoulos@cttc.es) received his Ph.D. degree from the Technical University of Catalonia (UPC) in 2012. He is currently a researcher with CTTC/CERCA. He has authored over 70 peer-reviewed publications on various topics, including energy-efficient network planning, 5G wireless networks, cooperative communications, and network economics. He received Best Paper Awards at IEEE GLOBECOM 2014 and EuCNC 2016, the Best Demo Award at IEEE CAMAD 2014, and the First Prize in the IEEE ComSoc Student Competition.

NIZAR ZORBA (nizarz@qu.edu.qa) received his B.Sc. degree in electrical engineering from Jordan University of Science and Technology, Irbid, in 2002, his M.Sc. degree in data communications and M.B.A. degree from the University of Zaragoza, Spain, in 2004 and 2005, respectively, and his Ph.D. degree in signal processing for communications from UPC in 2007. He has led and participated in over 25 research projects; and authored five patents, two books, seven book chapters, and over 100 peer-reviewed journals and international conferences. His research interests are quality of service/experience, energy efficiency, and resource optimization.

CHRISTOS VERIKOUKIS (cveri@cttc.es) (Ph.D., UPC, 2000) is a Fellow researcher at CTTC/CERCA and an adjunct associate professor at the University of Barcelona. He is a co-author of 4 books, 18 chapters, 2 patents, 108 journal papers, and over 180 conference papers. He has participated in more than 30 competitive projects and served as the principal investigator of national projects. He has supervised 15 Ph.D. students and 5 postdoctoral researchers. He received Best Paper Awards at IEEE ICC 2011, IEEE GLOBECOM 2014 and 2015, and EUCNC/EURACON 2016, and the EURASIP 2013 Best Paper Award for the *Journal on Advances in Signal Processing*. He is currently Chair of the IEEE ComSoc CSIM TC.

Ensuring that the security levels requested by OSPs and NSPs are maintained in NSC deployment can be an issue with high technical complexity. To this end, we believe that our study has shed some light on the OSPs' requirements and can motivate further investigation of NSC for OTT applications.

A Lifetime-Enhanced Data Collecting Scheme for the Internet of Things

Tie Qiu, Ruixuan Qiao, Min Han, Arun Kumar Sangaiah, and Ivan Lee

The authors propose an energy-aware and distance-aware data collecting scheme to enhance the lifetime of backpressure-based data collecting schemes. They propose an energy- and distance-based model that combines the factors of queue backlog, hop counts, and residual energy for making routing decisions. It not only reduces the unnecessary energy consumption, but also balances the residual energy.

ABSTRACT

A backpressure-based data collecting scheme has been applied in the Internet of Things, which can control the network congestion effectively and increase the network throughput. However, there is an obvious shortcoming in the traditional backpressure data collecting scheme for the network service chain. It attempts to search all possible paths between source node and destination node in the networks, causing an unnecessary long path for data collection, which results in large end-to-end delay and redundant energy consumption. To address this shortcoming of backpressure data collecting scheme in the Internet of Things, this article proposes an energy-aware and distance-aware data collecting scheme to enhance the lifetime of backpressure-based data collecting schemes. We propose an energy- and distance-based model that combines the factors of queue backlog, hop counts, and residual energy for making routing decisions. It not only reduces the unnecessary energy consumption, but also balances the residual energy. The experiment results show that the proposed scheme can reduce unnecessary energy consumption and end-to-end delay compared to the traditional and LIFO-based schemes. Meanwhile, it balances the energy of nodes and extends the lifetime of an Internet of Things.

INTRODUCTION

The Internet of Things (IoT) has a complicated structure that consists of many types of heterogeneous sensor networks. In recent years, the number of wearable and mobile devices in IoT has increased dramatically [1]. Thus, the smart network service chain becomes more complicated. The congestion problems due to huge data have become serious [2]. When some problems occur in local network topology, the cascading failures and chaining collapse are getting worse [3]. Therefore, controlling network congestion and guaranteeing the throughput become a focus for researchers to study in the field of IoT with high data load [4, 5].

There are many data collecting schemes for IoT, including the collection tree scheme [6], the ZigBee scheme [7], and others. However, these schemes cannot guarantee the throughput of networks. The backpressure data collecting scheme is a distributed and adaptive data collecting scheme that can effectively control the network congestion and guarantee the throughput of networks [8]. The backpressure architecture is a microservices architecture. All sensor nodes are distributed, and they are even programmed in different languages. It can achieve dynamic regulation in multihop networks by selecting the packet with the largest backlog difference between neighbor nodes. As a result, it increases the openness and user-centric capabilities of IoT. The backpressure data collecting scheme is also proved to be robust for the time-varied network environment. Also, the deployment of the backpressure data collecting scheme is not affected by the packet arrival rate and the channel state. In recent years, a lot of work has been proposed to improve the performance of the backpressure data collecting scheme in the network service chain.

Typical IoT architecture is shown in Fig. 1, which includes wireless sensor networks (WSNs), wireless Wi-Fi networks, mobile communications networks (third generation [3G]/4G/LTE/5G), WiMAX networks, and wireless mesh networks (WMNs). The data in these units will finally be collected to the Internet cloud platform. These heterogeneous units consist of large-scale low-power smart devices. The lifetime performance becomes a crucial issue while applying the backpressure data collecting scheme in IoT.

However, the backpressure data collecting scheme makes routing decisions only according to the queue backlog difference between neighbor nodes [9]. When the network load is low, it attempts to search all possible paths before the packets are delivered and consumes much energy, because there are not enough queue backlog difference gradients [10]. Besides, the backpressure data collecting scheme does not take energy balance into consideration. Redundant energy consumption and energy imbalance problems reduce the lifetime of IoT when applying the backpressure data collecting scheme. Therefore, how to deal with routing and enhance the lifetime for large-scale heterogeneous low-power sensor networks is a serious challenge faced by researchers in recent years.

In this article, we propose an energy-aware and distance-aware model (EDA). First, the weight of the distance is used in queue backlog difference calculation to reduce unnecessary data forwarding, which reduces both the end-to-end delay and energy consumption. The lifetime of the backpressure data collecting scheme is extended in sensor networks. Furthermore, we employ the

Digital Object Identifier:
10.1109/MCOM.2017.1700033

Tie Qiu and Ruixuan Qiao, and Min Han are with Dalian University of Technology; Arun Kumar Sangaiah is with VIT University; Ivan Lee is with the University of South Australia Mawson Lakes.

0163-6804/17/$25.00 © 2017 IEEE

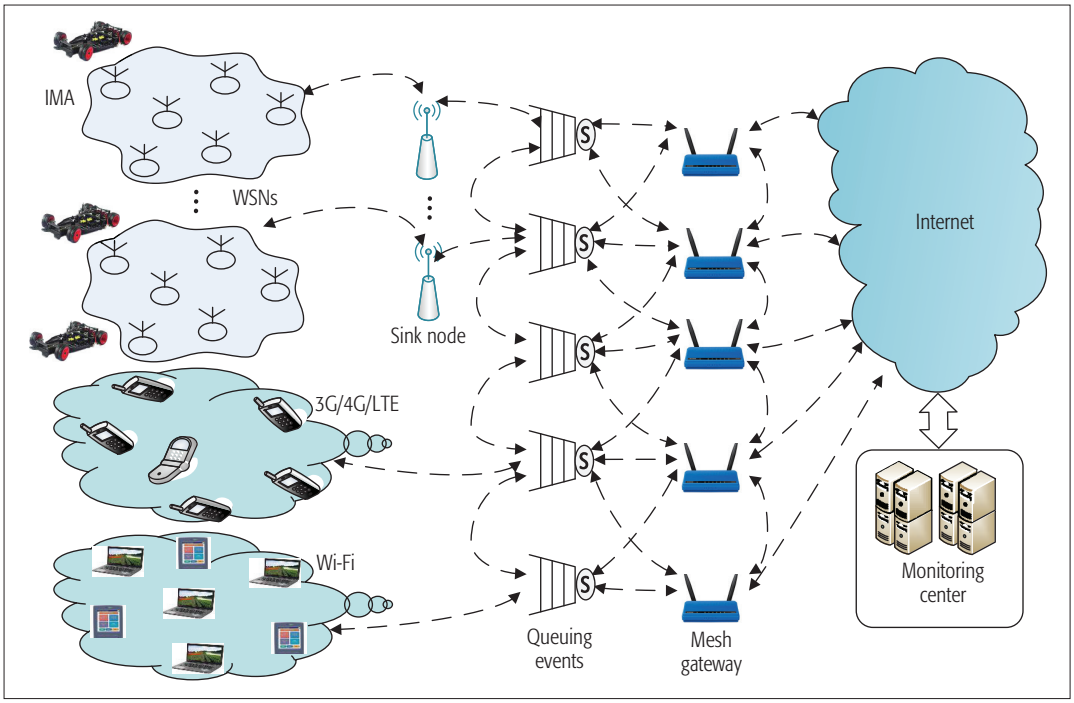


Figure 1. Typical IoT architecture.

Redundant energy consumption and energy imbalance problems reduce the lifetime of IoT while applying backpressure data collecting scheme. Therefore, how to deal with routing and enhancing the lifetime for large-scale heterogeneous low-power sensor networks is a serious challenge faced by researchers in recent years.

energy factor in next-hop node selection, which balances the energy of nodes and further extends the lifetime of sensor networks. We evaluate the proposed scheme and obtain better performance in terms of end-to-end delay, energy consumption, and lifetime through simulations compared to previous related work.

Energy Consumption Issue

The traditional backpressure routing strategy was first proposed by Tassiulas and Ephremides [8], and it is applied to WSNs [2]. The traditional backpressure data collecting scheme based on first-in first-out (FIFO) queue model is changed to last-in first-out (LIFO), reducing the end-to-end delay [10]. Further, Moeller *et al.* [6] described the application of the backpressure data collecting scheme in WSNs, which can achieve greater delivery ratio and throughput than the collection tree protocol. Huang *et al.* [11] analyzed the trade-off between LIFO and FIFO queue models. Liu *et al.* [12] used residual energy as the penalty parameter while calculating the queue backlog difference. It improves the traditional backpressure data collecting scheme in the field of energy balance for energy harvest sensor networks. A simple data collection scheme is proposed to increase the throughput of the network, while the average energy consumption converges to the energy consumption required to maintain the stability of the network [9]. Many researchers have used weighted queue and virtual queue methods to reduce the long path problem [13–15]. According to these ideas, improving the backpressure data collecting scheme becomes how to appropriately calculate the queue backlog difference. We need to consider both the energy and distance factors while improving the backpressure data collecting scheme in sensor networks.

The traditional backpressure data collecting scheme operates as follows:

- Select the best packets according to queue backlog of neighbor nodes (expressed by the network flow ID).
- Calculate the queue backlog difference between each pair of neighbor nodes for each link (m, n). (Each flow is expressed by the destination node.)
- Select the link rate between each node. The transmission rate of each link is determined at the beginning of each time slot t.
- Calculate weight and forward packets. Forward as many packets as possible according to the selected link rate and packets in time slot t.

In this article, we assume that all link rates are the same, and each node only forwards one packet in each time slot. Therefore, we make routing decisions by selecting the forwarding packets and next-hop nodes. The LIFO backpressure data collecting scheme changes the FIFO queue to the LIFO one [11] and reduces the end-to-end delay. However, it makes routing decisions based on the queue backlog difference. There is not enough queue backlog difference under low network load, resulting in a long path or even the loop path problem. We consider a sensor network as shown in Fig. 2.

$Q_{m,n}(t)$ represents the queue backlog of the node $N_{m,n}$ at time slot t. There is only one flow whose destination node is $N_{m,n}$ in the network. We assume that a packet has been forwarded to $N_{i+1,j+1}$ through multihop paths in the past time slots. Under the traditional backpressure data collecting scheme and the LIFO-based backpressure data collecting scheme, if $Q_{i+1,j+1}(t) - Q_{i+1,j}(t)$ is the maximum queue backlog difference, $N_{i+1,j+1}$ will forward the packet to $N_{i+1,j}$ at this time slot. However, $N_{i+1,j}$ is farther from the destination node $N_{m,n}$ compared to $N_{i+1,j+1}$. The packets are likely to be forwarded far from their destination nodes. Thus, the end-to-end delay increases. Furthermore, all nodes have limited energy storage in sensor networks. $E_n(t)$ represents the residual energy of node n at time slot t. For sim-

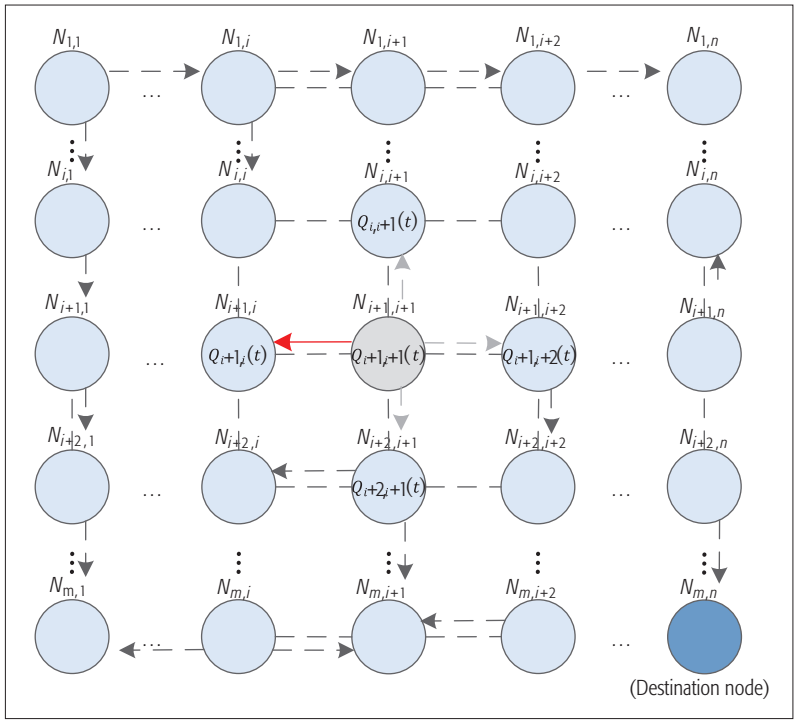


Figure 2. Long path problem for data collecting.

plicity, we assume that nodes consume energy only when forwarding and receiving packets. When the residual energy $E_n(t)$ equals 0, the energy of node n is used up. It is obvious that the lengthy paths cause increased energy consumption. Since most of the sensor nodes are powered by onboard batteries, the energy of sensor nodes is limited. Energy is an important resource for sensor networks. Redundant energy consumption shortens the lifetime of sensor networks.

A Lifetime-Enhanced Backpressure Data Collecting Scheme

The Distance-Aware Backpressure Data Collecting Scheme

In this article, the model of the sensor network can be represented by the graph G. V represents the set of nodes in the network; E represents the set of links. We assume that all nodes have the same transmission range, represented by R. $D_{m,n}$ represents the geographic distance of nodes m and n. If $D_{m,n}$ is less than or equal to R, (m, n) is in the set of $E(G)$ for each pair of nodes m and n. We use f to represent the queue number; f is used to represent the set of all queues in the network. The arrival process of packets in the network is assumed to follow a Poisson process with arrival rate λ (packets/slot). The time is slotted, represented by t. Node n calculates the queue backlog difference of each queue with neighbor nodes every time slot by using $Q_m^f(t) - Q_n^f(t)$. $Q_n^f(t)$ represents the backlog of queue f in node n at time slot t. Link (m, n) is used to forward packets with the largest queue backlog difference. Packets are forwarded according to the LIFO scheduling policy. In sensor networks, the packet can be generated by any node, and finally reach the destination node through the multihop path.

We attempt to improve the backpressure data collecting scheme in the step of making routing decisions, which does not affect the link transmitting process. Thus, our scheme can be directly extended to transmission rates in time-varying networks. To simplify our expression, we assume that the transmission rates are fixed. Making routing decisions only based on queue backlog is likely to cause packets to be forwarded far from the destination node. However, the energy is mainly consumed in the process of data forwarding and receiving. Therefore, reducing the forwarding times of packets can effectively save the energy of sensor nodes and extend the lifetime of the network. For this purpose, we make improvements in the step of selecting data packets and calculating weights for the backpressure data collecting scheme on the basis of the LIFO queue.

When calculating the link weights, the distance factor is introduced, which is based on hop counts from the neighbor nodes to the destination node. We use the quotient of H_n^f and the difference of H_n^f and H_m^f as a weight represented by w while calculating the queue backlog difference of node m and node n. If $H_n^f < H_m^f$, the weighted queue backlog difference is larger than the one under the LIFO backpressure data collecting scheme. If $H_n^f > H_m^f$, the weighted queue backlog difference is less than the one under the traditional backpressure data collecting scheme. If $H_n^f = H_m^f$, the queue backlog difference will be calculated without the distance weighted factor, as in the traditional backpressure data collecting scheme. After using the distance weight, the probability of a node being closer to the destination node increases. Then the packets have more probability to be forwarded to the node closer to the destination node, even if there are not enough queue backlog difference gradients. We can calculate the hop counts according to the Dijkstra algorithm after the whole sensor network has been deployed. Each node will obtain their minimum hop counts to every other node by the global broadcast once. We assume that all nodes in the networks are not mobile and the connections are stable; then the factor remains unchanged. We also assume that the transmission rates are fixed. Further, each node can broadcast the neighbor node list each time slot period. If the distance (hop count) changes, the neighbor node list will update. Thus, the result can easily be extended to mobile sensor networks.

The Energy-Aware and Distance-Aware Backpressure Data Collecting Scheme

As mentioned above, we use the distance factors as a weight while calculating the queue backlog difference. It will increase the weight of the nodes that are closer to the destination node, and the selected probability will increase. This can effectively alleviate the long path problem, but it will result in increased residual energy difference between nodes. The method does not balance the energy of nodes in the network. The energy distribution of the sensor nodes is not balanced, which leads to the energy depletion of sensor nodes at different times. Finally, it causes the problem of network partition, and the lifetime of sensor networks is shortened. Therefore, we calculate the weight according to the queue backlog, distance (hop count), and residual energy. We first select the queue $f_{m,n}^{sel}$ between node m and node n with the maximum weighted queue

backlog difference. Then we forward the packet in the selected queue $f^{sel}_{m,n}$ to the next-hop node.

We normalize $E_n(t)$ using the method of deviation normalization (min-max normalization) and use $E'_n(t)$ to represent $E_n(t)$ after normalization. We should give priority to choosing the node with more residual energy and smaller hop counts to the destination node. Thus, we calculate the energy and distance weight by using $E'_n(t)$ to multiply w, and choose the node with the maximum weighted queue backlog difference as the next-hop node.

The specific energy-aware and distance-aware backpressure data collecting scheme is as follows.

The Energy-Aware and Distance-Aware Backpressure Data Collecting Scheme: All nodes in the whole network broadcast their neighbor nodes list and calculate their shortest path with all other nodes at the first time slot. Next, at the beginning of each time slot, node m observes the queue backlog $Q^f_n(t)$, hop counts $H^f_n(t)$, and residual energy $E_n(t)$ of every neighbor node n.

Routing Decision:

- Calculate the weighted queue backlog difference of each neighbor node n. The packet in queue $f^{sel}_{m,n}$ is selected and waits to be forwarded.
- Select the next-hop node with the maximum forwarding weight.

Forwarding Decision: If the forwarding weight is positive, forward as many packets as possible in the capacity of the link (m, n) under a LIFO queue. Otherwise, the packet is not forwarded until the weight is recalculated.

Queue Backlog and Energy Updating: Update queue backlog $Q_n{}^{f(t+1)}$ and residual energy $E_n(t+1)$.

In order to analyze the influence of the energy and distance factors, we illustrate how to make routing decisions under the energy-aware and distance-aware backpressure data collecting scheme.

As shown in Fig. 3, there are five nodes in the networks (nodes 1, 2, 3, 4, and 5). We assume that there are only two flows in the network, whose destination node is node 4, and the other destination node is node 5. The backlogs of queue 4 in each node are shown in Fig. 3. For example, the backlog of queue 4 in node 1 is 6. We can also observe the hop counts between each node in Fig. 3 (e.g., H^4_1 equals to 2). Each node calculates the queue backlog difference with their neighbor nodes. As for link (1,2), the weighted queue backlog difference of queue 4 equals 9, and that of queue 5 equals 0. Thus, the maximum queue backlog difference is 9 after calculating the link weight, so queue 4 is selected. According to the traditional backpressure data collecting scheme, the queue with the maximum backlog difference is 5, and queue 5 will be selected to forward over link (1,2). We can also observe from Fig. 3 that node 2 is farther from node 5 than node 1. Therefore, choosing the packet in queue 5 to forward is not a good choice for link (1,2). It will forward this packet in a direction far from its destination, which will increase the end-to-end delay, waste much energy, and shorten the lifetime of the network. Thus, choosing the packet in queue 4 is more reasonable.

There are two neighbor nodes of node 1 (nodes 2 and 3), as shown in Fig. 3. Under the traditional backpressure data collecting scheme, the forwarding weight of link (1,2) is 3, and that of link (1,3) is 4. Then node 3 is selected as the next-hop node. We can observe from Fig. 3 that the residual energy of node 2 is 100 J, and the residual energy of node 3 is 99 J. $E'_2 = 1$ and $E'_3 = 0$ after normalization. Then we calculate the forwarding weight. The forwarding weight of link (1,2) is 9, and that of link (1,3) is 6 under our proposed scheme. Therefore, the node with the maximum forwarding weight is node 2. Thus, node 1 chooses node 2 to forward the packet in queue 4 at time slot t. It can be seen that the scheme is able to balance the energy of nodes while reducing the forwarding times.

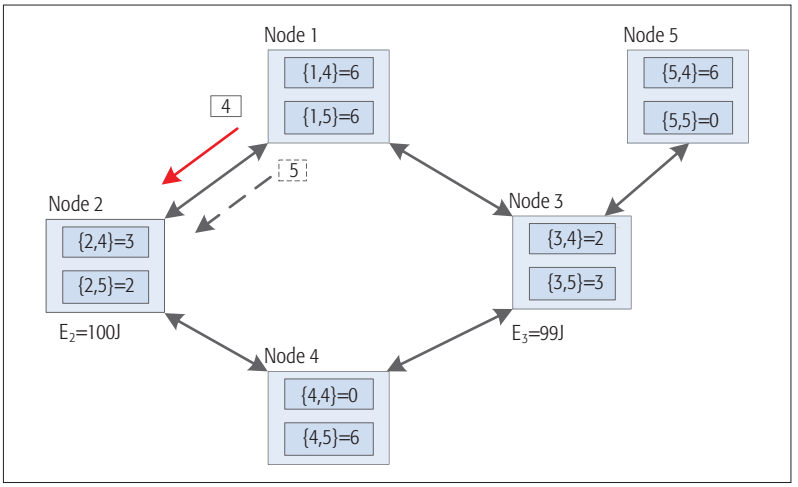


Figure 3. A topology of the improved backpressure data collecting scheme.

The energy consumption of sensor nodes in the network is reduced, and the energy is balanced because of the consideration of distance and energy factors. Taking energy and distance factors into consideration extends the lifetime of sensor networks while applying the backpressure data collecting scheme.

Simulation and Analysis

We use the NS-2 simulation platform to verify the proposed scheme and evaluate the performance. The simulation results of the proposed scheme (EDA-BP and DA-BP), traditional backpressure data collecting scheme (BP) [8], and LIFO-based scheme (LIFO-BP) [11] are compared.

Simulation Setup

We deploy a random network topology with 100 nodes in a 2200 m × 2200 m area. The transmission range of each sensor nodes is 250 m. The transmission capability is one packet in each direction for each time slot, and all links are bidirectional. We randomly create 20 network data flows in the network. The packet arrival rate of all network flows follows a Poisson process, and all network flows have the same packet arrival rate, donated by λ (packets/slot). The arrival rate λ varies from 0.1 to 0.8 packets/slot. We choose these parameters because the results are able to verify our analyses of the traditional backpressure scheduling scheme. The end-to-end delay will first increase, then decrease when the arrival rate varies from 0.1 to 0.8 packets/slot. We set the initial energy to 100 J for each node; and the transmit power is set to 0.690 W, the receive power to 0.395 W. The energy is mainly consumed in the process of

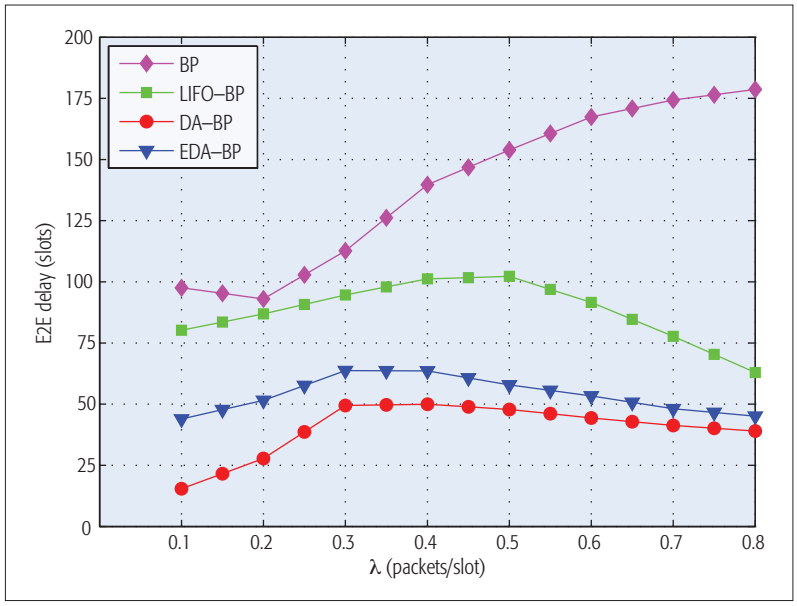


Figure 4. Average end-to-end delay.

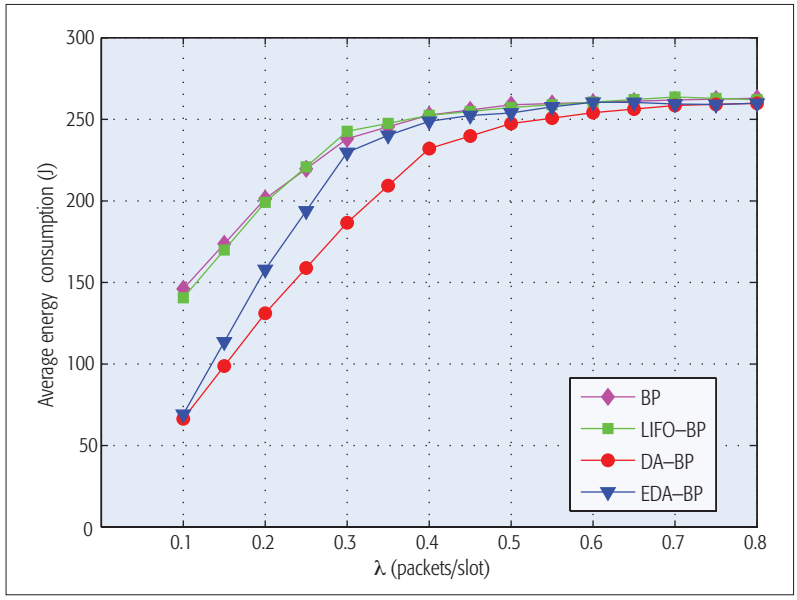


Figure 5. Average energy consumption.

data forwarding and receiving. We use the time of the first node running out of energy as the evaluation metric for network lifetime. We set the data packet arrival rate λ (packets/slot) as the horizontal coordinate in the simulation. We change the value of λ to observe the performance of the proposed scheme under different network loads. The simulation experiments are implemented in 500 time slots, each of which has been implemented for 10 iterations to avoid random error.

End-to-End Delay

Figure 4 shows the average end-to-end delay of the BP, LIFO-BP, and our proposed DA-BP and EDA-BP schemes. It is the average time from packet generation to delivery in the network. It can be seen that with the increase of the packet arrival rate, the end-to-end delay of the BP scheme first decreases, then increases. The reason is that the BP cannot generate enough pressure at low network load due to insufficient queue backlog difference gradients. The BP scheme will explore all possible paths, so the end-to-end delay increases. When the network load increases, the end-to-end delay of the BP scheme is reduced because the queue backlog difference gradient is formed. Subsequently, the queue backlog starts to increase, which leads to increasing end-to-end delay. The LIFO-BP scheme improves the BP scheme by changing the queue from FIFO to LIFO. With the formation of the queue backlog gradient, the delay reduces according to the LIFO-BP scheme when the network load increases.

The DA-BP scheme increases the weight of a node closer to the destination node on the basis of LIFO-BP. The distance factors are weighted so that the data packets can still be forwarded to the destination node when the queue backlog difference gradient has not been formed. The hop counts are reduced, so the average end-to-end delay is reduced, too. The EDA-BP scheme also considers the distance factor as does DA-BP; however, the energy factor increases the delay. Thus, the delay of the EDA-BP scheme is higher than that of the DA-BP scheme.

Average Energy Consumption

We show the average energy consumption of BP, LIFO-BP, and our proposed DA-BP and EDA-BP schemes in Fig. 5, which is the average energy consumption of each node in the network in 500 time slots. It can be seen that the average energy consumption of the BP scheme is higher than that of the LIFO-BP scheme. The reason is that the LIFO-BP scheme changes the queue model, which reduces the packet forwarding times. Thus, the average energy consumption is reduced. However, the energy consumption is still at a high level. The DA-BP and EDA-BP schemes reduce the forwarding times of packets using the weight of the distance factor. The packets are delivered to the destination node faster than the above two schemes. It avoids unnecessary forwarding of data packets, which reduces the energy consumption. Finally, the energy consumption is converged because every node forwards a packet at each time slot when the network load increases.

Lifetime

The lifetimes of the network under BP, LIFO-BP, and our proposed DA-BP and EDA-BP schemes are shown in Fig. 6. We assume that when the node with 100 J initial energy runs out, the node dies. We use the time that the first node runs out of energy as the lifetime of the network in this article. The DA-BP schene reduces the energy consumption compared to the BP and LIFO-BP schemes. The maximum energy consumption is lower, especially under low network load. The EDA-BP scheme balances the energy of each node on the basis of the DA-BP scheme, so the maximum energy consumption of the EDA-BP scheme is lower than that of the DA-BP scheme. We can see that the lifetimes of the DA-BP and EDA-BP schemes are longer than those of the BP and LIFO-BP schemes, especially under low network load. Furthermore, the EDA-BP scheme balances the energy of each node in the network, so the fluctuation is reduced, as is the maximum energy of a node. As a result, the lifetime of the EDA-BP scheme is longer than that of the DA-BP scheme.

Conclusion

In this article, we propose an energy-aware and distance-aware backpressure data collecting scheme (EDA) based on the LIFO queue model for the network service chain. First, the weight of the distance is used in queue backlog difference calculation, which reduces unnecessary packet forwarding and energy consumption. Then we use the energy-aware and distance-aware model for selecting the next-hop nodes. The proposed scheme reduces the forwarding times of packets. At the same time, it balances the energy of nodes in networks and extends the lifetime of large-scale low-power sensor networks. In applications, DA-BP is attractive for low-latency networks, and EDA-BP is suitable for networks that require long lifetimes. The simulation results show that our proposed energy-aware and distance-aware backpressure data collecting scheme effectively reduces the end-to-end delay and energy consumption. Furthermore, the EDA-BP scheme balances energy and extends the lifetimes of large-scale low-power sensor networks compared to other previous schemes.

With the increasing data, congestion problems occur in IoT. If we can consider the network congestion factors while calculating the weight in the EDA-BP scheme, making the calculation more dynamic, it will get better performance. In addition, there are different types of data packets in IoT, including emergency packets and regular packets. A backpressure data collecting scheme cannot guarantee the real-time demand of emergency packets. In future work, we will focus on the congestion control strategy and dealing with emergency packets while making routing decisions.

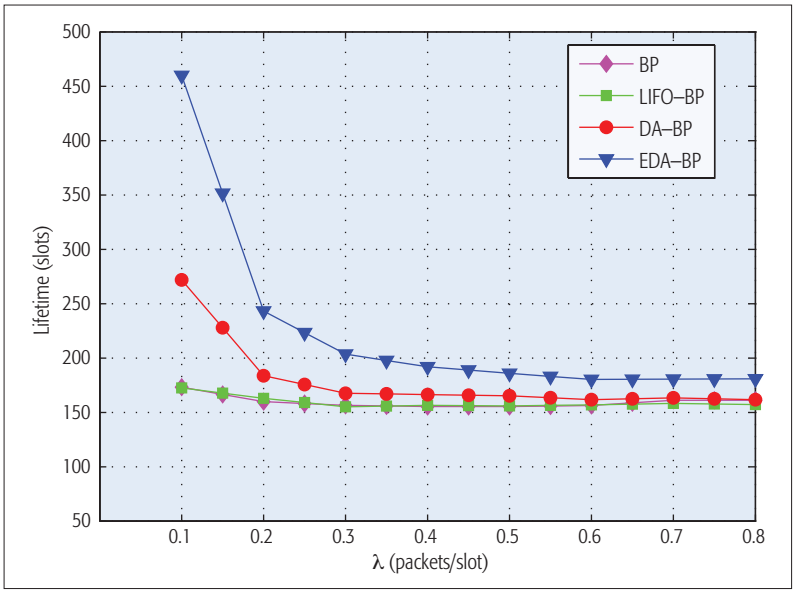


Figure 6. Lifetime.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (Grant Nos. 61374154 and 61672131) and the Fundamental Research Funds for the Central Universities (DUT17ZD216 and DUT16QY27).

References

[1] Y. Sun *et al.*, "Internet of Things and Big Data Analytics for Smart and Connected Communities," *IEEE Access*, vol. 4, 2016, pp. 766–73.
[2] T. Qiu, R. Qiao, and D. Wu, "EABS: An Event-Aware Backpressure Scheduling Scheme for Emergency Internet of Things," *IEEE Trans. Mobile Comp.*, 2017.
[3] V. P. Kafle, Y. Fukushima, and H. Harai, "Design and Implementation of a Dynamic Mobile Sensor Network Platform," *IEEE Commun. Mag.*, vol. 53, no. 3, Mar. 2015, pp. 48–57.
[4] O. Diallo *et al.*, "Distributed Database Management Techniques for Wireless Sensor Networks," *IEEE Trans. Parallel Distrib. Sys.*, vol. 26, no. 2, 2015, pp. 604–20.
[5] J. Teo, Y. Ha, and C. Tham, "Interference-Minimized Multipath Routing with Congestion Control in Wireless Sensor Network for High-Rate Streaming," *IEEE Trans. Mobile Comp.*, vol. 7, no. 9, 2008, pp. 1124–37.
[6] S. Moeller *et al.*, "Routing Without Routes: The Backpressure Collection Protocol," *Proc. ACM/IEEE Int'l. Conf. Info. Processing in Sensor Networks*, 2010, pp. 279–90.
[7] L. Parra *et al.*, "Design and Deployment of a Smart System for Data Gathering in Estuaries Using Wireless Sensor Networks," *IEEE CITS*, 2015, pp. 1–5.
[8] L. Tassiulas and A. Ephremides, "Stability Properties of Constrained Queueing Systems and Scheduling Policies for Maximum Throughput in Multihop Radio Networks," *Proc. IEEE Conf. Decision Control*, vol. 37, no. 12, 1992, pp. 1936–48.
[9] M. J. Neely and R. Urgaonkar, "Optimal Backpressure Routing for Wireless Networks with Multi-Receiver Diversity," *Ad Hoc Net.*, vol. 7, no. 5, 2009, pp. 862–81.
[10] S. Moeller *et al.*, "Backpressure Routing Made Practical," *Proc. IEEE INFOCOM*, 2010, pp. 1–2.
[11] L. Huang *et al.*, "LIFO-Backpressure Achieves Near-Optimal Utility-Delay Tradeoff," *IEEE ACM Trans. Net.*, vol. 21, no. 3, 2013, pp. 831–44.
[12] Z. Liu *et al.*, "Energy-Balanced Backpressure Routing for Stochastic Energy Harvesting WSNs," *Lecture Notes in Computer Science*, 2015, pp. 767–77.
[13] J. Lu *et al.*, "Distance-Weighted Backlog Differentials for Back-Pressure Routing in Multi-Hop Wireless Networks," *IEEE/CIC Int'l. Conf. Commun. China*, 2014, pp. 791–95.
[14] L. X. Bui, R. Srikant, and A. Stolyar, "A Novel Architecture for Reduction of Delay and Queueing Structure Complexity in the Back-Pressure Algorithm," *IEEE ACM Trans. Networking*, vol. 19, no. 6, 2011, pp. 1597–1609.
[15] Z. Jiao *et al.*, "A Virtual Queue-Based Back-Pressure Scheduling Algorithm for Wireless Sensor Networks," *Eurasip J. Wireless Commun. Networking*, vol. 2015, no. 1, 2015, pp. 1–9.

Biographies

Tie Qiu [M'11, SM'16] (qiutie@ieee.org) received his Ph.D. degree in computer science from Dalian University of Technology (DUT), China, in 2012. He is currently an associate professor at the School of Software, DUT. He is the founding director of the Smart Cyber-Physical Systems Laboratory (SmartCPS Lab). His research interests include the areas of embedded systems, cyber-physical systems, the Internet of Things, and mobile social networks.

Ruixuan Qiao (qiaoruixuan@mail.dlut.edu.cn) received his B.E. from DUT in 2014. He is a Master's student at the School of Software, DUT. He participated in the Open Source Hardware and Embedded Computing Competition and won the first prize. He is a member of the SmartCPS Lab. His research interests cover embedded systems and the Internet of Things.

Min Han [M'95, A'03, SM'06] (minhan@dlut.edu.cn) received her M.S. and Ph.D. degrees from Kyushu University, Fukuoka, Japan, in 1996 and 1999, respectively. Since 2003, she has been a professor in the Faculty of Electronic Information and Electrical Engineering, DUT. She is the author of four books and more than 200 articles. Her current research interests are neural networks and chaos, and their applications to control and identification.

Arun Kumar Sangaiah (sarunkumar@vit.ac.in) received his Ph.D. degree in computer science and engineering from VIT University, Vellore, India. He is presently working as an associate professor in the School of Computer Science and Engineering, VIT University. His areas of interest include software engineering, computational intelligence, wireless networks, bio-informatics, and embedded systems. He is also an Editorial Board member/Associate Editor of various international journals

Ivan Lee (Ivan.Lee@unisa.edu.au) received his Bachelor of Engineering (Hons), Master of Commerce, Master of Engineering by Research, and Ph.D. degrees from the University of Sydney, Australia. Since 2008, he has been a senior lecturer in the School of IT and Mathematical Sciences at the University of South Australia. His research interests include multimedia systems, medical imaging, data analytics, and the Internet of Things.

Bringing Computation Closer toward the User Network: Is Edge Computing the Solution?

Ejaz Ahmed, Arif Ahmed, Ibrar Yaqoob, Junaid Shuja, Abdullah Gani, Muhammad Imran, and Muhammad Shoaib

The authors highlight the significance of edge computing by providing real-life scenarios that have strict constraint requirements on application response time. From the previous literature, they devise a taxonomy to classify the current research efforts in the domain of edge computing. They also discuss the key requirements that enable edge computing. Finally, current challenges in realizing the vision of edge computing are discussed.

ABSTRACT

The virtually unlimited available resources and wide range of services provided by the cloud have resulted in the emergence of new cloud-based applications, such as smart grids, smart building control, and virtual reality. These developments, however, have also been accompanied by a problem for delay-sensitive applications that have stringent delay requirements. The current cloud computing paradigm cannot realize the requirements of mobility support, location awareness, and low latency. Hence, to address the problem, an edge computing paradigm that aims to extend the cloud resources and services and enable them to be nearer the edge of an enterprise's network has been introduced. In this article, we highlight the significance of edge computing by providing real-life scenarios that have strict constraint requirements on application response time. From the previous literature, we devise a taxonomy to classify the current research efforts in the domain of edge computing. We also discuss the key requirements that enable edge computing. Finally, current challenges in realizing the vision of edge computing are discussed.

INTRODUCTION

Rapid advancements in computing technologies have enabled a wide range of applications, usually categorized as future Internet applications. Examples of these applications are road traffic and smart surveillance. However, the majority of these emerging applications are compute-intensive in nature and impose stringent requirements on delay. Their compute-intensive nature makes such applications difficult to run on resource-constrained mobile devices. The limited capabilities of mobile devices are augmented by leveraging the resources of cloud servers. The cloud provides virtually unlimited resources and a wide range of services to enable such compute-intensive applications on resource-constrained mobile devices. The delay-sensitive compute-intensive applications still suffer from long wide area network (WAN) latency that transforms euphoria into a problem because of the strict delay requirements. The current cloud computing paradigm augments the capabilities of resource-constrained devices, but it cannot fulfill the requirements of location awareness, mobility support, and low latency [1].

In this context, researchers introduced the vision of edge computing to enable applications on billions of smart connected devices to run directly at the network edge. Similar to the cloud, edge computing also assists the user by providing compute, data, storage, and application services [2]. The distinguishing characteristics of edge computing include its dense geographical distribution, support for mobility, and proximity to end users. Edge computing aims to provide location awareness, maintain low latency, support heterogeneity, and ameliorate quality of service (QoS) for real-time applications, such as transportation, industrial automation, networks of actuators and sensors, and real-time big data analytics. Services are deployed at edge devices, such as access points or set-top boxes. Table 1 highlights the differences between cloud computing and edge computing.

Current edge computing architectures are modeled as three-level hierarchies, as illustrated in Fig. 1. In these hierarchies, smart things can connect to edge servers, these devices can interconnect with each other, and each edge device is also connected to the cloud.

The contributions of this article are as follows:

- We present possible use cases for edge computing.
- The study creates a classification of the existing literature by devising a taxonomy.
- Key requirements to enable edge computing are identified.
- The current challenges in realizing the vision of edge computing are highlighted.

The article also enables edge computing application engineers and service providers to leverage on the relevant features that can minimize communication and computation latencies while providing edge services to users. The identified requirements can serve as a guide for framework designers in incorporating specific features to efficiently execute the application in the edge computing paradigm. Similarly, these identified challenges highlight future research directions. These contributions are provided in separate sections.

POSSIBLE USE CASES

This section presents the application scenarios in the domain of edge computing. Edge computing can resolve several issues in various scenarios, such as real-time image processing, gaming, smart

Ejaz Ahmed, Ibrar Yaqoob, and Abdullah Gani are with the Centre for Mobile Cloud Computing Research, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia; Arif Ahmed is with the National Institute of Technology Silchar, India; Junaid Shuja is with COMSATS Institute of Information Technology Abbottabad, Pakistan; Muhammad Imran and Muhammad Shoaib are with King Saud University, Saudi Arabia.

Digital Object Identifier:
10.1109/MCOM.2017.1700120

0163-6804/17/$25.00 © 2017 IEEE

grid, smart traffic lights and connected vehicles, smart building control, and the smart health environment.

Real-Time Image Processing: Peter, a foreign visitor in Japan, wants to know the details on the food listed in a hotel menu written in Japanese. Thus, he takes a snapshot of the menu to process it using a character recognition application. The execution of the application is compute-intensive. Therefore, he is unable to run the application on his mobile phone. In this context, he can execute his application on the edge server available inside the hotel network. He registers with the edge server and migrates the application to the server. The results are then sent back to him upon completion of the task.

Gaming: Peter's son, who is fond of computer games, wants to play a high definition 3D game from his mobile phone on a train while traveling from Seoul to Pohang-si, South Korea. He finds an available edge server on the train that provides the services to passengers of the train to run their applications. The game engine is migrated to the edge server where the actual business logic of the game will run. The game interface is the only component that runs on the mobile device. The execution of the business logic on the edge server extends the battery life of the mobile device and enables game execution on the resource-constrained mobile device.

Smart Grid: Energy load balancing applications running on smart meters and microgrids enable automatic switching to alternative energies, such as wind and solar energy after considering the lowest price, energy demands, and availability. The data generated by grid devices and sensors are processed by the edge collector at the edge network. Edge devices process the delay-sensitive data in the edge network and send the rest of the data to the cloud server. In a smart grid environment, the lowest tier stores temporary data, whereas the highest tier stores the semi-permanent data.

Smart Transportation: Edge computing can also contribute to improving the functionalities and services provided by smart transportation. Smart lights that can serve as edge devices can take sensing information of the flashing lights of an ambulance from a video camera and consequently open lanes for the ambulance. Similarly, smart street lights take information from sensors to detect the presence of bikers and pedestrians and then turn the lights on or off depending on when movement is detected or traffic has passed.

Real-Time Big Data Analytics: Big data processing has been a hot research area for computer science researchers. Real-time big data analytics in cloud computing is a challenging process because of the enormous volume of data and large WAN latency. Edge computing furnishes on-demand resources for processing huge amounts of real-time big data. Processing big data in the edge network reduces the traffic in the network and the workload on the cloud server [3]. Edge computing can complement the services provided by cloud computing. For instance, in a large-scale environment monitoring system, regional and local data can be collected and mined at edge servers, thereby enabling timely responses in emergency cases. Compute-intensive tasks such as detailed analysis can be performed in the cloud server. Such edge computing-based big data analytics can be useful for the Internet of Things (IoT) and smart cities where sensor devices continuously generate enormous amounts of data [4, 5].

Parameters	Cloud computing	Edge computing
Service location	Within the Internet	In the edge network
Distance (number of hops)	Multiple hops	Single hop
Latency	High	Low
Jitter	High	Very low
Location awareness	No	Yes
Geo-distribution	Centralized	Distributed
Mobility support	Limited	Supported
Data en route attacks	High probability	Very low probability
Target user	General Internet users	Mobile users
Service scope	Global	Limited
Hardware	Scalable capabilities	Limited capabilities

Table 1. Differences between cloud computing and edge computing.

Taxonomy

A taxonomy on the edge computing paradigm is shown in Fig. 2. This taxonomy is classified broadly into the following attributes:

- Access technologies
- Computing devices
- Computing paradigms
- Objectives
- Enabling technologies
- Computational hierarchy
- Applications

The rest of the section discusses each attribute of the taxonomy by providing a definition of each term.

Access Technologies

The end user can access the edge services using both wireless and wired access technologies. Wired communication (i.e., Ethernet) is used to connect within an office network and in the server room to provide a link between short distances. Telecommunication network operators with the assistance of third-party application developers can deploy new services for consumers and enterprise business services rapidly at the base station. Mobile users can subscribe to third generation (3G), 4G, and 5G networks to access the applications [6]. Deploying computational offloading services at the WiFi access point to handle a large compute-intensive application targeting a few users is also possible. IoT devices, such as small embedded devices and smart bulbs, use WiFi networks to perform machine-to-machine (M2M) communication.

Computing Devices

Unlike traditional cloud data centers, edge computing devices are deployed closer to the end user in a distributed geographical location; however, such devices have certain weaknesses, such

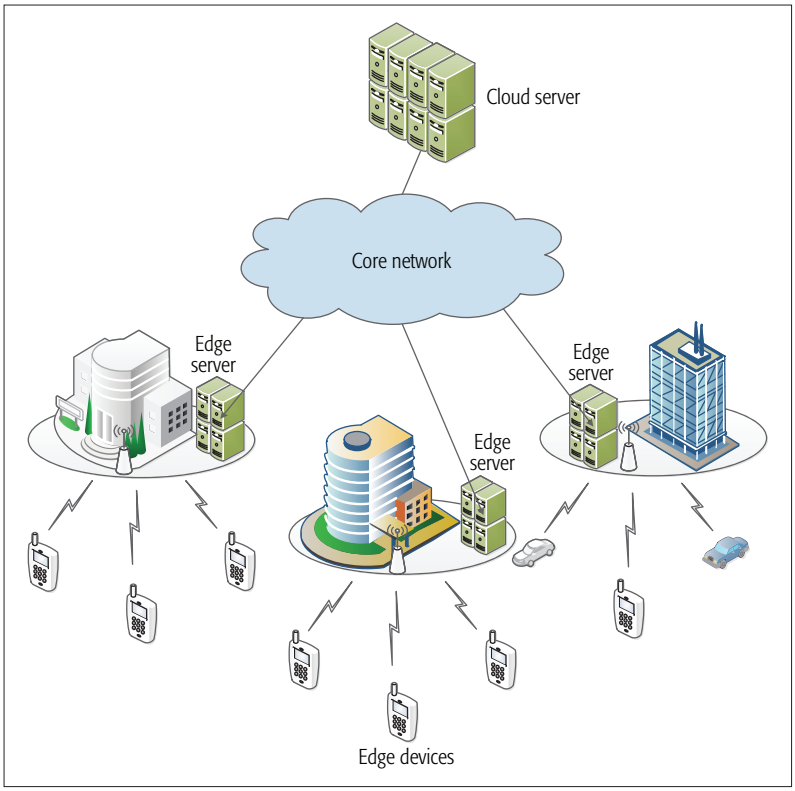


Figure 1. Edge-fog-cloud computing architecture as an ecosystem of different cyber-physical systems.

as computational, storage, and network bandwidth. The service applications can be deployed in the telecommunication network (base station), network devices (e.g., router and switches), and computer systems.

Computing Paradigms

The sole purpose of edge computing is to bring cloud computational resources and services within close proximity to the user. Three emerging paradigms that aim to achieve this goal are fog computing [7], cloudlet [8], and mobile edge computing (MEC) [9, 10]. Fog computing and MEC represent the commercialization of the idea of the cloudlet. These paradigms provide an intermediate layer between the end device and the central cloud to reduce the network latency. In MEC, the virtualized server provides all the services and deploys them in the telecommunication network. In fog computing, the main computational services are available at network routers and switches.

Objectives

Edge computing brings the centralized computing platform of the core network to the edge network. The objectives of bringing the computational part from the core network to the edge network are as follows.

Utility Maximization: The quality of experience (QoE) of users can be improved significantly. The end users are in the proximity of the service provider, the location and context of the user can be accessed easily from the low-level network information (i.e., dynamic network environment). This information can be utilized by cloud providers to provide dynamic services.

Carbon Footprint Minimization: The distributed data center also consumes considerable energy and emits a huge amount of carbon to the surrounding atmosphere. Energy consumption and the amount of carbon emission are challenges of the current data center. Joint resource allocation can minimize the carbon footprint of the distributed nodes of the edge servers [11]. Tiny IoT devices require a considerably small amount of energy for computing, and energy can be harvested from renewable energy sources.

Number of Executed Tasks Maximization: Efficient utilization of hardware resources in edge computing is one of the prime objectives. Distributed computing in the edge environment helps to allocate the task jointly to the edge nodes. The distributed task allocation not only reduces the idle state of the devices, but also maximizes the average number of executed tasks on the device.

Backhaul Traffic Reduction: Edge computing can provide advanced caching services at the base stations to reduce backhaul traffic. Such edge computing architecture can enable the design of a cost-effective backhaul by transmitting voice and data from the radio cell site to an edge switch.

Latency Minimization: Network latency is critical for many applications, such as augmented reality and m-game, if the application is required to connect to the centralized cloud server in the core network. The network latency of the core network is longer than that of the edge network [12]. The QoE of users improves if the computational part is deployed in the edge network to reduce the latency of delay-sensitive applications.

Bandwidth Saving: Several applications, such as video streaming, consume a considerable amount of bandwidth. The edge network can apply mechanisms such as web caching to minimize the bandwidth consumption of such applications.

Enabling Technologies

The realization of edge computing can be attributed to the recent development in telecommunications and distributed computing. Key enablers of edge computing are presented here.

Networking Technologies: The development of telecommunication networks has resulted in the increased bandwidth capacity of wireless links. Mobile subscribers use wireless communication, such as 3G or 4G telecommunication networks, to access edge servers. These networking technologies provide highly reliable data communication between an edge server and subscribers.

Software Development Kit: Many software designers provide software development kits (SDKs), which consist of an application programming interface (API) that assists in the seamless integration of new services into the existing software package. They enhance the usability of the software package and foster the development of new edge applications. Most SDKs are open source packages and are available to the developers.

Mobile Technologies: The development of mobile technologies has become notable with the invention of smartphones and portable devices such as tablets. The number of mobile subscribers has been increasing exponentially in recent years.

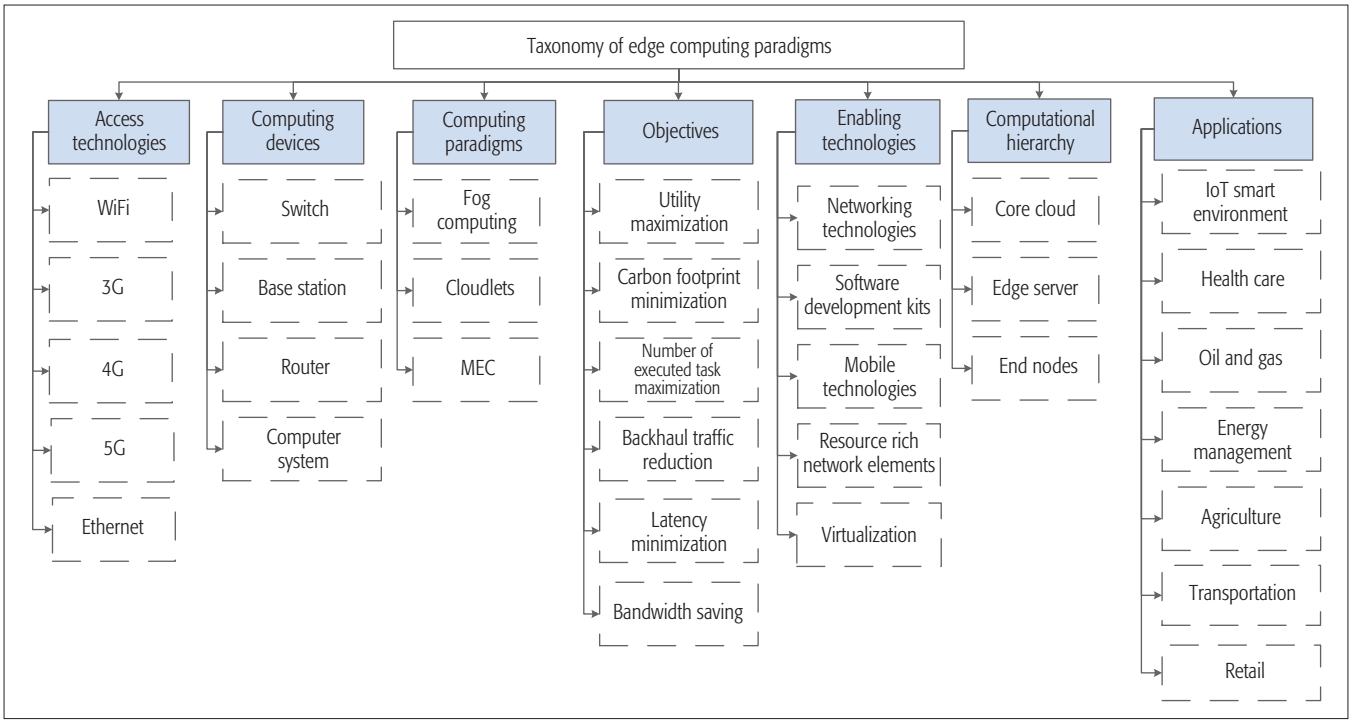


Figure 2. Edge computing taxonomy.

Because the majority of users use mobile devices to access cloud services, mobile technologies are key enablers of edge computing.

Resource-Rich Network Elements: The recent development of computer hardware has enabled telecommunication networks to deploy a resource-rich network infrastructure.

Virtualization: Virtualization creates virtual or logical infrastructure above the same physical hardware. Virtualization can be in both hardware and software. Virtual resources are highly efficient in terms of usability of the physical resources, and provide flexible and effective resource management. Edge computing services can be deployed in virtual machines and containers to increase efficiency in resource utilization.

Computational Hierarchy

Computational hierarchy shows that the execution in edge computing can be performed at different levels, such as core cloud, edge server, and end nodes. Although the goal of edge computing is to execute the compute-intensive delay-sensitive part of an application in the edge network, some applications in the edge server communicate with the core cloud for synchronization of data with the global application. It is also notable that the hierarchy represents the computing capacity of edge computing elements and their characteristics. The lowest element comprises end nodes that have less computational capability and subscribe mostly to edge services. The intermediate node hosts the edge computing services closer to the end user. The edge server sometimes accesses the core cloud located at a far distance from the end nodes.

Applications

Application attributes define the range of edge computing applications hosted by edge servers. In the IoT environment, the edge server hosts services for the intelligent data processing of the collected IoT data [4]. The IoT devices are controlled and managed directly by the edge server. In transportation, the edge server can control road traffic conditions intelligently by acquiring data from roadside cameras and sensor devices. The edge server can host online vehicle parking for smart parking of city dwellers. In healthcare, a person's health data can be collected from wearable devices, and the edge server can monitor the data for any emergency health assistance. Agriculture services, such as monitoring the environmental parameters (i.e., temperature, soil, and weather) can be deployed at the edge server. In retail, manufacturing and food processing can be automated with the help of edge computing [13].

Requirements for Edge Computing

Edge computing migrates the utility services hosted in the centralized cloud data centers to decentralized edge network devices. An edge network that has relayed data packets to and from users can also act as a computing entity. In this manner, cloud services can be provided by the edge network with minimal latency. However, the edge computing paradigm must ensure several requirements, such as scalability to workload, to penetrate the IT utility market. Cloud computing services have captured most of the IT utility market. Edge computing must add incentives to the utility model of cloud computing. The listed requirements also apply to both cases of the standalone functioning of edge computing as a mini-cloud and interplay with the cloud resources. We list the requirements of edge computing below.

Reliability

How will edge computing provide reliable services in instances of edge server failure and high user mobility?

While cloud interplay can lead to far greater scalability for both entities, the standalone functioning of edge devices requires pre-planning because of the limited resources of edge devices. Edge should ensure redundancy of resources in case it acts as an independent mini-cloud. Intelligent management of the scalability options in different scenarios is critical to the functioning of edge networks.

Edge devices and architectures are often dynamic and evolving. Moreover, most devices connected to the edge have high mobility. Therefore, a failover plan must be presented for user services in case of dynamic changes in the foggy environment. Reliable edge computing ensures failover mechanisms in scenarios such as the following:

- Failure of individual node/server/end user
- Lack of network coverage because of the limitations of the access network and mobility of end devices
- Failure of edge network and service platform

In cloud computing, the reliability of services is ensured by duplication/replication of servers and data center nodes. However, due to the decentralized nature of edge networks, new nodes cannot be added immediately to the infrastructure. Protocols similar to the packet path finding need to be implemented to find services elsewhere in a failover plan. Moreover, client session data stored on edge devices need a consistent backup in case of device/network failure.

In general, three techniques can be applied to improve the reliability of edge computing:

- Checkpointing, that is, periodically saving the state of end devices and user services
- Replication of edge servers and services in multiple geographical locations
- Rescheduling of failed tasks

Scalability

How will edge computing scale to the number of mobile end users, diverse applications, and low-bandwidth access networks?

Edge computing is scalable if it can penetrate diverse networks, dynamic sensors, end-user devices, and provide steady performance in case of a rapid increase in the number of end users and applications. Edge computing can be scaled by:

- Adding a new point of services (geographic expansions)
- Adding new service nodes to existing points of service
- Utilizing cloud interplay

While cloud interplay can lead to far greater scalability for both entities, the standalone functioning of edge devices requires pre-planning because of the limited resources of edge devices. The edge should ensure redundancy of resources in case it acts as an independent mini-cloud. Intelligent management of the scalability options in different scenarios is critical to the functioning of edge networks.

Security

Which technologies can enable the sandboxing of user data on edge networks and secure edge devices from malicious users?

Security in edge computing has two aspects:

- Isolation of data paths and memory as multiple users share edge devices
- Security of edge devices from malicious users because their addresses are advertised to end devices for service discovery

Application and session data from sensors and IoT devices are often cached on edge facilities. Moreover, multiple users contend for the limited resources of edge servers. While cloud computing provides isolation in the form of virtualized devices, edge computing needs to define sandboxes for user applications to ensure data isolation and monitor resource usage. A candidate for the fulfillment of the security requirement in the edge is network functions virtualization (NFV). NFV can be applied to network nodes to ensure security and isolation of user domains. However, NFV is still a new technology, and not all network vendors adhere to NFV standards. Moreover, as network devices advertise their available resources, they become more vulnerable to rogue attacks.

Resource Management

How will resources be federated by an edge network in the presence of multiple players and decentralized geo-dispersed resources?

Resource management comprises several activities, including resource allocation, reallocation, load balancing, and monitoring. The requirement of resource management is magnified in edge computing because of the decentralized and geo-dispersed nature of resources along with the interplay of cloud computing. The cloud interplay also adds complexity to the resource usage price model. Moreover, network resources have to be monitored because they largely dictate the execution of applications over edge and cloud infrastructure. Resource management has to be performed within an edge network server and between edge networks, thereby making the task more complicated. Therefore, multi-level resource management techniques need to be adopted in edge computing. These resource management techniques can be applied independently and locally at the edge network level or in coordination with multiple edge networks and remote cloud providers. In this regard, resource discovery and synchronization requirements also arise because of the high mobility of user devices. NFV and software-defined networking (SDN) are potential technologies that can enable ease of management of edge resources.

Interoperability

Given the magnified heterogeneity of edge devices, protocols, and techniques, how will edge players interact and interoperate with each other?

Edge computing needs to provide interoperability and interactivity among multiple heterogeneous devices and service architectures. Heterogeneity is a major characteristic of edge computing that arises from the differences of device architectures, communication protocols, and network configurations. However, the edge should support heterogeneity. In the long term, edge computing has to define interoperability standards for applications and data exchange between its players. Moreover, M2M communication is an integral part of edge computing, and standard protocols that apply to all participating devices need to be devised. Interoperability is also essential in attracting sensors and IoT devices to the edge computing paradigm.

Business Model

How will fair sharing be ensured among edge players in the act of distributed resource utilization?

Multiple players are involved in the business model of edge computing: Internet service pro-

viders (ISPs) who own edge servers and network devices, cloud service providers that may or may not interplay, and end users/devices that may act as both client and server for edge services in an ad hoc environment. Edge computing needs a business model similar to that of the pay-as-you-go model of cloud computing. However, unlike cloud computing, the decentralized location and utilization of resources make the task difficult. To create a complete business model, how resources will be accounted and monitored need to be determined. Moreover, how edge players will divide the payment among themselves and how user devices who participate in the edge by advertising their spare resources will gain incentives also have to be determined.

Open Research Challenges

In this section, we highlight some of the most important challenges that impede the success of edge computing paradigms. The discussion of these challenges provides research directions for further investigation in the edge computing paradigms.

Seamless Edge Execution Handover

Seamless edge execution handover is vital in enabling the uninterrupted migration of execution between different edge servers when the mobile user is on the move. Mobility between different edge networks disrupts the execution when a mobile device moves across two different network coverage areas. In edge computing, if a mobile user moves away from the connected edge computational platform, the performance of the application degrades because of the increased communication latency. Also, local service providers do not allow the mobile user access to and use of their resources from outside of the network because of security concerns. This situation leads to a number of edge execution handovers. Seamless edge execution handover becomes a challenging task because of the non-deterministic mobility behavior of the user and intrinsic limitations of the wireless medium. Seamless edge execution handover is a vital research problem that needs to be addressed for the success of the edge computing paradigm. The seamless handover solutions from wireless networks, such as the one reported in [14], can be applied in designing mechanisms for seamless edge execution migration.

Monitoring, Accounting, and Billing

Monitoring of edge computing resources, accounting, and billing are necessary to ensure satisfactory QoS and properly charging the user for the offered services by the edge computing service provider. However, monitoring, accounting, and billing require a sustainable business model for the edge computing service providers. Designing such a business model is challenging from the research perspective because of the mobile nature of the user and limited scope of the service. Usually, a user accesses the edge services for a limited time, such as during lunch hours in a university cafe; this short-time access makes the business model relatively complex. Further, when the user moves and execution migrates from one edge platform to another, the division of the charges among the involved edge computing service providers raise new challenges for monitoring, accounting, and billing. These challenges require a business model that incorporates the various levels of charging granularity.

Lightweight Security and Privacy

Security and privacy are paramount issues that impede the success of edge computing [15]. The offloading of an application from the end user device to the edge server requires the transfer of data from the mobile device to the edge server. The data are exposed to intruders through security breaches. Lightweight security and privacy mechanisms are necessary because of the battery-powered nature of the end users' devices. Furthermore, the security and privacy solutions should also be agile to support the aim of edge computing by reducing execution time. The diversity of the environment and complexity of the problem make designing a reliable lightweight security solution a challenging research task.

Lightweight security and privacy mechanisms are necessary because of the battery-powered nature of the end users' devices. Furthermore, the security and privacy solutions should also be agile to support the aim of edge computing by reducing over execution time. The diversity of the environment and complexity of the problem make designing the reliable lightweight security solution a challenging research task.

Real-Time Data Processing

Providing the best services at the edge of a user network in a certain environment where load and data are growing at tremendous rates has become a real challenge. Recent statistics revealed that 20.8 billion devices will be connected to each other by the end of 2020.[1] This fast rate of connectivity among devices can cause different scalability issues in terms of functionality, administration, and load. Connectivity among a large number of devices results in a flood of data production that can make it difficult for the edge node to perform real-time processing. Moreover, the large amount of data can increase the load at the network edge, which can degrade the performance of applications that require low latency in terms of execution. To enable the scalable platform, research efforts have been carried out; however, these efforts are in their infancy, and considerable attention should be given to solving the scalability issues in the near future.

Social Collaboration

To enable social collaboration among different edge nodes belonging to different vendors to achieve a common goal regarding efficient analytics has become difficult. The main hurdles in social collaboration are standardization and competition. In a market where different companies offer different edge devices as service providers, social collaboration is not possible because of heterogeneous device architecture and contentious issues. The enabling of social collaboration among edge nodes is a challenging task that must be solved to enable efficient data analytics.

Conclusion

Edge computing extends cloud computing by bringing the services closer to the end user at the network edge. Although edge and cloud computing paradigms use similar attributes (multi-tenancy and virtualization) and mechanisms, the extension is a challenging task that brings several new challenges. This study is conducted with the aim of exploring the edge computing paradigm.

In this article, we highlighted the significance of edge computing by providing real-life scenarios

[1] http://www.gartner.com/newsroom/id/3165317

Although the edge computing paradigm has some distinguishing characteristics, such as dense geographical distribution, mobility support, location awareness, proximity, low latency, and context awareness, the paradigm is in its early stages of development. Hence, considerable attention should be given to addressing the challenges to facilitate the adoption of edge computing.

requiring strict constraints on application response time. Moreover, we categorized and classify the edge computing literature and devised a taxonomy based on relevant parameters. We identified and outline the requirements that need to be met to enable edge computing. Furthermore, several challenges that remain to be addressed are discussed as future research directions. Finally, we conclude that although the edge computing paradigm has some distinguishing characteristics, such as dense geographical distribution, mobility support, location awareness, proximity, low latency, and context awareness, the paradigm is in its early stages of development. Hence, considerable attention should be given to addressing the challenges to facilitate the adoption of edge computing, which will be a core component of the future computing landscape.

Acknowledgments

This work is supported by the Deanship of Scientific Research at King Saud University through Research group No. (RG # 1435-051). This work is also funded by the High Impact Research Grant from the University of Malaya under references UM.C/625/1/HIR/MOE/FCSIT/03 and RP012C-13AFR.

References

[1] E. Ahmed and M. H. Rehmani, *Mobile Edge Computing: Opportunities, Solutions, and Challenges*, 2017, pp. 59–63.
[2] M. Satyanarayanan, "The Emergence of Edge Computing," *Computer*, vol. 50, no. 1, Jan 2017, pp. 30–39.
[3] N. Kamiyama *et al.*, "Analyzing the Effect of Edge Computing on Reduction of Web Response Time," *IEEE GLOBECOM 2016*, Dec. 2016, pp. 1–6.
[4] X. Sun and N. Ansari, "Edgeiot: Mobile Edge Computing for the Internet of Things," *IEEE Commun. Mag.*, vol. 54, no. 12, Dec. 2016, pp. 22–29.
[5] M. Satyanarayanan *et al.*, "Edge Analytics in the Internet of Things," *IEEE Pervasive Computing*, vol. 14, no. 2, Apr 2015, pp. 24–31.
[6] B. P. Rimal, D. P. Van, and M. Maier, "Mobile Edge Computing Empowered Fiber-Wireless Access Networks in the 5G Era," *IEEE Commun. Mag.*, vol. 55, no. 2, Feb. 2017, pp. 192–200.
[7] F. Bonomi *et al.*, "Fog Computing and Its Role in the Internet of Things," *Proc.1st Ed. ACM MCC Wksp. Mobile Cloud Computing*, 2012, pp. 13–16.
[8] U. Shaukat *et al.*, "Cloudlet Deployment in Local Wireless Networks: Motivation, Architectures, Applications, and Open Challenges," *J. Network and Computer Applications*, vol. 62, 2016, pp. 18–40.
[9] A. Ahmed and E. Ahmed, "A Survey on Mobile Edge Computing," *2016 10th Int'l. Conf. IEEE Intelligent Systems and Control (ISCO)*, 2016, pp. 1–8.
[10] N. Kumar, S. Zeadally, and J. J. Rodrigues, "Vehicular Delay-Tolerant Networks for Smart Grid Data Management Using Mobile Edge Computing," *IEEE Commun. Mag.*, vol. 54, no. 10, Oct. 2016, pp. 60–66.
[11] Y. Mao *et al.*, "Power-Delay Tradeoff in Multi-User Mobile-Edge Computing Systems," *IEEE GLOBECOM 2016*, Dec. 2016, pp. 1–6.
[12] W. Hu *et al.*, "Quantifying the Impact of Edge Computing on Mobile Applications," *Proc. 7th ACM SIGOPS Asia-Pacific Wksp. Systems*, 2016, p. 5.
[13] C. Liu *et al.*, "A New Deep Learning-Based Food Recognition System for Dietary Assessment on an Edge Computing Service Infrastructure," *IEEE Trans. Services Computing*, vol. PP, no. 99, 2017, pp. 1–1.
[14] S. Fu *et al.*, "A Game Theory Based Vertical Handoff Scheme for Wireless Heterogeneous Networks," *Proc. 10th Int'l. Conf. Mobile Ad-Hoc and Sensor Networks*, 2014, pp. 220–27.
[15] I. Stojmenovic and S.Wen, "The Fog Computing Paradigm: Scenarios and Security Issues," *2014 Federated Conf. Computer Sci. and Info. Systems*, Sept 2014, pp. 1–8.

Biographies

Ejaz Ahmed [S'13, M'17] (ejazahmed@ieee.org) received his Ph.D. (computer science) from the University of Malaya, Kuala Lumpur, Malaysia, in 2016. He is an Associate Editor of *IEEE Communications Magazine*, *IEEE Access*, *KSII TIIS*, and *Elsevier JNCA*. His areas of interest include mobile cloud computing, mobile edge computing, the Internet of Things, and cognitive radio networks. He has successfully published his research in more than 40 international journals and conferences.

Arif Ahmed (arifch2009@gmail.com) received his M.Tech. degree in computer science and engineering (CSE) from the National Institute of Technology Silchar, India, in 2014, and his B.Tech. in information technology (2nd rank in the class) from Assam University, Silchar, India, in 2012. He worked as a visiting scientist at the Centre for Development of Advanced Computing, Mumbai, India, from August 2014 to February 2015. After returning from Mumbai, he taught as an assistant professor (contractual) in the CSE Department of the National Institute of Technology from July 2015 to May 2016.

Ibrar Yaqoob (ibraryaqoob@siswa.um.edu.my) received his Ph.D. degree in computer science from the University of Malaya, Malaysia, in 2017. He worked as a researcher at the Centre for Mobile Cloud Computing Research, University of Malaya. His research experience spans over more than three and a half years. He has published a number of research articles in refereed international journals and magazines. His numerous research articles are very famous and among the most downloaded in top journals. His research interests include big data, mobile cloud, the Internet of Things, cloud computing, and wireless networks.

Junaid Shuja (junaidshuja@ciit.net.pk) works as an assistant professor at Comsats Institute of Information Technology (CIIT) Abbottabad. He completed his Ph.D. from the University of Malaya in 2017 and his M.S. from CIIT Abbottabad in 2012. His primary research interest is code offload in heterogeneous execution architectures. Other research interests encompass topics like energy-efficient data centers, sustainable cloud computing, and the emerging edge computing paradigm.

Abdullah Gani (M'01, SM'12] (abdullahgani@ieee.org) is a full professor in the Department of Computer Systems and Technology, University of Malaya. He received his Bachelor's and Master's degrees from the University of Hull, United Kingdom, and his Ph.D. from the University of Sheffield, United Kingdom. He has vast teaching experience due to having worked in various educational institutions locally and abroad: schools, teaching college, the Ministry of Education, and universities. His interest in research started in 1983, when he was chosen to attend a scientific research course in RECSAM by the Ministry of Education, Malaysia. More than 150 academic papers have been published in conferences and respectable journals. He actively supervises many students at all levels of study — Bachelor, Master, and Ph.D. His research interests include self-organized systems, reinforcement learning, and wireless-related networks. He worked on mobile cloud computing with a High Impact Research Grant for the period of 2011–2016.

Muhammad Imran (cimran@ksu.edu.sa) is an assistant professor in the College of Computer and Information Science, King Saud University. His research interests include mobile ad hoc and sensor networks, WBANs, IoT, M2M, multihop wireless networks, and fault-tolerant computing. He has published a number of research papers in peer reviewed international journals and conferences. His research is financially supported by several grants. He is serving as a Co-Editor-in-Chief for *EAI Transactions on Pervasive Health and Technology*. He also serves as an Associate Editor for the *Wireless Communication and Mobile Computing Journal* (Wiley), the *Inderscience International Journal of Autonomous and Adaptive Communications Systems*, *Wireless Sensor Systems* (IET), and the *International Journal of Information Technology and Electrical Engineering*. He has served/serves as a Guest Editor for *IEEE Communications Magazine*, *IJAACS*, and the *International Journal of Distributed Sensor Networks*. He has been involved in a number of conferences and workshops in various capacities, such as Program Co-Chair, Track Chair/Co-Chair, and Technical Program Committee member. These include IEEE GLOBECOM, ICC, AINA, LCN, IWCMC, IFIP WWIC, and BWCCA. He has received a number of awards such as an Asia Pacific Advanced Network fellowship.

Muhammad Shoaib (muhshoaib@ksu.edu.sa) received his Ph.D. degree in communication and information systems from Beijing University of Posts and Telecommunications, China (2010). His areas of research include video compression techniques, multilayer video coding, commercial data center facilities, and IP packet-based networks, infrastructure, and security. He is currently working as an assistant professor in the College of Computer and Information Sciences (Information Systems Department) of King Saud University.

Cognitive Radio Network and Network Service Chaining toward 5G: Challenges and Requirements

Ioanna Kakalou, Kostas E. Psannis, Piotr Krawiec, and Radu Badea

ABSTRACT

Cognitive radio is a promising technology that answers the spectrum scarcity problem arising from the growth of usage of wireless networks and mobile services. Cognitive radio network edge computing will enhance the CRN capabilities and, along with some adjustments in its operation, will be a key technology for 5G heterogeneous network deployment. This article presents current requirements and challenges in CRN, and a review of the limited research work on the CRN cloud, which will take off CRN capabilities and 5G network requirements and challenges. The article proposes a cognitive radio edge computing access server deployment for network service chaining at the access layer level.

INTRODUCTION

Cognitive radio is a promising technology that answers the spectrum scarcity problem arising from the growth of usage of wireless networks and mobile services. Based on software defined radio (SDR), which can reconfigure its parameters (modulation, frequency, etc.) , it adds a cognitive cycle [1] in order to observe the environment, orient, plan, design, act, and learn from past experiences. Cognitive radio senses the spectrum for vacancies, so called "spectrum holes," for cognitive radio network (CRN) users to transmit. In the case of the licensed spectrum, when licensed users (primary users, PUs), vacate the spectrum, CRN users, also called secondary users (SUs), can access it. There are limitations to the interference SUs can cause to PU. On the other hand, the underutilized spectrum has resulted in an immense need for dynamic spectrum access, which exploits spectrum opportunistically.

Dynamic spectrum access includes, among other factors, sensing, spectrum management, spectrum sharing, and spectrum mobility. For spectrum sensing — PU detection — cognitive radio uses filter detection, energy detection, and feature detection. The spectrum management includes characterization, selection, and reconfiguration of the spectrum (channel, modulation, bandwidth, power, and transmission time). On appearance of the PU, the SU has to vacate the channel immediately and continue transmission in another vacant channel. Spectrum sharing is essential to avoid overlapping of multiple cognitive radios as well as handoff (loss of connection for a mobile SU or poor quality of service, QoS).

CRN uses machine learning, genetic algorithms, game theory techniques, knowledge representation, and optimization techniques for efficient resources allocation. Further, the CRN learns the network conditions [2] and encompasses past experiences to its cognitive cycle.

The cognitive radio uses the first open systems interconnection (OSI) layer (SDR) and second OSI layer (cognitive medium access control, MAC) basically, but actually relies on the whole OSI stack, and the decisions made in the CRN have to meet the whole network's needs. A high degree of interaction takes place within the CRN to achieve optimal network performance. Thus, this article considers cognitive radio cloud and proposes a cognitive radio edge computing architecture to expand the CRN's capabilities and performance while placing network service chaining at the access level as a key technology for increasing cognitive radio access diversity.The proposed solution would support 5G heterogeneous networks, and an analysis of challenges and requirements of 5G networks is provided to justify the former considerations and proposals within 5G. Although current research work on CRN cloud (CRNC) has started to emerge, this article goes beyond and bypasses the existing limitations in CRN with enhancements of radio access capabilities to respond to the vast needs of future wireless/mobile networks. The 5G example is presented.

The article is organized as follows. An introduction to mobile cloud and mobile edge computing is given; then we present the CRN requirements and challenges, the CRNC, and current research work in this field, and propose a server-based architecture for cognitive radio access for network service chaining at the access layer level. We give a brief discussion on the 5G requirements and challenges, while we combine the cognitive radio access network (RAN) service chaining solution to the 5G heterogeneous network deployment.

This article presents current requirements and challenges in CRN, and a review of the limited research work on the CRN cloud, which will take off CRN capabilities and 5G network requirements and challenges. The article proposes a cognitive radio edge computing access server deployment for network service chaining at the access layer level.

Ioanna Kakalou and Kostas E. Psannis (corresponding author) are with the University of Macedonia; Piotr Krawiec is with Warsaw University of Technology; Radu Badea is with Politechnica University of Bucharest.

Digital Object Identifier: 10.1109/MCOM.2017.1700086

0163-6804/17/$25.00 © 2017 IEEE

Mobile edge computing provides a highly distributed computing environment that can be used to deploy services and delay-sensitive and context-aware applications to be executed in close proximity to mobile users. This creates an ecosystem where new services are developed in and around the base station.

Mobile Cloud and Mobile Edge Computing

Mobile Cloud

The Mobile Cloud Computing Forum introduced cloud computing leveraging to the mobile network. "Mobile Cloud Computing at its simplest refers to an infrastructure where both the data storage and the Data Processing happen outside of the mobile device. Mobile cloud applications move the computing power and at a storage away from mobile phones and into the cloud, bringing applications and mobile computing to not just smart phone users but a much broader range of mobile subscribers" — Mobile Cloud Computing Forum (MCC-Forum, 2011).

There are several existing definitions of mobile cloud computing" and different concepts of the mobile cloud: applications run as thin clients to powerful remote servers on one hand, and on the other hand mobile devices may establish peer-to-peer connections locally with other powerful devices providing resources without the cost of latency and bandwidth issues. These systems are self-organized [3] and could offload jobs on local mobile resources. A cloudlet may be a cluster of multi-core computers connected to the cloud, and if it is not available, the mobile device will have to be served by the cloud. A virtual machine is built in the cloudlet to which the mobile devices connect as thin clients. Open issues are the distribution of processing, storage and networking capacity, the trade-off between QoS and cost for cloudlet providers, and security. The CloudClone is another implementation of local service infrastructure that creates a clone of an application. CloudClones do not virtualize native resources.

Mobile cloud has to address, besides the basic requirements of the cloud (i.e., scalability, availability, and self-awareness), the loss of connectivity, mobility, and power issues.

Cloud computing can serve mobile cloud in many aspects [3]:

- Extend battery life. Actually, remote application execution can save energy up to 45 percent for numerical computations.
- Improve data storage and processing power.
- Improve reliability.

Mobile Edge Computing and Network Service Chaining

Mobile edge computing (MEC) provides a highly distributed computing environment that can be used to deploy services and delay-sensitive and context-aware applications to be executed in close proximity to mobile users. This creates an ecosystem where new services are developed in and around the base station.

The work by the European Telecommunications Standards Institute (ETSI) on MEC-RAN aims to provide IT and cloud computing capabilities within the RAN. The key element of MEC is the MEC application server, which is integrated at the RAN element. The MEC-RAN provides computing resources, storage capacity, connectivity, and access to user traffic, radio, and network information.

Mobile edge computing allows cloud application services to be hosted alongside mobile network elements, and also facilitates leveraging of the available real-time network and radio information. The MEC-RAN delivers information from the radio network relating to users and cells, and is based on network-layer signaling messages. MEC-RAN also provides measurement and statistics information related to the user plane.

Multiple virtual machines (VMs) can be deployed in a single platform to share the hardware resources. Traffic can be routed to a VM from a physical interface and from a VM back to the physical interface. Cloud and virtualization technologies (network functions virtualization, NFV) are key enablers for MEC.

The ETSI MEC-RAN covers network layer signaling only and does not infiltrate to the lower layers. As a consequence, the traffic shaping service is a basic service.

Network service chaining is a key technology enabling automated provisioning of network applications with different characteristics. The "chain" in service chaining represents the services that can be connected across the network using software provisioning. New services can be instantiated as software-only, running on commodity hardware.

Network service chaining capabilities mean that a large number of virtual network functions can be connected together in an NFV environment. Because it is done in software using virtual circuits, these connections can be set up and torn down as needed with service chain provisioning through the NFV orchestration layer.

Cognitive Radio Network Cloud

The current challenges in CRN are storing large amounts of data and processing them in real time, and the exchange of nodes' current status on the fly. These challenges are in contrast to the limited storage and processing ability (plus battery lifetime) of cognitive devices; thus, the need for additional capabilities arises. Cognitive radio network cloud (CRNC) is an infrastructure consisting of mobile nodes and the cloud whose primary goal is to keep up-to-date status of the spectrum availability in the network for all (PUs and SUs) to access. The network status will be maintained in the cloud and updated by the network nodes. The need for intense and accurate sensing makes multiple-input multiple-output (MIMO) technology appropriate for cognitive radio. The large amount of sensing data and processing of MIMO antennae as well as the signal intelligence as a whole can be mitigated to the cloud. Current research on cognitive radio mobile cloud is limited. In the following paragraphs, a review of the existing research work in this field is presented along with the arising CRNC challenges and requirements.

A CRNC prototype, as proposed in [4],collects sensing data, processes them in real time, and provides the results to all nodes. Hence, CRNC should also be capable of running the cognition cycle for the network. The nodes will continually report their status to the cloud, and store and process their data and plan. Thus, the control message exchange between the mobile nodes will be eliminated, and only data transfer will occur.

In [4] the data transfer between mobile nodes A and B will occur after the cloud has reserved the resources in the multihop cognitive network path: $A \rightarrow C_i \rightarrow B$ (where C_i denotes all the rest of

the cognitive nodes in the path), or the data transfer will take place directly between node *A* and node *B* as soon as the necessary resources reservation has been made by the cloud. Another issue that will be answered by the cloud architecture is that there will be no data loss upon PU arrival.

Actually, there are two options for implementing data transmission:

- The cloud reserves the resources along the transmission path, and then transmission occurs between the wireless nodes without the cloud's interception.
- The data is sent to the cloud and then copied to the destination node.

In the latter case, there will be no data loss upon PU arrival as they will be stored in the cloud instead. The SUs' requests for spectrum access will arrive to the cloud in first come first served (FCFS) order, but policies can be applied on the queue for implementing QoS classes.

Overlapping,the hidden terminal node problem, and the exposed terminal problem will be avoided [5] as the cloud keeps the geolocation position of each node — the overlapping nodes will be well known for a given data transmission — while the handoff will be seamless. Common control channel (CCC) was the solution in ad hoc networks to handle the coordination and resource management between the nodes as well as the hidden terminal problem. When all the control messages of the network are transmitted via one channel, the network is vulnerable to congestion and attacks (there are protocols [2] that deal with this problem though). The cloud overcomes the CCC problem.

CRNC should cover both cognitive radio infrastructure networks and cognitive radio infrastructure-less networks (Fig. 1). Cognitive radio infrastructure-less networks, although they are self-organized and implement distributed resources allocation, still suffer from limited storage, processing ability, and power supply. Demanding tasks (e.g., signal intelligence) could be offloaded to powerful nodes locally, allowing the local network to be self-organized. A critical issue in CR Infrastructure-less network deployment on the cloud would be the standard interface operability for cognitive radio users to connect to the cloud or local powerful nodes.

The CRNC should accommodate databases for past experience information and databases for the sensing data. Cognitive radio uses the past experience to learn its environment and plan.The cognitive nodes will connect to cloud front devices playing the broker's role to provide their sensing reports as proposed in [4] or data for processing. Those devices will split data and the processing load to the intermediate cloud computers. There is a trade-off between the degree of parallelism and the data exchange. In [4] they use a scalable method to partition the geographical area according to the SUs' density in order to eliminate the processing time and then call the Map/Reduce; the time and location are the keys for Map and location the key for Reduce.The Sparse Bayesian Learning Algorithm is used in [4] to estimate the cooperative sensing outcome. The CRNC architecture in [5] includes the interface, the controller, the query processor, the database, and the knowledge database. A game theoretic resources allocation in the CRNC is presented in [6], where

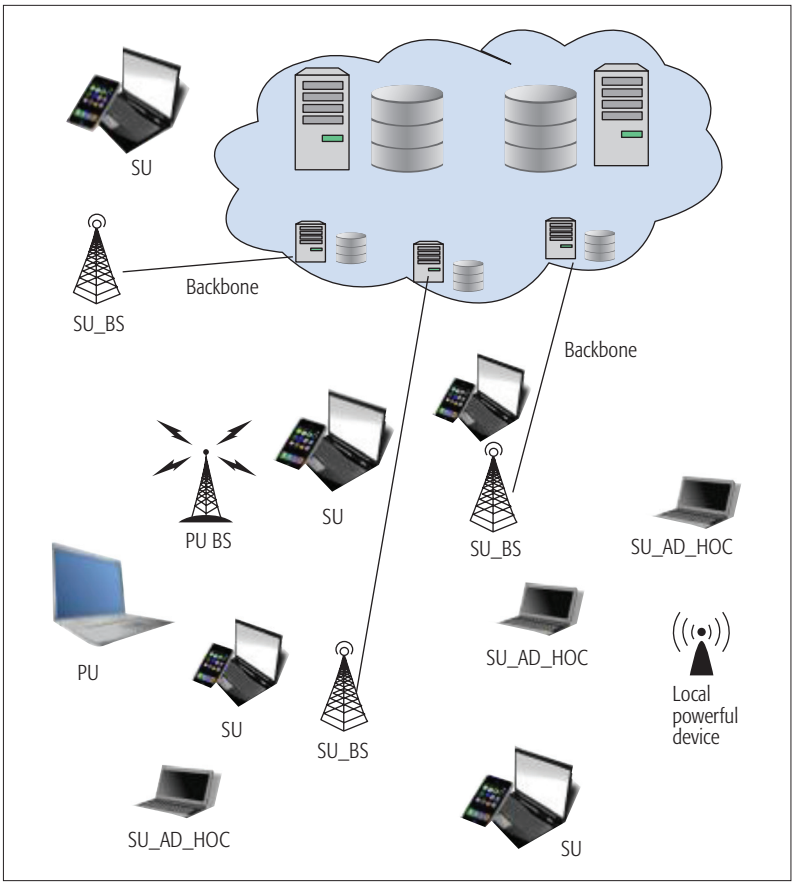


Figure 1. The cognitive radio network cloud.

the SUs adapt their power in a distributed manner and the greedy behavior is controlled by the cloud manager.

In [7] the geolocation of idle bands and the SUs' transmission requirements, such as data rate and timestamp, are reported to the cloud server. The decision regarding channel availability is made at the energy detection threshold and the bandwidth threshold. The cloud server reports the available channels to the nodes, which then select the channel that satisfies their transmission requirements best. The authors consider both device-to-device and device-to-infrastructure communication.The authors in [8] propose a cloud architecture for CRNs where the SUs are equipped with a GPS; sensing by the SUs is completely avoided.

The authors in [9] introduce powerful mobile devices that act as resource providers serving the CRN when the application data size and complexity is below a threshold; otherwise, they are served by the cloud. They have also developed a technology called MapReduce on Opportunistic Environments or Opportunistic Cloud to ensure job completion and good performance of MapReduce [10] by building a private cloud where dedicated nodes in the cloud supplement volatile wireless nodes (e.g.,in terms of jamming). Cooperative sensing and localization for power map reconstruction are proposed in [11, 12].

MIMO systems are capable ofachieving a capacity gain and/or increasing link robustness in CRN, but they increase processing time, energy

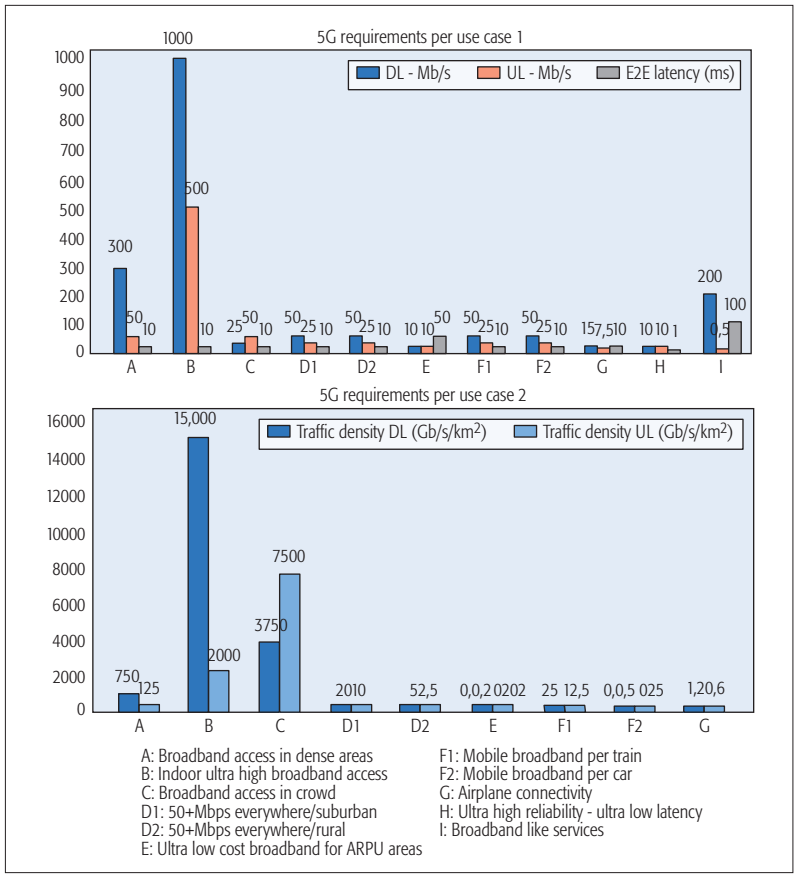


Figure 2. 5G requirements per use case.

consumption, and processed data amount. In [13] they propose a sub-optimal solution with parallel QR-factorization algorithms to establish an adaptive transmitter system by dynamically selecting the antennae — making use of the parallel computing of the cloud.

The necessity for a cognitive MAC on the cloud arises: as the PUs can utilize the spectrum any time,continuous sensing and storage of the huge amount of sensing data as well as real-time processing are required. An MEC architecture and CRN virtualization would make feasible lower-layer services and applications that could decrease latency, and increase the quality of experience (QoE) and security. Lower-layer cognitive radio services and applications on the edge computing would increase capabilities not only within the RAN but within the local mobile network on a peer-to-peer basis for access and backhauling. In the latter case, ad hoc networks or vehicles would leverage powerful local nodes, allowing them to be self-organized. A proposal for such an architecture is presented in Fig. 3 with the cognitive radio edge computing (CREC) server to offload storage and processing at the RAN or at the powerful local nodes in the form of access services provisioned for the CRN by virtualizing the lower layer's functionality and resources and leveraging the connection to the core network and the cloud connection.

We can distinguish three parts of the CREC access server:

- The basic one covers the lower layers' functionality and resource virtualization (i.e., SDR and resources, which are infrastructure-oriented).
- The application platform supports services such as local network access control and handover for the applications that would respond to, say,QoS.
- Each VM with an application will run at the SU node.

Network service chaining allows the creation of new access capabilities and is performed at the application platform. The server may run at a RAN or at a powerful local/wireless node to which other nodes can connect on a peer-to-peer basis (device-to-device communication), reducing latency. The cognitive MAC is available as a service and adapts to the wireless nodes' requirements. The SU node will run the corresponding VM for each cognitive MAC service and its application. Virtual machines communicate via the application platform that runs on the server. The server connects to the cloud for further support. The proposed architecture is flexible, reduces latency, and is easily adaptable as more services and applications are adjusted in a simple way.

Services can connect through the network composing powerful network service chaining for CRNs, controlling access and providing high-level QoE to the network users.

In Fig. 4 we can see a CREC server being part of the RAN and connected to the core network and the cloud, and in the second case being part of powerful nodes operating locally, for example, for an Internet of Things (IoT) network. In Fig. 5a we can see the service, application registration, and data sending for computations and processing to the CREC server operating at a powerful node of local network A. Later, the CREC server decides to initiate the handover process for the SU and sends a handover request to local network B. The SU is notified and registers with network B. In Fig. 5b the server operates at the RAN, and the SU registers its service and application and sends data to the server for processing. The server processes the amount of data that are not computationally intensive, and the rest are passed to the cloud for processing. Later, a handover process is initiated, and the request is passed to the cloud (e.g.,to update the network topology database of the CRN).

5G: Challenges and Requirements

Mobile networks will become the primary means of network access for person-to-person and person-to-machine connectivity where access to information and data sharing are possible anywhere, anytime. An increasingly diverse set of services, applications, and users with diverse requirements and flexible spectrum use of all non-contiguous spectrum will also characterize 5G technology. A vastly diverse range of things (IoT) will be connected, which implies new functions to be developed. Millions of low-cost connected devices and sensors that need to operate on batteries would require low energy consumption reduced by a factor of 1000 to improve connected device battery lifetime. A thousand times today's traffic volume will be supported in an affordable, sustainable way, cost- and energy-efficiently.

Next generation wireless access networks will need to support fiber-like data rates at 10 Gb/s to make possible ultra-high definition visual com-

munications and immersive multimedia interactions and support mobile cloud service; 100 Mb/s should be generally available, while 1 Mb/s should be the baseline everywhere (Fig. 2).

Ultra large data rates, latency of 1 ms, always-on users per cell reaching several millions, and signaling loads to almost 100 percent will be included in performance requirements. Other challenges for a 5G network are: less than 1 ms latency for real-time mobile applications and communications, and maximum 10 ms switching time between different radio access technologies for seamless delivery of services. On the quest for efficient usage of radio link, modulation techniques like nonlinear multiple-user precoding, joint modulation and coding, physical network coding, and advanced physical layer adaptation are being tested. For example, non-orthogonal multiple access (NOMA), which is an intra-cell multi-user multiplexing scheme using the power domain, and faster than Nyquist (FTN) are included in research efforts. Air interface and RAN will accommodate massive capacity, extremely large amount of connections, and high speeds for new network deployments. Latency reduction will improve user experience, so techniques such as pre-scheduling, local gateway, local breakout, local server, local cache, shortened transmission time interval (TTI), faster decoding, and QoS will control network delay, backhaul delay, radio access delay, and terminal delay. The new radio access technology with new numerology — wider subcarrier spacing — will achieve shortened TTI and thus reduced latency to 1 ms.

Advanced antenna solutions with multiple elements (massive MIMO) including beamforming and spatial multiplexing will achieve high data rates and capacity. Massive MIMO technologies experience small interference and consequently higher throughput.

5G, unlike previous mobile network technologies will have to not only proceed to flexible and efficient use of available non-contiguous spectrum, but also extend the operation range for wireless access into higher frequencies above 10 GHz (the spectrum from 10–100GHz, i.e., the millimeter-wave range is considered so that multi-gigabits-per-second data rates are feasible). Advances in waveform technologies, multiple access, coding, and modulation would improve spectral efficiency so as to support scalability of massive IoT connectivity and decrease latency. Computationally intensive and adaptive new air interfaces are necessary. Single-frequency full-duplex radios will increase spectrum efficiency, reduce network cost, and increase energy efficiency. Plug-and-play will be essential in deployment, allowing nodes to self-organize spectrum blocks for access and backhauling.

The extension of mobile devices' capabilities will be necessary for device-based on-demand mobile networking for services like device-to-device communications. Advanced device-to-device communication would enhance spectrum efficiency and reduce latency as the offloading of network data locally will minimize processing cost and signaling. A single radio resource could be reused by different groups of users of the cellular network if the interference occurring within those groups is tolerable. Advanced small cell technology will utilize higher frequency bands; taking advantage of the vast bandwidth makes it suitable for dense small cell deployment, where massive MIMO will be essential. Furthermore, user-centric virtual cells that consist of a group of BSs are introduced for 5G. In-band wireless backhaul can be used between the BSs for cooperative communication, reducing cost and complexity for the network backhaul.

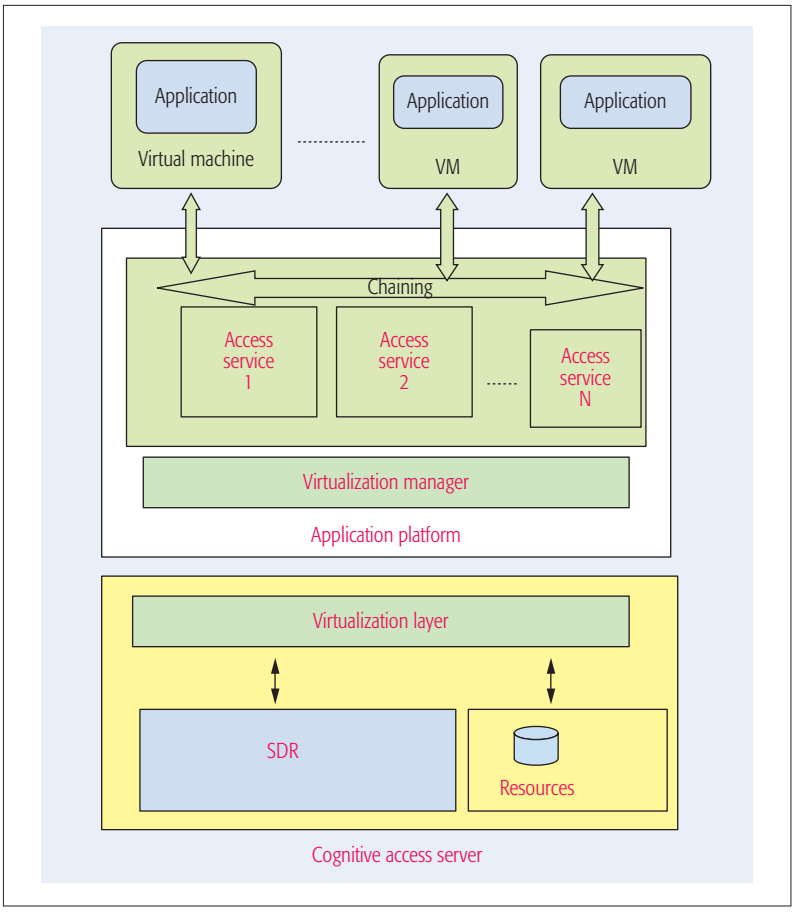


Figure 3. The cognitive radio access server.

Service requirements need to be mapped to the best combinations of frequency and radio resources by spectrum access and programmable air interface technologies. Software defined networking and cloud architectures will enable customization of mobile technologies and QoS guarantees. Cloud computing will allow leveraging of new services and applications and provide on demand processing, storage, and network capacity. The cloud will enable seamless connections between people and between humans and machines, and will coordinate network resources for inter-RAT, inter-frequency, inter-site radio access for efficient network management. Virtualization and SDN are technologies that can simplify and optimize the 5G network. Multi-radio access technologies (RAT) convergence and intelligent management will lie on the cloud. Not only that but in the 5G, network capabilities such as bandwidth, latency, and QoS will be configurable, allowing access to a wider range of services. The 5G network will also integrate existing and heterogeneous networks with diverse requirements.

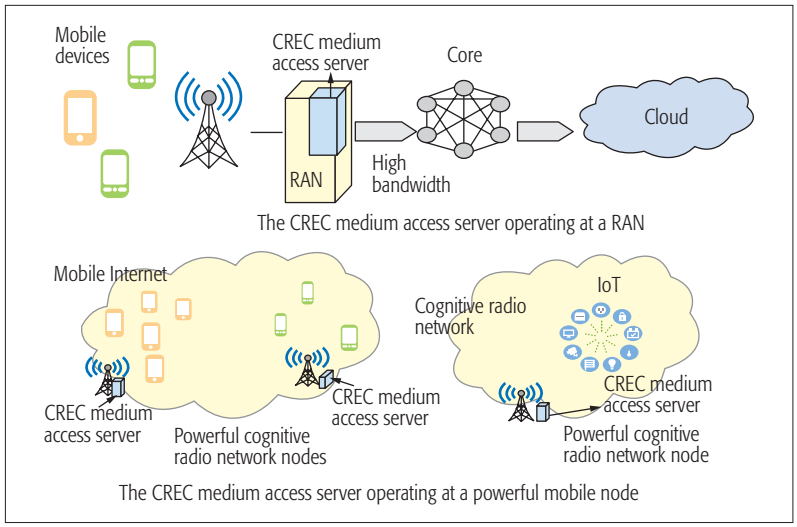


Figure 4. The CREC medium access server deployed on the RAN and locally

CRNC for 5G Heterogeneous Networks

In general, massive MIMO is an evolving technology of next generation networks, which is energy-efficient, robust, secure, and spectrum-efficient [14]. According to [14], massive MIMO technology would:

- Improve energy efficiency by 100 times and capacity on the order of 10 and more
- Be put together with the help of low-power and less costly components
- Decrease latency on the air interface
- Encompass a simple multiple access layer
- Increase the signal strength against interference

The proposed cognitive radio access server (Fig. 3), which is supported by the cloud, would be an appropriate architecture for fast processing of the computational load of massive MIMO technology.

There are mainly two spectrum sharing techniques that enable mobile broadband systems to share spectrum in 5G: distributed solutions and centralized solutions. Distributed spectrum sharing techniques are more efficient as they can take place in a local framework. Besides the centralized and distributed spectrum sharing considerations, cognitive radio with dynamic spectrum management will enhance the network and application performance in 5G [14]. The proposed cognitive radio access server would accommodate centralized solutions and distributed solutions at local powerful nodes (Figs. 4 and 5). Access and backhauling convergence would easily be deployed. Furthermore, full-duplex cognitive radios will be empowered to support 5G.

If a device links directly to another device or apprehends its transmission through the support of other devices, it will be on the device level (device-to-device communication). Thus, the combined resources of the numerous mobile devices and other available stationary devices in the local area will be exploited. This method supports user mobility and identifies the potential of mobile clouds to perform collective sensing [14]. Cognitive radio access as a platform service will enable the 5G network to accommodate heterogeneous networks with diverse requirements, such as small cell dense environments when the cognitive radio access server runs the adaptive access services in the local vicinity and wireless backhaul between base stations for cooperative communication. Network service chaining will be realized, enabling end users to make the best choices, introducing high-level QoE based on the enhanced air interfaces and capabilities of 5G, signifying a new era in network infrastructures.

Conclusion

This article makes a revision of cognitive radio network requirements and challenges — including cognitive radio network cloud, mobile edge computing, and network service chaining — and provides a review of current research work on CRNC that will support all the rest. Distributed resource/spectrum management (devices as resource providers), centralized resource/spectrum management (cloud), and processing offloading would be easily feasible with the proposed cognitive radio edge computing access server paradigm. Furthermore, a cognitive radio access server will

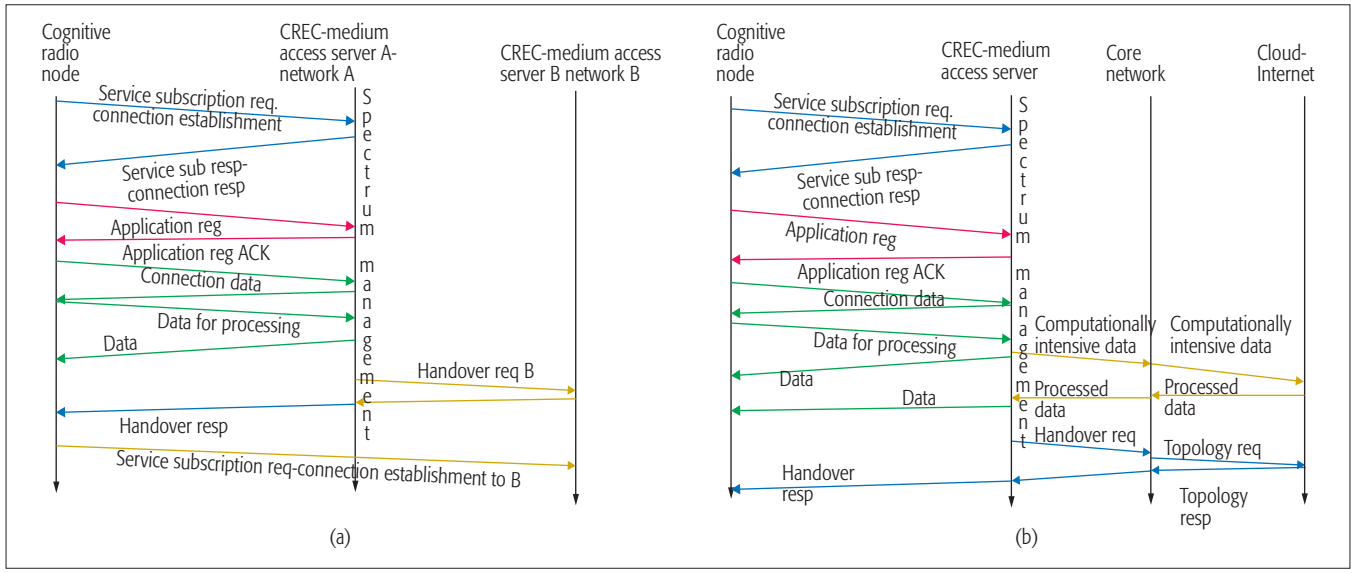


Figure 4. CREC— medium access server operating: a) at a powerful local node; b) on the RAN.

support the 5G heterogeneous network operating as a platform service providing radio access in the case of a high-capacity backhaul when cloud support is needed. Otherwise, resource/spectrum management will be performed locally, effectively enabling network service chaining in end-user-oriented mode in the diverse wireless environment of the 5G network. Thus, bypassing its limitations, the cognitive radio network will respond to the vast needs of future wireless/mobile networks; the 5G example has been presented.

References

[1] J. Mitola and G. Q. Maguire, Jr., "Cognitive Radio: Making Software Radios More Personal," *IEEE Personal Commun.*, vol. 6, no. 4, Aug. 1999, pp. 13–18.

[2] I. Kakalou *et al.*, "A Reinforcing Learning-Based Cognitive MAC Protocol," *IEEE ICC 2015*, London, U.K., 2015, pp. 5608–13.

[3] H. D. Thai *et al.*, "A Survey of Mobile Cloud Computing: Architecture, Applications and Approaches," *Wireless Commun. and Mobile Computing*, vol. 13, no. 18, 2013, pp. 1587–1611

[4] C. H. Co, D. H. Huang, and S.-H. Wu, "Cooperative Spectrum Sensing in TV White Spaces: When Cognitive Radio Meets Cloud, Workshop on Cloud Computing," *IEEE INFOCOM 2011*, 2011, pp. 683–88.

[5] B. Y. Reddy, "Solving Hidden Terminal Problem in Cognitive Networks Using Cloud Technologies," *Proc. 6th Int'l. Conf. Sensor Technologies and Applications*, 2012, pp. 235–40.

[6] D. B. Rawat, S. Shetty, and K. Raza, "Game Theoretic Dynamic Spectrum Access in Cloud-Based Cognitive Radio Networks," *Proc IEEE Int'l. Conf. Cloud Engineering 2014*, 2014, pp. 586–91.

[7] D. B. Rawat, "ROAR: An Architecture for Real-Time Opportunistic Spectrum Access in Cloud-Assisted Cognitive Radio Networks," *Proc.13th IEEE Annual Consumer Commun. & Net. Conf. 2016*, Las Vegas, NV, 2016.

[8] D. B. Rawat *et al.*, "Cloud-Assisted GPS-Driven Dynamic Spectrum Access in Cognitive Radio Vehicular Networks for Transportation Cyber Physical Systems," *Proc. IEEE Wireless Commun. and Networking Conf. 2015*, New Orleans, LA, 2015.

[9] F. Ge *et al.*, "Cognitive Radio Rides on the Cloud," *Proc. MILCOM 2011*, 2011, pp. 1448.

[10] H. Lin *et al.*, "MOON: MapReduce on Opportunistic Environments," *Proc. ACM Int'l. Symp. High Performance Distributed Computing 2010,* Chicago, IL, 2010, pp. 95–106.

[11] S.-H. Wu *et al.*, "A Cloud Model and Concept Prototype for Cognitive Radio Networks," *IEEE Wireless Commun.*, vol. 19, no. 4, 2012, pp. 49–58.

[12] D. Huang, S.-H. Wu, and P.-H. Wang, "Cooperative Radio Source Positioning and Power Map Reconstruction: A Sparse Bayesian Learning Approach," *IEEE Trans. Vehic. Tech.*, vol. 64, no. 6, 2015, pp. 2318–32.

[13] S. Y. Chang and H. C. Wu, "Adaptive Antenna Selection by Parallel QR-Factorization for Cognitive Radio Cloud Network," *Proc. IEEE GLOBECOM 2014, Cognitive Radio and Networks Symp.*, Austin TX, 2014, pp. 882–87.

[14] A. Gupta and R. K. Jha, "A Survey of 5G Network:Architecture and Emerging Technologies," *IEEE Access*, vol. 3, 2015, pp. 1206–32.

[15] J. M. Batalla *et al.*, "ID-Based Service-Oriented Communications for Unified Access to IoT," *Computers & Electrical Engineering*, Elsevier, vol. 52 , no. C, 2016, pp. 98–113.

Biographies

IOANNA KAKALOU received her Diploma of Computer Engineering and Informatics from the Computer Engineering and Informatics Department of the Engineering School of the University of Patras, Greece. She received her Master's degree in communications systems and technologies from the Computer Science Department of Aristotle University, Thessaloniki, Greece. She is currently working on her Ph.D. at the University of Macedonia, Thessaloniki, Greece. Her research interests cover cognitive radio and cognitive radio network cloud.

KOSTAS E. PSANNIS [M] (kpsannis@uom.edu.gr) received a degree in physics from Aristotle University of Thessaloniki and his Ph.D. degree from the Department of ECE of Brunel University, United Kingdom. In 2001 he was awarded the British Chevening scholarship, and in 2006 a research grant by IISF, Japan. He is an assistant professor in the Department of Applied Informatics, University of Macedonia. He is serving as an Associate Editor for *IEEE Access* and *IEEE Communications Letters*.

PIOTR KRAWIEC received his M.Sc. and Ph.D. degrees in telecommunications from the Warsaw University of Technology in 2005 and 2011, respectively. Since 2012, he has been an assistant professor with the Department of Internet Architectures and Applications, National Institute of Telecommunications, and the Institute of Telecommunications, Warsaw University of Technology. His research areas include IP networks (fixed and wireless), future Internet architectures and applications, and prototyping and testbeds.

RADU BADEA is with the Telecommunications Department, Electrical Engineering at Politechnica University of Bucharest, Romania.

Cognitive Radio Access as a platform service will enable 5G network to accommodate heterogeneous networks with diverse requirements e.g., for small cell dense environments when the Cognitive Radio Access Server runs the adaptive access services on the local vicinity and wireless backhaul between BSs for cooperative communication.

Computing, Caching, and Communication at the Edge: The Cornerstone for Building a Versatile 5G Ecosystem

Evangelos K. Markakis, Kimon Karras, Anargyros Sideris, George Alexiou, and Evangelos Pallis

The authors present a unified communication, computing, and caching (3C) solution for the upcoming 5G environment that will allow service, content and function providers to deploy their services/content/functions near the end users; to allow network providers to virtually deploy their connectivity services over commodity hardware; and to enable end users to renounce their role as passive 5G stakeholders and become active ones by offering their 3C resources to the 5G ecosystem.

ABSTRACT

This article presents a unified computing, caching, and communication (3C) solution for the upcoming 5G environment that will allow service, content, and function providers to deploy their services/content/functions near the end users (EUs); to allow network providers to virtually deploy their connectivity services over commodity hardware; and to enable end users to renounce their role as passive 5G stakeholders and become active ones by offering their 3C resources to the 5G ecosystem. In this direction, we foresee the exploitation of a peer-to-peer-like middleware/app solution that upon installation will enhance the end user devices with the ability to form virtual fogs capable of providing their 3C resources to the 5G ecosystem. Additionally, we propose the introduction of heterogeneous nodes (e.g., FPGAs and GPUs) at the networks edge, which will boost the processing capabilities without paying a premium in power consumption. This will enable efficient and thorough filtering of the information that makes it all the way up to the cloud. In summary, this article proposes an architecture that exploits and advances the edge and extreme edge 3C paradigms toward enabling the 5G ecosystem to meet its own criterion for low end-to-end latencies and, as such, enable it to provide and sustain high QoS/QoE levels.

INTRODUCTION AND CONTEXT

The forthcoming emergence of the Internet of Things (IoT) in everyday life, the ongoing explosion of media services, and the continuously increasing number of connected users pose a significant challenge to current information and communications technologies (ICT) architectural and operational paradigms, especially when considering the need to handle and accommodate the colossal amount of data generated at the network edge. This challenge is further exacerbated by the current highly heterogeneous and immensely fragmented network environment.

In this context, 5G is viewed as the key network technology that will meet the aforementioned challenge and allow for the realization of a "hyper-connected society" where billions of users, via their respective end devices (e.g., wearables, smart home appliances, connected cars, smart phones, laptops), and machines, will be able to enjoy services characterized by high data rates at the network edge (1–10 Gb/s) and ultra-low end-to-end latency (~1 ms).

These ambitious requirements are expected to be at least partially satisfied by exploiting next-generation radio access technologies, faster network hardware, and densification of the edge networks, primarily of the wireless ones (e.g., macro/small cell dense wireless networks). However, these alone may not be enough, especially when taking into account the plethora of data expected to be generated at the network edge. These data often require processing and storing, actions that when done exclusively in the cloud can congest the edge's backhaul links, increase end-to-end latency, and thus have a severe impact on the overall quality of experience (QoE).

To mitigate this, edge computing and edge caching have been proposed as solutions as with these approaches both data processing and storing takes place near the edge where the data will be predominately generated and consumed. An example of these is the installation of caching (FemtoCaching) and processing units in base stations and access points (APs). However, the installation and maintenance of these hardware units incurs extra costs to mobile operators. A more promising alternative derives from the concept of "edge cloudification" that envisages the realization of "small" clouds, named fogs, operating at the network edge, able to accommodate the computing and storing needs of the local area. In this way, the sharing of the fog's computing and caching resources can yield significant gains in terms of capital and operating expenses, flexibility in ownership models, and statistical multiplexing.

In addition to the fog, two other concepts are anticipated to play a vital role in achieving a unified networking infrastructure, an important requirement for the fifth generation's (5G's) success: software-defined networking (SDN) and network functions virtualization (NFV). NFV, which can be viewed as the "cloudification" of the network itself, has the power to utilize the edge's communication resources by virtualizing entire classes of network node functions. This, along with SDN's capabilities for efficient network management, can enable future 5G operators to set up services quickly

Digital Object Identifier:
10.1109/MCOM.2017.1700105

Evangelos K. Markakis, George Alexiou, and Evangelos Pallis are with the Technological Educational Institute of Crete; Kimon Karras and Anargyros Sideris are with Future Intelligence Ltd.

0163-6804/17/$25.00 © 2017 IEEE

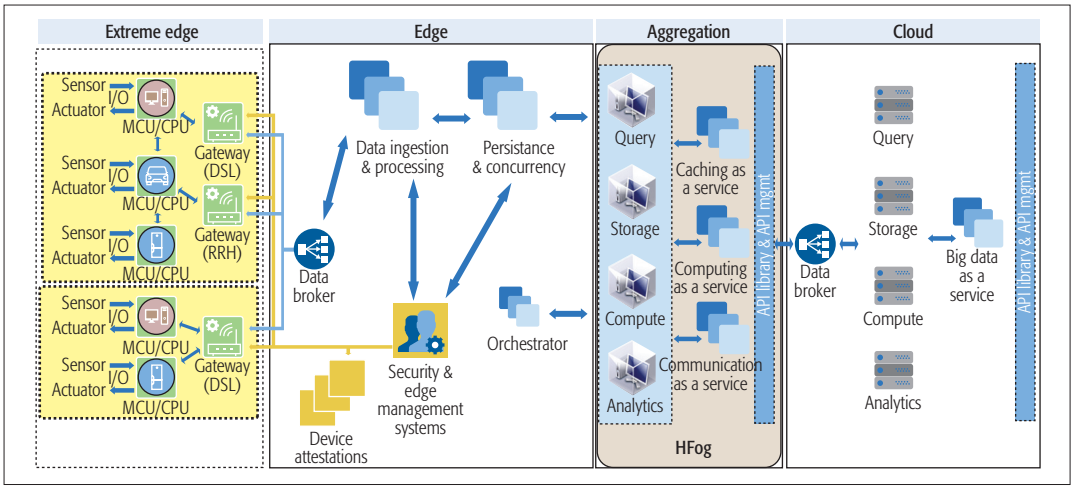


Figure 1. High-level system architecture.

and move them around as virtual entities (e.g., virtual machines (VMs), containers, unikernels) in response to dynamic network demands.

Even so, edge computing, caching, and communication (3C) could face a hard challenge in coping with the requirements of 5G-grade applications and services (augmented/virtual reality, 3D gaming, video processing, etc.) that demand extensive computation, short proximity caching, and fast transfer of the collected data. On the other hand, it is hard to miss that the vast number of end devices — such as smartphones, desktops, laptops, and smart TVs, most of them utilizing some form of processing power, storage space, and network connectivity — could constitute a pristine "ocean" of 3C resources. These resources remain unexploited so far and can form the backbone of a new architecture that can radically reshape the next generation network (NGN) field by enabling a solution that allows for the exploitation of these "extreme edge" resources toward meeting 5G's low-latency and high-throughput criteria.

In this context the architecture presented within this article aims to design, develop, and deploy a unified 3C solution for the 5G environment that will decidedly alter the existing network paradigm by extending service function chaining both horizontally and vertically to encompass all layers of a multi-tier architecture as well as span across the same layer to different subnetworks. Furthermore, it will enhance service provisioning and orchestration by introducing heterogeneous compute resources across the various layers. These resources will be based on programmable logic, which will enable them to attain higher performance with significantly lower power consumption. To accomplish this, it will introduce cross-layer, heterogeneous orchestration, a new multi-tiered network architecture based on a minimum of three layers (edge, fog, and cloud), and a mechanism to build isolated fog private subnetworks called virtual fogs, which can deliver specific services and tie them together into hyper fogs if such a need arises. The result is a highly pliable architecture, which can be sliced and diced to meet the demands of 5G services. The work presented herein is based on the EXEGESIS architecture, which was introduced in [1] and has been significantly refined and augmented with architectural elements and simulation, proving the efficacy of the proposed paradigm.

The architecture envisages a 5G ecosystem where the exploitation of fog installations at the network edges will set the basis for low end-to-end latency by providing the necessary 3C resources needed to deploy and maintain the 5G grade services close to their consumers. Moreover, identifying the ongoing need for extra processing, storage, and bandwidth capacity, it goes a step further and allows for the exploitation of the extreme edge resources by enabling their "owners" to cooperatively form virtual fogs (vFogs); in other words, it empowers the end users with the ability to become 3C resource providers to the 5G ecosystem. In this direction, the architecture foresees the exploitation of a peer-to-peer (P2P)-like middleware/app solution that upon installation will enhance the end-user devices with the ability to coordinate and form a vFog. More specifically, the solution will follow a hybrid P2P approach where one of the devices will undertake the role of a super node (SN) for performing the required managerial tasks.

To facilitate the use of the resources, each fog, based on its physical infrastructure capacity, and each vFog, based on the resources offered from the participating end users, creates an abstract pool of the available 3C resources, which in turn can be accessed from any other stakeholder. Furthermore, each (v)Fog will fully exploit the SDN and NFV paradigms for managing its resources. In order to address the complexity of having to interact with many different resource pools in the process of locating and reserving resources, we introduce the afore mentioned hyper fog (hFog) aggregation plane, which aims to unify the underlying (v)Fog's 3C resources in one common pool.

Background and Related Work

Concept

The proposed paradigm empowers 5G stakeholders to become 3C resource providers by enabling them to cooperatively form vFogs. Figure 1 illustrates the architecture, where the edge and fog devices compose a vFog with the main goal to collect the offered resources from the devices' owners and create an abstract pool of them. This pool can in turn be accessed by other participants. In this way, the architecture enables the entities operating at the edge of the 5G network to access and exploit the so far unutilized 3C resources of the edge devices.

The forthcoming emergence of IoT in everyday life, the ongoing explosion of media services, and the continuously increasing number of connected users pose a significant challenge to current ICT architectural and operational paradigms, especially when considering the need to handle and accommodate the colossal amount of data generated at the network edge.

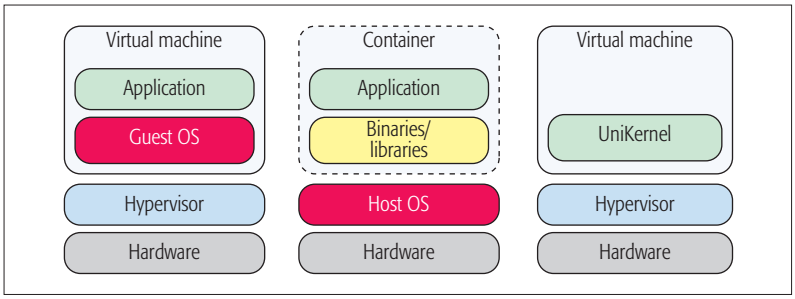


Figure 2. VMs vs. containers vs. UniKernels.

Virtual fogs will be the workhorse for carrying computationally intensive tasks, supporting higher caching sizes, and providing high-capacity communication links between the nodes and the cloud. In this configuration, the data generated at the extreme edge can be sent to the associated fog(s) for processing and storing. A data broker inside the fog assigns data, as well as the operations to be performed to them, to specific virtual instances (VIs). When the demand for 3C resources exceeds the available fog capacity, the responsible orchestration entity will delegate these demands to other v- or hFogs or the cloud. Finally, the edge assumes the responsibility for validating the end devices comprising the vFog in order to avoid malicious attacks, frauds, and so on.

Technical Approach

Communication: Fog networks are highly heterogeneous by nature. Therefore, their management or even implementation can be a challenging task, especially for large-scale deployment of IoT devices and applications. Emerging techniques, such as SDN and NFV, can be exploited toward easing implementation and management of fog networks, increasing network scalability and reducing operational costs. Even so, there are still pending issues to be resolved, including the following:

1 How do we design a distributed SDN system that meets the strict requirements of fog computing for latency, scalability, and mobility?

2 How do we achieve high performance of virtualized network appliances in vFog networks? More specifically, how can the vFog network provide the required throughput and latency to the deployed virtual appliances, and how can efficient instantiation, placement, and migration of virtual appliances in a dynamic network be achieved?

This work tries to answer these questions by proposing a novel architecture consisting of multiple, interweaved planes. At the lowest level, the extreme edge plane tries to provide an open solution that will conform to the hybrid P2P approach where a peer can be *primus inter pares*.[1] P2P is primarily an end-user technology that fosters self-deployment and self-organization while it achieves optimized resource utilization for the deployed applications and services even in large-scale deployments. In other words, P2P technology has succeeded where network management mechanisms have failed.

Extreme edge nodes together with intermediate-level gateways form the fog layer. Fog layer nodes can be grouped together into virtual clusters called vFogs. A vFog network relies on a number of heterogeneous devices, named vFog-nodes, that can vary from low-end devices, such as set-top boxes and APs, to high-end ones such as Cisco's IOx.

The fog network nodes can be classified into two sets of peers[2] (Fig. 2), regular extreme edge nodes (EENs: RENs) and super EENs (SENs). A REN can be any end device having a "pinch" of processing and communication capabilities that will allow the deployment of the P2P-like middleware on it and thus transform the device to a fully operational vFog node. A REN is able to interact with its corresponding SEN, first to inform it about the device's available resources and second to receive and carry out the assigned networking tasks.

A SEN is a REN on steroids that possesses the ability to:

- "Manage" and "manipulate" the vFog network
- Record the node churn ability
- Locate reliable peers that are able to provide information as a service upon request

A SEN plays two roles inside the ecosystem, namely the role of the EEN's intra-manager and the role of the (v)Fog orchestrator. As an intra-manager, a SEN:

- Oversees the formation of the EEN network by performing operations such as the (de) registering of EENs
- Queries the registered EENs about their state and their available resources
- Creates a logical topology of the EEN network along with a virtual pool of the RENs' available resources
- Enforce SDN policies in SDN-enabled EEN nodes

Following this hybrid P2P paradigm, a SEN is elected from the currently running RENs taking into account several attributes, including processing and memory capabilities, network capacity, power level/type, and so on. Acknowledging that the uncontrolled participation of EENs in the election process could pose security threats, the architecture will provide a means of "screening" the candidates list based on the stakeholder's policies.

The tremendous number and vast heterogeneity of the devices living at the edge of the network pose a significant challenge toward forming manageable and efficiently operating EEN networks. To handle this challenge, a middleware solution that will sit on top of each device's operating system (OS) will be developed and deployed. This middleware will utilize a southbound application programming interface (API) for interacting with the OS and acquiring access to the device's actual resources and a northbound API for communicating with its (v)Fog orchestrator. A hypervisor will be exploited for deploying in containerized form — it reduces the system's footprint and increases services deployability — the REN/SEN module, and, if assigned from the (v)Fog orchestrator, other software units that carry out computational tasks or realize a service.

Caching: This work anticipates that caching at the (extreme) edge will be an enabler for meeting 5G's low latency requirement, an important factor for provisioning and preserving high quality of service (QoS) and QoE levels to all the 5G stakeholders. There are three spatial categories of caching in our architecture: extreme edge caching (EEC), edge caching (EC), and cloud caching (CC). This work focuses on the first two

[1] First among equals.

[2] A peer can be an either an IoT or a regular end device such as a machine, drone, or vehicle.

categories, whereas the last one is viewed as a complementary option to the former two. From a temporal point of view, caching can be classified in reactive (caching data based on request) and proactive (prefetching data) with the latter being more suitable for achieving higher spectrum efficiency (SE) when prediction errors are kept close to zero. This architecture will support both approaches, while giving emphasis to the proactive ones for their better SE.

EEC takes place at the nodes of the vFogs, namely the RENs and SENs. In more detail, each vFog's SEN collects from its associated RENs the offered caching resources (i.e., available space) along with information about the cache type (e.g., solid state drive, hard disk drive). The collection process could be performed by both pushing (REN-initiated) and pulling (SEN-initiated) methods via the APIs that will be provided by the middleware. In this way, a SEN has the necessary information for creating an abstract pool of resources (APR). EEC is about storing services data (e.g., video content) in the vFog node(s) that offer storage capacity. This is not a trivial issue, as many parameters have to be taken into consideration, like the content's specific characteristics (e.g., size and encoding type), the content's popularity, the consumer's profile, the user mobility pattern, and, of course, caches' explicit attributes like size, type, and proximity, along with implicit ones like the available network capacity between the cache and the origin of the service/content. Taking all these into account, substantial research effort is being put on finding optimal ways to exploit both reactive and proactive caching schemes for caching data to the vFogs. In this direction, big data analytics at the fog and cloud levels is anticipated to play a crucial role in minimizing the prediction uncertainty and thus allowing the selection of the most spectrum-efficient proactive caching method.

The implementation of the edge caching layer will be based on the exploitation and enhancement of existing frameworks and technologies, the exact nature of which will depend on the technologies used to implement the fog itself. For example, if Openstack software is used, the obvious choices to deploy for caching are SWIFT (object storage) and even CINDER (block storage) services.

Compute: One of the core concepts of this work is to facilitate and enable a diversified approach to task offloading and acceleration by weaving the resources of the cloud, the fog, and the extreme edge. While conceptually similar, the deployment of these tasks to the different layers of the architecture might differ for efficacy reasons. Today, there are three competing technologies in these areas, all at different areas of maturity. The first and most well known of these, the VM, has been around for a long time and is widely used in the cloud. Ancillary functions like orchestration have also been developed on top of them, and thus the flow is well understood and mature. Containers have been the talk of the town for the past few years as they promise similar benefits to VMs but with a leaner, swift, low overhead flair to them. As with VMs, container-based service chains can be set up through systems like Apache's Mesos and Google's Kubernetes. Finally, UniKernel is the next big thing, which is poised to take over in the next five years or so. UniKernels are the least mature technology, but they promise to provide a leap forward in security and deployment efficiency by creating OSs that contain the absolutely minimal amount of code required. UniKernels have been shown to implement complex tasks like domain servers with a footprint of just 400–500 kB. That being said, due to their novelty, their integration into the tool chain is almost nonexistent. Figure 2 shows all three concepts arrayed to facilitate comparison.

The challenge is to evaluate all of these technologies in light of the multi-tier architecture proposed and select the most appropriate one for the devices in each tier while keeping in mind their diverging feature set and capabilities, while keeping in mind that orchestrating a function chain which mixes and matches all these technologies at will might present significant issues.

EENs are the least powerful from a processing capability perspective and the most power-constrained devices in this architecture. At the same time, they are the most numerous, which makes it tempting to be able to harvest their resources for use in the wider system. In order to do so, the most lightweight solution must be adopted, making UniKernels and containers especially amenable. At the same time, it is important to investigate how these technologies can be harmonized with the existing software stack on these nodes. Devices at the edge run on processors and OSs with very limited capabilities; thus, containers or UniKernel deployment on them might present inherent difficulties. At the same time, additional challenges like mobility and unreliable wireless links are issues also need to be weighed.

Fog nodes are located one level up from the EENs and facilitate the local interconnect of these nodes. They can be expected to possess superior processing and storage capabilities in comparison to the EENs, without this being a prerequisite. However, fog nodes in this architecture have two distinctive traits compared to the EE ones. It is anticipated that a substantial amount of the processing tasks will be deployed as virtual instances running over the fog nodes, and thus a portion of their resources must be devoted to this task. This comes with significant implications, since processing resources are scarce even at this tier, so frugality is important, but also due to security concerns, since it is critical to maintain strict separation between the user tasks and the management software. In light of this, the UniKernel presents an especially promising option since it ticks both of the boxes. Containers can also be leveraged if security concerns can be assuaged. This will be one of the core challenges in the fog's computing layer.

In addition to this, fog nodes might contain heterogeneous compute resources in the form of programmable logic to accelerate certain tasks. These nodes will be based on previous VM-based work done in the T-NOVA project [2, 3], but will need to be updated accordingly to be in line with the architecture herein. Thus, investigating and determining how to integrate them into the virtualization and deployment framework will be one of the key questions to answer.

Virtualizing and deploying tasks on cloud resources is by far the most mature of the three categories. Hence, off-the-shelf solutions will be a major part in this tier. The exception is again the integra-

The implementation of the edge caching layer will be based on the exploitation and enhancement of existing frameworks and technologies, the exact nature of which will depend on the technologies used to implement the fog itself.

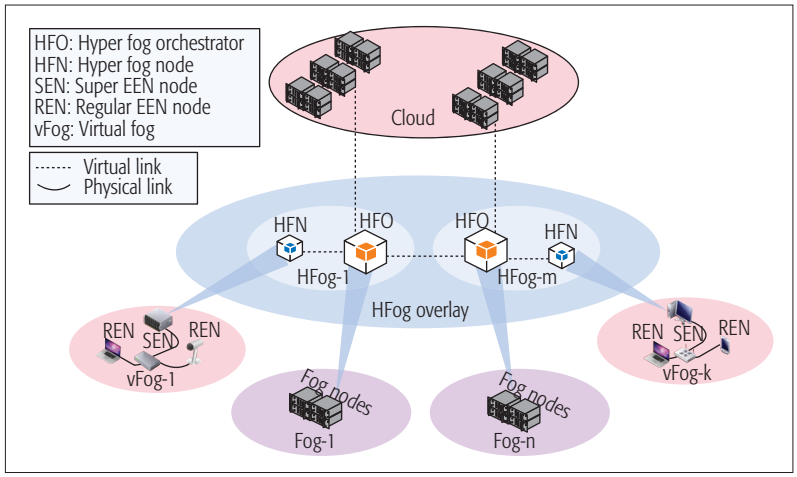


Figure 3. System architecture with active HFog overlays.

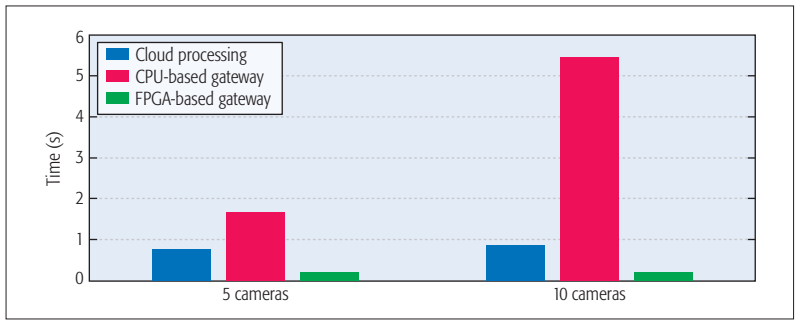


Figure 4. Processing time for the various scenarios.

tion of programmable resources into the cloud, a topic where considerable expertise is already available [13] that is currently being integrated into our architecture in order to reach a solution where these resources can be correctly and efficiently identified by the cloud platform and the orchestrator and tied into the rest of the system.

Hyper Fog: Hyper fogs are constellations of vFog networks, which can be grouped together in one entity to facilitate processing and data exchange that requires resources from more than one vFog. Each HFog will be formed from a number of the underlying fogs and vFogs following a P2P approach. In this configuration, a fog (vFogs are relegated to a secondary choice, mainly because of their potential volatility) will be selected as the HFog's orchestrator based on several criteria such as its processing capacity. The HFog Orchestrator (HFO) will be responsible for registering the other fogs and vFogs in the HFog. In addition to this, the HFO will build a common API representing the available 3C resources and handle the necessary managerial tasks (e.g., delegate demands for a processing task to another HFog or to the cloud when the task's requirements for 3C resources exceed those available at the HFog).

In more detail, the core element of the infrastructure management plane is the virtualized infrastructure manager (VIM), which is the functional entity responsible for controlling and managing the infrastructure (compute, storage, and network) resources. The management scope of the VIM is generally restricted within a single NFV infrastructure point of presence (NFVI-PoP) even in a single-plane cloud architecture. Thus, in a full deployment architecture, multiple VIMs may operate across operator data centers, providing multiple NFVI-PoPs that can operate independently or cooperatively as required under the control of a federator/orchestrator. In contrast, a VIM, in general, can potentially offer specialization in handling certain NFVI resources; hence, the VIM is seen to encompass all management and control functionalities needed for the proper administration of the infrastructure, as well as the virtualized services running on top of it. The VIM will run as a service in the (v)Fog orchestrator. It will comprise a combination of existing VIM technologies (e.g., Openstack, Docker) in order to utilize resources of various processing capabilities without adding significant computational overhead.

Preliminary Evaluation

This section provides a preliminary investigation into the benefits and functionality of the proposed edge-centric architecture. This is accomplished by simulating a scenario where a number of cameras collect data and feed them to a processing entity. That entity can be either the cloud, meaning a remote data center, or a CPU-based gateway located in the fog, or finally, the solution propounded in this article, namely a field programmable gate array (FPGA)-based fog gateway. In all scenarios, a decision is returned to a set of actuators located at the edge of the network to perform some functionality, depending on the result of the processing. More specifically, the processing performed attempts to detect the traffic load on a busy street intersection and to adjust traffic lights accordingly.

In all scenarios, the cameras generate a constant stream of data, which is uploaded over the network to the cloud or gateway. There are two important metrics that we track in these simulations. The first is the time the processing and decision making takes, and the second is the traffic reduction achieved between the edge and the cloud when the processing takes place in the cloud. Furthermore, we provide plots for scenarios including 5 and 10 cameras, respectively, to show the effect that utilizing additional edge nodes has on the aforementioned two parameters.

Figure 4 illustrates the total processing time for 5 and 10 cameras. In each case, the aforementioned three scenarios are depicted. It is obvious that FPGA acceleration is beneficial as it accelerates decision making in both scenarios. On the other hand, edge processing on its own yields worse results than its cloud equivalent. This is because in edge processing a low-power, wimpy CPU is used, which does not have enough horsepower to process the required data quickly enough.

Indeed, a detailed look at the results as provided in Fig. 5 indicates that while both edge processing scenarios show significant gains in video transfer times, the use of a weak, low-power CPU negates this advantage due to the much higher processing time required. The FPGA-based edge node, on the other hand, outperforms both alternatives by a very significant margin.

In terms of network throughput saved by performing the computation at the end, Tables 1a and 1b provide the respective data for both scenarios. It is plainly evident that edge computation whittles down the data that need to be sent to a remote data center by almost four orders of magnitude and

thus provides massive benefits in terms of reducing network traffic in hyper-scale data centers.

Conclusions

This article introduces a novel architecture that aims to reshape the network landscape in order to allow it to meet the upcoming 5G requirements. This architecture is based on two pillars: a multi-tiered, flexible layering of network elements, where resources can be logically grouped together to form vFogs that take over specific tasks, and the introduction of heterogeneous compute resources into the edge and cloud in order to speed up processing while retaining acceptable power consumption. These heterogeneous resources are integrated into a comprehensive cross-layer orchestration software that manages the vFogs.

We use extensive simulation to provide a preliminary investigation of the efficacy of our concept. Our results prove the validity of our approach by offering 0.7-0.8 s improvement in decision making latency as well as reducing the network traffic to the cloud by four orders of magnitude.

References

[1] E. K. Markakis *et al.* , "EXEGESIS: Extreme Edge Resources Harvesting for A Virtualised Fog Environment," *IEEE Commun. Mag.*, vol. 55, no. 7, July 2017.

[2] K. Karras *et al.* , "A Cloud Acceleration Platform for Edge and Cloud," *EnESCE: Wksp. Energy-Efficient Servers for Cloud and Edge Computing*, 23–25 Jan. 2017.

[3] Y. Rebahi *et al.* , "Virtual Network Functions Deployment between Business Expectations and Technical Challenges: The T-NOVA Approach," *Recent Advances in Commun. and Networking Technology* (formerly *Recent Patents on Telecommun.*), vol. 5, no. 1, 2016, pp. 49–64.

[4] "Fog Computing and the Internet of Things: Extend the Cloud to Where the Things Are"; http://www.cisco.com/c/dam/en_us/solutions/trends/iot/docs/computing-overview.pdf, accessed July 2016.

[5] G. L. Klas, *Edge Cloud to Cloud Integration for IoT*, 2016.

[6] L. M Vaquero and L. Rodero-Merino, "Finding Your Way in the Fog: Towards a Comprehensive Definition of Fog Computing," *ACM SIGCOMM Computer Commun. Review*, 2014, vol. 44, no. 5, pp. 27–32.

[7] http://www.openfogconsortium.org/news, accessed July 2016.

[8] A. Poenaru, R. Istrate, and F. Pop, "AFT: Adaptive and Fault Tolerant Peer-to-Peer Overlay — A User-Centric Solution for Data Sharing," *Future Generation Computer Systems*, 2016.

[9] A. Sfrent and F. Pop "Asymptotic Scheduling for Many Task Computing in Big Data Platforms," *Info. Sciences J.*, vol. 319, Oct. 2015, pp. 71–91.

[10] J.F. Riera, *et al.* , "TeNOR: Steps Towards an Orchestration Platform for Multi-PoP NFV Deployment," *2016 IEEE NetSoft*, Seoul, Korea, 2016, pp. 243–50. DOI: 10.1109/NETSOFT.2016.7502419.

[11] H. Gupta *et al.* ,"iFogSim: A Toolkit for Modelling and Simulation of Resource Management Techniques in Internet of Things, Edge and Fog Computing Environments," *CoRR*, vol. abs/1606.02007, June 2016.

[12] J. Weerasinghe *et al.* , "Enabling FPGAs in Hyperscale Data Centers," *2015 IEEE 12th Int'l. Conf. Ubiquitous Intelligence and Computing, 2015 IEEE 12th Int'l. Conf. Autonomic and Trusted Computing, and 2015 IEEE 15th Int'l. Conf. Scalable Computing and Commun. and Its Associated Wksps.*, Beijing, China, 2015, pp. 1078–86.

[13] J. Mongay Batalla *et al.* , "A Novel Methodology for Efficient Throughput Evaluation in Virtualized Routers," *IEEE ICC*, London, U.K., June 2015. DOI: 10.1109/ICC.2015.7249425

[14] A. Beben, *et al.* , "Content Aware Network Based on Virtual Infrastructure," *Proc. 13th ACIS Int'l. Conf. Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Kyoto, Japan, Aug. 2012. DOI: 10.1109/SNPD.2012.68

Biographies

Evangelos Markakis [M] (markakis@pasiphae.eu) holds a Ph.D. from the University of the Aegean. Currently he acts as a senior research associate for the Technological Educational Institute (TEI) of Crete, and he is the Technical Manager for the HORIZON 2020 DRS-19-2014 EMYNOS project. His research interests include fog networking, P2P applications, and NGNs. He has more than 30 refereed publications in the above areas. He is acts as Workshop Co-Chair for the IEEE SDN-NFV Conference.

Kimon Karras received his Ph.D. in embedded systems design from the Technical University of Munich, and has spent four years at Xilinx Research Labs working on data center acceleration through FPGAs and innovative high-level synthesis applications. For the past year, he has been with Future Intelligence Ltd., where he is responsible for the development of the company's Programmable Cloud Platform.

Anargyros Sideris holds a Ph.D. from the University of the Aegean, Department of Information & Communication Systems Engineering. He joined Research & Development of the Telecommunications Systems Laboratory at TEI of Crete. His current research activities and interests are in the fields of network programming, digital interactive television, fog computing, and IoT.

George Alexiou received his Bachelor's degree from the Applied Informatics and Multimedia Department of TEI of Crete in 2014. He has worked as a full stack developer in the web hosting industry. Currently he is a research associate at PASIPHAE Laboratory working on various European funded projects. Additionally, he is doing his Master's degree on informatics and multimedia in the Department of Informatics Engineering of TEI of Crete.

Evangelos Pallis [M] holds an M.Sc. and a Ph.D. in telecommunications from the University of East London, United Kingdom. He currently serves as an associate professor at TEI of Crete in the Department of Informatics Engineering and director of PASIPHAE Lab. His research interests are in the fields of wireless and mobile networking. He has more than 200 refereed publications. He is a member of IEE/IET and ComSoc, and a Distinguished Member of the Union of Regional Televisions in Greece.

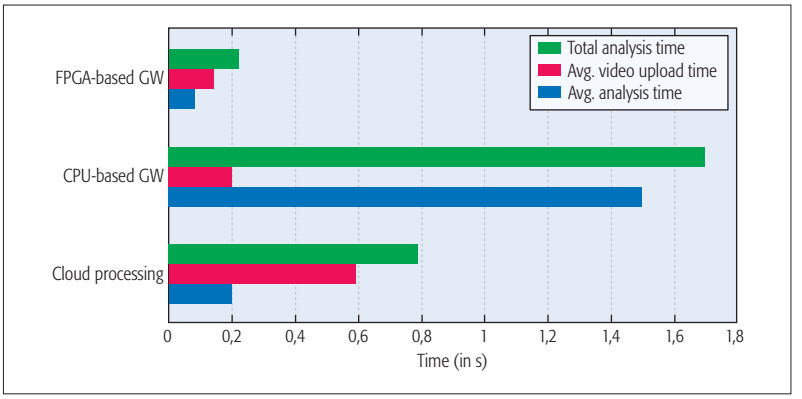


Figure 5. Results breakdown for 5 cameras.

(a) Network traffic for 5 cameras	
Total traffic produced (MB)	Traffic sent on to cloud (MB)
211	N/A
258	0.06
230	0.049

(b) Network traffic for 10 cameras	
Total traffic produced (in MB)	Traffic sent on to cloud (MB)
431	N/A
286	0.047
399	0.04

Table 1. Network traffic for 5 and 10 cameras.

COMMUNICATIONS EDUCATION & TRAINING: THE SCHOLARSHIP OF TEACHING AND LEARNING


Dave Michelson


Peter Ostafich


Carolyn Ottman

The Scholarship of Teaching and Learning (SoTL) encourages teachers and educators to examine their own classroom practice, record their successes and failures, and ultimately share their experiences in a formal and scholarly way so that others may reflect on their findings and build upon existing teaching and learning processes.

SoTL acknowledges that concerns for privacy and other ethical issues associated with studies involving human subjects place limits on the types of research that can be conducted in the classroom setting. Nevertheless, SoTL provides a mechanism for raising the standard of discussion concerning teaching and learning in the literature. The Scholarship of Research and Supervision (SoRL) is a related concept that invites the same reflective approach to improving the quality of training through research, especially that conducted at the postgraduate level.

This Feature Topic on SoTL is intended to hasten the incorporation of SoTL and SoRL into communications engineering curricula by providing educators and researchers with an opportunity to share their experience, best practices, and case studies. The articles in this Feature Topic run the gamut from case studies in experiential learning and research-oriented courses to a comprehensive study of student understanding of key concepts to a novel framework for conducting lab-based courses. The work presented provides a fascinating glimpse at the worldwide effort to reflect upon and improve the quality and effectiveness of university teaching.

In "Combining Teaching and Research through Barcode Experiments," Xinbin Wang and colleagues from Shanghai Jiaotong University in China introduce an experimental research-oriented course that they developed for junior-year students that uses original research tasks rather than conventional lecture and laboratory approaches to develop student skill and insight. Based on a cutting-edge research project, the course progresses through introduction, demonstration, research, and evaluation phases and concludes with a three-dimensional topic, work achieved, and written report evaluation framework. Outcome and experiential evaluations show that students developed important research skills through this process.

In "Incorporating Experiential Learning in Engineering Courses," Atousa Hajshirmohammadi from Simon Fraser University in Canada presents the results of her efforts to incorporate experiential learning into a common core, first-year course on logic circuits. The experiential approach may be considered to be an intermediate step between conventional lab assignments and full projects that encourages development of higher-level skills than conventional assignments but with a far smaller commitment of time and effort than a full project. She presents the results of students' feedback on this method of learning compared to conventional approaches and reflects on how the lessons learned may be applied to courses in analog and digital communications.

In "Insights into Students' Conceptual Understanding of Operating Systems: A Four-Year Case Study in Online Education," Sonia Pamplona and her colleagues from Universidad a Distancia de Madrid and Universidad Politecnica de Madrid address the need to know where and why students are experiencing gaps or misconceptions in their conceptual understanding of technical concepts. In particular, their SoTL-based study provides insights into why students have conceptual understanding gaps or misconceptions in an online operating systems course. They present the results of a four-year qualitative case study of 78 online students in order to identify misconceptions and their root causes. Their results indicate that the natural-language meaning of technical terms can be a significant barrier to good understanding. This has led to their development of a methodology for discovering misconceptions and its causes. Finally, they consider how such a study could benefit developers of communications engineering courses.

In "The iLab Concept: Making Teaching Better, At Scale," Marc Oliver-Pahl from the Technical University of Munich in Germany shows how lab courses can be managed in a way that efficiently supports learners while significantly reducing the workload of teachers. His iLab concept consists of a blended learning teaching methodology and the lab system eLearning platform that was especially designed for supporting the teaching methodology. The concept reduces administrative and organizational overhead so that the focus of both instructors and students can return to teaching and learning. In this manner, the iLab concept enables teaching more content in less time. More than 1500 students have benefited from the iLab concept between 2004 and 2017. Assessments and evaluations have confirmed the effectiveness of the approach.

The next Feature Topic on Education and Training will focus on Humanitarian Engineering and Community Engagement in Education and will appear in May 2018. See the Call for Papers on the *IEEE Communications Magazine* website for additional details.

BIOGRAPHIES

DAVID G. MICHELSON (davem@ece.ubc.ca) leads the Radio Science Lab at the University of British Columbia, where his research focuses on wireless communications. He completed the UBC Faculty Certificate Program on Teaching and Learning in Higher Education in 2011, served as ComSoc's Director of Education and Training from 2012 to 2013, and has been a member of the ComSoc Educational Services Board from 2012 to present. He is also an elected Member-at-Large of the ComSoc Board of Governors (2013–2015, 2017–2019).

PETER OSTAFICHUK (ostafich@mech.ubc.ca) is a professor of teaching in the Department of Mechanical Engineering at the University of British Columbia. His research interests focus on aerodynamics and hydrodynamics. His teaching interests focus on team-based learning, active learning, outcomes-based assessment, and team dynamics. In 2015, he was awarded a fellowship from the Society for Teaching and Learning in Higher Education, in partnership with 3M Canada. The fellowship recognizes his leadership in engineering curriculum advancement.

C. KELLY OTTMAN (ottman@msoe.edu) is a professor in the Rader School of Business at MSOE University. She was a Teaching and Learning Center Scholar at the University of Wisconsin-Milwaukee. There, in 2005, she was awarded the Outstanding Teaching Award. Her SoTL research addresses understanding participation and learning in traditional and online environments, the use of teams in large lectures, and international short-term immersions. She served as the SoTL Chair for the Team Based Learning Collaborative (2009–2014).

Combining Teaching and Research through Barcode Experiments

Xinbing Wang, Jiaqi Liu, Zhe Yang, Junfa Mao, Luoyi Fu, Xiaoying Gan, and Xiaohua Tian

ABSTRACT

A core mission of university education is to train students' problem solving abilities, or in some cases, students' research skills. To achieve this, educators assign students certain experiment courses, where students complete a task following preset procedures. However, such courses often employ outdated teaching materials that are far from cutting-edge problems. Meanwhile, a preset experiment procedure cannot evoke students' innovations. In this article we introduce an experimental course we opened at Shanghai Jiao Tong University for junior-year students — the Research-Oriented Course (ROC). ROC is based on a cutting-edge research project and includes an Introduction — Demonstration — Research — Evaluation four-phase course as well as the T — W — R three-dimensional evaluation criteria. Evaluations showed that students finished their research tasks satisfactorily and they improved their research skills through this process.

INTRODUCTION

Educators often provide students with experiment courses to help them gain knowledge in a certain domain or acquire specific abilities, in order to train them to be prepared for future research. However, these courses suffer from two drawbacks: outdated experiment contents and inability to evoke innovative and independent thinking. Educators want students to get in touch with real research in their experiment courses. Nevertheless, the applied materials are often outdated, and some of them are used every year, which makes them far from cutting-edge research. On the other hand, if educators do employ their current research as experiment materials, it is hard to control the difficulty, for students who take these experiment courses often have no experience in research before. The second drawback is that the existing experiment courses often require students to follow a certain procedure, and the results of these experiments are fixed, which only enhances students' manipulative ability and does not require students' creative thinking.

Therefore, how to design a course that has a close relationship to cutting-edge research problems, which does not require much knowledge or many related skills that students do not possess, and which can effectively improve students' research skills, remains a crucial problem. To tackle this problem, we opened an experimental course in Shanghai Jiao Tong University, the Research-Oriented Course (ROC).

One important question is, which research topic should we select to let students work on? An ideal case is one that requires prior knowledge and possible solutions within students' curricula. In ROC we employed recent research results of a group of Shanghai Jiao Tong University researchers on barcode as the basis of ROC. The barcode technology we applied is ARTcode [1], short for Adaptive Robust doT matrix barCode. It is a barcode design that combines an image and a barcode, enabling dual-channel communication for both human eyes and smartphone cameras. The reasons that we select ARTcode as our course basis are threefold. Firstl, it is an interesting design, and students are able to easily observe what they can get. Second, the required prior knowledge is not complicated and all within the coverage of undergraduate curricula. Finally, ARTcode is now a prototype. If researchers want to put it into public use, there are still many improvements to be made.

Another important question is, how should we organize the course to develop students' research abilities? The most important skills for students to do research are to find problems independently, to have innovative solutions to problems, as well as to collaborate with colleagues and to present their work to others. To develop these research skills, in ROC, we designed an Introduction — Demonstration — Research — Evaluation four-phase course setting, and the selected Topic — Work achieved — written Report (T — W — R) three-dimensional evaluation criteria. Through the Introduction and Demonstration phases, we give students necessary background on barcodes and ARTcode. In the Research phase, we encourage students to think of their own research directions and implement their proposed approaches. In the Evaluation phase, we assess students' performances.

Many traditional project-based courses often follow a regular teaching strategy, that is, introduction — research — evaluation. Different from the traditional ones whose projects have been completely developed so that students can easily find many related materials, our course gives students cutting-edge research projects. It means that students can hardly learn by themselves, and consequently we should provide them with more instructions.

The authors introduce an experimental course they opened at Shanghai Jiao Tong University for junior year students: the Research-Oriented Course (ROC). ROC is based on a cutting-edge research project and includes a four-phase course as well as the T – W – R three-dimensional evaluation criteria. Evaluations showed that students finished their research tasks satisfactorily, and they improved their research skills through this process.

The authors are with Shanghai Jiao Tong University.

Digital Object Identifier:
10.1109/MCOM.2017.1700283

0163-6804/17/$25.00 © 2017 IEEE

ARTcode is an adaptive robust dot matrix barcode that carries both human-friendly information (mostly, artistic images and layouts) and machine-friendly information (coded data bits). It has an encoder and a decoder. The encoder is a computational device that takes an image and a short message as input and outputs an image with information embedded.

To this end, the following two improvements are included in our course. First, we add the Demonstration phase, where the teacher demonstrates to students how ARTcode works and responds to students' questions. The goal of this part is to help students effectively comprehend the implementation of ARTcode. Second, we offer instant help to students in the Research phase. One of the teaching assistants comes from the group that designed ARTcode, and thus students can obtain professional suggestions. In summary, we opened an experimental course, ROC, at Shanghai Jiao Tong University. The primary goal of ROC is to train students with the important research skills of independent thinking, problem solving, collaboration, and work presentation. The secondary goal of ROC is to crowdsource students' thinking on ARTcode improvement. In ROC, we first give students background knowledge on barcodes and ARTcode. Then students are free to select their own research directions. With the aid of our instructors, students realize their own research goals through experiments. The last step is to evaluate students' work and performance. The contributions of ROC are twofold:

- We used cutting-edge research problems as our course materials to help students get in touch with real research scenarios.
- We used an Introduction — Demonstration — Research — Evaluation four-phase course as well as a T — W — R three-dimensional evaluation criterion in the ROC setting, which significantly enhanced students' research skills and also achieved pertinent improvement for ARTcode.

Related Work

In this part we review some literature related to ROC. We first introduce some theoretical studies that discussed the possibilities of establishing a link between research and teaching; then we examine some concrete courses that tried to equip the students with research skills and some courses that applied innovative teaching methods.

Linkage between Research and Teaching: There have been a number of studies focusing on the theoretical linkage between research and teaching. In [2], M. Healey surveyed the different perceptions of relationships between research and teaching. In [3], the authors concluded that student centered learning can lead to a positive connection between teaching and research, of which one example is problem-based learning. In [4], the authors explained how to strengthen the teaching-research nexus in policy formulations. The above literature launched theoretical studies on the linkage between research and teaching, and provide us with a theoretical background.

Project-Based Learning: An educational concept that is close to our ROC is project-based learning (PBL), which refers to a student-centered pedagogy that involves a dynamic classroom approach in which students acquire a deeper knowledge through active exploration of real-world challenges and problems. Some works focus on theoretical studies of PBL. In [5], the authors studied how PBL stirred up students' motivations toward learning. In [6], J. S. Krajcik *et al.* studied the theoretical background and five features of PBL, which are driving questions, situated inquiry, collaborations, using technology tools to support learning, and creation of artifacts. Additionally, in computer science and electrical engineering courses, PBL has been widely applied to help teachers improve education efficiency. In [7], the authors gave courses using an e-learning model based on blending learning, as a combination of independent learning, online discussions, and problem-based learning. In [8], Antonio Carpe *et al.* analyzed several key factors in achieving greater active and cooperative learning efficiency. These works differ from our proposed course in that we apply cutting-edge research problems into course design, and the research results of students have positive feedback on the cutting-edge research.

Innovations in Networking Courses: Apart from PBL, there are also other forms of innovations in networking related courses [9–12]. To tackle the complexity of networking education, educators applied virtual laboratories in [9, 11]. In [12], the authors analyzed the effects of a remote lab on several courses. These cases differ from our case in that they update the teaching equipment while we use new teaching contents and materials. In [10], the authors proposed a novel tool called Packet Tracer, which can help teachers develop students' creativity. This course is different from ROC in that ROC applies novel teaching methodology, while this course applies a new tool to assist with the teaching process.

Preliminary on ARTcode

In this section we introduce the realization and functionality of ARTcode, on which the research-oriented training course is based.

ARTcode is an adaptive robust dot matrix barcode that carries both human-friendly information (mostly artistic images and layouts) and machine-friendly information (coded data bits). It has an encoder and a decoder. The encoder is a computational device that takes an image and a short message as input and outputs an image with information embedded. On the decoder side, a smartphone camera scans an ARTcode and decodes the contained information. The architecture of ARTcode is shown in Fig. 1. To generate an ARTcode, we first use the colored dot matrix to form the original image. Then we embed data into a block. Finally, we select encoding colors. On the receiver side, we perform code detection and then recover the codes within the background image noise. To explain the core algorithm of ARTcode, designers answered two questions:

- Where should data be embedded?
  Embedding position is determined with an algorithm called Shuffling.
- How should data be embedded?
  To minimize visual distortion, data is encoded to bit change with a data hiding technique. Then bit change is mapped to colors using an adaptive color palette.

Research-Oriented Course Setting

ROC is held in the summer semester. It includes eight two-hour classes. The implementation of the course follows four phases:

- **Introduction:** Giving students an introduction lecture on existing barcodes and the technical details of ARTcode
- **Demonstration:** Giving students the MATLAB source code of the ARTcode project and introducing its major functions
- **Research:** Organizing students in groups of three or four to freely alter ARTcode source code
- **Evaluation:** Evaluating the projects of each group

Phase 1: Introduction

The first phase of ROC is the introduction, which takes one class. First, students are shown some of the currently popular barcodes, including Data Matrix (a two-dimensional barcode that is often used to mark small items) and QR code (one of the most common barcodes in consumer advertising). In addition to the black and white barcodes that have been widely applied in practice, colored barcodes have drawn much attention in recent years. Therefore, we also introduce students to two typical colored barcodes: COBRA [13] and PiCode [14]. The introduction to various barcodes is aimed at demonstrating the traditional or classical methodology for barcodes to encode or decode data.

Another important component of this phase is an overview of ARTcode. We give students a brief introduction on the encoding scheme of ARTcode, and focus on demonstration of the decoding part, because this is the major part that we expect to improve in ARTcode design. The decoding approaches proposed in [1] achieve a decoding accuracy of 96 percent, but the whole decoding process takes 34 to 40 s, which is unacceptable in real-world applications. Therefore, one important expectation for ROC is to crowdsource students' thinking on how to improve the performance of ARTcode, especially its decoding efficiency.

Phase 2: Demonstration

In the next class, students are organized into groups of three according to their student identification numbers. Twenty-two students participated in this course in total, so we partitioned them into seven groups (one group consisted of four students). The original MATLAB source code of ARTcode is given to each group, with a document explaining each function. The teacher demonstrates to students how each function in the ARTcode project corresponds to encoding and decoding methods mentioned previously. Then students have one hour to go through the code. The teacher responds to students' questions. The goal of this part is to help students effectively comprehend the implementation of ARTcode so that they are prepared to work on their projects.

Phase 3: Research

In the following five classes, students are required to work in groups and find a research topic to improve the performance of ARTcode. The research topics are unconstrained for students. The recommended research topic is to improve the decoding efficiency while keeping as much of the decoding accuracy as possible. However, students can also think of their own research topics. To encourage students to think of innovative topics that may improve the performance of ARTcode, we bring in an incentive mechanism that rates the topics students choose. An innovative topic that reveals some latent problems of ARTcode and is worth working on brings bonuses to students who come up with them. Evaluations of these topics are done by instructors. It is highly recommended that each group selects only one topic so that they can fully explore the topic.

Students register their selected topics anytime before the end of the second class of the group research section. Once their topic is validated, students can continue to work on their selected topics. If a group's topic is not validated or they fail to come up with a topic by the end of the second class, they work on the decoding efficiency topic.

The research experiments are arranged in a laboratory equipped with MATLAB-installed computers. Students can also do the research on their own computers. They are encouraged to utilize their own mobile devices as testing platforms, for this can guarantee the robustness of ARTcode under different circumstances. For groups who do not own any mobile devices, we provide them with Nexus 4 E960 smartphones. For each of the research section classes, two teaching assistants stay in the laboratory to accompany the students and provide them with instant help. One teaching assistant comes from the group that designed ARTcode, and the other teaching assistant is familiar with programming in MATLAB.

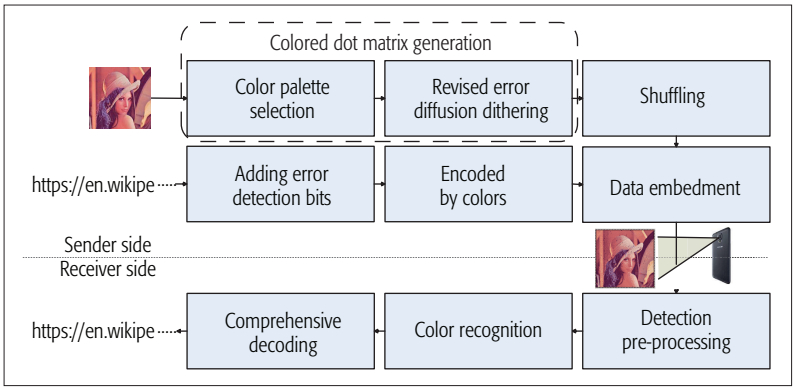


Figure 1. ARTcode architecture.

Phase 4: Evaluation

In the last class, we evaluate students' projects. Specifically, each group should make a presentation no more than 15 minutes long. In this presentation, students should include the following aspects:

- Students should mention their selected topics, explain why they select those topics, and describe their expected outputs.
- Students should describe the major challenges they encounter, and elucidate their approaches to tackle these challenges.
- Students should present the outputs of their work, including their revised ARTcode structure if they change the ARTcode encoding scheme, and the decoding efficiency, accuracy, and robustness of their decoding scheme.

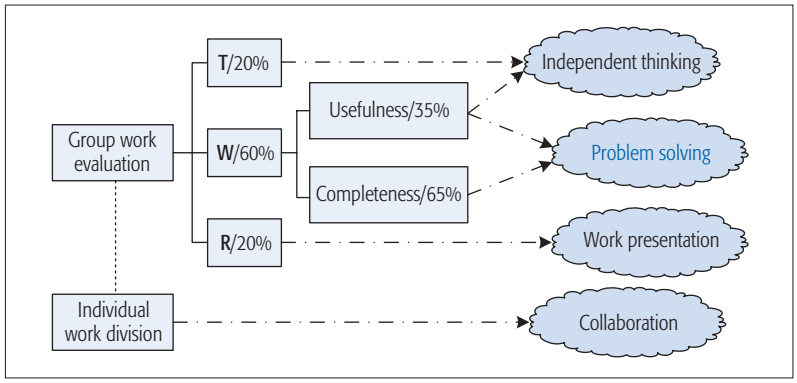


Figure 2. Illustration of ROC evaluation criteria and how they reflect students' research skills.

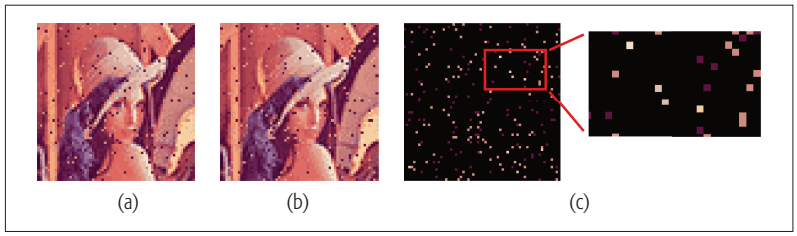


Figure 3. These images illustrate the effect of multiple encoding colors, which is the research result of group 2: a) ARTcode without multiple encoding colors; b) ARTcode with multiple encoding colors; c) the left image shows all encoding positions replaced by encoding colors; the right image is a zoom-in. We can see clearly that multiple encoding colors are employed here.

- Students should explain their work division within their groups.

The professor determines their score of W according to criteria below.

Evaluation Criteria

The evaluation criteria of groups' performances were released to students in the first class, and they are of high importance as they measure students' performances in ROC, which directly relate to their research skills development. The scoring scale is 100. We evaluate the following three items of each group as criteria of their course performance: selected topic (T), work achieved (W), and written report (R).

T: As mentioned above, students are provided with a recommended research topic — to improve decoding efficiency with keeping decoding accuracy. If they choose the recommended topic, they will obtain 90 percent in **T**. Otherwise, their research topic is evaluated according to its prospect and feasibility. Prospect means whether the expected result may help improve ARTcode. Feasibility means whether the idea can be realized. The scores of **T** range from 60 to 100 percent. Topics with scores lower than 60 percent are considered unacceptable, and these groups are forced to choose the recommended topic. The reasons to have such a setting for **T** are threefold. First, we do not fix the research topic, so students can develop their creative thinking to come up with their own research topic; second, as crowdsourcing students' intelligence in improving the performance of ARTcode — especially its decoding efficiency — is an important goal of ROC, we set a high score (90 percent) for improving ARTcode efficiency; third, we have to consider that some students may not have participated in any research projects before, so a recommended topic can lower the course difficulty to some extent. The **T** part constitutes 20 percent of the final score.

W: The **W** part is the most important criterion. We inspect the projects of each group according to their usefulness and completeness. Usefulness means whether the project leads to pertinent improvements to ARTcode, which is evaluated according to students' selected research direction. Completeness means whether students achieved complete and solid work, which is evaluated according to the effects of students' proposed encoding approaches or the robustness and efficiency of students' proposed decoding approaches. Note that the **T** criterion and usefulness both reflect students' creative thinking abilities, but **T** focuses on research topic selection (a bigger picture), while usefulness focuses on research direction selection (a smaller research angle, a detailed approach). Research direction is totally determined by students themselves through their own studying of the existing ARTcode work. Meanwhile, completeness reflects students' problem solving abilities. We give usefulness a weight of 35 percent and completeness a weight of 65 percent. The **W** part constitutes 60 percent of the final score.

R: At the end of the course, each group is required to hand in an electronic version of their project report. The report has to be in English, with at least five pages. It should contain students' understanding of barcodes and ARTcode, and the details of each group's challenges and approaches. Students are encouraged to hand in reports written with LaTeX, while other formats are also acceptable. The report is evaluated based on their structure, clarity of problem explanation, and language fluency. We evaluate **R** because presenting one's own work clearly is an essential skill for a researcher, and we can also evaluate what the students have learned about barcodes through the report. This part constitutes 20 percent of the final score.

In general, the evaluation criteria are illustrated in Fig. 2, which shows the weight of each criterion and how they are correlated to students' research skills development. It is worth mentioning that the above criteria determine a single group's score. To determine each individual student's score, we refer to the work division within each group as mentioned above and give individual scores accordingly.

Research Results and Evaluations

In this part we present the research results of each group. Afterward, we show some selected research results of certain groups that can promisingly and effectively improve the performance of ARTcode. Then we present the evaluation results.

Research Results

Among the seven groups, six groups chose to improve the decoding efficiency, and the other group chose to improve the image preservation quality of ARTcode. In the six groups who proposed their approaches to improve the performance of the decoding speed, groups 1 and 6 conducted some minor changes to the existing

scheme (group 1 changed the data structure of a variable in the pre-processing step; group 6 studied the effect of color erosion in the color recognition process); group 4 initially targeted increasing decoding efficiency, but in the end added a redundancy algorithm in the module localization part, which actually increased the decoding time but enhanced the robustness of decoding; groups 3 and 5 made major changes to the ARTcode decoding scheme that can promisingly improve the decoding efficiency of ARTcode; and group 7 proposed a partial thresholding binarization algorithm, but they did not finish implementing their proposed approach. The group that did not select the default research direction, group 2, proposed an approach that improves the performance of the color selection mechanism in the encoding part.

We present the research results of groups 2, 3, and 5 in detail. Group 2 targeted improving the appearance of ARTcode. They found that in the "encoded by colors" step in Fig. 1, encoding colors were selected from a set of two to three colors. Meanwhile, the colored dot matrix was generated with more colors (normally 16). They used multiple colors as encoding colors so that the encoded data will appear less abrupt. The results of group 2 showed that there were improvements in appearance, although they were limited. Still, we consider this an interesting and useful improvement. The results are in Fig. 3. To better demonstrate the effect of multiple encoding colors, we decrease ARTcode resolution. In Fig. 3c, we can see that more than three colors are manipulated to generate ARTcode, and based on the comparison between Figs. 3a and 3b, the improvement exists but is limited.

Group 3's topic was to improve the decoding efficiency. They analyzed the running time of each algorithm in the ARTcode decoding part and found that local thresholding binarization took up a large part (42.7 percent) of ARTcode decoding. They proposed that the local thresholding binarization algorithm can be replaced by a normal binarization algorithm (a fixed threshold for the whole image). Group 3 revised the localization algorithm to fit the change of binarization. Running time of binarization decreased from 11.0 s to 0.884 s, which is a decrease of 92.0 percent.

Group 5 studied the module localization algorithm, which corresponds to color recognition in Fig. 1. The function of this algorithm is to find the positions of each module inside the code with the help of the alignment pattern as reference, considering camera shooting distortions. In MATLAB this algorithm was originally written with *for* loops. A common scenario for MATLAB is that its matrix calculation efficiency is much higher than its loops efficiency. Therefore, group 5 rewrote the loops with matrix calculations, and they also used a MATLAB function *sub2ind* that resolved an index problem. The unning time of the module localization algorithm decreased from 9.48 s to 0.11 s, which is 98.8 percent.

Evaluations

In this part we evaluate students' performances. We list the scores of each group in Table 1, where U means Usefulness and C means Completeness as defined above. Note that the scores of individual students are determined by their roles within their groups on the basis of group performances.

Group #	Research direction	T	W		R
			U	C	
1	Data structure improvement in pre-processing	90%	70%	92%	90%
2	Image preservation improvement with multiple encoding colors	95%	93%	96%	93%
3	Replacing local thresholding binarization with fixed threshold	90%	100%	100%	85%
4	Enhancing decoding robustness with redundancy	82%	80%	91%	92%
5	Improvement of MATLAB loops	90%	100%	100%	98%
6	Studying the impact of color erosion on accuracy	90%	82%	96%	96%
7	Partial thresholding binarization	90%	100%	62%	87%

Table 1. Evaluations of each group.

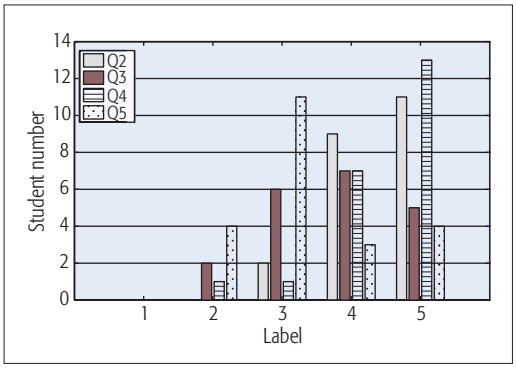


Figure 4. Students' evaluation of their gain through this course.

For ***T***, groups 1, 3, 5, 6, and 7 all selected improving ARTcode decoding efficiency as their research topic. Group 2's topic showed their original insights on ARTcode encoding. It is both feasible and promising. In light of the consideration of encouraging students to put forward creative thoughts, we give them 95 percent. Group 4's topic is helpful and feasible, but it appears to be incremental. *For* ***W***, the usefulness is, in a way, a complement to **T**, especially when students selected the same topic but different research directions. Groups 2, 3, 5, and 7 independently proposed interesting schemes that might lead to pertinent improvement of ARTcode; thus, we give them relatively high marks. Groups 1 and 6 focused on a helpful but trivial point, while group 4's proposed approach is incremental. In light of completeness, groups 3 and 6 implemented their approach completely, and they launched convincing experiments under variant illuminance conditions that testified to the robustness of their proposed approaches. Group 5 implemented their approach and stated that no additional experiments were needed to testify to their approach. Group 2's encoding approach was solid. Group 1's improvement required not as

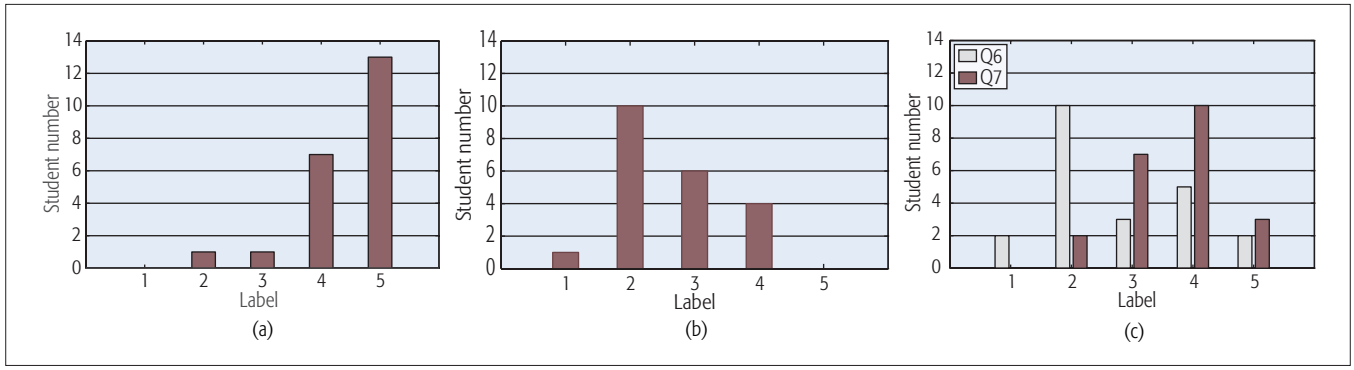


Figure 5. Students' feedback on ROC: a) overall satisfaction; b) course difficulty; c) ARTcode improvement.

much workload. Group 7 failed to finish the project. *For* ***R***, groups 1 and 4 handed in reports written with Microsoft Word, and the other groups used LaTeX. Reports of groups 5 and 6 were well organized, clearly explained, and in fluent English, while other groups suffered from different problems.

As for work division, groups 2, 3, and 6 worked in a democratic way, meaning that everyone contributed in each phase. In groups 1 and 5, one student was relatively inactive. Groups 4 and 7 had clear partitions in their tasks; for example, in group 4, one student worked on report writing, and the other two students worked on algorithm design and implementation.

Student Feedback

We examine the students' perceptions of ROC. We sent a questionnaire to each student at the last class and required them to return the questionnaires within a week. In order to guarantee candor, we published our questionnaire through SurveyMonkey, a free online survey software and questionnaire tool [15], which allows the students to give online feedback anonymously. By the end of the deadline, we had successfully collected 22 questionnaires. There are eight questions in one questionnaire:

- Evaluate your overall satisfaction with this course.
- Rate your knowledge gain on barcodes via this course.
- Evaluate how much you think you have improved your problem finding ability.
- Evaluate how much you think you have improved your problem solving ability.
- Evaluate how much you think you have improved your collaboration ability.
- Do you think you successfully improved the performance of ARTcode?
- Do you think that the combined results of all seven groups have a positive impact on the improvement of ARTcode?
- Do you find this course to be difficult (compared to a normal experiment course where you follow instructions and complete the tasks step by step)?

The responses should be on a scale from 1 to 5, that is, response 1 means not satisfied at all/no improvements/no improvement/extremely difficult, 3 means neutral/sufficient improvements/sufficient improvement/neutral, and 5 means very satisfied/great improvements/great improvement/extremely easy. We cluster these eight questions ($Q1$ – $Q8$) into three clusters: overall evaluation ($Q1$ and $Q8$), students' gain through this course ($Q2$–$Q5$), and ARTcode's improvement ($Q6$ and $Q7$). The results are in Figs. 4 and 5.

Discussion

All seven groups provided independent thinking on how to improve the performance of ARTcode, while three of them are considered to have achieved pertinent improvement. Their work will be adopted by researchers of ARTcode into the design of an upgraded version of ARTcode. Meanwhile, according to our survey distributed to students, they are satisfied with their knowledge gain and research skills improvement through this course.

In our future work, we plan to implement ARTcode using some low-level computer programming languages such as C. Since MATLAB sometimes involves unknown overhead, using such codes could help students better understand and predict the effects of a proposed change. Besides, since many students labeled ROC as difficult or extremely difficult, we plan to provide students with more assistance in the group research phase and increase the intervention of instructors. In our current setting, we only provide students with a major research topic. In future courses, we plan to provide some specific research directions as well. In addition, we intend to compare the student feedback with some data to make the result more convincing. Particularly, the feedback data can be compared to that from other related courses at the university or that from the same course running in different years.

Conclusion

In this article we present a novel experiment course, named Research-Oriented Course (ROC). The contributions of the course are twofold: enabling students to get in touch with real research scenarios and achieving improvement for the research project. Furthermore, this work provides some implications for teaching practice at universities. It demonstrates the combinability of teaching and research, and additionally presents a systematical teaching method, that is, the four course phases, Introduction — Demonstration — Research — Evaluation, and the **T — W — R** three-dimensional evaluation criteria. Based on the case described in this work, educators can apply the teaching method to other appropriate research projects, which can greatly benefit both teaching and research.

Acknowledgment

This work was supported by NSF China (No. 61532012, 61325012, 61521062, 61602303, and 91438115).

References

[1] Z. Yang *et al.*, "ARTcode: Preserve Art and Code In Any Image," UbiComp, 2016.
[2] M. Healey, "LinkinG Research and Teaching Exploring Disciplinary Spaces and the Role of Inquiry-Based Learning," *Reshaping the University: New Relationships Between Research, Scholarship and Teaching*, 2005, pp. 67–78.
[3] L. Elton, "Research and Teaching: Conditions for a Positive Link," *Teaching in Higher Education*, vol. 6, no. 1, 2001, pp. 43–56.
[4] A. Jenkins and M. healey, "Critiquing Excellence: Undergraduate Research for All Students," *International Perspectives on Teaching Excellence in Higher Education*, 2007, pp. 117–32.
[5] P. C. Blumenfeld *et al.*, *Motivating Project-Based Learning: Sustaining the Doing, Supporting the Learning*, vol. 26, no. 3-4, 1991, pp. 369–98.
[6] J. S. Krajcik and P. C. Blumenfeld, *Project-Based Learning*, 2006, pp. 317–34.
[7] N. H. Bozic, V. Mornar, and I. Boticki, "A Blended Learning Approach to Course Design and Implementation," *IEEE Trans. Education*, vol. 52, no. 1, 2009, pp. 19–30.
[8] A. Carpeño *et al.*, "The Key Factors of an Active Learning Method in a Microprocessors Course," *IEEE Trans. Education*, vol. 54, no. 2, 2011, pp. 229–35.
[9] M. Wannous and H. Nakano, "NVLab, A Networking Virtual Web-Based Laboratory that Implements Virtualization and Virtual Network Computing Technologies," *IEEE Trans. Learning Technologies*, vol. 3, no. 2, 2010, pp. 129–38.
[10] Y. Zhang, R. Liang, and H. Ma, "Teaching Innovation in Computer Network Course for Undergraduate Students with Packet Tracer," *IERI Procedia*, no. 2, 2012, pp. 504–10.
[11] L. Xu, D. Huang and W. T. Tsai, "Cloud-Based Virtual Laboratory for Network Security Education," *IEEE Trans. Education*, vol. 57, no. 3, 2014, pp. 145–50.
[12] M. A. Marques *et al.*, "How Remote Labs Impact on Course Outcomes: Various Practices Using VISIR," *IEEE Trans. Education*, vol. 57, no. 3, 2014, pp. 151–59.
[13] T. Hao, R. Zhou, and G. Xing, "COBRA: Color Barcode Streaming for Smartphone Systems," *Mobisys*, 2012.
[14] W. Huang and W. H. Mow, "PiCode: 2D Barcode with Embedded Picture and ViCode: 3D Barcode with Embedded Video," *Mobicom*, 2013.
[15] SurveyMonkey; https://www.surveymonkey.com/home/, 2017.

Biographies

Xinbing Wang received his B.S. degree from Shanghai Jiao Tong University, China, in 1998, and his M.S. degree from Tsinghua University, Beijing, China, in 2001. He received his Ph.D. degree from the Department of Electrical and Computer Engineering, North Carolina State University, in 2006. Currently, he is a professor in the Department of Electronic Engineering, Shanghai Jiao Tong University. He has been an Associate Editor of *IEEE/ACM Transactions on Networking* and *IEEE Transactions on Mobile Computing*.

Jiaqi Liu received her B.E. degree in electronic engineering from Shanghai Jiao Tong University in 2014. She is currently pursuing a Ph.D. degree in electronic engineering at Shanghai Jiao Tong University. Her research interests are in the area of wireless networks, social networks, and evolving networks.

Zhe Yang received his B.E. degree in electronic engineering from Shanghai Jiao Tong University in 2014. He is currently pursuing an M.E. degree in electronic engineering at Shanghai Jiao Tong University. His research interests are in machine learning and the intelligent Internet of Things.

Junfa Mao received his B.S. degree in radiation physics from the University of Science and Technology of National Defense, Hunan, China, in 1985, his M.S. degree in experimental nuclear physics from the Shanghai Institute of Nuclear Research, China, in 1988, and his Ph.D. degree in electronics engineering from Shanghai Jiao Tong University in 1992. His current research interests include interconnect and package problems of integrated circuits and systems, and analysis and design of microwave circuits.

Luoyi Fu received her B.E. degree in electronic engineering from Shanghai Jiao Tong University in 2009 and her Ph.D. degree in computer science and engineering from the same university in 2015. She is currently working with Prof. Xinbing Wang as a postdoctoral researcher in the Department of Electronic and Engineering at Shanghai Jiao Tong University. Her research interests are in the area of scaling laws analysis in wireless networks, connectivity analysis, sensor networks, and social networks.

Xiaoying Gan received her Ph.D. degree in electronic engineering from Shanghai Jiao Tong University in 2006. Currently, she is an associate professor in the Department of Electronic Engineering, Shanghai Jiao Tong University. From 2009 to 2010, she worked as a visiting researcher at the University of California San Diego. Her current research interests include network economics, social aware networks, heterogeneous cellular networks, multiuser multi-channel access, and dynamic resource management.

Xiaohua Tian received his B.E. and M.E. degrees in communication engineering from Northwestern Polytechnical University, China, in 2003 and 2006, respectively. He received his Ph.D. degree in the Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, in 2010. He is currently an assistant professor in the Department of Electronic Engineering at Shanghai Jiao Tong University. His research interests include application-oriented networking, the Internet of Things, and wireless networks.

Incorporating Experiential Learning in Engineering Courses

Atousa Hajshirmohammadi

The author presents an example of an assignment given to students that incorporates experiential learning, and the results of students' feedback on this method of learning compared to conventional methods are summarized. A similar approach can be applied to other subjects in engineering, even without requiring a laboratory component in the course.

ABSTRACT

In recent years, "experiential learning" as a method of delivering course material has become more recognized in university education. This article is a report on a project conducted in the School of Engineering Science at Simon Fraser University that incorporates experiential learning in a lower division course. The course is a first-year, common core course in logic circuits. At the time that this project was implemented, the course did not have any lab components attached to it; therefore, incorporating experiential learning into the course was even more challenging. In this article, an example of an assignment given to students that incorporates experiential learning is presented, and the results of students' feedback on this method of learning compared to the conventional method are summarized. A similar approach can be applied to other subjects in engineering, even without requiring a laboratory component in the course. This article also provides example applications for analog and digital communication courses.

INTRODUCTION

"Experiential learning" as a method of delivering course material is becoming more recognized in university education in recent years. Many North American, as well as international, universities have formed research teams dedicated to incorporating experiential education in their curricula. Simon Fraser University (SFU) in Canada, with which the author is affiliated, is no exception. The Experiential Education Project at SFU began in late 2010 with a focus on documenting SFU's use of credit-bearing experiential education [1].

Experiential learning is the process of making meaning from direct experience, that is,"learning from experience" [2]. Aristotle once said, "For the things we have to learn before we can do them, we learn by doing them" [3]. A definition of experiential education by the Association of Experiential Education states, "Experiential Education is a process through which the learner constructs knowledge, skill, and value from direct experience" [4]. The idea of "Experiential Learning" was first reflected in the writings of John Dewey [5] and later popularized by many educational experts, such as David A. Kolb. Some researchers have distinguished between "experiential learning" and "experiential education" in that experiential learning relies solely on the individual and does not necessarily require a teacher, whereas many others have used the two terms interchangeably [2]. Obviously, it is the second school of thought (integration of experiential learning and education) that can potentially be applied to educational settings, such as degree programs at the university level.

The four critical steps associated with experiential learning are Action, Reflection, Abstraction, and Application (Fig. 1). This circular procedure is initiated with an experience and followed by reflective observation. Observations help people form general rules and modify their experience [2].

An enormous amount of our everyday knowledge is obtained through experiential learning. In other words, we experience various situations before understanding the rules governing them. For example, the process of language learning that a child experiences vividly expresses the steps necessary for experiential learning. The revolution in foreign language education, in which students experience various concepts before acquiring any sense of grammar, might have originated from experiential learning theory.

Research on experiential and community engagement learning shows that the concept of experiential learning has been mostly applied in the arts and humanities, has been applied to some extent in natural sciences, and has been applied minimally in engineering. This may, at first glance, seem rather intuitive considering the closeness of humanities subjects to human experiences. However, the role of engineering concepts in our modern lives should not be overlooked either.

Incorporating experiential learning in subjects such as electrical and computer engineering is more easily achieved in upper-division courses through course projects, internships, and capstone projects. In fact, many universities refer to their capstone projects as an example of providing experiential learning opportunity to their students. Examples are engineering programs at Purdue University, Oregon State University, and the University of British Columbia [6–8].

However, students need to understand the purpose and application of the theoretical knowledge they are gaining as early as possible, and this early integration of theory and practice can be achieved if they have experience with real-life examples related to what they learn.

Digital Object Identifier:
10.1109/MCOM.2017.1700373

The author is with Simon Fraser University.

0163-6804/17/$25.00 © 2017 IEEE

For lower-division courses, experimental work in the laboratory while learning theoretical background in conventional classes is often considered as experiential learning. While the importance of "experimental" work or hands-on experiments in education is undeniable, it should not be referred to as "experiential learning." As previously explained, experiential learning is a methodology where students experience different scientific experiments even before obtaining the theoretical concepts governing them.

A true implementation of experiential education is not single experiences in individual courses, but rather it is a task that must be organized at the program/school level that involves more effort of the trainers compared to conventional educational methods. In the engineering discipline, however, most courses are of a technical nature, and students require knowledge of theoretical concepts, usually based in math and/or physics, to learn new and larger concepts in engineering. A complete overhaul of the education system from the traditional model to an experiential method may thus not be possible, or desired, at the program level or even for a single course in its entirety. The author, however, believes that it is possible to successfully deliver components of courses taught within the discipline of engineering in the form of experiential learning.

This article explains how experiential learning can be incorporated into parts of engineering courses, with or without hands-on laboratory components. The concept is explained by giving an example of an experiential learning assignment given to students in ENSC-150: *Introduction to Computer Design*. At the time of the implementation of this initiation, this course was part of the first-year curriculum of the Engineering Science program at SFU and did not include a laboratory component.

The rest of this article is organized as follows. In the following section, I present the experiential learning example used in ENSC-150. Then a summary of a survey conducted on students' experience is provided. Finally, I explain how a similar method to the example provided in this article can be applied to courses in the communication engineering discipline, and how further research can take the study of the effectiveness of this method to the next level.

Example of Experiential Learning in ENSC-150

This section provides an example of how experiential learning was incorporated in the form of an assignment as part of the course delivery for ENSC-150.

In this course, students are expected to learn about basic logic circuits and how logic blocks can be designed and employed to construct simple digital devices. Combinational and sequential logic designs are the main topics covered in this course. The course has large enrollments (above 100 students), and it includes lectures and tutorials but no laboratory component.

An example of a digital device that can be introduced to students in the advanced stages of this course is a digital clock. In the conventional method of teaching, the building blocks of this digital clock, that is, basic logic gates as well as flip flops, are first introduced to the student, followed by the design procedure for counters (which will count the seconds, minutes, and hours). The student is then taught the details of how each counter is controlled and is reset after going through one cycle. For example, the "seconds" counter is reset to zero after reaching the number 59, and this reset should trigger the "minutes" counter to increment by 1. In other words, in the conventional teaching method, the student is taught the individual components and the design procedure of a digital clock without being able to discover how this device works by first experimenting with it.

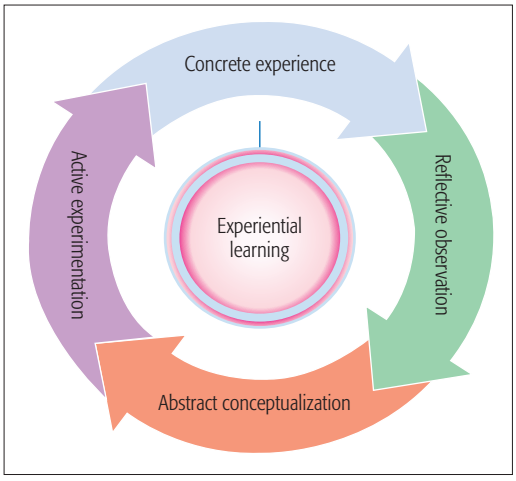


Figure 1. Experiential learning, circular diagram.

If, instead of following the conventional method, we were to use the concept of experiential learning by adhering to its true definition, each student would be asked to start by experiencing an actual digital clock, analyzing the circuit inside it, figuring out the building blocks of the clock, and understanding the interaction of the logic signals within the clock. Obviously, this approach would not be practical considering the scope and timeline of the course!

The challenges in ENSC 150 are the large enrollment in the class, the lack of a laboratory component in the syllabus of the course, and finally, the fact that it is not possible to see the circuit inside an integrated digital circuit! The approach we took here was to adopt a combination of theoretical (conventional) and experiential learning methods, as explained below.

Because the course did not have access to any hardware design lab, we used a logic simulation program, called DesignWorks, to enable the students to see the function of various digital circuits, and be able to make changes and modifications to circuits and see the results.

We presented this experiential learning example in the form of an assignment. The students were given a simplified block diagram of a digital clock. This clock consists of three parts, shown in Fig. 2. Part A receives a clock pulse of frequency 1 Hz and counts and displays the "seconds." Part B receives a clock pulse of frequency 1 pulse per minute from Part A and counts and displays the "minutes." Part C receives a clock pulse of 1 pulse per hour from Part B and displays the hours.

Students were also provided with the schematic of the circuits inside parts A and B. Figure 3

In this course, students are expected to learn about basic logic circuits and how logic blocks can be designed and employed to construct simple digital devices. Combinational and sequential logic designs are the main topics covered in this course. The course has large enrollments (above 100 students), and it includes lectures and tutorials but no laboratory component.

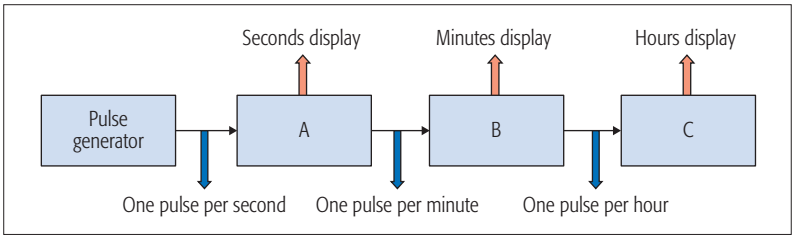


Figure 2. Simplified block diagram of a digital clock provided in the assignment.

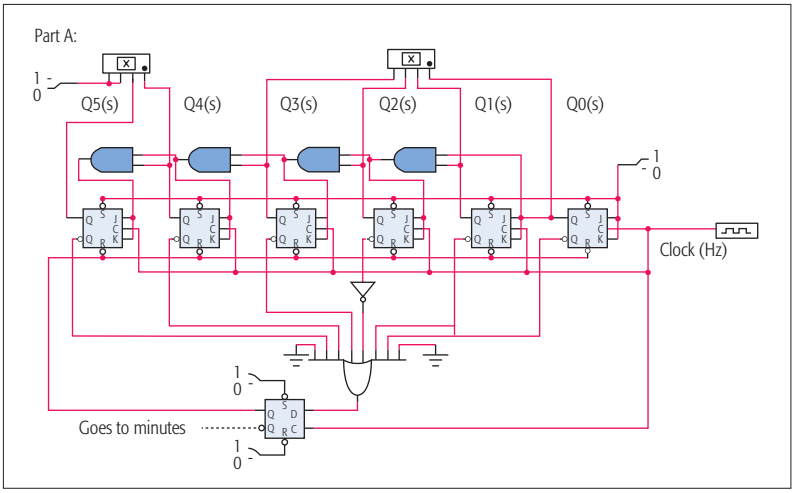


Figure 3. Digital circuit inside of part A; this circuit counts the seconds.

shows the circuit for part A (i.e., counter of the seconds). DesignWorks simulation files were also provided, so students could run the simulations and observe how parts A and B of the digital clock function, and how they interact with each other. Note that at this point in the course, the students had already learned about JK-flip flops, but they had not yet seen the structure of the counters used in this example.

For their assignment, the students were asked to run the simulations for parts A and B, to explain the function of these parts, and to demonstrate their understanding of how the counters for seconds and minutes work. Based on what they learned from the function of each of these parts and their interaction, the students were then asked to design part C of the clock and add it to the simulated circuit, thereby creating a complete clock that counts and displays hours, minutes, and seconds. They were also asked to design and include a seven-segment display showing the letters A and P representing a.m. and p.m., respectively.

As you can see, this assignment is quite different from the assignments engineering students are used to completing. Conventionally, students learn new concepts in a classroom setting during lectures or tutorials. Then they work on assignments where they are asked to apply the learned concepts to solve relevant problems. In the case of this assignment, students are first given access to an electronic device (in this case a partially functioning digital clock) in a lab environment (in this case a simulation program). By examining the components of this clock and observing the input and output signals to each component, they are expected to learn the operation of the counters used in a digital clock. They are then asked to demonstrate their learning, not only by explaining it in an article, but also by designing and adding the last counter to the clock, and, as a result, simulating a fully functioning digital clock.

After submitting their assignment, students were given a survey about their learning experience. The survey questions and results are explained in the next section.

Students' Feedback

A short anonymous survey was given to students at the end of this assignment. It was made clear to the students that participation in the survey was voluntary.The main questions on the survey were:

1. On a scale of 1 to 5, rate how you prefer the experiential method over the conventional method (1: not preferred, 2: less preferred, 3: equally preferred, 4: more preferred 5: strongly preferred)
2. Comment on the differences between the two methods.

Sixty-seven students (about 40 percent of the class) completed the questionnaire. Figure 4 shows the diagram corresponding to their answers to question 1. More than 37 percent of the students favored the experiential learning method compared to 28 percent who preferred the conventional method. The rest of the students did not favor one method over the other.

As for the students' comments in response to question 2, many found this experiment to be interesting, and they mentioned that they looked forward to more experiential learning approaches in future courses. The experiential method being more challenging and more time consuming was among the drawbacks of this method mentioned by some of the other students.

Other types of interesting and useful feedback received from students include students' rating of their level of understanding of subjects to which they were first exposed experientially compared to their understanding of subjects they learned with the traditional approach. However, it is important that topics at similar levels of difficulty be chosen for comparison. Questions can be posed independently for each topic, or can directly ask students to compare the experiential topics with traditional topics.

Possible Application to Communication Courses

This article uses an example to show how the concept of experiential learning can be applied to courses within the engineering discipline. The course considered here is a course in computer design. The course is a large enrollment course with no laboratory component.

The method explained in this article can be adopted and adjusted for various other courses in the field of electrical and communications engineering. In fact, in courses that include a hardware laboratory component, the opportunity to implement experiential learning is more readily available. A very simple example can be given for an analog circuit course. Here, students can be first introduced to a fundamental concept such as Kirchhoff's Voltage Law (KVL) through a laboratory experiment. Students can construct a simple resistive circuit including two or three loops

to measure the voltage across each branch of each loop, and to find the relationship that exists among the measured voltage.

More sophisticated examples can be implemented in third- and fourth-year courses in communications engineering. For example, in a classic analog communication course, the students can first be introduced to the amplitude modulation (AM) system in the laboratory, where modulation and demodulation systems are available in modular format. Knowing that students have the pre-requisite background in signals and systems (i.e., concepts of the time and frequency domains), the instructor will ask students to experiment in the lab with the AM modulator and to observe the output signal on an oscilloscope and on a spectrum analyzer. Students will vary the frequency and amplitude of the input signals (message and carrier) and try to guess how the AM signal carries the information about the message signal. They can also change the DC offset of the message signal to observe the role of the modulation index and over modulation in AM. The instructor will then refer to students' observation in the lab, and provide them with the mathematical formulae that explain the observed results. If a hardware lab is not available, simulation programs such as Matlab can easily replace an experiment like this.

Another example for a digital communication course can be applied to teaching the concept of additive white Gaussian noise (AWGN) and its effect on the error rate of binary signaling. Students need to first learn about the nature of AWGN, that it can be modeled with a normal distribution, and that it is additive. In the lab, students can vary the level of the noise added to a baseband binary phase shift keying (BPSK) signal and measure the error rate at the output of the BPSK receiver (demodulator). They can then plot their measured error rate vs. noise level and try to make predictions about how changing the noise level or changing the modulation scheme would vary their results. The instructor will then guide students to reach the formulae for error calculations.

Finally, it should be noted that engineering educators who are interested in evaluating the effectiveness of these types of assignments may want to measure the learning outcome and level of understanding of students who have learned a concept through this method of experiential learning compared to a control group of students who have learned the same concept through the conventional educational method. This can be achieved if several experiential learning assignments are designed for a course so that students can have equal opportunity of being in a control group (conventional learning method) and a test group (experiential learning group).

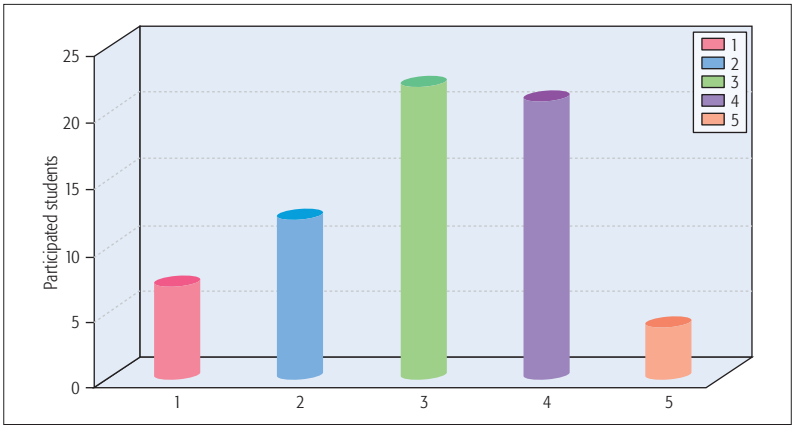


Figure 4. Students' feedback on experiential learning assignment.

Acknowledgments

The author would like to thank the anonymous reviewers and the editors for their valuable remarks. She also thanks Mike Sjoerdsma for his help in proofreading the manuscript, and Dr. Lila Torabi, Mr. Amir Kassaian, Mr. Rio Li, and Ms. Nilgoon Zarei for their collaboration in earlier stages of this project. This work has received funding from the Teaching and Learning Grant Office of SFU.

References

[1] http://www.sfu.ca/experiential/, accessed May 8, 2017.
[2] C. M. Itin "Reasserting the Philosophy of Experiential Education as a Vehicle for Change in the 21st Century," *J. Experiential Education*, Fall 1999, pp. 91–98.
[3] W. R. Bynum and R. Porter, *Oxford Dictionary of Scientific Quotations*, Oxford Univ. Press, 2005.
[4] C. Luckmann, "Defining Experiential Education," *J. Experiential Education*, vol. 19. no. 1, 1996, pp. 6–7.
[5] J. Dewey, "Democracy and Education by John Dewey 1916," *The University of Chicago Press Journals*, vol. 5, no. 1/2, 2008, pp. 87–95.
[6] "Experiential Learning & Research," Electrical and Computer Engineering, Purdue University; https://engineering.purdue.edu/ECE/Academics/Undergraduates/UGO/Enrichment_Opportunities/Experiential_Learning_Research, accessed May 8, 2017.
[7] "Experiential Learning," College of Engineering, Oregon State Univ., http://engineering.oregonstate.edu/experiential-learning, accessed May 8, 2017.
[8] "Experiential Learning and Research," Univ. British Columbia, Electrical Engineering at UBC's Okanagan Campus; http://you.ubc.ca/ubc_programs/electrical-engineering-okanagan/, accessed May 8, 2017.

Biography

Atousa Hajshirmohammadi (atousah@sfu.ca) is a senior lecturer in the School of Engineering Science at Simon Fraser University, British Columbia, Canada. Prior to that, she was a senior design engineer at LSI Logic Inc., California (2001–2004). She received her B.Sc. and M.Sc., in electrical and computer engineerin, from Isfahan University of Technology, Iran, and her Ph.D. in communications engineering from the University of Waterloo, Ontario, Canada. Her research interests are in multimedia communications, cognitive radio, and engineering pedagogy.

Insights into Students' Conceptual Understanding of Operating Systems: A Four-Year Case Study in Online Education

Sonia Pamplona, Isaac Seoane, Javier Bravo-Agapito, and Nelson Medinilla

The authors provide insights into why students have misconceptions in an online course on operating systems. Specifically, this study presents a four-year qualitative case study of 78 online students in order to identify misconceptions and the causes that generate them. Their results indicate that students experienced misconceptions with the concept of interrupt.

ABSTRACT

For decades, instructors and researchers have been trying to improve or enhance the learning process of students. In this process, it is important to know whether students have misconceptions in their conceptual understanding. The study of these elements is becoming a relevant research area in science and engineering education. This article provides insights into why students have misconceptions in an online course on operating systems. Specifically, this study presents a four-year qualitative case study of 78 online students in order to identify misconceptions and the causes that generate them. Our results indicate that students experienced misconceptions with the concept of interrupt. In fact, this study reveals that the natural-language meaning of the term interrupt is a hindrance to understanding this concept. In addition, a methodology for discovering misconceptions and their causes is developed.

INTRODUCTION

Student misconceptions and conceptual understanding represent an important research area in science and engineering education. The development of this area initially began in the last century in physics with the design and development of the Force Concept Inventory (FCI) [1]. The main objective of a concept inventory (CI) is to identify possible student misconceptions through multiple-choice questions. Although Hestenes *et al.* produced promising results, relatively little research has been conducted focusing on engineering curricula [1]. Specifically, research into conceptual understanding can be found in the communications curriculum. For instance, Bristow *et al.* developed a CI in control systems [2], and Goncher *et al.* evaluated conceptual understanding and identified possible student misconceptions in signal processing [3]. In addition, Webb *et al.* developed a CI to explore students' misconceptions of operating systems [4]. Even though these studies analyze conceptual understanding and student misconceptions, they do not perform in-depth analyses of the causes that generate difficulties in student understanding. In fact, a key question that remains largely unanswered is what makes some concepts so difficult to learn and some misconceptions so difficult to repair [5].

The main goal of our article is to provide insights into this question. Indeed, this work looks at why students lack conceptual understanding or have misconceptions in the syllabus of an operating system undergraduate course. The aim of this study overlaps with the goal of an important area of engineering education research: identification of threshold concepts. We are looking for troublesome concepts, and troublesomeness is one of the characteristics of a threshold concept according to Meyer and Land [6]. In particular, these authors identify five characteristics of a threshold concept: troublesome, transformative, irreversible, integrative, and bounded. Therefore, our results also make a contribution to the field of threshold concept research.

E-learning provides new ways to transmit, organize, and present educational content, but the adoption of e-learning is slow by many universities. A reverse trend is observed, however, in online universities, since educational content must be integrated into virtual classrooms. Online education does not provide a direct interaction between teachers and students, and students learn at their own pace. Moreover, instructors are not fully aware of the student learning process. For this reason, these learning difficulties could be higher in online education than in face-to-face education. Our work tries to find the misconceptions of 78 online students about operating systems. For this purpose, a qualitative case study was carried out to discover the hidden understanding difficulties of students, and to identify the causes that produce these difficulties.

The work in this article is a longitudinal study started in 2012. Our initial results were presented at the Koli Calling Conference [7]. This initial work only provided results for 14 online students and focused on finding difficulties of understanding, and so did not include in-depth analysis of causes. The current work continues this analysis with a larger sample and provides insights into what might generate student misconceptions.

Although this study has been carried out in a

Digital Object Identifier:
10.1109/MCOM.2017.1700362

Sonia Pamplona, Isaac Seoane, and Javier Bravo-Agapito are with Universidad a Distancia de Madrid; Nelson Medinilla is with Universidad Politécnica de Madrid.

0163-6804/17/$25.00 © 2017 IEEE

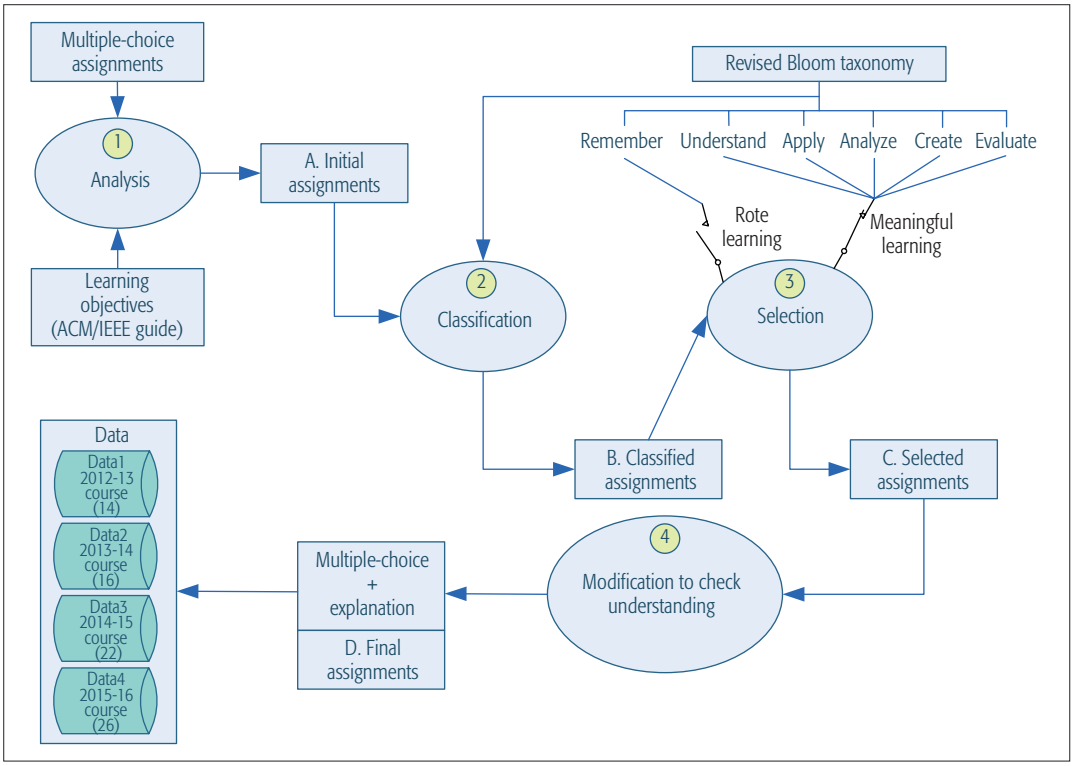


Figure 1. Flow diagram of the assignment design processes.

computer science course, the approach, experience, and results will be valuable to educators in communications engineering. Communications engineering is a multidisciplinary field of study. Knowledge and skills related to design, implementation, and programming of digital systems and devices are provided by a variety of courses across the undergraduate curriculum in order to acquire skills in design and operation of telecommunication networks for communication services [8]. In the case of this study, concepts about basic operating principles of digital programmable devices are included in the first part of the syllabus of the Operating Systems undergraduate course. The main results of the study are focused on misconceptions about the concept *interrupt*, which is an important concept in the communications curriculum.

This article is organized as follows. The next three sections describe the methodology followed in this research. Then we show an example of analysis according to the research methodology. The next section contains the main results and discussion. The final section sets out the main conclusions and future lines of research.

Methodology Description

The research methodology is problem-driven. Exploring misconceptions needs to uncover student thinking at a deeper and more detailed level. Therefore, the nature of the problem implies the use of a discovery-driven research methodology [9]. The methodology chosen for this research is the qualitative case study, which allows us to discover processes that would probably be overlooked if we used other more superficial research methods.

In order to discover the misconceptions and their causes, we have performed a qualitative analysis of the written explanations obtained through the assessment tests designed for this study. This method is based on the first stage of the development of the FCI [1]. The methodology used in this research can be seen in Figs. 1 and 2. The shapes used in the flow diagrams are described below to facilitate the understanding of these figures:

- Circular shapes are used to highlight processes that obtain or transform the information.
- Rectangular shapes are used to mark the inputs and outputs through the process flow.
- Some processes have configuration criteria to parameterize each process operation. These criteria have been drawn as a flag switch to indicate whether they are active.

This research has been carried out in two main stages. The first (Fig. 1) focuses on the design of the assignments used in the discovery process. The second (Fig. 2) consists of analyzing the students' answers to these assignments.

Assessment Design

In order to discover misconceptions, we need to probe student thinking. Therefore, we require a set of assignments that trigger cognitive processes beyond remembering, that is, assignments whose main objective is to foster meaningful learning. Meaningful learning is based on transference, which is the ability to use what was learned to solve new problems, answer new questions, or facilitate learning new subject matter. In contrast, rote learning is based on retention, which is the ability to remember material at some later time in much the same way it was presented during instruction [10]. In our design, we have used Blooms' taxonomy to distinguish between rote learning and meaningful learning, as we explain later.

In order to discover misconceptions, we need to probe student thinking. Therefore, we require a set of assignments that trigger cognitive processes beyond remembering, that is, assignments whose main objective is to foster meaningful learning.

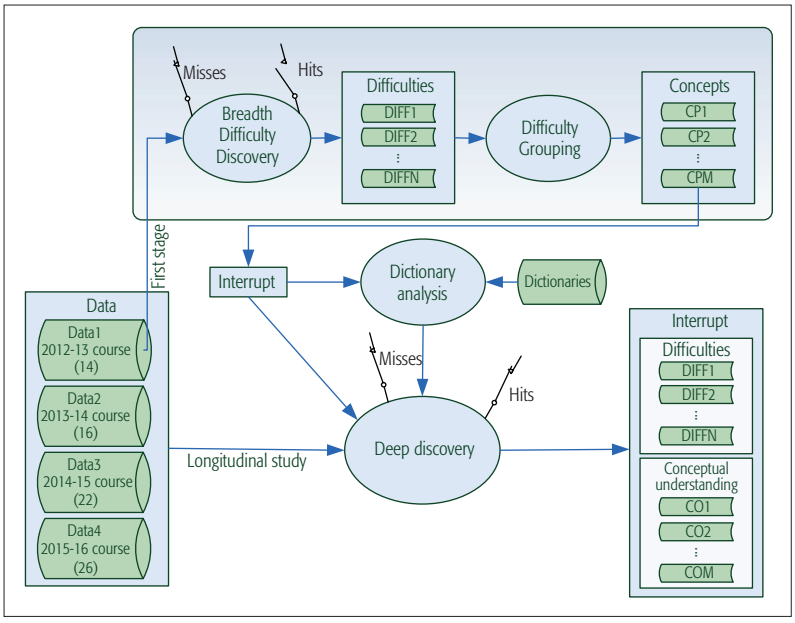


Figure 2. Flow Diagram of the Deep Discovery process design.

What code do you think needs to be run with interrupts inhibited? A. None, because the interrupts could be lost. B. All operating system code. C. Certain critical parts of the operating system code, such as context switching. Justify your answer

Table 1. Final assignments. Example of a question.

We have chosen the revision of Bloom's taxonomy (RBT) [10] because classifying an educational objective according to the revision of this taxonomy is easier than using the original framework. Each of the six major categories of RBT is associated with two or more specific cognitive processes. For instance, the category "understand" is associated with the cognitive processes interpreting, exemplifying, classifying, summarizing, inferring, comparing, and explaining. Consequently, classifying an objective into the appropriate category is facilitated by focusing on the cognitive processes rather than on the larger categories. Nonetheless, any other taxonomy could be used to distinguish between rote learning and meaningful learning in a replication of this study.

Figure 1 shows the flow diagram for the design of the final assignments. In these assignments, every student has to answer 10 multiple-choice questions and explain why the right answer is right and/or the wrong answers are wrong. Table 1 provides an example of this particular assignment consisting of a multiple-choice question and an explanation.

Final assignments are obtained by following these steps. The first is to analyze (process 1 in Fig. 1) multiple-choice questions and learning objectives in order to select those assignments aligned with the desired learning objectives. Multiple-choice questions are extracted from common operating systems textbooks [11–13]. Learning objectives are selected from the professional society guidance (ACM/IEEE-CS Joint Review Task Force). The output of this process is set **A. Initial assignments**, which will be classified in the classification process (process 2).

In the classification process (2), the RBT is used to classify the A output into six categories. These categories are: "remember," "understand," "apply," "analyze," "create," and "evaluate." The result of this second process is the same set of assignments (**B. Classified assignments**), but classified following Bloom's six cognitive process categories [10].

After the classification process, the third task is the selection (3) of only those assignments that assess meaningful learning. This action is shown in Fig. 1 using a switch flag, indicating that only categories related to meaningful learning will be chosen. According to the RBT these categories are "understand," "apply," "analyze," "create," and "evaluate." Hence, the result of this process is set **C. Selected assignments**.

Finally, to ensure the discovery of misconceptions, the final task (4) modifies set **C. Selected assignments** in order to check understanding. This set consists of a list of modified multiple-choice tests where students were asked to explain why the right answer is right and/or the wrong answers are wrong. These explanations would help to reveal students thinking in order to uncover students' misconceptions and check their understanding.

These tests were translated into Spanish. Regarding the language involved in the study, the students were from Spain and took the tests in Spanish. However, most of the reference texts of the course are in English or are translations of textbooks whose original language is English.

In conclusion, three tests were designed with 10 multiple-choice questions plus an explanation in each (**D. Final assignments**). The tests correspond to the course content as follows:

- Questionnaire I. Introduction to operating systems and process management.
- Questionnaire II. Process scheduling. Process communication and synchronization.
- Questionnaire III. Memory management. I/O.

Qualitative Case Study

The second part of the study is shown in Fig. 2. The flow diagram shows the processes involved in the discovery of misconceptions and their root causes. Data analysis was carried out in two stages, which are described separately below. We performed a qualitative analysis [14] of the students' explanations included in the assignments using ATLAS.ti qualitative analysis software. In particular, we analyzed the answers to the second part of the questions (justify your answer) as shown in Table 1.

First Stage

The aim of the first stage is to identify concepts difficult for students to understand. The study began by researching the set of tests of a cohort of 14 students from 2012–2013. The process is shown in the upper part of Fig. 2. The first step consists of looking for the difficulties shown in the students' answers to the final

assessments. (Fig. 1, **D. Final assignments**). We have called this process "breadth difficulty discovery" because we search for any type of conceptual difficulty.

To perform this first discovery process, the incorrect multiple-choice questions are used to analyze its corresponding textual explanation. This is shown in Fig. 2 as two input switches, one closed (for the mistakes) and the other open (for the hits), indicating that correct answers are not being analyzed in this initial process. In particular, if the student's explanation was incorrect, a code [15] was created whereby the justification and the answer to the multiple-choice questions were combined (e.g. "I think that all operating system code needs to be run with *interrupts* inhibited because..."; see question from Table 1). The result of this process is a set of difficulties obtained from the students' textual explanations.

Finally, the codes obtained from the previous process related to the same concept were grouped. Hence, these groups show the concepts that are involved in misconceptions and are the results of the first stage. These results are detailed in the next section.

Results Obtained After the First Stage

At this stage of the research, and after the cohort corresponding to the academic year 2012–2013, the main result is that virtually all learning difficulties found were represented in the concept *interrupt*. This fact suggests that there might be a problem with the conceptual understanding of *interrupt*. Therefore, the second part of the analysis will focus on the difficulties arising from this concept and their causes.

The fact that almost every learning difficulty was related to the concept of *interrupt* should not be surprising because *interrupt* is a key concept in the knowledge of operating systems, since an operating system is a kind of software assisted by *interrupts* [11]. An operating system wakes up at certain moments to attend to several kinds of events: key strokes, software signaling to the hardware and vice versa, and so on. Each time the operating system needs to run an operation, it pauses whatever tasks the computer might be involved in to attend to the event properly, performing whatever task routines should be done depending on the nature of the event.

An important result from this first stage analysis is that students' understanding of the natural-language meaning of interrupt appeared to interfere with their adoption of the technical meaning of the term. For this reason, in-depth analysis of the meanings of the entries for *interrupt* in two baseline dictionaries was undertaken.[1] The results showed that *interrupt* has six different meanings, as seen in Fig. 3. In other words, students could understand the concept of *interrupt* as any of these six meanings.

The next part of the analysis studies the meanings of *interrupt* to which students are referring in their answers. The six codes shown in Fig. 3 are used: "1.-Signal," "2.-Feature," "3.-Act," "4.-State," "5.-Anything," and "6.-Intermission." The answers provided by students are tagged and grouped by these codes.

Second Stage: Longitudinal Study of the Dataset and Deep Discovery Process

The aim of this stage is to identify the difficulties that students have with the concept interrupt and the meaning that students associate with the term in each question ("conceptual understanding" in Fig. 2). These data led us to discover misconceptions about the concept *interrupt* and their possible causes.

In order to contrast this process with the previous one ("breadth discovery"), we have called it "deep discovery" because although we only search for explanations related to the term *interrupt*, we take into account any type of answer (correct and incorrect). This can be seen at the bottom of Fig. 2. The reason for searching within all answers is that we wish to know the meaning understood by students in each answer, correct or otherwise, in order to know if this meaning varies over the questions or stays constant.

The longitudinal analysis was conducted over four academic years of an undergraduate online course on operating systems. Four student cohorts were tested with a total number of 78 students. Each cohort belongs to an academic year. Cohort 1, from 2012–2013, had 14 students, Cohort 2, from 2013–2014, had 16 students, Cohort 3, from 2014–2015, had 22 students, and finally, Cohort 4, from 2015–2016, had 26 students.

An example of this process is described in the following section.

Example of the Question Analysis Process

In order to illustrate the analysis process of the second stage, some student answers to the second question from Table 2 are analyzed in the following sections. Table 2 contains the questions included in the final assignments and are related to the concept of interrupt. Each question belongs to a category and fosters a cognitive process following the RBT [10].

In this question, students should infer whether *interrupts* would or would not improve processor utilization. Given that the lecture notes used do not make an explicit statement, students should infer the answer by connecting the information they have read. From the readings, they could infer the meaning of concepts such as polling input/output (I/O), and interrupt-driven I/O. By using *interrupts*, the processor is able to run other processes and routines while an I/O operation is ongoing. Therefore, it can be said that, indeed, *interrupts* help to take advantage of the waiting time in I/O transactions. Consequently, the correct answer is "a) True." This question is important because it forces the student to consider deeply the advantages of the *interrupt* mechanism.

Discussion about the Meanings of Interrupt in the Example Question

Reviewing the answers to the second question of Table 2, there appears to be a misconception related to one of the given meanings of *interrupt:* the state of being interrupted (fourth meaning shown in Fig. 3). A student could understand the meaning of the term *interrupt* in this way, giving

An important result from this first stage analysis is that students' understanding of the natural-language meaning of interrupt appeared to interfere with their adoption of the technical meaning of the term. For this reason, in-depth analysis of the meanings of the entries for interrupt in two baselines dictionaries was undertaken.

[1] Entries of interrupt retrieved from: Collins, https://www.collinsdictionary.com/es/diccionario/ingles/interrupt, accessed Apr. 14, 2017, and Merriam-Webster; https://www.merriam-webster.com/dictionary/interrupt, accessed Apr. 14, 2017.

After reviewing the answers given by students to the question, and the comments written by them to justify their answers, the main results of the analysis show that they usually understand one of these two meanings of the concept interrupt and use it for their answer.

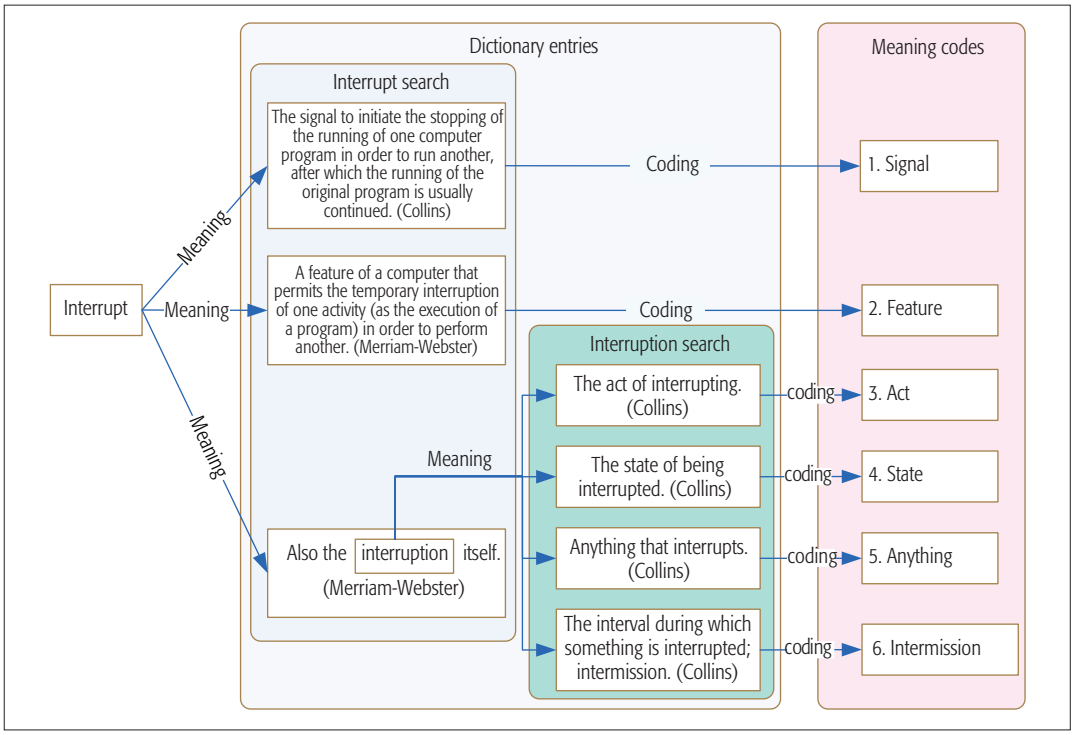


Figure 3. Analysis of the meaning of interrupt in dictionary entries.

a negative interpretation to the *interrupt* mechanism's effect on processor performance because it causes the process to stop.

The students should understand the concept *interrupt* in another way in order to answer the question correctly. For instance, meanings (Fig. 3) related to the signal (Code 1.- Signal), the intrinsic feature (Code 2.- Feature), or even the action (Code 3.- Act) would be more desirable. The key to a proper understanding of the question is to think about how the interrupt occurs, not about the effect it has. *Interrupts* improve processor performance because they free up the processor for a program that is continuously asking whether an I/O process has finished or when there is a hardware problem.

Review Process of the Students Answers

The dataset analyzed consists of the explanations offered by students to the proposed question. The following aspects have been studied:

- Different meanings given to the term *interrupt* by the students
- Difficulties in the process of understanding the concept *interrupt*

After reviewing the answers given by students to the question, and the comments written by them to justify their answers, the main results of the analysis show that they usually understand one of these two meanings of the concept *interrupt* and use it for their answer.

Meaning Number 2: "*A feature of a computer that permits the temporary interruption of one activity (such as the running of a program) in order to perform another.*"

Meaning Number 4: "*The state of being interrupted.*"

With regard to the difficulties, only one difficulty has been found: students think that interrupts are not good for processor performance.

Example with Regard to Meaning 4: "The State of Being Interrupted"

The following example consists of the answer of a student who has interpreted the term *interrupt* with the meaning of a state. The student says that if many interrupts are generated, the process running gets slower. In this answer, the student associates the meaning of *interrupt* with the state of being interrupted and not with the mechanism that enables the *interrupt*.

Example: "I guess it is false because an *interrupt* is a programmed temporal cutoff, it does not improve the use of the processor but it stops it, and if the degree of *interrupt* is high, it could cause the process execution to be even slower as the processor might be fully focused on attending the *interrupts* but not the rest of the processes."

The incorrect answer seems to be caused because of the meaning the student has assigned to the question. The association of the term *interrupt* with neither meaning 1.-SIGNAL or with meaning 2.-FEATURE does not enable the student to select the correct answer.

Example with Regard to Meaning 2: "An Intrinsic Feature"

In the example, the student has assigned the term *interrupt* to the following meaning:

Meaning 2.-FEATURE: "*a feature of a computer that permits the temporary interruption of one activity (such as the execution of a program) in order to perform another.*"

It is important to highlight that every student who has understood the *interrupt* as a mechanism answered correctly. The answer of the student was as follows:

Example: "[...], peripherals that would use hardware *interrupts* to communicate are more efficient than the main program, using fewer clock cycles, allowing other processes to run while *interrupts*

Source	Question	Category	Cognitive process
Testbank. Chapter 1 Multiple-choice questions Question 6 from [11]	1. In a uniprocessor system, multiprogramming increases processor efficiency by: A. Taking advantage of time wasted by long wait interrupt handling. B. Disabling all interrupts except those of highest priority. C. Eliminating all idle processor cycles. Justify your answer	Understand	Interpreting
Testbank. Chapter 1 TRUE/FALSE questions Question 8 from [11]	2. Interrupts are provided primarily as a way to improve processor utilization. A. True. B. False. Justify your answer	Understand	Inferring
Instructor Companion Site. Testbank. Chapter 3 Question 7 from [12]	3. What code do you think needs to be run with interrupts inhibited? A. None, because the interrupts could be lost. B. All operating system code. C. Certain critical parts of the operating system code such as context switching. Justify your answer	Understand	Inferring
Page 9. Question 1.2.4 from [13]	4. A process switch: A. is performed by the scheduler. B. modifies the entry in the process table of the process evicted. C. is always caused by a clock interruption. D. occurs whenever a process leaves the waiting process queue and enters in the ready process queue. Justify your answer	Understand	Inferring

Table 2. Questions according to the concept of interrupt used to build the assignment tests.

Based on the evidence analyzed and the methodology applied, it can be said that the misconceptions around interrupt are not only caused by the complexity of the subject itself, but because of other meanings which are completely accepted by society, and may keep coming up in every course in this area.

are attended. *Interrupts* may notify the processor about I/O from a peripheral, preventing the main processor from periodically controlling the input so as to know if the data is already available or not, rendering the use of the main processor more efficient [...]"

Results on Understanding of the Concept Interrupt

The following conclusions can be drawn after analyzing the data obtained from the answers considered in the research. Four different meanings of *interrupt* can be found from the students' answers from its six different meanings:

- Meaning number 1.-SIGNAL: The signal to initiate the stopping of the running of one computer program
- Meaning number 2.-FEATURE: A feature of a computer that permits the temporary interruption of one activity
- Meaning number 4.-STATE: The state of being interrupted
- Meaning number 5.-ANYTHING: Anything that interrupts

The meanings that students assign to *interrupt* change over time and are different from student to student. This has two consequences:

- Students vary their interpretation depending on the context and not always in an appropriate way, because sometimes they assign a meaning that does not allow them to answer the question correctly.
- Students have not detected any ambiguity in the term *interrupt*, as they have not written any argumentation about it, and they have not asked any question in this regard. Therefore, they think there is no misconception in the meaning they apply in their answer, and they are not conscious of the existence of several meanings, which interferes with their correct knowledge.

With regard to the difficulties in the learning process, seven difficulties have been detected during the research:

- Difficulty 1: They do not define the term *interrupt* fully.
- Difficulty 2: They become confused about the source of *interrupts*.
- Difficulty 3: They state that *interrupts* do not improve the processor performance.
- Difficulty 4: They are not able to justify the reason why some routines for operating systems must be run in a disabled *interrupts* mode.
- Difficulty 5: They think that a change in the running processes is always caused by a clock *interrupt*.
- Difficulty 6: They usually think that *interrupts* are used to make a context switch.
- Difficulty 7: They think that every time an *interrupt* appears, a context switch is made.

Finally, with regard to the possible causes of the difficulties described above, the following conclusions are set out here:

- The first difficulty, an incomplete definition of the term *interrupt*, can be explained by the fact that students are unaware of the several meanings of the term *interrupt*, and they try to explain their own interpretation using the first interpretation they think of when answering.

Future work might be undertaken with the purpose of overcoming the difficulties discovered, and finding new sets of difficulties and misconceptions around other concepts, in order to draft a concept inventory for subjects involved in training in digital systems and design and programming of devices for communications engineering, such as Operating Systems courses.

- The second difficulty, misconceptions about the source of the *interrupts*, might arise from the association between the term *interrupt* with the meaning of state. Because of this, they know the state that *interrupts* cause, but they ignore the sources that cause them. They only think about the consequences, not about the origins.
- The statement that the *interrupts* do not improve processor utilization (difficulty 3) is also related to the misconception of thinking about the *interrupts* as a state. The association of *interrupt* with the effect of pausing or stopping and not with the mechanisms prevent them from inferring the correct answers. As an example, an answer was found saying that "a lot of *interrupts* will cause the execution of processes to become slower."
- Difficulty 4 is not yet clear at the time of writing, and is still a part of the ongoing longitudinal research. The data obtained are not enough for the authors of this research to infer a cause for this difficulty, because the explanation part of the question is left blank in the answers analyzed.
- Finally, the association of *interrupt* with "context switching" (difficulties 5, 6, and 7) seems to originate again with mistaking the term *interrupt* with the meaning of state. They seem to think that context switching originates an *interrupt* in the running processes, and they assign this interpretation to the term *interrupt*.

In conclusion, one of the most important problems in understanding the concept *interrupt* consists of interpreting the term *interrupt* according to its colloquial meaning of "an effect" (meaning number 4 in Fig. 3) instead of the technical noun meanings (meaning numbers 1 and 2 in Fig. 3).

Conclusions

In this article, a qualitative methodology approach, not often used in engineering education research, has led to successfully discovering misconceptions and their causes, during a longitudinal study of four cohorts of students from 2012 to 2016.

The misconceptions discovered in this study involve the term *interrupt*, a key concept within the fields of communications engineering and computer science because of its relevance to the design, implementation, and programming of the digital systems and devices involved in telecommunications systems and networks and for the understanding and design of an operating system.

Based on the evidence analyzed and the methodology applied, it can be said that the misconceptions around *interrupt* are not only caused by the complexity of the subject itself, but because of other meanings that are completely accepted by society and may keep coming up in every course in this area. This should be a key aspect to be taken into account in order to improve the design of learning experiences around this topic in the future.

Implications for Instruction

The first implication of our study is that teachers should be aware of the misconceptions discovered. Hence, they should not assume students attach the same meaning to the concept *interrupt* that teachers do.

Second, although addressing students' misconceptions is always a challenge, teachers and curriculum developers can build learning experiences that challenge misconceptions to promote conceptual change. In particular, the multiple-choice questions about the concept *interrupt* that formed part of our discovery process can be considered as a model when designing these learning experiences.

Regarding the methodology we have used to discover misconceptions, it may serve as a reference for engineering teachers in order to assess conceptual knowledge, foster deep learning, and deal with student misconceptions in any knowledge area. Moreover, although our study was carried out in an online context, it can be used equally well in face-to-face scenarios.

Implications for Education Research

Interrupt is a troublesome concept involved in several engineering disciplines such as communication engineering and computer science. Our results suggest that it can be a threshold concept as well. Evidence is needed to support this statement, that is, the concept *interrupt* has the other four characteristics of a threshold concept: transformative, integrative, irreversible, and bounded.

Future work might be undertaken with the purpose of overcoming the difficulties discovered, and finding new sets of difficulties and misconceptions around other concepts, in order to draft a concept inventory for subjects involved in training in digital systems and design and programming of devices for communications engineering, such as operating system courses.

References

[1] D. Hestenes, M. Wells, and G. Swackhamer, "Force Concept Inventory," *Phys. Teach.*, vol. 30, no. 3, Mar. 1992, pp. 141–58.
[2] M. Bristow *et al.*, "A Control Systems Concept Inventory Test Design and Assessment," *IEEE Trans. Educ.*, vol. 55, no. 2, May 2012, pp. 203–12.
[3] A. M. Goncher, D. Jayalath, and W. Boles, "Insights into Students' Conceptual Understanding Using Textual Analysis: A Case Study in Signal Processing," *IEEE Trans. Educ.*, vol. 59, no. 3, Aug. 2016, pp. 216–23.
[4] K. C. Webb and C. Taylor, "Developing a Pre- and Post-course Concept Inventory to Gauge Operating Systems Learning," *Proc. 45th ACM Technical Symp. Comp. Sci. Educ.*, 2014, pp. 103–08.
[5] R. A. Streveler *et al.*, "Learning Conceptual Knowledge in the Engineering Sciences: Overview and Future Research Directions," *J. Eng. Educ.*, vol. 97, no. 3, July 2008, pp. 279–94.
[6] J. H. F. Meyer and R. Land, "Threshold Concepts and Troublesome Knowledge: Linkages to Ways of Thinking and Practising Within the Disciplines," *Improving Student Learning Ten Years On*, C. Rust, Ed., Oxford, 2003, pp. 412–24.
[7] S. Pamplona, N. Medinilla, and P. Flores, "Exploring Misconceptions of Operating Systems in an Online Course," *Proc. Koli Calling '13*, 2013, pp. 77–86.
[8] T. S. El-Bawab, "Telecommunication Engineering Education (Tee): Making the Case for a New Multidisciplinary Undergraduate Field of Study," *IEEE Commun. Mag.*, vol. 53, no. 11, Nov. 2015, pp. 35–39.
[9] M. C. Wittrock, *Handbook of Research on Teaching: A Project of the American Educational Research Association*, Macmillan; Collier-Macmillan, 1986.
[10] L. W. Anderson *et al.*, *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives, Abridged Edition*, Allyn & Bacon, 2000.
[11] W. Stallings, *Operating Systems: Internals and Design Principles*, 7th ed., Prentice Hall, 2011.
[12] A. Silberschatz, P. B. Galvin, and G. Gagne, *Operating System Concepts with JAVA*, Wiley, 2011.

[13] A. Casillas and L. Iglesias, *Sistemas Operativos: Problemas y Ejercicios Resueltos*, Pearson, 2004.
[14] M. B. Miles and A. M. Huberman, *Qualitative Data Analysis: An Expanded Sourcebook*, SAGE Publications, 1994.
[15] J. Saldaa, *The Coding Manual for Qualitative Researchers*, SAGE Publications, 2012.

Biographies

Sonia Pamplona (sonia.pamplona@udima.es) is a Ph.D. associate professor at the Universidad a Distancia de Madrid, UDIMA. She is a computer engineer with a degree from the Universidad Politécnica de Madrid (UPM) and a Ph.D. in computer science, also from UPM. She currently teaches undergraduate courses in operating systems and human-computer interaction, and a postgraduate course in mobile learning. Her area of research is engineering education, with a special interest in online learning.

Isaac Seoane [M'01] (isaac.seoane@udima.es) is a Ph.D associate professor and researcher in telecommunication engineering at UDIMA. He obtained his Ph.D. in telematic engineering in 2012 and his Telecommunications degree in 2004 from the University Carlos III de Madrid, Spain. He is currently involved in research projects related to innovation in engineering education, educational technology for STEM courses, and action to address the technology gender gap.

Javier Bravo-Agapito (javier.bravo@udima.es) holds a Ph.D. in computer science and telecommunications from the Universidad Autónoma de Madrid. He has collaborated with researchers from recognized institutions in France and the United States. He collaborated with Prof. Serge Garlatti at Télécom Bretagne, Brest, France, and with Prof. Peter Brusilovsky at the University of Pittsburgh, Pennsylvania. Currently, he is an associate professor at UDIMA, and his research interests focus on e-learning, adaptive educational hypermedia systems, and data mining.

Nelson Medinilla (nelson@upm.es) is a Ph.D. associate professor at UPM. He is an electrical engineer with a degree from the Universidad de la Habana, with a Ph.D. in information technology from UPM. He has worked for 20 years as a professor of electrical engineering. He currently teaches software design courses at the Information Technology Faculty, UPM. His areas of research include software design and engineering education.

The iLab Concept: Making Teaching Better, at Scale

Marc-Oliver Pahl

Lab courses are a great setting to teach. However, to result in a successful learning experience, they often require teachers to spend a significant amount of time. The author reports on how lab courses can be implemented for efficiently supporting learners while significantly reducing the workload of teachers.

ABSTRACT

Lab courses are a great setting to teach. However, to result in a successful learning experience, they often require teachers to spend a significant amount of time. This article reports about how lab courses can be implemented for efficiently supporting learners while significantly reducing the workload of teachers. The presented iLab concept consists of a blended learning teaching methodology and the labsystem eLearning platform that was especially designed for supporting the teaching methodology. Applying the concept results in students and teachers not having to spend time on surrounding tasks that produce overhead, but instead being able to focus on learning and teaching. The iLab concept enables teaching more content in less time. It especially reduces the workload on teachers, making lab courses scale. The iLab concept shows very good learning results with more than 1500 students between 2004 and 2017. The iLab concept was originally developed for teaching students about computer networks and distributed systems. In the meantime, it was successfully used in other domains such as training future teachers.

INTRODUCTION

Lab courses are a great setting to teach. However, to result in a successful learning experience, they often require teachers to spend a significant amount of time.

This article reports on how lab courses can be implemented for efficiently supporting learners while significantly reducing the workload of teachers:

- How can a high amount of knowledge be transferred efficiently for teachers and learners?
- How can students be most efficiently supported in their (self-) learning process?
- How can lab courses become highly scalable?

The presented *iLab concept* consists of a *blended learning teaching methodology*, and the *labsystem* eLearning platform that was especially designed for supporting it.

Together with changing colleagues, including, in historical order, Uwe Bilger, Heiko Niedermayer, Stephan Günther, Benjamin Hof, and Lukas Schwaighofer, the author has been developing the iLab concept since 2004. It has been continuously applied for teaching four computer science courses on computer networks and distributed systems at the Technical University of Munich (TUM) and the University of Tübingen. For details about the curricula see [1, 2]. In 2014 the concept was successfully implemented at the TUM School of Education to prepare future high school teachers.

Until 2017, over 1500 students learned, following the teaching methodology. Their feedback is continuously positive as recent comments exemplify: "Perfectly organized lab course with a good balance of team work, self-study and lecture," "[I like the] syllabus and the way the assignments are organized. The course content, paradigm, and the learning curve," "Good insight into various technologies. Comprehensive exploration of the topics at hand. Nice e-Learning system!"

SHORTCOMINGS OF A TRADITIONAL LAB WORKFLOW

The motivation for developing the iLab concept was a computer networking lab course at the chair of Prof. Carle at the University of Tübingen in 2003. The course was based on the book *Mastering Networks: An Internet Lab Manual* [3]. Its workflow is representative for many lab courses. It starts with a meeting that presents the assignment's topics. Students get printouts including preparatory reading material (*prelab*) and instructions for the hands-on lab. In the following week teams of two to three students meet for the practical part (*lab*) in the lab room. The teams take notes that one team member compiles to a written report afterward. The (paper) journal is handed in together with a digital medium for tool output and so on. Finally, the teaching assistant grades the hand-ins and returns them to the teams.

The described traditional lab workflow has several shortcomings. Important ones that led to the iLab concept are presented and labeled for <s.0> *students* and <T.0> *teachers*.

The *prelab* consists of information that is typically new to *students*. The amount of information and its novelty make it <s.1> *difficult to assess the relevance of the presented material*. The traditional workflow requires high intrinsic motivation as it only provides <s.2> *few motivational elements*. It <s.3> *does not ensure that all team members prepare in the prelab*, resulting in disappointment for

Digital Object Identifier:
10.1109/MCOM.2017.1700394

The author is with Technical University of Munich.

0163-6804/17/$25.00 © 2017 IEEE

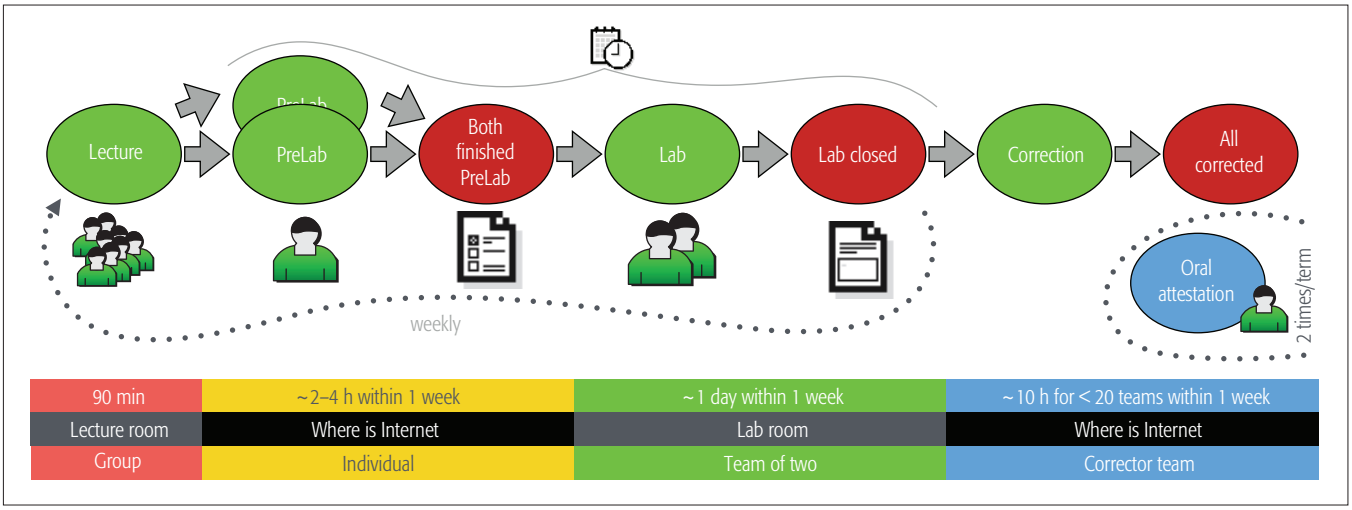


Figure 1. The workflow of an iLab assignment with expected time spans, locations, and settings.

the unprepared students and their team members. Unprepared students have difficulties in understanding the actions of their fellows. Their team members miss valuable support and discussion when solving the exercises together.

In the *lab*, collecting tool output and creating reports from scratch produce <s.4> *significant time overhead*. The time spent on organizational duties cannot be used for learning. Missing tools <s.5> *hinder collaboration between team partners*. The risk occurs that strong team members pull their entire teams through the exercise. As a result, weaker students miss individual tutoring while stronger students miss challenging questions.

<s.6> *Remotely exchanging about specific parts of an assignment with team members or teachers is difficult*. The medium is unclear. Referring to a certain context is difficult and produces overhead: "in question 3 of section 6.4."

Toward students, the *correction* wants to enable learning from mistakes. Mixed media and the missing context in the answer sheets require <s.7> *high effort to understand and learn from reported errors*. <s.8> *Getting an up-to-date overview on the current personal performance is difficult*.

Manual correction and limited instructor resources lead to <s.9> *significant time distance between handing in a report and receiving correction feedback*. Relating the feedback to a student's own actions becomes difficult.

For teachers it is <T.1> *difficult to get insights regarding the progress and performance of their students during an assignment*. It requires physical presence.

<T.2> *Continuously collecting and processing student feedback* is important but work-intensive in the traditional workflow. A standard communication channel is missing. Obtaining inquiry contexts consumes time (<s.6>).

Traditional hand-ins make <T.3> *the correction of prelab and lab very time consuming*. <T.3.1> *Non-standardized hand-in formats* are problematic. Each team has its own tooling and formatting for the hand-in. Different media are involved including paper, USB sticks, and email. Deciphering <T.3.2> *handwritten hand-ins* slows the correction down. Individual presentations <T.3.3> , such as nice handwriting vs. bad handwriting, *bias the correction*.

Cross-correction helps make the correction fairer since answers of the entire class are directly compared. In addition, cross-correction *speeds the correction up significantly*. Correctors only have to get into the answer context once. The traditional workflow <T.4> *lacks cross-correction support*. When correcting per team, a corrector has to go through the entire exercise repeatedly for each team, including many context switches. <T.5> *Not having the instructions and questions inside the lab report* slows the correction down. It requires switching between instructions, questions, and student reports frequently.

<T.6> *Keeping track of the student performance* by summing up credits and manually logging the performance takes time.

The iLab Teaching Methodology

In an ideal world students listen attentively to lectures and interact with their teachers in order to transform presented information into their own knowledge. After a lecture they deepen the learned concepts on their own by reading additional material and discussing it with their fellow students. Driven by the interest to apply their newly obtained knowledge, they meet others to experiment with what they learned. Finally, they are able to reproduce, apply, and extend what they were taught [4, 5]. This last stage ideally lasts for their entire lives.

Often students do not behave as described. However, suitable guidance can help create excellent learners. In a nutshell, the iLab didactic concept implements teaching methods that result in students following the utopian workflow described above and becoming better learners.

The iLab concept fits best for courses that consist of multiple assignments. Students get familiar with the methodology within one to two assignments and learn even better subsequently.

Each assignment follows the same workflow shown in Fig. 1. From left to right, the stages are *Lecture*, individual self-study in the *PreLab*, practical hands-on in the *Lab*, distributed *Correction*, and individual *Oral Attestation*. The bottom of the figure shows the expected time span for each stage, the timeframe a student is allowed to work

The PreLab is entirely done within the specialized labsystem eLearning environment. Students have individual accounts and log in via the Internet. The PreLab consists of learning material that deepens the lecture content. This individual stage allows studying at one's own needs, pace, and depth. To support different experience levels and interests, the PreLab covers diverse aspects of an assignment.

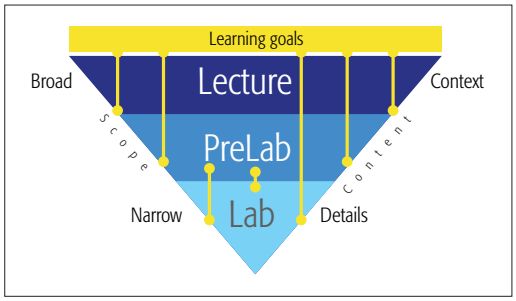


Figure 2. Teaching pyramid of the iLab.

on each part, the location of the activity, and the collaboration setting.

Next, the stages of the iLab concept are detailed with a description, a discussion of the learning theoretical meaning and role within the concept, experiences with the teaching element, and an explanation why and how the element contributes to overcome the challenges identified earlier. The challenges are referenced as <S.0> *I describe how the challenge is solved.*

Lecture

An iLab assignment starts with a lecture of typically 90 minutes. Lecture attendance is mandatory because of its important role as the first and last stages of an assignment in the iLab concept.

Figure 2 shows the distribution of learning content to the stages Lecture, PreLab, and Lab. The Lecture starts with a *discussion of the last assignment*. It then covers the current assignment's context including real-world motivations and applications, live demos, group discussions, and best practices for the Lab. It closes with a teaser on the practical part, motivating students to continue working on the assignment.

From a learning theoretic point of view the Lecture fulfills different purposes. The Lecture is either the students' first contact with or a refresher of an assignment's topics [4, 5]. It gives context, framing the assignment and its learning goals [6]. It provides the basic *structure* for reaching the learning goals. The teaching of content related to the learning goals is depicted as vertical lines in Fig. 2. The lines do not always go from the top to the bottom of the flipped pyramid — some teaching formats are more suitable to teach certain aspects than others. For example, software tools are typically only presented in the PreLab and applied in the Lab, while complex application scenarios are better suited for Lecture and PreLab.

The Lecture is the *community event* of the iLab concept. Perceiving oneself as a member of a group that works together enables better learning [7]. The group setting *encourages active discussion* about the learning content.

Interaction is motivating [4]. We stimulate interaction by regular discussion with the group. Typically, two to three meetings where we encourage discussions and create a positive error tolerant atmosphere break the ice, enabling discussion supported teaching. Besides improving the learning experience, the discussion leads to valuable feedback for further improving the learning material. *Continuous improvement* is central in the iLab concept.

The Lecture should be scheduled at the beginning of a week, giving students the rest of the week to continue with their individual preparation.

PreLab

With the input from the lecture in mind, students have about one week for their individual preparation (Fig. 1). Going through the PreLab typically takes two to four hours. As shown in Fig. 2, the PreLab deepens the assignment's core topics.

The PreLab is entirely done within the specialized *labsystem* eLearning environment. Students have individual accounts and log in via the Internet. The PreLab consists of learning material that deepens the lecture content. This individual stage allows studying at one's own needs, pace, and depth. To support different experience levels and interests, the PreLab covers diverse aspects of an assignment.

It starts with a repetition of the learning goals. This helps students to <s.1> *identify which are the most relevant parts of the learning material*. Most other prelab shortcomings are overcome by introducing *self-correcting multiple choice questions*.

Knowing the intended learning goals and the upcoming Lab, teachers could point out the relevance of each presented aspect in direct interaction. In the distant eLearning setting <s.1> *multiple choice questions are placed within the study material*. They *highlight relevant aspects* and *stimulate interaction* with the learning material resulting in active *learning* [4, 5]. An example is asking, after the corresponding reading sections, to which address type a given IPv6 address belongs.

Via the labsystem, students get immediate feedback. Each question allows three answering attempts. The immediate feedback with hints enables learning from mistakes.

The multiple choice questions are a <s.2> *motivational element*. Students get a score depending on how many tries they needed. This score is only informational since not having to worry about grades fosters explorative knowledge acquisition [5].

Students can see their relative performance in the group (Fig. 3). They can technically only continue with the next Lab stage when their entire team answered all PreLab multiple choice questions. Such gamification elements <s.2> *increase the motivation* [8]. By looking at the reached scores, teachers get a <T.1> *timely and accurate overview on the progress of the students*.

For collecting feedback, <T.2> *the labsystem implements multiple direct feedback channels* including an email form. The coupling with <T.2> *a support ticket system* increased the timeliness and quality of our feedback notably.

The PreLab has three main roles in the iLab. First, it *fosters active learning* by requiring interaction with and thinking about the learning material. Second, it enables students to *independently reach a similar state of preparation*. Third, the PreLab *ensures that both team partners are sufficiently prepared for the practical part*.

Virtual Barrier

The practical hands-on is done in teams. To foster success in the practical part, all team members have to be prepared with the necessary theoret-

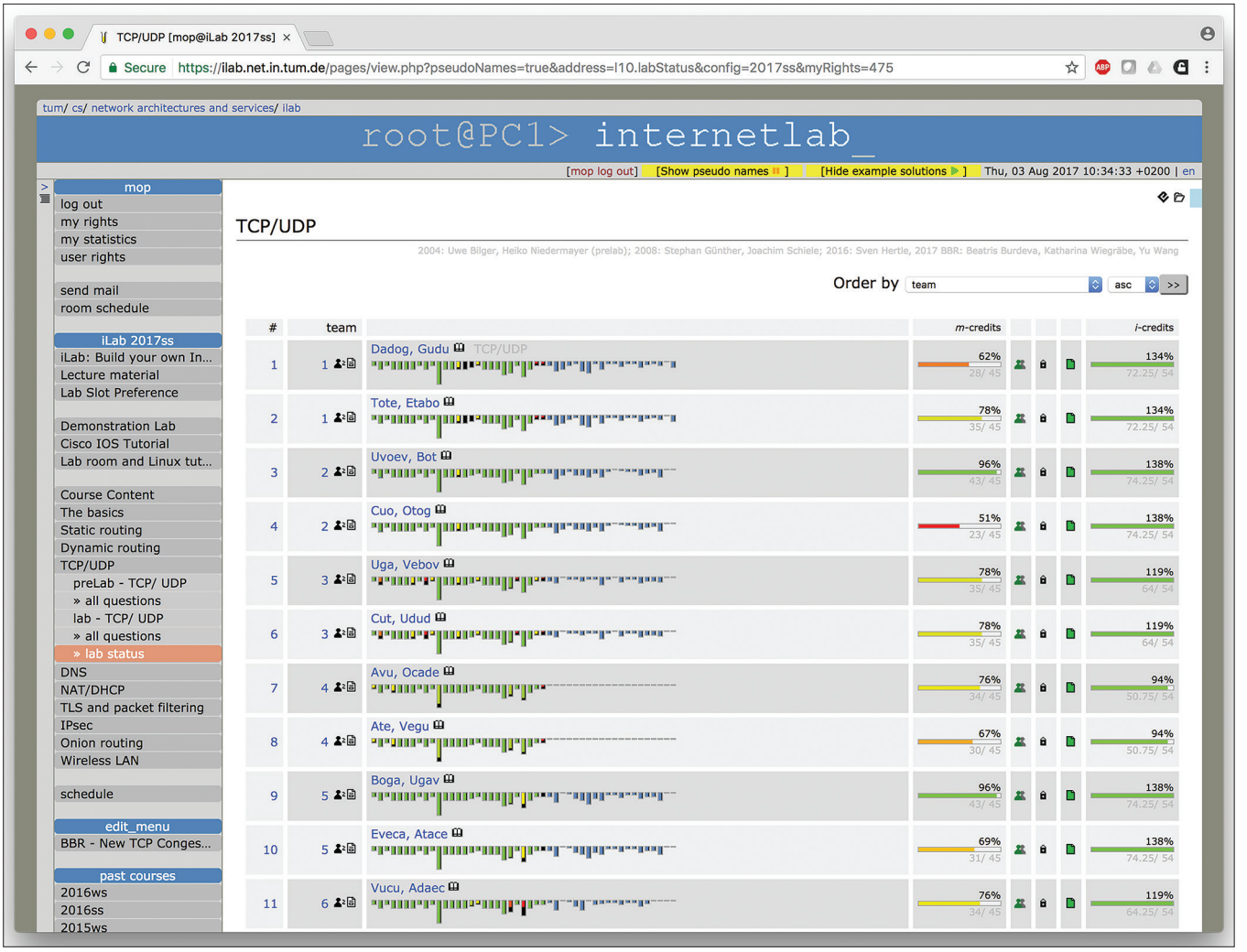


Figure 3. The labsystem.

ical background. For ensuring such preparation the <s.3> *labsystem provides a virtual barrier*: only when all team members have finished all PreLab multiple choice questions can they see the Lab content.

A system-enforced barrier turned out to be highly effective. It is perceived as fair and not questioned. However, with the barrier, the new challenge occurs to prevent cheating by a dumb exchange of answer vectors, *<correct, wrong, wrong>*.

The labsystem prevents such cheating by individually shuffling the order of the multiple choice answers per student. Cheating students now have to communicate the actual answers. Cheating becomes active learning [4].

Lab

Behind the virtual barrier, students see the Lab. The teams get a dedicated Lab workplace called "iLab isle" (Figs. 4 and 5) for an entire day. This enables attending other lectures and so on during a lab day.

The Lab consists of instructions and <T.3.1> <T.3.2> <T.3.3> *questions within the labsystem*. Coming from the book [3], we started having the Lab instructions on paper and the questions online. This media disruption was problematic. Students started guessing the intended activities by only looking at the online questions. As a result, all content is online now.

Each lab has a real-world story that supports the learning. A complex challenge, such as "We share a flat with five others. Our Internet connection seems to be saturated often. How can we change this?" is solved step by step by the students following the given lab instructions. The continuous story guides the students. Being able to continue tells them immediately that they are on the right track.

The implicit confirmation targets <s.2> *becoming proud and satisfied* of having found a solution to a complex problem. In the case of getting stuck, it allows going back with the insight that things might be different than originally thought. At the end of a lab, all teams have solved all tasks. Still, the quality of their answers differs, resulting in different scores. The <s.2> *complexity of the real-world tasks and the immediately perceived success stimulate the learning*.

The lab is not "tutorial style" but more a "guide at the side" [4] of engineers almost independently solving real-world problems. With enough experience in exercise design, it is possible to provide just enough detail not to get lost while giving students enough freedom to solve challenging problems on their own.

Figure 4. Two students collaboratively working on a Lab.

Around 30 questions with free text inputs are part of a Lab story. They allow students to gain insights by applying their previously acquired knowledge, and to create new knowledge by solving the challenges. A typically Lab question is, "Why does X not work?" Active learning happens [4, 5].

For the practical part students are paired in fixed teams of two. Both <s.5> *team partners share the free text answering fields* in their labsystem accounts, encouraging cooperation. A *team lock* feature ensures that only one question is answered at one time. This effectively prevents parallel work. Instead, the collaborators discuss the answers to their shared questions, resulting in *active learning*. Ideally, both switch the writer role between each question.

Working in teams is done for different reasons. The team members *support* each other in solving the Lab challenges, enabling *faster progress*. Working with a team partner is <s.2> *motivating and enables collaborative learning*, increasing the learning success [7].

A team size of two is chosen as it is the most interactive setting. <s.5> *Each isle is designed to support exactly two people*. It has two monitors, keyboards, and mice so that both team partners can research material and do configurations independently (Fig. 5). The <s.5> *exercises are designed to provide tasks for both partners*. This prevents team members always taking a passive part throughout an exercise.

During the lab, students <s.6> <T.2> can send feedback within the eLearning system via tickets and emails. As for the PreLab, <T.1> *teachers can continuously monitor the progress and get detailed feedback statistics* (Fig. 3).

After answering all questions, a team is done with the exercise. <s.4> *The additional effort of switching between media, and, most important, creating a lab report, is removed*. We use the time gained for teaching more content.

Either students close an assignment or a Lab closes automatically based on its schedule (Fig. 1, top). Like with the virtual barrier students perceive a system-enforced deadline as fair and accept it.

Correction

Enabling fast and fair correction at scale was the original motivation for developing the labsystem. <T.3> <T.4> <T. 5> <T.6> The correction happens entirely inside the eLearning system. Correctors connect over the Internet.

The correction starts when all questions of an assignment are closed. This is necessary for <T.4> *cross-correction:* correcting per exercise and not per team. Cross-correction makes the <T.3> *correction faster and enhances the correction quality*.

We apply a four-eye principle with two correction rounds. The first corrector proposes credits, provides feedback to students, and makes comments for the second corrector. The second corrector oversees the proposals and sets the correction visible for the students.

Each corrector is responsible as first and second corrector for a fixed, previously agreed part of an assignment. To reduce biasing effects, beyond having standardized hand-ins, correctors can enable "pseudo names". The labsystem replaces all student names with pseudonyms (Fig. 3). *Standardized hand-ins, distributing the load between correctors, and the anonymization functionality make the correction fair and fast.*

<T.5> *The correctors have exactly the same view as the students* with the full context of each question. This increases correction speed and quality. Three tutors typically correct 10 teams (20 students) within < 10 hours. The labs typically close on Sundays, and the <s.9> *correction is done until the following Thursday*. To keep this tight schedule, the labsystem automatically sends emails to the correctors once a Lab closes.

Having only a short time between answering and getting feedback is ideal for students to learn from their mistakes. The labsystem shows credits and corrector comments immediately to students when an answer is set to the status "corrected," resulting in very fast (partial) feedback. When an entire Lab is corrected, a team receives an automated mail. Getting a timely notification and seeing the corrections directly within the Lab context <s.7> *ease learning from mistakes.*

<T.6> *The labsystem handles all additional management tasks including summarizing the given credits and informing the students about finished corrections*. <s.8> <T.6> It continuously gives students and teachers detailed up-to-date information about the performance including the currently reached credits.

Group Discussion

An iLab term of 14 weeks has about 10 assignments. Each assignment cycle is followed by a lecture that kicks off the next assignment. The first part of this lecture is dedicated to the previous assignment. Based on common mistakes known from the correction, a discussion is started. We clarify remaining open issues and collect feedback.

Oral Attestation

Over the term, each student has two individual oral attestations. During the 15-minute attestation, we ask about relevant learning content from the corrected assignments.

The oral attestation has multiple purposes. Regarding grading, it allows distinguishing

Figure 5. The iLab laboratory at TUM. Each table is one iLab isle.

between both team partners. Regarding the learning, the attestation makes students study the topics once again with a time distance. Relevant topics are thereby covered at least four times with Lecture, PreLab, Lab, and learning for the attestation (Fig. 2).

The oral attestations provide a good indication of whether the learning materials reach their intended goals. They allow raising topics again in the Lecture and adjusting the learning material.

Monitoring the Learning Success

The iLab concept targets efficient self-learning. A major tool is (self-)monitoring [5].

For students, the iLab concept provides various monitoring metrics: the multiple-choice results, the ranking in the group, the discussions with others, the credits from the lab, the correction comments, the discussions in the Lecture, the attestation feedback, and the attestation marks. Teachers can use the same metrics to assess their learning materials and to assess their students.

Grading

For the overall grading we linearly map the reached percentage in all Lab free text questions to a mark that counts for 60 percent. Each oral attestation counts for 20 percent.

Communication

Communication is essential for learning in teams and groups [5]. The labsystem includes several communication channels. Students can send *emails* to other students and the teachers. Links to specific places within the learning material can be included, enabling <s.6> <T.2> context-specific exchange. Teachers can use the functionality to contact the students.

The labsystem <s.6><T.2> interfaces with a standard ticket system for channeling the student-teacher interaction, including requesting help from the Lab room, sending inquiries, and giving feedback for improvements. It automatically adds a link to the exact student context to each ticket, making <T.2> *tutors directly see the corresponding student context*.

We use different ticket categories to prioritize our responses. Help tickets have highest priority, increasing the feeling of *always available support*. Using such tickets significantly increased our answering speed and quality. It made presence in the lab room almost unnecessary, increasing the scalability.

Additional Learning Theoretical Discussion

The iLab didactic concept implements blended learning. It follows a holistic approach covering diverse aspects of learning, including a special methodology, customized tools, and specially designed lab rooms (https://youtu.be/SPOkrnKQ09c). Mixing online and traditional classroom settings already affects learning in a positive way [7].

However, compared to regular blended learning offers, the iLab teaching methodology notably increases the variation in teaching. In contrast to traditional professor-up-front lecture setups where knowledge is primarily received and memorized, the iLab fosters knowledge analysis, synthesis, and personal evaluation. The instructors accompany learners [4].

The iLab concept fosters *active learning* [4, 5] by *varying learning modes*, *settings*, and *locations*. The different stages of the teaching methodology (Fig. 1) feature different teaching methods. The learning *modes* vary from passive knowledge consumption in a lecture setting toward active application and vital discussion of previously acquired knowledge. Students acquire, deepen, apply, and reproduce knowledge. The learning *settings* are group learning, individual learning, and learning in teams of two. The learning *locations* are a classroom, everywhere with Internet access, and the laboratory.

The iLab concept applies a *constructivist learning approach* [4–6]. It targets active involvement of students [9]. After each Lecture, students create knowledge: They read and reason about possible answers to multiple choice questions in the PreLab, and apply the previously (partly) acquired knowledge to new situations in the Lab. They answer thought-provoking questions in the Lab, actively challenging their new knowledge by discussing it with their team partner. Finally, they have to reproduce and apply it in the oral attestation. [4]

Key tools in the iLab methodology are continuous *repetition* and *constructive alignment* [6]. Constructive alignment in the iLab includes being aware of the learning goals and aligning the teaching methods to them (Fig. 2).

The various forms of presentation and the continuous (self-) assessment guide students to reach the learning goals. The concept points out learning goals explicitly and implicitly via guiding the students, for example, with the content selection, the multiple choice questions, and the frame in the Lab. Instead of repeating the same learning

Compared to regular blended-learning offers, the iLab teaching methodology notably increases the variation in teaching. In contrast to traditional professor-up-front lecture setups where knowledge is primarily received and memorized, the iLab fosters knowledge analysis, synthesis, and personal evaluation. The instructors accompany learners.

The iLab laboratory room.

Using the lab system as a teacher.

Using the lab system as a student.

Labsystem on GitHub.

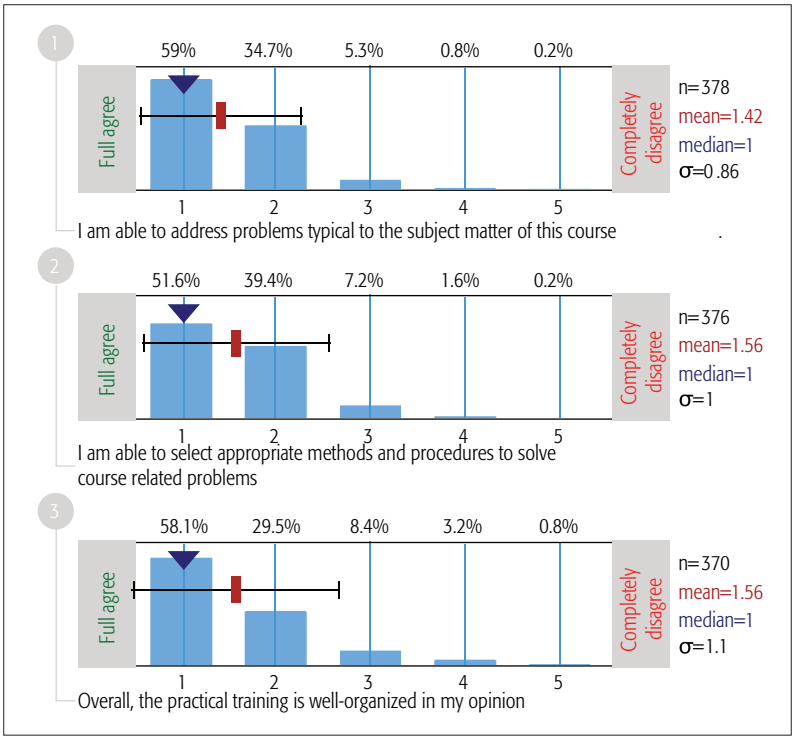


Figure 6. Teaching evaluation results from TUM (2012-2017).

materials multiple times, the iLab concept fosters the *presentation of knowledge from different perspectives*.

Repetition and variation stimulate the interaction with the subjects. The repeated interaction with the learning content from different perspectives stimulates the creation of *cognitive connections*. Ideally such connections are created in or before the first stage, the Lecture. Throughout the other phases of the concept, these connections are ideally strengthened. [4, 5]

The iLab concept creates long lasting knowledge. A student reported that after having been to industry for several years, the iLab content would be among the few things he would still actively remember from university.

The Labsystem eLearning Platform

The *labsystem* is the eLearning system that enables the described iLab concept workflows. The open source system [10] evolves continuously with the teaching methodology.

Unlike in 2004, many eLearning systems exist today. However, they are typically not developed around a specific teaching methodology. Instead, they are general-purpose, such as Moodle (http://moodle.org/). Consequently, implementing the iLab concept with them is complex, and results in continuous configuration overhead and usability inconvenience.

Built for the described methodology, the labsystem provides native support for the presented teaching methodology. Its core functionality was detailed earlier. Summarizing, the labsystem provides content, course, and learning management. The key features that increase the learning experience and enable the iLab concept to scale are barriers, gamification, fixed deadlines, cross-correction, tickets, and course management. Figure 3 shows the course management functionality with detailed student results for an assignment.

Differentiating features compared to other existing eLearning systems include the virtual barrier, the tight integration of a ticket system, automated email sending, and privacy-aware data handling including anonymization, decreasing the correction bias.

A video on how the labsystem is used by students and teachers can be found in [11, 12] together with many video tutorials.

Evaluation

An important part of the iLab concept is continuously collecting feedback. The quotes in the introduction are exemplary of the continuously encouraging feedback.

Until 2017, more than 1500 students participated in courses applying the iLab concept. These courses are competing for students with other offers. The relatively high number of participants indicates the popularity of the concept and the topics.

At Technical University of Munich the iLab courses participate in the faculty-wide teaching evaluation. Figure 6 shows results from the past five years. The diagrams show the feedback to the questions:

- I am able to address problems typical of the subject matter of this course.
- I am able to select appropriate methods and procedures to solve course-related problems.
- Overall, the practical training is well organized in my opinion.

The histogram bars show how the student votes are distributed. The sample size (n) is given on the right. The arrow shows the median, the thick indicator the mean, and the candles the standard deviation (σ) to each side of the mean.

The very good results in the first two questions show that the teaching methodology is successful. Central didactic goals of a lab course are reached. The third question explicitly asks for the organization of the course where the iLab concept is a major part. The answer shows that the iLab concept provides high-quality teaching.

Conclusion

The presented iLab teaching methodology improves the learning success in university lab courses, while it significantly reduces the workload on teachers. The iLab concept enables learning by doing in a way that turns out to be fun for the students and very efficient for teachers. The described use of blended learning allows focusing on content, while the described methodology ensures high-quality, time-efficient teaching. Established workflows are used, and most of the organization is automatically done by the labsystem.

The online parts of the labsystem introduce high flexibility for learners and teachers. The concept's workflow support enables even inexperienced teachers to create high-quality teaching experiences.

References

[1] M.-O. Pahl *et al.*, TUM, lab courses IN0012, IN2106, IN4060, IN8016, "ilab – Build Your Own Internet," https://ilab.net.in.tum.de/, 2004–present, accessed Aug. 16, 2017.

[2] M.-O. Pahl and L. Schwaighofer, TUM, Lab Courses IN0012, IN2106, IN4097, IN8018, "ilab2 — You Set the Focus," https://ilab2.net.in.tum.de/, 2011–present, accessed Aug. 16, 2017.
[3] J. Liebeherr and M. E. Zarki, *Mastering Networks: An Internet Lab Manual*, Addison-Wesley Longman, Aug. 2003; online material, http://www.tcpiplab.net/, accessed Aug. 16, 2017.
[4] A. King, "From Sage on the Stage to Guide on the Side," *College Teaching*, vol. 41, no. 1, Jan. 1993, pp. 30–35.
[5] K. P. Cross, "Learning Is about Making Connections, The Cross Papers Number 3," Jan. 1999.
[6] J. Biggs, "Enhancing Teaching through Constructive Alignment," *Springer Higher Education*, vol. 32, no. 3, Oct. 1996, pp. 347–64.
[7] A. P. Rovai and H. Jordan, "Blended Learning and Sense of Community: A Comparative Analysis with Traditional and Fully Online Graduate Courses," *Int'l. Review of Research in Open and Distance Learning*, vol. 5, no. 2, Aug. 2004.
[8] A. Iosup and D. H. J. Epema, "An Experience Report on Using Gamification in Technical Higher Education," *Proc. 45th ACM Tech. Symp. Comp. Sci. Educ.*, Mar. 2014.
[9] L. C. McDermott, "Millikan Lecture 1990: What We Teach and What Is Learned — Closing the Gap," *American J. Physics*, vol. 59, no. 4, Apr. 1991, pp. 301–15.
[10] M.-O. Pahl and L. Schwaighofer, The Labsystem Elearning System Open Source Project on GitHub, https://github.com/m-o-p/labsystem/, 2004–present, accessed Aug. 16, 2017.
[11] M.-O. Pahl, Video Tutorial on How Students Use the Labsystem, https://youtu.be/KsmLexPe1FQ, 2017, accessed Aug. 31, 2017.
[12] —, Video Tutorial on How Instructors Use the Labsystem, https://youtu.be/zCMDSEMEKVM, 2017, accessed Aug. 31, 2017.

Biography

Marc-Oliver Pahl (pahl@net.in.tum.de) leads the IoT Smart Space Orchestration team at the Chair for Network Architectures and Services at Technical University of Munich, Germany (http://s2labs.org/). He has been doing educational research since 2004. He received the Ernst Otto Fischer teaching prize in 2013. He received awards for excellence in teaching in 2014 and 2015. In 2016 he received a grant for creating a Massive Open Online Course about the Internet (http://s2labs.org/?site=mooc4masters).

ACCEPTED FROM OPEN CALL

A Withered Tree Comes to Life Again: Enabling In-Network Caching in the Traditional IP Network

Kaiping Xue, Tingting Hu, Xiang Zhang, Peilin Hong, David S.L. Wei, and Feng Wu

The authors propose in-network caching in IP-based networks by adding a content identifier into a newly defined IPv6 extension header, where the new architecture is named CAIP. CAIP abandons the complicated name-based forwarding table in ICN, and instead integrates IP routing lookup with cache index lookup, which is compatible with the IP network and also inherits the proven advantages of ICN.

ABSTRACT

This article presents our work proposing in-network caching in IP-based networks by adding a content identifier into a newly defined IPv6 extension header, where the new architecture is named CAIP. CAIP abandons the complicated name-based forwarding table in ICN, and instead integrates IP routing lookup with cache index lookup, which is compatible with the IP network and also inherits the proven advantages of ICN. Cache index exchanging and cooperative caching are implemented between one-hop CAIP enabled neighboring routers, which is simple but efficient. Moreover, for per-chunk caching, as an extension, bitmap is introduced to merge multiple request packets into one. Performance analysis shows that CAIP gains significant performance improvement in terms of access delay and traffic load.

INTRODUCTION

Recently, Internet access from mobile devices has grown dramatically in popularity, and according to Cisco's VNI report [1], 80 percent of all IP traffic will be represented by video traffic in 2019. Hence, content retrieval applications will contribute to most Internet traffic. However, while the demand for multimedia contents has increased tremendously in recent years, the capacity growth of the wireless link, mobile radio network, and mobile core network cannot practically cope with the explosively growing bandwidth demand. Moreover, the IP network is designed and treated as dump pipes, which makes the network carry a large number of duplicate data.

In fact, consumers are usually interested in the contents themselves rather than where they are located. Meanwhile, in the current Internet, a significant proportion of the tremendous increasing network traffic is from duplicate requested contents. Therefore, a feasible method should be integrating content cache and delivery as a legacy network feature, which means the contents can be cached by any network entities equipped with high-performance storage. By means of some new network technologies, such as software defined networking (SDN) and locator/identifier separation, some research work has introduced in-network caching into the architecture design, such as [2, 3]. However, these schemes usually require centralized servers to dynamically maintain the mapping of cached contents and the corresponding locations, which will lead to the scalability issue. Based on the consideration of fully distributed processing, a number of innovative new Internet infrastructures shifting from the current host-to-host communication model to a receiver-driven content retrieval model have been proposed. These innovative network design schemes for the future Internet architecture are uniformly called information-centric networking (ICN)[4], which is currently being investigated and developed in several projects, such as Named Data Network (NDN) [5] and Data-Oriented Network Architecture (DONA) [6]. In spite of some distinctive differences between them (e.g., content naming, security mechanisms, routing strategies, cache management), they share a common property of a receiver-driven data exchange model based on content names (or identifiers). The primary goal of ICN is to facilitate in-network caching for universal content caching in every internal network node, and it enables routers to cache passing-by data to satisfy subsequent requests. It can effectively reduce the distance between the consumers and the content data, and the caching policy can adapt to any dynamic traffic without any specific deployment.

However, this type of clean-slate approach has created a trajectory that is to replace the current IP-based Internet. ICN comes with some significant drawbacks and is complicated to execute [7]. For instance, replacing IP by ICN as the main Internet protocol comes with burdens of not only tedious standardization procedures, but also agreements among many stakeholders involved in the current Internet, such as operators, vendors, and policymakers. Moreover, although the idea of hierarchical name structure is introduced in NDN, the huge volume of contents in the Internet still requires a huge number of content prefixes, which can result in the size of the forwarding information base (FIB) in NDN usually being several orders of magnitude larger than that of the IP routing table in the current Internet. Meanwhile, routers in NDN need to handle routing updating due to content publishing or deletion, and caching update as well as caching policy, which results in the FIB being updated much more frequently than the traditional IP routing table.

Furthermore, some aspects of ICN have not yet been recognized and still require practical

Digital Object Identifier:
10.1109/MCOM.2017.1600487

Kaiping Xue, Tingting Hu, Xiang Zhang, Peilin Hong, and Feng Wu are with the University of Science and Technology of China; David Wei is with Fordham University.

0163-6804/17/$25.00 © 2017 IEEE

solutions, such as content name resolution, and efficient name-based forwarding table maintenance and lookup (e.g., the longest prefix matching of a variable-length hierarchical name). In addition, it is worth mentioning that a certain percentage of traffic (about 30 percent now, and it takes up a higher proportion of the session number) is still generated by end-to-end sessions and relies on host addresses (e.g., voice calls, emails, instant messages), in which involved contents are not repeated or requested by multiple customers. For these end-to-end sessions, compared to the IP network, ICN has no any advantages.

Therefore, an awkward situation arises. On one hand, with the progress of technologies, core network devices can have strong computing power and large enough storage capacity, in addition to transmitting capacity, but not be well utilized to reduce redundancy. How to leverage core network elements' growing computing and storage capacity to reduce redundancy is a pressing issue that demands prompt solutions. On the other hand, the future Internet architecture aims to replace the current network, but it is still debating, and more research is ongoing. Thus, no one else can completely replace the existing network architecture within a predictable time.

Therefore, the appropriate approach is to gradually improve and evolve the current IP network architecture, which requires the features of optional in-network caching, supporting legacy TCP/IP-based applications, and being capable of processing traditional IP packets. We define an IPv6 extension header to make legacy IP packets content-aware. Also, we still use traditional IP routing and use longest prefix match (LPM) for IP forwarding. Recently, Detti *et al.* [8] defined an IP option to make IP packets content-aware; their work is named CONET. They do a measurement to show that the added IP option handling is not a critical performance bottleneck. However, it is essentially still an ICN scheme over an IP network, which has complicated name-based forwarding tables, and a source routing mechanism is introduced to guarantee bidirectional communications to traverse the same routers. The main purpose of CONET is to achieve the coexistence of traditional IP and ICN, and finally transit the network architecture from IP to ICN. However, the goal of our scheme is to enable traditional IP to have in-network caching with no need for complicated name-based forwarding. Like the Network Address Translation (NAT) technology, CAIP can make traditional IP to come to life again.

Our contributions in this article can be summarized as follows:

•We define a new IPv6 extension header to carry the content identifier, which can make legacy IP packets content-aware and enable the routers to have the capacity of in-network caching and content retrieval. For routers in the network, the processing of the extension header is optional, which means that it does not require all routers to support this newly designed extension header and the corresponding function processing. The existing legacy routers can still work well and just treat the IP packets with the new extension header as the legacy ones so that the new network architecture can be deployed step by step. The more CAIP enabled routers, the better the performance of the system.

•We enable each CAIP enabled router to implement local lookup in a cache index table and legacy IP-based routing with no need for complicated name-based routing and forwarding. This way, it can still deal with in-network caching and fast forwarding. The cache index exchanging and cooperative caching are implemented between one-hop neighboring CAIP-enabled routers, which can increase the local cache hit ratio without much interaction complexity.

•Aiming to reduce the number of request packets, we introduce the bitmap structure into the newly defined IPv6 extension header.

The rest of this article is organized as follows. We state our motivation and describe the proposed CAIP framework in the following section. Then operations in a router are described. Neighboring CAIP router discovery and cooperative caching policy are given. Moreover, for per-chunk caching, an extension of introducing bitmap in CAIP is proposed to reduce the number of requested packets. Then performance evaluation is shown, and the final section draws conclusions about our work.

Our Proposed CAIP Framework

The Main Design Motivation

CAIP is a new architecture to achieve in-network caching in the current IP network, which needs no change of the current host-to-host communication model; also, it is content-aware for content retrieving services. On one hand, because CAIP is intrinsically supported by the underlying IP routing, such end-to-end communication sessions could continue to be supported, and CAIP can achieve fast IP forwarding. On the other hand, CAIP facilitates in-network caching in core network so as to ensure that the content services such as video streaming can be fetched much more closely and quickly.

Basic Components of CAIP

For achieving in-network caching in the current IP network, a content identifier is carried in the extension header of an IPv6 packet, which makes the IP packet content-aware, so routers can quickly process the packet without deep packet inspection to fetch the content message. In addition, CAIP enabled routers can handle the processing of the new defined extension header in IP packets, manage content record entries, and have the capability of caching. Figure 1 shows the process of establishing a service session in CAIP, and also gives the format of the novel IP extension header. Such an architecture enables the following three main functionalities related to the life cycle of a session with the newly defined IPv6 extension header support.

IPv6 Extension Header Pre-Process (Box1): A consumer can send out one of two types of request messages: a legacy IP packet without the newly defined IPv6 extension header support, or an IP packet with the newly defined IPv6 extension header support. Whether it has the IPv6 extension header with the content identifier or not is determined by the type of service (e.g., for

For achieving in-network caching in the current IP network, a content identifier is carried in the extension header of an IPv6 packet, which makes the IP packet content-aware, so routers can quickly process the packet without deep packet inspection to fetch the content message.

Harnessing the content identifier in a novel IPv6 extension header can help the current IP network be content-aware, which would benefit from ICN specific in-network caching policy. This approach can be implemented with good support of current IP network without making substantial changes.

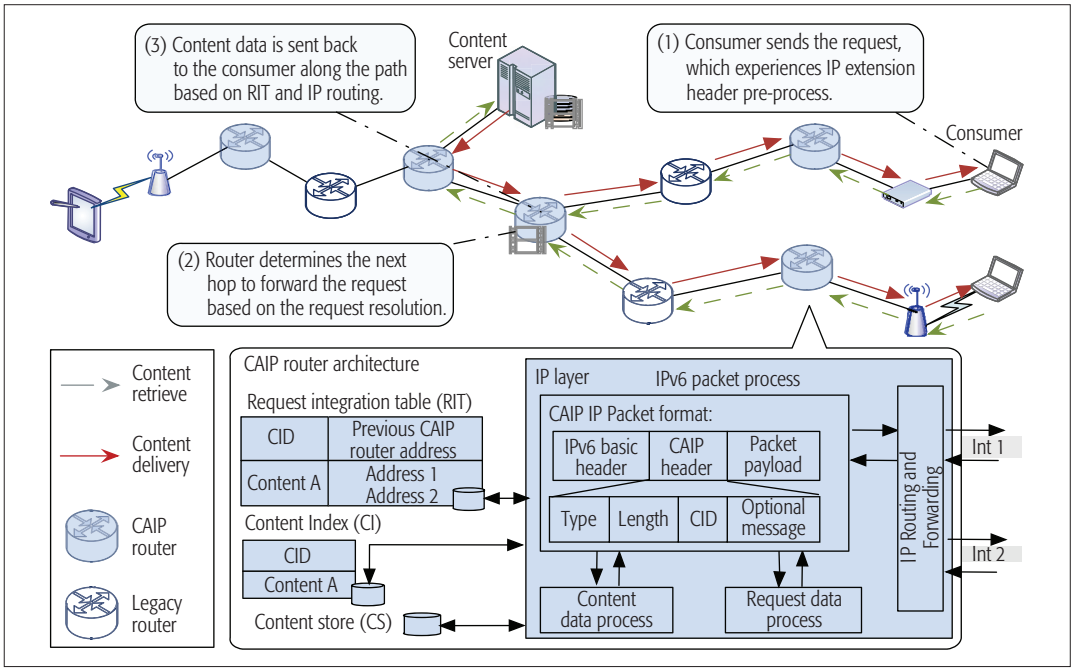


Figure 1. CAIP framework.

video service, this extension header is required, while for email service, it is not).

Request Resolution (Box2): For a packet with IP extension header support, the CAIP enabled router checks whether it has the requested content related to the identifier in the new defined extension header of the IP packet. If not, it will then determine the next hop toward the destination server based on IP routing. The request routing decision is then carried out hop by hop at each CAIP router until the content is reached.

Delivery (Box3): Content delivery is along the reverse of the request path based on forwarding states created during the associated request phase. In order to realize reliable transmission hop by hop, the packets should be reassembled first as in ICN. For sake of simplicity, in this article, we consider a packet carrying a small chunk without fragments.

Harnessing the content identifier in a novel IPv6 extension header can help the current IP network be content-aware, which would benefit from an ICN-specific in-network caching policy. This approach can be implemented with good support of the current IP network without making substantial changes. For data transmission of multimedia services, we define the IPv6 extension header format as shown in Fig. 1. The Type field indicates whether the IP packet is upstream (a request packet) or downstream (a content data packet),which should be further allocated by the Internet Assigned Numbers Authority (IANA). The Length field gives the variable length of the IP extension header in bytes. The CID field specifies the content identifier to identify the carried or requested content data.

Routers in CAIP can be classified as legacy routers, which do not support CAIP handling, and CAIP routers, which support CAIP handling. The deployment of CAIP routers is managed by the infrastructure providers (e.g., network operators) according to the consumers' geographical distribution, historical traffic load statistics, operators' will, and so on. From the perspective of performance effect, for consumer-concentrated areas and aggregated traffic areas, the more CAIP routers are deployed, the better the performance.

Each CAIP router contains three more content management related components than the legacy ones: a request integration table (RIT), a content index (CI), and a content store (CS). The forwarding model of a CAIP router is the same as the legacy IP router, which is based on the IP routing table and forwarding table. The CS is a temporary cache of chunks in accordance with cache strategies. In order to facilitate querying the presence of the requested content in the CS, we consider using a CI, which maps the chunk identifier to the cache position that points to the local CS. The RIT contains forwarding state information for each unacknowledged Interest packet, that is, maintaining the addresses of the former-hop neighbor routers from which each individual request comes, thereby ensuring that the data can be responded to correctly. A CAIP router basically receives one or more requests for the same content identifier forwarded from its neighboring routers, and the router only needs to forward it once so as to avoid duplication of the request forwarding. In other words, if the requested content identifier exists in the RIT, forwarding the subsequent request for the same content would be no longer performed in a CAIP router, and just one copy of content traverses the reverse path to this CAIP router, which in turn significantly reduces the traffic load.

Operations

The corresponding process that is required to be implemented within each CAIP router is best illustrated by the example in Fig. 2. The retrieval of content data involves a sequence of packet processing phases, in which the upstream and downstream processes are discussed separately.

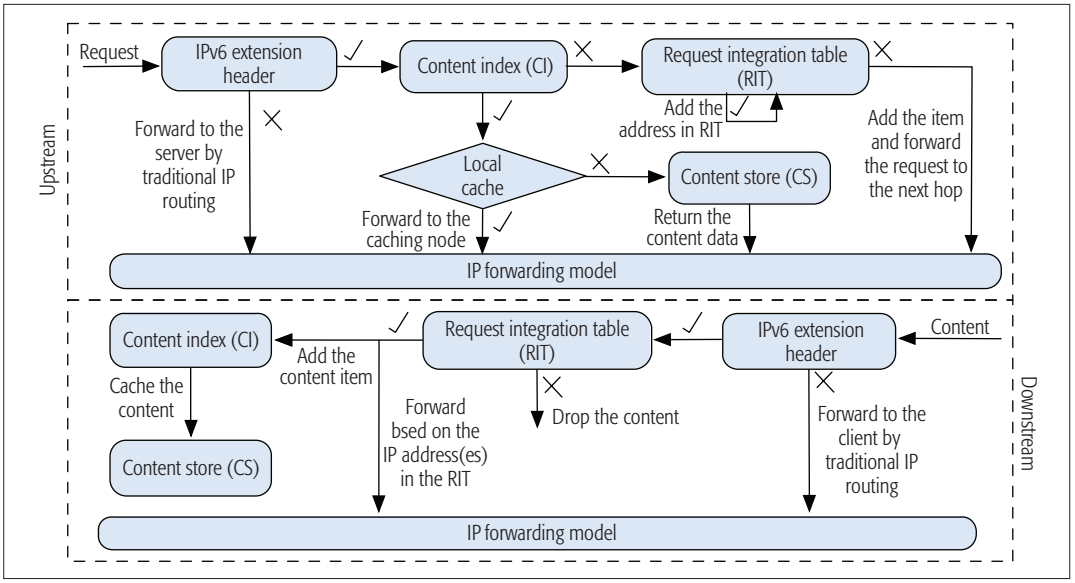


Figure 2. Forwarding process at a CAIP router.

Upstream Process: A consumer requests the service by issuing a request, which includes the content ID in the IP extension header. Meanwhile, the consumer should know the destination IP of the service and include it in the IP header. Legacy routers treat these types of request packets and legacy packets the same, and just forwards packets to the next hop according to the IP routing. For a CAIP router, each packet will experience a series of the following processes. The CAIP router first confirms whether there is a new defined extension header in the packet. For a request packet without the extension header, the router will simply forward it based on IP routing. Otherwise, the router will implement the following steps to reflect the benefit of in-network caching in CAIP. First, the CAIP router checks the CI for matching data. If there is an entry for the given content ID, the router will return the content data to the former CAIP router, whose IP address is recorded in the fourth field in the IP option of the received request packet. Otherwise, the router checks whether the content ID exists in its RIT. If a matching entry exists, it only needs to add the IP address of the former CAIP router in the matched RIT entry. If there is no matched entry, the router will add a new record in the RIT entry, which includes the previous CAIP router address. Then the CAIP router forwards the request to the next hop based on the IP routing. Moreover, the same operation is performed continuously in the path to the destination, until a suitable intermediate caching node is reached that has cached the corresponding content, or the destination server to retrieve the corresponding content in the worst case. The content server in CAIP should also be modified to process the IPv6 extension header to get the content identifier and then return the corresponding content data, which is also carried in packets with the new defined extension header.

Downstream Process: When receiving content packets with a CID, for a legacy router, it just forwards the packet to the next hop based on the IP routing. For a CAIP router, it will look up the RIT and forward the data packet to all nodes with the IP addresses listed in the entry with this content ID. As there are several legacy routers between the two neighboring CAIP routers, we can use an IP-in-IP tunnel mechanism (or other suitable means) to achieve direct transmission between the two neighboring CAIP routers. Subsequently, it wipes the matching RIT entry, and caches the content into the CS, and also updates the index table in the CI.

Content packets always go through the reserve path of a corresponding request packet. This means that each content packet can pass in reverse the CAIP routers the related request packet has passed. This mechanism ensures the feasibility of request aggregation, and the content data delivery can be made only one in some paths and separated to multiple ones to the downstream nodes according to RIT entries.

Neighboring CAIP Router Discovery and Cooperative Caching Policy

An important issue behind our architecture is deploying a simple but effective caching policy to make the best use of in-network caching and reduce redundancy. In this article, we do not adopt a complicated cooperative caching mechanism in our architecture, but only introduce a cache index exchange and one-hop neighbor cooperative caching mechanism named NCC.

At first, the CAIP router needs to discover its neighboring CAIP routers. In general, if two routers are at a distance of one hop physically, they can be defined as neighboring routers. However, there might be some legacy routers lying between the two neighboring CAIP routers in our architecture, as shown in Fig. 1. Thus, we cannot use the usual way of broadcasting to discover the neighboring router, which may induce large-scale network traffic. Hence, we need a simple but effective strategy to achieve neighboring router discovery.

As previously mentioned, we justify that due to the request integration, each content packet should reversely pass the CAIP routers through which the related request packets have already passed. The simple method is that the latter

The content server in CAIP should also be modified to process the IPv6 extension header to get the content identifier, and then returns the corresponding content data, which is also carried in packets with the new defined extension header.

We found that in this basic operation mechanism, the IP addresses of the previous CAIP routers have been recorded during the upstream and downstream. We thus can also use this mechanism to find neighboring routers with no need of much more extra overhead.

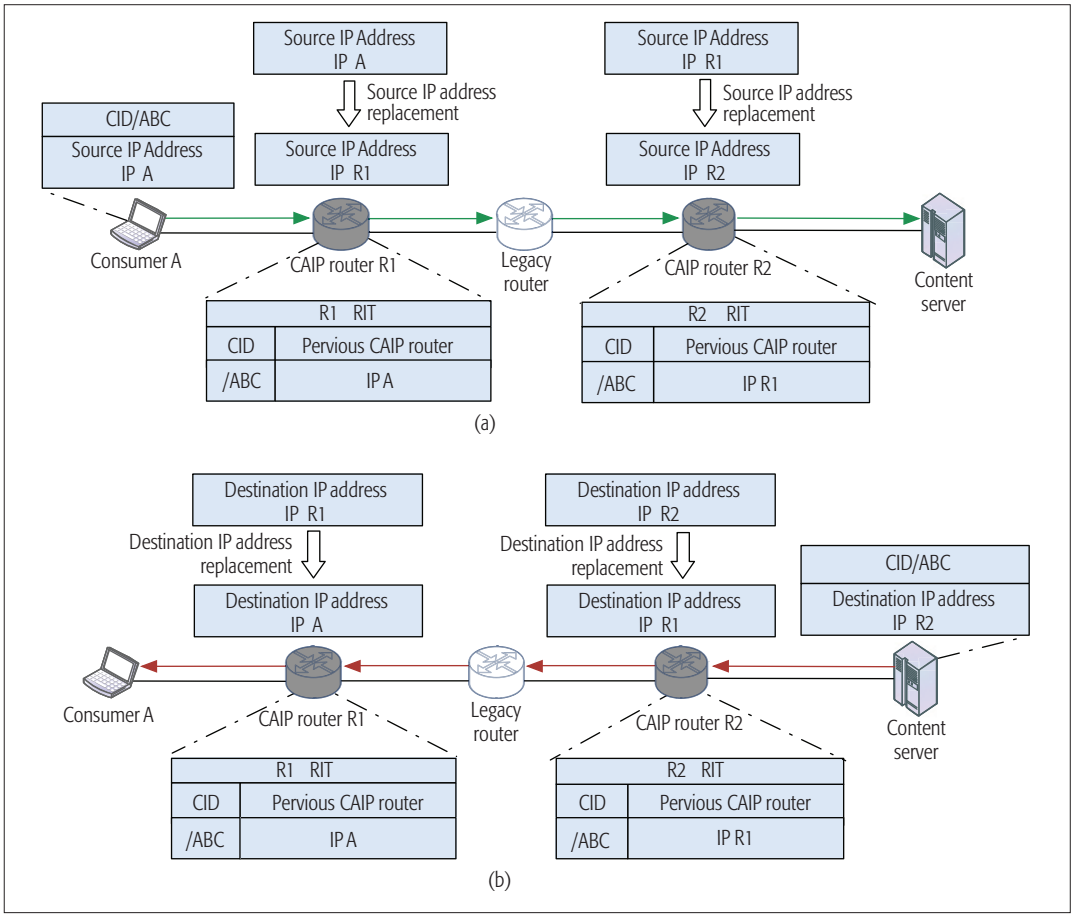


Figure 3. IP address replacement scheme: a) upstream IP address replacement; b) downstream IP address replacement.

CAIP router maintains IP addresses of the previous CAIP routers during the upstream process. However, there might be some legacy routers lying between the two neighboring CAIP routers. Therefore, we propose an IP address replacement scheme, shown in Fig. 3, to guarantee that content packets always go through the reverse path of the corresponding request packets. We have found that in this basic operation mechanism, the IP addresses of the previous CAIP routers have been recorded during the upstream and downstream. Thus, we can also use this mechanism to find neighboring routers with no need for much more extra overhead.

The IP address replacement scheme works as follows. As Fig. 3a shows, when the CAIP router R1 cache misses and the RIT in the CAIP router R1 has no corresponding item, the source IP address will be recorded into the RIT. Then R1 should let the next CAIP router or the content server know its IP address due to the request integration. Therefore, it replaces the source IP address with its own IP address. Then the CAIP router R2 that receives this request would know about the IP address of R1. Since there are no other CAIP routers between these two CAIP routers, we can define these two CAIP routers as neighbors. For the downstream, as shown in Fig. 3b, R2 can deliver the content to R1 by means of replacing the destination IP address with R1 based on RIT. This IP address replacement scheme not only guarantees the path symmetrically, but also achieves neighboring router discovery simultaneously.

When the neighbor relationship is established, routers are still not aware of what contents neighboring routers have cached when making caching decisions. To eliminate this unawareness, each CAIP router exchanges its own cache index with its neighbors. With neighbors' cache indices and their own cache indices, nearby routers can implicitly cooperate to serve each other's requests. NCC is a real-time policy. When receiving a new content object, the router should determine whether to cache it or not. The router first checks its neighbor table. If the new content object is already presented in the neighbor table, this indicates that at least one of its neighbors has cached the content, and the router will not cache it again. Otherwise, the router adds the new content object to its own CS.

Furthermore, other cooperative caching policies for ICN [9, 10] can also be further considered for our proposed architecture. We investigate simpler but more effective cooperative caching mechanisms and give a detailed performance evaluation.

Extension: Per-Chunk Caching

In the vast majority of ICN studies, due to the limitation of the underlying maximum transmission unit (MTU), a large chunk must be fragmented to several fragments and re-assembled at each router. For the re-assembling operation, all the frag-

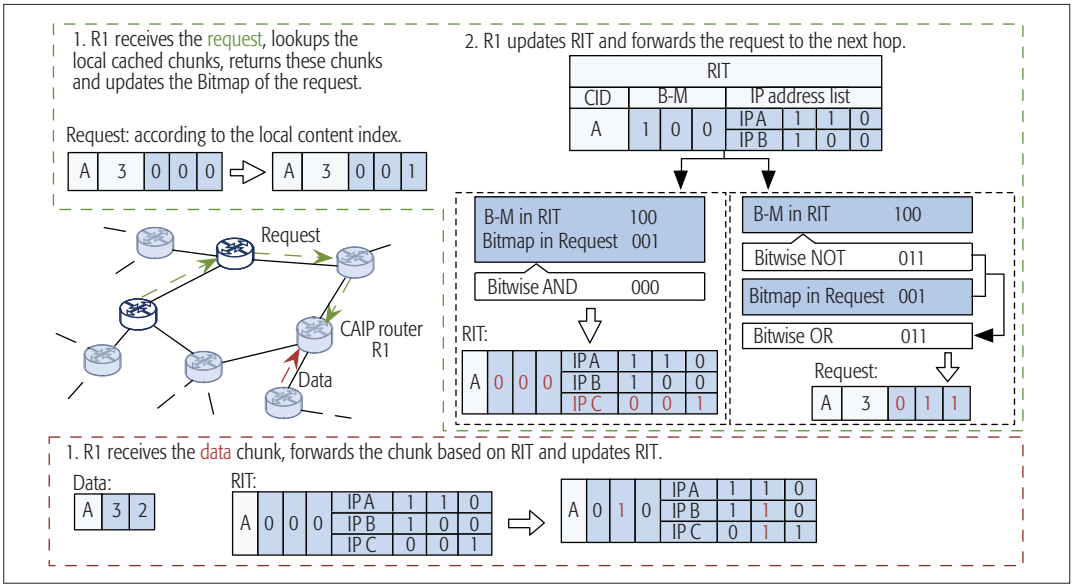


Figure 4. Per-chunk caching.

ments must be gathered, which leads to spending much time on gathering. Small size chunks can avoid being fragmented into too many fragments and are flexible for being cached in the router, but the data consumer has to send Interest packets frequently, especially for high definition video delivery. Therefore, we adopt small chunk size for transmission, and group multiple sequential chunks as a big chunk (named Content).

Here, we further describe how we use a bitmap structure in the defined IP extension header to reduce the number of request packets. We update the CID field in the request packet to the triple <CID, bitmap length, bitmap> and update the CID field in the data packet to the triple <CID, total number of chunks, Chunk No.>. Each content consists of multiple sequential chunks, which are identified by sequential sequence numbers as 1, 2, ... Each request packet is embedded with a CID with an optional bitmap structure, which can be utilized to request all or part of the whole content. The bitmap length indicates the number of chunks in a content that the user wants to request. In an *n*-bit bitmap, each bit represents a chunk and sets to 0 or 1, where 1 indicates that the chunk with the corresponding sequence number has been received, and 0 indicates that the corresponding chunk has not yet been received and is still in request. RIT should also be modified in such a way. For each CID, there are multiple entries, where each value of "previous CAIP router address" is followed by a corresponding bitmap. We also define a new bitmap structure named B-M following each CID in RIT, which is the result of the bitwise operation of AND on all bitmaps corresponding to the same CID.

As shown in Fig. 4, during the upstream process, when each CAIP router along the path receives a request packet, it first checks whether it has the chunks corresponding to the bits of the bitmap with 0. (If some chunks are confirmed to be cached in one of the neighboring CAIP routers, it will redirect the request to fetch the chunks.)Then this router reversely provides the matched chunks and modifies the corresponding bits of the bitmap in the request to 1. If the updated bitmap is not full of 1s, the router stores it together with the source IP in the IP header of the packet (as the "the previous CAIP router" value) in RIT. Then the router further updates the bitmap in the packet by implementing the bitwise OR operation on the bitmap in the request and the result of the NOT operation of the corresponding B-M in RIT. The router further forwards the updated request packet to the next hop if there are still one or more 0s in the bitmap. Finally, the router updates B-M. During the downstream process, if when receiving a content data chunk, the router finds all the previous one-hop addresses with the specific CID and having 0 in the specific bit of the bitmap, it forwards the chunk to all these routers. The router then updates 0 to 1 in the above-mentioned bits and updates B-M. If there is a bitmap full of 1s, the corresponding IP address and bitmap can be removed from the specific CID.

Performance Evaluation

The performance of CAIP was evaluated using NS3. We use the BRITE topology generation tool to generate the test topology [11]. Based on Waxman's probability model [12], the topology consists of 100 routers (setting CAIP routers randomly), 20 end hosts, and 20 original servers. The data access pattern is Zipf distribution [13], which states that the relative probability of a request for the *i*th most popular content is proportional to $1/i^{\alpha}$ with the shape parameter α. As our work is focused on how to achieve in-network caching in an IP network and improve cache utilization, we mainly evaluate the performance through two metrics:

- Cache hit ratio, which is defined as $(N_a - N_s)/N_a$. N_a is total number of requests generated, and N_s is the number of server hits.
- Traffic load, which is calculated by $\Sigma_{all\ requests}(RequestContentSize \cdot HopCount)$. We use percentages on the vertical axis for this metric, which represents the traffic load ratio of one specific scheme to the scheme of no caching.

For each CID, there are multiple entries, where each value of "previous CAIP router address" is followed by a corresponding bitmap. We also define a new bitmap structure named "B-M" following each CID in RIT, which is the result of the bitwise operation of "AND" on all bitmaps corresponding to the same CID.

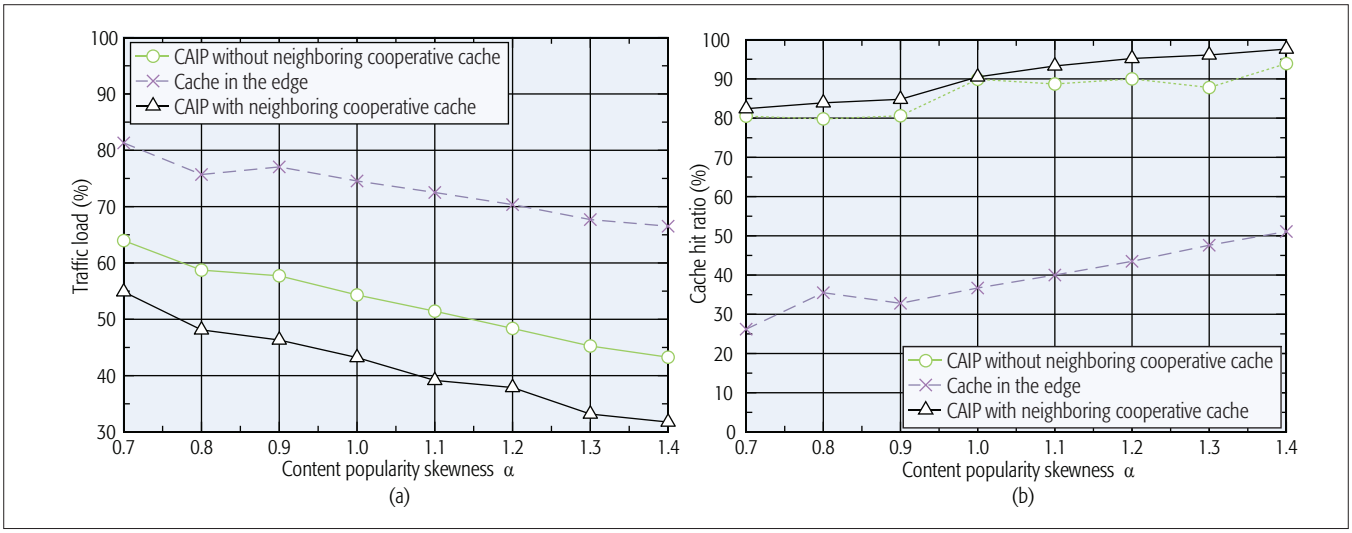


Figure 5. Performance evaluation of CAIP with neighboring cooperative cache, CAIP without neighboring cooperative cache, and cache in the edge: a) traffic load (the ratio of the experimental results to the results of using traditional IP without caching); b) cache hit ratio.

(a) Performance for content index lookup							
Size of content index (*K*)	8	16	32	64	128	256	512
Lookup throughput (MSPS)	55.6	47.6	38.5	32.2	26.3	20.4	14.5
(b) Performance for IPv6 RIB lookup							
IPv6 RIB size (1*1000)	5	10	15	20	25	30	35
Lookup throughput (MSPS)	3.87	3.84	3.78	3.72	3.68	3.66	3.63
(c) Performance for NDN FIB lookup							
NDN FIB size (1*100000)	1	2	3	4	5	6	7
Lookup throughput (MSPS)	1.06	0.96	0.93	0.80	0.69	0.66	0.65

Table 1. Performance evaluation of content index lookup, IPv6 RIB lookup, and NDN FIB lookup.

As shown in Fig. 5, CAIP makes performance remarkably better in cache hit rate and traffic load than that of caching in the edge and no caching (the case of traditional IP without caching). As CAIP achieves in-network cache in which each CAIP router can cache any passed content, consumers can retrieve the content from a closer router rather than that server. Moreover, with the popularity skewness α increasing, more requests are associated with the popular contents that have been cached already, so CAIP can better improve the overall network performance. Moreover, CAIP with neighboring cooperative cache policy can reduce the content redundancy and improve the utilization of cached contents, so it can further decrease the total traffic load, as shown in Fig. 5a, and improve the cache hit ratio, as shown in Fig. 5b.

Furthermore, we conduct an experiment to evaluate the performance of our CAIP's forwarding engine. We have implemented our design of the CAIP router's process engine via programming Linux Kernel 4.4.0-51-generic on a GB-BSi-7HA-6500 mini-pc platform, which running OS Linux 16.04. The platform hardware configurations are CPUs with Intel Core i7-6500U and 32 GB of DDR4 2133. We first evaluated the CI lookup throughput by varying CI target sizes from 8K to 512K entries. The CI is implemented using a chained hash table of size 512K. The size of a queried content label is about 20–60 bytes. Table 1a shows the lookup throughput by increasing the CI size. Then we compare our CAIP IPv6 forwarding engine with the NDN forwarding engine. According to the BGP Routing Table Analysis Report (https://bgp.potaroo.net/, accessed Mar. 25, 2017), we collect several real IPv6 Border Gateway Protocol (BGP) tables of different sizes. The experimental results of IPv6 lookup speed on the real-life BGP tables are shown in Table 1b. The NDN forwarding engine is implemented by deploying an NDN Forwarding Daemon (NFD) [14] based on the same platform as CAIP's. We generate an NDN FIB (around 1 million) from a URL dataset named URLblacklist (http://urlblacklist.com/, accessed Mar. 25, 2017). Name traces are generated to mimic the content labels by appending randomly generated directory paths to name prefixes in the FIB as the method in [15]. Table 1c shows the lookup speed as FIB size grows.

To put the numbers in context, the line rate of forwarding translates to 1.488 Mpackets/s with 64-byte frames on a 1 Gb/s link. We can see that the CI lookup speed is 14.5 MSPS (million searches per second) when the load factor (size of inserted CI/size of hash table) is 1. The CI search has only a little effect on the overall forwarding engine. According to the BGP Routing Table Analysis Report, the size of BGP forwarding table entries is about 35,000 in 2017. According to our IPv6 routing lookup experiment, the lookup throughput is about 3.63 MSPS when the size is about 35,000. Thus, the IPv6 routing lookup can satisfy the requirement of the line rate. But for the NDN FIB lookup, the lookup throughput is far lower than the IPv6 routing lookup, and the real size of FIB is far larger according to our experiment.

Conclusion

Along with the development of mobile Internet, the current network can hardly bear an explosive increase of global multimedia traffic. In order

to improve network resource utilization, content distribution and duplicated data utilization in IP networks are regarded as important issues. For the new proposed ICN, just as "far hydrolyze, not close thirsty," there is still a long way to go. In this article, based on the consideration of the forward compatibility, we propose a novel content-aware IP-based architecture, named CAIP, which can not only draw the advantage of in-network caching from ICN, but also guarantee the simplicity of the end-to-end model. CAIP enables the traditional IP network to have better vitality, and in turn results in a significant impact on network development.

Acknowledgment

This work is supported in part by the National Natural Science Foundation of China under Grant No. 61379129, the Key Research Program of the Chinese Academy of Sciences (CAS) under Grant No. ZDRW-KT-2016-2-5, Youth Innovation Promotion Association CAS under Grant No. 2016394, and the Fundamental Research Funds for the Central Universities.

References

[1] C. V. N. Index, "Forecast and Methodology, 2014–2019 White Paper," tech. rep., Cisco, 2015.
[2] Y. Cui *et al.*, "SDN-Based Big Data Caching in ISP Networks," *IEEE Trans. Big Data*, 2017; https://doi.org/10.1109/TBDATA.2017.2651901, accessed June 27, 2017.
[3] A. Venkataramani *et al.*, "MobilityFirst: A Mobility-Centric and Trustworthy Internet Architecture," *ACM SIGCOMM Comp. Commun. Review*, vol. 44, no. 3, 2014, pp. 74–80.
[4] B. Ahlgren *et al.*, "A Survey of Information-Centric Networking," *IEEE Commun. Mag.*, vol. 50, no. 7, July 2012, pp. 26–36.
[5] L. Zhang *et al.*, "Named Data Networking," *ACM SIGCOMM Comp. Commun. Review*, vol. 44, no. 3, 2014, pp. 66–73.
[6] T. Koponen *et al.*, "A Data-Oriented (And Beyond) Network Architecture," *ACM SIGCOMM Comp. Commun. Review*, vol. 37, no. 4. ACM, 2007, pp. 181–92.
[7] G. Carofiglio *et al.*, "From Content Delivery Today to Information Centric Networking," *Computer Networks*, vol. 57, no. 16, 2013, pp. 3116–27.
[8] A. Detti *et al.*, "CONET: A Content Centric Inter-Networking Architecture," *Proc. ACM SIGCOMM Wksp. Information-Centric Networking*, 2011, pp. 50–55.
[9] G. Zhang, Y. Li, and T. Lin, "Caching in Information Centric Networking: A Survey," *Computer Networks*, vol. 57, no. 16, 2013, pp. 3128–41.
[10] M. Zhang, H. Luo, and H. Zhang, "A Survey of Caching Mechanisms In Information-Centric Networking," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 3, 2015, pp. 1473–99.
[11] A. Medina *et al.*, "BRITE: An Approach to Universal Topology Generation," *Proc. IEEE MASCOTS 2001*, 2001, pp. 346–53.
[12] B. M. Waxman, "Routing of Multipoint Connections," *IEEE JSAC*, vol. 6, no. 9, 1988, pp. 1617–22.
[13] L. Breslau *et al.*, "Web Caching and Zipf-Like Distributions: Evidence and Implications," *Proc. IEEE INFOCOM 1999*, vol. 1, 1999, pp. 126-34.
[14] A. Afanasyev *et al.*, "NFD Developer's Guide," tech. rep. NDN-0021, rev. 6, NDN Project, 2016.
[15] H. Dai *et al.*, "BFAST: High-Speed and Memory-Efficient Approach for NDN Forwarding Engine," *IEEE/ACM Trans. Networking*, vol. 25, no. 2, 2017, pp. 1235-48.

Biographies

Kaiping Xue (kpxue@ustc.edu.cn) received his B.S. degree from the Department of Information Security, University of Science and Technology of China (USTC), Hefei, in 2003 and received his Ph.D. degree from the Department of Electronic Engineering and Information Science (EEIS), USTC, in 2007. Currently, he is an associate professor in the Department of Information Security and Department of Electrical Engineering and Information Science (EEIS), USTC. His research interests include next-generation Internet, distributed networks, and network security.

Tingting Hu (hutingt2@mail.ustc.edu.cn) received her M.S. degree from the Department of EEIS, USTC, in 2017. Her research interests include future Internet architecture design and network functions vitalization.

Xiang Zhang (mm1201@mail.ustc.edu.cn) received his B.S. degree from the Department of Information Security, USTC in July 2015. He is currently a graduate student in communication and information systems in the Department of EEIS, USTC. His research interests include next-generation Internet and network security.

Peilin Hong (plhong@ustc.edu.cn) received her B.S. and M.S. degrees from the Department of EEIS, USTC, in 1983 and 1986. Currently, she is a professor in the Department of EEIS, USTC. Her research interests include next-generation Internet, policy control, IP QoS, and information security. She has published 2 books and over 150 academic papers in several journals and conference proceedings.

David S. L. Wei (wei@dsm.fordham.edu) received his Ph.D. degree in computer and information science from the University of Pennsylvania in 1991. He is currently a professor in the Computer and Information Science Department at Fordham University. He has authored and co-authored more than 100 technical papers in various archival journals and conference proceedings. Currently, his research interests include cloud computing, big data, IoT, and cognitive radio networks.

Feng Wu (fengwu@ustc.edu.cn) received his M.S. and Ph.D. degrees in computer science from the Harbin Institute of Technology in 1996 and 1999, respectively. He is currently a professor with USTC, and the Dean of the School of Information Science and Technology. He has authored or coauthored more than 200 high-quality papers. His research interests include image and video compression, media communication, and media analysis and synthesis.

CAIP makes performance remarkably better in cache hit rate and traffic load than that of caching in the edge and no caching (the case of traditional IP without caching). As CAIP achieves in-network cache in which each CAIP router can cache any passed content, consumers can retrieve the content from a closer router rather that the server.

ACCEPTED FROM OPEN CALL

Performance Study on Seamless DA2GC for Aircraft Passengers toward 5G

Michal Vondra, Ergin Dinc, Mikael Prytz, Magnus Frodigh, Dominic Schupke, Mats Nilson, Sandra Hofmann, and Cicek Cavdar

The authors assess the capacity limitations of DA2GC links provided to an aircraft by employing currently available 4G radio technologies. They also study different capacity enhancement techniques toward 5G such as an implementation of MU-MIMO or coordinated beamsteering to improve DA2GC capacity, and they compare DA2GC performance with connectivity offered by satellites.

ABSTRACT

Mobile users seek ubiquitous broadband connectivity even during a flight above the clouds. As many passengers are expected to be connected to high-speed Internet, robust and high-capacity direct air-to-ground communication (DA2GC) to connect the aircraft with the ground cellular network is particularly attractive. In this article, we assess the capacity limitations of DA2GC links provided to an aircraft by employing currently available 4G radio technologies. Further, we study different capacity enhancement techniques toward 5G such as an implementation of Multi-User Multiple Input Multiple Output (MU-MIMO) or coordinated beamsteering to improve DA2GC capacity. We also compare DA2GC performance with connectivity offered by satellites. According to our results, a DA2GC network extended with techniques toward 5G can increase the capacity available per aircraft by nearly 100 Mb/s in comparison with the 4G DA2GC network, assuming 20 MHz bandwidth is exploited. Although the maximum capacity provided by the 5G DA2GC network and satellite communication is the same, satellite communication exhibits only 40 Mb/s for 10000 deployed satellites on average. To motivate future investigation in this field, this article also outlines open challenges and research directions toward more efficient DA2GC.

INTRODUCTION

In the past decade, the worldwide volume of flight traffic and the number of transported passengers have risen significantly every year. With rising demands of mobile users on the capacity of wireless networks, a missing connection on board an aircraft can cause major dissatisfaction of passengers [1]. Therefore, airline companies have started providing Internet access to passengers. Several initiatives and project consortiums dealing with air-to-ground (A2G) communication were established in recent years (e.g., [2]). Such projects mainly focused on two connectivity solutions regarding how to deliver Internet connections to users and the aircraft's systems represented by a machine type communication (MTC) on board: direct air-to-ground communication (DA2GC) and satellite communication (SC). The main objective of this article is to assess and compare different technologies used for A2G communication.

In the case of DA2GC, Internet connection is supplied in the same way as a connection for regular terrestrial mobile users, meaning the aircraft is connected via a direct link to direct air-to-ground eNodeBs (DA2G-eNBs), as shown in Fig. 1. The main advantage of the DA2GC solution is relatively easy and cheap deployment of DA2G-eNBs. DA2GC can be quickly introduced to airspace above the ground but cannot be easily provided above the sea to achieve worldwide service. In the case of SC, the communication between a ground station and an aircraft is relayed by satellites (Fig. 1). This approach provides global coverage. However, the main constraints of SC compared with DA2GC consist in higher communication delay and limited achievable data rates for the overall system. According to [3], currently only 36 percent of all aircraft are equipped with onboard Internet connection provided mostly via satellites. Available satellite A2G solutions exploit mostly the S (2–4 GHz), Ku (12–18 GHz) and Ka (26.5–40 GHz) bands and offer maximally tens of Mb/s/aircraft. Since throughput available via satellite is insufficient, only 6 percent of aircraft are able to deliver high-speed connectivity enabling video streaming and other highly demanded services [3]. A possible way to overcome this capacity limitation is the exploitation of DA2GC.

One of the existing DA2GC solutions is the GoGo network. This network offers DA2GC in the U.S. and Canada through more than 225 base stations [4]. The GoGo network operates at 850 MHz with 2 MHz bandwidth (BW) for uplink and 2 MHz BW for downlink by using the 3G Code Division Multiple Access Evolution-Data Optimized (CDMA EvDO) standard. The network provides DA2GC capacities up to 10 Mb/s [4].

One of the first research papers dealing with DA2GC [5] focuses on the propagation aspects of air-to-ground communication including antenna design. However, only the signal level is evaluated without assessing the signal quality or capacity. Moreover, the analysis covers only relatively low frequencies of up to 1 GHz, which is not considered in our article. Current research dealing with DA2GC focuses mainly on performance analysis of different technologies. In [6], the authors prove the feasibility of Multiple Input Multiple Output (MIMO) in A2G communication. The authors present field measurements of low-altitude A2G MIMO channels. The propagation characteristics indicate that the spatial properties of the A2G channel will support MIMO communication. DA2GC can also be exploited by unmanned aerial vehicles, as introduced in [7].

Digital Object Identifier:
10.1109/MCOM.2017.1700188

Michal Vondra, Ergin Dinc, Mats Nilson, and Cicek Cavdar are with KTH, Royal Institute of Technology; Mikael Prytz and Magnus Frodigh are with Ericsson Research; Dominic Schupke and Sandra Hofmann are with Airbus.

0163-6804/17/$25.00 © 2017 IEEE

The concept of low-cost microsatellites is introduced in [8]. Although the proposal is not primarily aimed at A2G communication, aircraft can also exploit available capacity through micro-satellites. The advantage of small low Earth orbit (LEO) satellites is the low delay compared with geostationary Earth orbit (GEO) satellite services mostly used by aircraft today. While some LEO constellations already exist, several new initiatives aim to build high-throughput LEO satellite networks. Particularly interesting are the concepts of OneWeb, SpaceX and LeoSat planning to deploy between 100 (LeoSat) and over 4000 (SpaceX) LEO satellites in the next few years [9].

DA2GC can be the primary system used for communication with aircraft over the mainland due to its relatively low-cost, simple, and quick installation advantages [10]. However, to provide a complete solution, the future evolution of Internet connectivity for aircraft requires hybrid solutions where cooperation of satellites together with the DA2GC system is present. Therefore, this article aims to assess and compare of two types of DA2GC networks and connectivity provided via satellites.

The main contribution of this article can be listed as:

- Performance assessment of DA2GC based on current 4G LTE standards and on envisioned 5G mobile networks that utilize LTE together with advanced radio techniques such as antenna arrays enabling coordinated beamsteering, multi-user MIMO (MU-MIMO) and higher-order modulations.
- Performance analysis of SC based on the LTE standard with advanced techniques toward 5G.
- Comparison of different radio technologies in terms of signal quality and capacity available per aircraft under static and dynamic scenarios.
- Propose a new mobility model based on real traffic behavior of aircraft.
- Outline new challenges toward future DA2GC networks.

The rest of this article is structured as follows. The following section gives an overview of challenges toward future DA2GC systems. Following that, our aircraft mobility model exploited for dynamic analysis is introduced. We then describe radio access technologies and simulation settings. Following that, results of performance comparisons are presented and discussed. The last section summarizes major conclusions and outlines future work.

Challenges toward Future DA2GC

In this section, challenges toward future DA2GC are introduced and discussed. First, we describe challenges related to spectrum and regulations. The second subsection defines challenges associated with advanced technologies and techniques of next generation wireless networks. In the last subsection, business challenges, such as business models and new business opportunities, are outlined.

Spectrum and Regulations

Dedicated spectrum bands for exclusive DA2GC use are only available in some regions of the world, such as North America, where 4 MHz BW is exploited by GoGo. However, the peak capacity of the GoGo system is 10 Mb/s [4] even if a single aircraft is connected. To provide sufficient quality of service, more spectrum is needed.

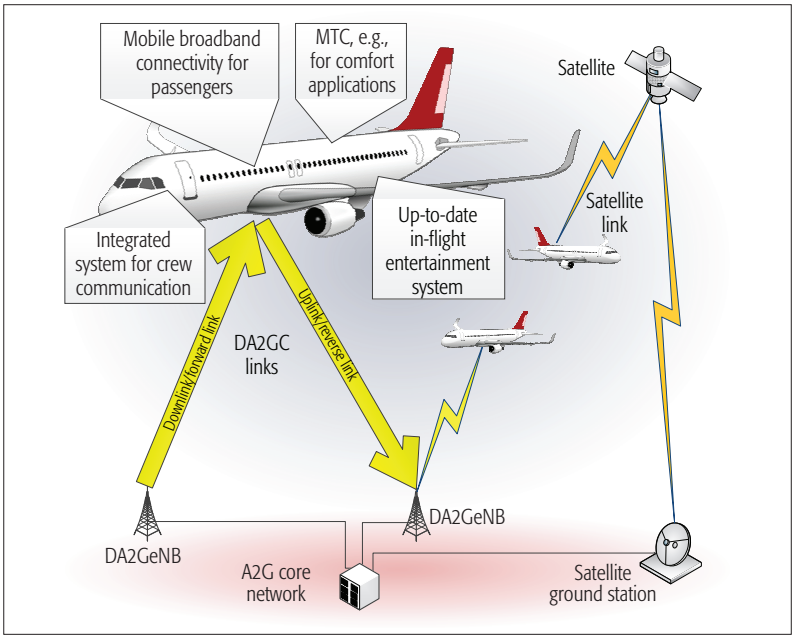


Figure 1. Structure of A2G communication including DA2GC and SC with services available for onboard passengers, crew and MTC.

For efficient DA2GC, a strong need for spectrum harmonization among many countries is required. However, spectrum utilization in DA2GC may be perceived low if compared with terrestrial networks. Another challenge is the authorization of dedicated spectrum, which requires coordinated processes involving all participating countries.

Sharing spectrum with other radio systems such as satellite or terrestrial systems can be a feasible option. The deployment geometry and airspace coverage requirements lead to difficult interference problems, which, however, can be mitigated or solved with system coordination techniques.

As part of its ongoing work on spectrum harmonization for DA2GC, the European Commission mandates two of the unpaired 2 GHz bands only to allocate one of the bands in TDD mode, which leads to an own recommendation at The European Conference of Postal and Telecommunications Administrations (CEPT) Electronic Communications Committee (ECC) level [11].

The Mobile Satellite System (MSS) spectrum at 2 GHz in Europe, which was awarded in 2008 and which was supposed to have satellite deployment in 2011, permits a called Complementary Ground Component (CGC). Recently one of the rights holders, Inmarsat, teamed up with Deutsche Telekom under the name European Aviation Network (EAN) to deploy an air-to-ground system [2].

In the U.S., several industry stakeholders have been putting forth arguments to the FCC to make available 500 MHz of spectrum for DA2GC in the 14 GHz range. Moreover, current research trends focus on even higher frequencies such as millimeter waves where even wider spectrum bands can be used for DA2GC.

There are many open issues with respect to spectrum bands that will be available for DA2GC communications and their harmonization. Since spectrum will always be a scarce resource, dynamic spectrum management techniques need to be

To ensure the provision of Internet connection to passengers on board, the emerging DA2GC business environment requires the creation of new business models with new roles for current players and the introduction of completely new entities participating in DA2GC.

Parameter	Value
Max altitude of aircraft	43000 feet (13.1 km)
Average climb time	1500 s
Variance of climb time	75 s
Average descent time	1800 s
Variance of descent time	90 s
Number of take-offs	20 starts/minute
Total length of simulation	2 hours
Number of background aircraft (always present)	10
MIMO	2×2 (polarization diver.)
Thermal noise PSD	–174 dBm/Hz
Beamforming gain	5 dB
Coordinated beamsteering gain	20 dB [14]
Number of symbols/ms	168
LTE modulation	From QPSK to 256QAM (2–8 bits/symbol)
LTE overhead (for all links)	25%
Inter-site distance (ISD)	40–200 km
Total number of DA2G-eNBs	300–7500
Total number of satellites	1000–10000
Number of satellites above Europe	20–200

Table 1. List of simulation parameters.

developed as the aircraft moves between countries for interference cancellation and re-use.

Advanced Communication Technologies and Techniques toward 5G

5G is currently being developed by the 3GPP mobile community to provide a leap in mobile system capabilities. It is expected that early trial systems could be in operation as early as 2018, with large-scale deployments starting around 2020. One part of 5G is the continued evolution of LTE, and with several aggregated carriers and new multi-antenna functionalities, LTE radio access is a clear candidate for future DA2GC systems.

Another part of 5G is the new radio (NR) for which 3GPP standardization started in March 2016. It is a new radio access built around key technology features that are particularly interesting for DA2GC:

- Large bandwidths, 100 MHz and more.
- Very large antenna arrays.
- Beamforming/beamsteering.

The impact of large bandwidths is immediate on capacity while very large antenna arrays (thousands of antenna elements) allow improving antenna gain. Beamforming/beamsteering avoids interference caused by other DA2G-eNBs and is even able to accomplish multi-user spectrum reuse from within the same DA2G-eNBs. While not ready yet, the 5G NR should be viewed as a strong candidate for future DA2GC systems.

Besides NR, new techniques such as coordinated multipoint (CoMP) with coordinated scheduling/beamforming or joint transmission are investigated as open research issues in this field. One of the main challenges of CoMP employed in DA2GC is relatively high delay caused by sparse deployment of DA2G-eNBs. Moreover, DA2GC opens a field to exploit advanced communication techniques such as non-orthogonal multiple access (NOMA), which can maximally exploit specific DA2GC environments with a dominant line of sight (LOS) component.

Business Challenges

To ensure the provision of Internet connection to passengers on board, the emerging DA2GC business environment requires the creation of new business models with new roles for current players and the introduction of completely new entities participating in DA2GC. New roles should be adopted by current players such as airlines or terrestrial operators to accept new business opportunities. Besides these, new entities such as cabin system operators or A2G network operators could emerge.

The A2G network operator is a consortium of all involved terrestrial and satellite operators, which provides the connection for cabin system operators. The A2G operator acts like a virtual operator, who buys network capacity from terrestrial and satellite operators and sells it to cabin operators who are responsible for the provision of network services for passengers on board [12].

Besides the regulatory environment, the value chain and business ecosystem for DA2GC are quite complex. There are many players involved to provide this technology to the end user. Business models need to be studied that clearly define the roles of partners and leave room for different players to stimulate the market (such as Europe), differing from monopolistic solutions. Novel research directions are viable business models for all players participating in DA2GC. Besides this, modeling the costs of connections should be formulated differently including CAPEX and OPEX, presenting another open research direction.

Aircraft Mobility Model

In order to model the movement of aircraft in our analysis, we have developed an aircraft mobility model in MATLAB. The developed model simulates the movement of aircraft in the European airspace. Each aircraft has its origin and destination point and a predefined path between two points. The set of origin and destination for the aircraft consists of real locations of 40 major airports in Europe and five entry/exit points to/from Europe. An origin and destination airport is selected for each aircraft based on real traffic density in each airport with uniform distribution.

All flights have a climb phase, a cruise phase and a descent phase with predefined characteristics different for each flight. Characteristics of each flight, particularly maximum speed, maximum altitude, climb and descent times, are derived based on observation of real aircraft [13]. In particu-

lar, each flight has a target altitude and a maximum flight speed used through its entire cruise phase. During the climb and descent phases, the speed and altitude increase and decrease linearly, respectively. The average climb time is 25 minutes and average descent time is 30 minutes randomized with the normal distribution (Table 1).

In the simulation, 20 new taking-off aircraft are considered on average per minute. An aircraft starts at a randomly selected time and flies directly to the predefined destination. Although aircraft are usually forced into corridors during the cruise flight, direct paths between two airports are taken on short-haul flights [13]. Besides these aircraft with predefined paths, 10 randomly flying aircraft are always present in the simulation as background aircraft that require connection.

We assume that exact positions of aircraft, the serving DA2G-eNBs, and the serving satellites are known. It enables coordination between DA2G-eNBs and elimination of Doppler shift by opposite shifting of transmitting/receiving frequency.

System Parameters

This section provides a description of the comparative analysis of A2G communication systems. In two subsections, the radio access technologies and the simulation settings are introduced.

The performance is evaluated in terms of signal quality and communication capacity available for one aircraft. Since most onboard communication traffic is passengers downloading, the analysis is made for the forward link (from DA2G-eNBs to aircraft). The full buffer data traffic model is employed in the simulation.

Regarding signal propagation, the forward link is dominated by the LOS component during all flight phases. Because of this, the classic free-space path loss model affected only by distance d and frequency f is used for modeling path loss (PL). This model is used in both cases, between DA2G-eNBs and the aircraft as well as between satellites and the aircraft:

$$PL = \left(\frac{4\pi df}{c}\right)^2. \quad (1)$$

Radio Access Technologies

To provide a fair comparison of different types of A2G communications, we exploit LTE technology with the same BW equal to 20 MHz for both 4G and 5G DA2GC as well as for SC. If wider BW is employed, total capacity can be proportionally increased in all networks. In the case of the DA2GC forward link, wider BW can be obtained, for example, by reusing frequencies exploited for terrestrial networks. A wide unused BW is available in higher frequency bands where hundreds of MHz can be exploited especially by SC. However, in the case of SC, such a large number of deployed satellites cannot be used only for one purpose, such as provisioning only A2G communication. The connectivity should also be offered for other terrestrial applications. Therefore, BW dedicated to A2G communication would be only part of total BW available from SC.

4G DA2GC Network: The 4G DA2GC network is based on the 3GPP LTE standard currently widely implemented in terrestrial networks. The selected frequency band for the 4G DA2GC network is 3.4 GHz [11]. To suppress interference between two neighboring DA2G-eNBs, beamforming is exploited in the three horizontal 120-degree sectors of each DA2G-eNB. Within each sector, the available BW is split between all connected aircraft. The maximum modulation available in terms of bits per Hz for this 4G network is 64 Quadrature Amplitude Modulation (64QAM). Note that LTE employs adaptive modulation and coding rate derived based on signal quality represented by signal-to-interference and noise ratio (SINR).

5G DA2GC Network: The 5G DA2GC network characterizes future wireless networks based on LTE with advanced techniques toward the 5G network, such as antenna arrays enabling coordinated beamsteering, multi-user MU-MIMO, and higher-order modulations. In particular, with the coordinated beamsteering technique, the antenna pattern and maximum antenna gain can be set in the direction of the served aircraft while minimizing interference toward other aircraft. Together with MU-MIMO, where a high number of antennas are employed in a DA2G-eNB, one DA2G-eNB is able to serve more aircraft simultaneously. Moreover, each aircraft can have its own dedicated beam with full BW. This enables Space Division Multiple Access (SDMA). SDMA allows reaching higher network capacity available for each aircraft. Moreover, the specific environment with ensured LOS between DA2G-eNBs and the aircraft and antennas with higher gain allows using higher-order modulations, such as 256QAM. We assume that the 14 GHz band will be exploited for the 5G DA2GC network, as suggested in [11].

Satellite Communication Network: To provide a more comprehensive analysis of potential solutions, we also assess the performance of A2G connection provided via satellites. In this study, the satellite connection is represented by the link between an aircraft and an LEO satellite that is located 1500 km above ground level, using the carrier frequency of 100 GHz [8]. The main advantages of satellites located in LEO in comparison with satellites in GEO (35786 km) are significantly lower communication delay (20 ms for LEO compared with 250 ms for GEO [8]) and higher capacity available per user. Since LEO satellites have a narrower beam coverage area, each satellite beam serves a lower number of users. For this reason, LEO satellites provide higher throughput compared with GEO satellites.

Since the satellites are located in LEO, their orbital speed needs to be higher than the Earth's rotation to stay in the same orbit. Because of the higher speed and since Europe covers only 2 percent of the Earth's surface, each satellite remains above Europe for approximately ten minutes maximally. For example, around 1000 satellites in total are needed around the Earth to ensure minimally 20 satellites simultaneously above Europe. In the case of 200 satellites above Europe, 10000 satellites are required in LEO, which exceeds the number of satellites of all planned projects together [9].

For SC, the high antenna gain is typically used on both communication sides, i.e., satellite and aircraft. As in the case of 5G DA2GC, 256QAM and MU-MIMO (represented by spot beam technolo-

Since the satellites are located in LEO, their orbital speed needs to be higher than the earth's rotation to stay in the same orbit. Because of the higher speed and since Europe covers only 2 percent of the earth's surface, each satellite remains above Europe for approximately ten minutes maximally.

gy in SC) are employed. For the sake of simplicity, we assess only the communication link between the satellite and the aircraft. The link between the satellite and the ground station is not considered in this analysis. For this link, we assume a dedicated channel with enough capacity.

Simulation Settings

The performance of A2G connections is assessed in static link budget analysis and in dynamic simulations. Static link budget analysis shows edge capacity reachable by one aircraft, while dynamic performance comparison evaluates the influence of other aircraft in a dynamic environment with moving aircraft.

Static Simulation: The link budget calculations describe the static case of the A2G network. For both 4G-based and 5G-based DA2GC networks, two cases, i.e., the best and worst cases, are examined. The worst case represents an aircraft placed at a cell edge, equidistant to three DA2G-eNBs (one serving and two interfering). Considered inter-site distance (ISD) between the two DA2G-eNBs is equal to 200 km, which means the longest distance between the aircraft and the DA2G-eNBs is around 115 km. In the best case, the aircraft is placed directly above the serving DA2G-eNB where the distance is only 10 km. For SC in link budget calculations, we consider 200 satellites above Europe.

Dynamic Simulation: In dynamic simulations, the influence of aircraft movement on available SINR and capacity is assessed by using our aircraft mobility model. For a more detailed comparison, ISDs and a different number of satellites are taken into account. The deployment of DA2G-eNBs is assumed as always to be hexagonal and uniform with the same ISD for all of Europe. The deployment and movement of satellites are random with uniform distribution. The parameters of the dynamic model, as well as the wireless channel, are listed in Tables 1 and 2.

Parameters	4G network		5G network		SC [8]
	Cell center	Cell edge	Cell center	Cell edge	–
Transmit power [dBm]	45				30
Tx and Rx antenna distance [km]	10	115	10	115	1500
Frequency band [GHz]	3.4		14		100
Bandwidth [MHz]	20				
Propagation loss [dB]	123.07	144.28	135.36	156.75	195.96
Tx antenna gain [dBi]	20		30		53
Rx antenna gain [dBi]	10				53
Received power [dBm]	–48.07	-69.28	–50.36	–71.58	–59.96
Noise [dBm]	–100.99				
SNR [dB]	52.92	31.71	50.63	29.41	41.03
Interference + noise [dBm]	–76.07	–71.27	–92.68	–88.32	–89.62
SINR [dB]	28.00	1.99	42.32	16.75	29.66
Shannon maximum capacity [Mb/s]	97.16	31.56	108.74	82.99	98.77
Shannon spectral efficiency [b/s/Hz]	4.86	1.58	5.44	4.15	4.94
Used LTE-type modulation	64QAM	QPSK	256QAM	64QAM	256QAM
Used LTE-type coding rate	948/1024	449/1024	948/1024	772/1024	948/1024
LTE-type capacity (2×2 MIMO) [Mb/s]	139.86	22.10	186.73	113.99	186.73
LTE-type spectral efficiency [b/s/Hz]	6.99	1.11	9.34	5.70	9.34

Table 2. Static link budget analysis.

Results

In this section, numerical results for the analysis of 4G and 5G DA2GC and of SC are presented. In the first subsection, static link budget analysis is calculated, while in the second and third subsections we compare dynamic performance.

Static Link Budget Calculations

Table 2 introduces detailed link budget calculations. As expected, the highest PL is exhibited by SC. Compared with DA2GC networks, the difference is nearly 40 dB. This is caused by the combination of higher frequency band and longer distance between transmitting and receiving antennas, which mainly influence PL, as stated in Eq. 1. However, thanks to higher antenna gain in the case of SC, experienced signal-to-noise ratio (SNR) is comparable with the cell edge cases of DA2GC networks.

If interference and noise are taken into account, the SC shows lower SINR than the 5G DA2GC network but higher SINR than the 4G DA2GC network in the cell center. The reason for the low SINR of the 4G DA2GC network is low transmitting antenna gain and mainly interference caused by close DA2G-eNBs. The highest theoretical Shannon capacity of ~109 Mb/s holds for the 5G DA2GC network in the cell center with BW of 20 MHz. The 4G DA2GC network exhibits the lowest available capacity at the cell edge.

By employing an LTE network with 2 x 2 MIMO and adaptive modulation and coding rate, the maximum capacity for a 5G DA2GC network and SC is the same, equal to ~187 Mb/s. The equality is caused by maximum modulation, which is 256QAM in both cases. It means that even if the 5G DA2GC network is able to provide a SINR of nearly 43 dB in the cell center, 256QAM is not able to fully exploit SINR higher than 28 dB [15]. In other words, if SINR is higher than 28 dB, the highest modulation and coding scheme with the maximal efficiency is always selected and thus no improvement in capacity can be seen for SINR levels higher than 28 dB. The solution could be to employ an even higher-order modulation scheme, which can help improve maximum capacity. On the other hand, a 4G DA2GC network can achieve only ~140 Mb/s and ~22 Mb/s in the cell center and at the cell edge, respectively. Due to high interference at the cell edge, the 4G DA2GC network has the lowest capacity results. In addition, the peak capacity is limited due to maximum modulation, which is only 64QAM.

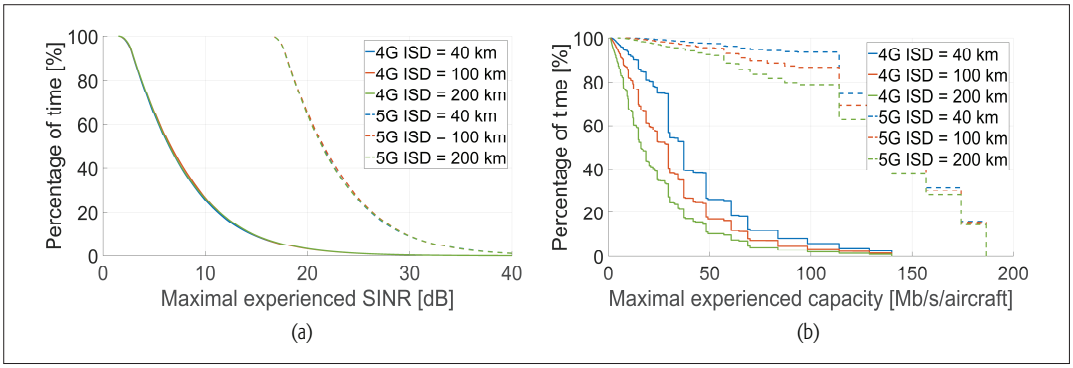


Figure 2. a) Maximal average SINR; b) capacity experienced by aircraft over flight duration for 4G and 5G DA2GC networks.

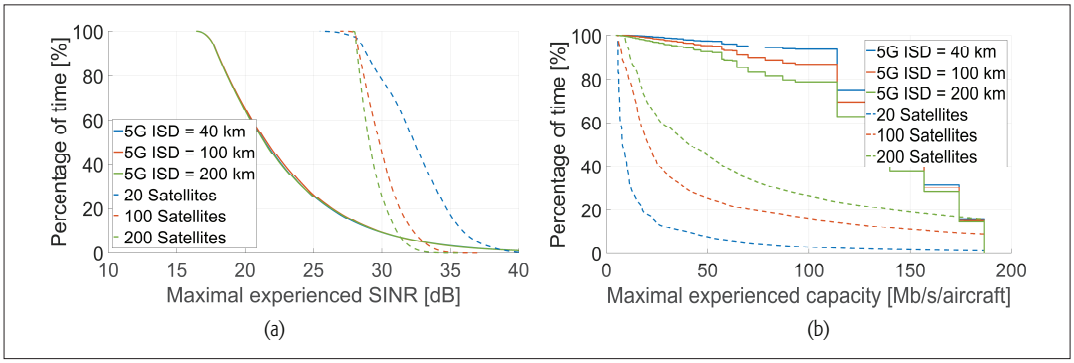


Figure 3. a) Maximal average SINR; b) capacity experienced by aircraft over flight duration for 5G DA2GC network and satellite communication.

Note that these results are valid only for one static aircraft served by one antenna. The evaluation of the total available capacity for one aircraft in the dynamic environment is the objective of the following subsections.

DA2GC Capacity Analysis under Dynamic Scenario: 4G vs. 5G

In this subsection, the dynamic performance analysis of 4G and 5G DA2GC networks is provided. The results of the simulations are shown in Fig. 2.

The subplots of Fig. 2 show the distribution for experienced SINR and experienced capacity over flight duration for a 4G and 5G DA2GC network, respectively. In other words, the depicted curves show the percentage of time spent by the aircraft experiencing at least a given SINR or capacity. The curves express a complementary cumulative distribution function (CCDF). As can be seen in Fig. 2a, the experienced SINR is nearly the same regardless of ISD for a single DA2GC network. The difference between 4G-based and 5G-based DA2GC networks is given mainly by coordinated beamsteering employed in the 5G DA2GC network. Results show that an aircraft employing a 4G DA2GC network can guarantee at least 2.5 dB SINR for 95 percent of flight time, while an aircraft using a 5G DA2GC network can have SINR 17.5 dB or better for 95 percent of flight time. The 5G network allows fully exploiting higher modulation schemes and reaching higher capacity.

Figure 2b indicates the maximum available capacity per aircraft. As expected, a lower ISD means higher available capacity per aircraft for both networks. Although the positions of deployed DA2G-eNBs are the same for each ISD-curve for both networks, the 5G DA2GC network significantly outperforms the 4G DA2GC network for all settings. The higher capacity of the 5G DA2GC network is achieved mainly by employing MU-MIMO, which enables SDMA. It allows serving nearly each aircraft by its own signal beam, thus exploiting full BW. In the case of the 4G DA2GC network, each DA2G-eNB has only three sectors and all aircraft connected to the same antenna have to share the same resources. Therefore, in the case of ISD equal to 200 km, the capacity is larger than 130 Mb/s for nearly 50 percent of flight time if the 5G DA2GC network is employed. In contrast, in the case of the 4G DA2GC network with the same ISD, the aircraft spent 50 percent of flight time with a guaranteed capacity of about 16 Mb/s or better.

Comparison of DA2GC and Satellite Communication under Dynamic Scenario

This section compares the performance of a 5G DA2GC network and SC. As in the previous case, the subplots of Fig. 3 depict maximal experienced SINR and maximal experienced capacity over flight duration for 5G DA2GC network and for SC, respectively.

As Fig. 3a shows, the SINR experienced by aircraft connected via SC is higher than in the case of DA2GC for more than 95 percent of flight time. This is especially due to the higher gain of transmitting and receiving antennas. DA2GC can attain higher SINR only in the case of the short distance between a DA2G-eNB and an aircraft.

However, as Fig. 3b shows, 5G DA2GC outperforms SC in total capacity available for one aircraft. The main reason is the distance between

Results show that an aircraft employing a 4G DA2GC network can guarantee at least 2.5 dB SINR for 95 percent of flight time, while an aircraft using a 5G DA2GC network can have SINR 17.5 dB or better for 95 percent of flight time. The 5G network allows fully exploiting higher modulation schemes and reaching higher capacity.

In the case of satellite communication, although satellites provide a high-quality signal link, the capacity is divided among more aircraft, resulting in the reduction of capacity available for one aircraft, reaching maximally 40 Mb/s for 50 percent of flight time, even if 10000 satellites are employed globally.

the satellites and the aircraft. Although we assume antennas with very high gain and with MU-MIMO (spot beam technology) on satellites, the width of the beam close to the ground is always higher than the width of a beam produced by 5G DA2GC network antennas due to the distance. It results in more aircraft placed in the same beam needing to share the same resources. As the results show, even if there are 200 satellites placed above Europe (i.e. 10000 globally), 50 percent of flight time is at maximum capacity of only ~40 Mb/s.

Conclusion

In this article, we analyze the capacity limitation of direct air-to-ground communication (DA2GC) networks based on 4G and 5G technology and of air-to-ground communication provided by satellites. Besides numerical and simulative analyses, this article also outlines a set of challenges with DA2GC. The highest capacity is achieved by the 5G DA2GC network, mainly by coordinated beam-steering in combination with MU-MIMO. Aircraft connected to a 5G DA2GC network experience more than 130 Mb/s for nearly 50 percent of flight time if 200 km inter-site distance and 20 MHz bandwidth are exploited. Conversely, the DA2GC network based on 4G suffers from interference, which significantly limits the signal quality and thus the total available capacity. Therefore, only 16 Mb/s is available for 50 percent of flight time if the same setting is employed. In the case of satellite communication, although satellites provide a high-quality signal link, the capacity is divided among more aircraft, resulting in the reduction of capacity available for one aircraft, reaching maximally 40 Mb/s for 50 percent of flight time, even if 10000 satellites are employed globally.

Acknowledgment

This work is supported in part by the EIT Digital ICARO-EU (Seamless Direct Air-to-Ground Communication in Europe) Project.

References

[1] Honeywell, "Honeywell Survey Explores What Passengers Demand from In-Flight WiFi: Constant Connectivity and Speed," 2014; available: http://www.honeywell.com/newsroom/pressreleases/2014/07/honeywell-survey-explores-what-passengers-demand-from-in-flight-wi-fi-constant-connectivity-and-speed, accessed: 20 Sept. 2016.
[2] Inmarsat & Deutsche Telekom, "The European Aviation Network," 2015; available: https://www. telekom.com/static/-/288318/2/150921-product-sheet-si, accessed: Sept. 2016.
[3] Routehappy, "Annual global state of in-flight wi-fi 2016" 2016; available: https://www. routehappy.com/insights/wi-fi, accessed: 20 Sept. 2016.
[4] GoGo, "Gogo ATG-4 — What Is it, and How Does it Work?" 2014; available: http://concourse.gogoair.com/technology/gogo-atg-4-work, accessed: 20 Sept. 2016.
[5] R. Kirby, J. Herbstreit, and K. Norton, "Service Range for Air-to-Ground and Air-to-Air Communications at Frequencies above 50 Mc," *Proc. IRE*, vol. 40, no. 5, 1952, pp. 525-36.
[6] T. J. Willink *et al.*, "Measurement and Characterization of Low-Altitude Air-to-Ground MIMO Channels," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, 2016, pp. 2637-48.
[7] Y. Saleem, M. H. Rehmani, and S. Zeadally, "Integration of Cognitive Radio Technology with Unmanned Aerial Vehicles: Issues, Opportunities, and Future Research Challenges," *J. Network and Computer Applications*, vol. 50, 2015, pp. 15-31.
[8] F. Khan, Mobile Internet from the heavens, 2015, Aug., arXiv:1508.02383; available: https://arXiv:1508.02383.
[9] J. Foust, "The Return of the Satellite Constellations," 2015; available: http://www.thespacereview. com/article/2716/1, accessed: 15 Jan. 2017.
[10] Y. A. Nijsure *et al.*, "Adaptive Air-to-Ground Secure Communication System Based on ADS-B and Wide-Area Multilateration," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, June 2016, pp. 3150-65.
[11] ECC Report 214, Broadband Direct-Air-to-Ground Communications (DA2GC), Approved 30 May 2014.
[12] E. Dinc *et al.*, "In-Flight Broadband Connectivity: Architectures and Business Models for High Capacity Air-to-Ground Communications," to appear in *IEEE Commun. Mag.*, 2017.
[13] Live flight tracker; available: www.flightradar24.com, accessed: 15 Jan. 2017.
[14] V. Rabinovich and N. Alexandrov, "Typical Array Geometries and Basic Beam Steering Methods," *Antenna Arrays and Automotive Applications*, Springer New York, 2013, pp. 23-54.
[15] S. O. Elbassiouny and A. S. Ibrahim, "Link Level Performance Evaluation of Higher Order Modulation in Small Cells," *Proc. Int'l. Wireless Commun. Mobile Computing Conf. (IWCMC)*, 2014.

Biographies

Michal Vondra (mvondra@kth.se) received his B.Sc. degree in electronic engineering and M.Sc. and Ph.D. degrees in telecommunication engineering from the Czech Technical University in Prague in 2008, 2010, and 2015, respectively. At present he is a postdoc researcher with the Department of Communication Systems, KTH Royal Institute of Technology (Wireless@KTH), Stockholm, Sweden. He has been actively involved in several national and international projects funded by the European Commission such as FREEDOM, or TROPIC. In 2014, he spent six months on an internship with University College Dublin, Ireland. His area of research interest covers topics toward 5G mobile networks, such as mobility management in wireless networks, direct air-to-ground communication, and intelligent transportation system.

Ergin Dinc (ergind@kth.se) received his B.Sc. degree in electrical and electronics engineering from Bogazici University, Istanbul, Turkey, in July 2012. He received his Ph.D. degree in electrical and electronics engineering from Koc University, Istanbul, Turkey in June 2016. After his Ph.D., he started working as a postdoctoral researcher at KTH-The Royal Institute of Technology, Stockholm, Sweden. Currently, he is a research associate in molecular communications and nanonetworks at The University of Cambridge, Cambridge, UK. His research interests include direct-air-to-ground communication, beyond-line-of-sight communication, 5G wireless communication, nanonetworks and molecular communication.

Mikael Prytz (mikael.prytz@ericsson.com) is research leader and head of the Network Control and Management unit at Ericsson Research, focusing on 4G and 5G mobile network architectures and protocols (radio and core), network control and automation, and end-to-end performance. He has more than 15 years of experience in fixed and mobile communications research and engineering from Ericsson's research and services units, as well as from international research programs, including EU projects Ambient Networks, E3, and QUASAR. He holds a Ph.D. in optimization and systems theory from KTH Royal Institute of Technology in Stockholm, Sweden, and an M.S. in operations research from Stanford University, Palo Alto, CA, US.

Magnus Frodigh (magnus.frodigh@ericsson.com) is the research area director for network architecture and protocols at Ericsson Research, responsible for research in network architecture and protocols covering radio networks, transport networks and core networks including network management. He joined Ericsson in 1994 and has since held various key senior positions within Ericsson's Research & Development and Product Management focusing on 2G, 3G, 4G and 5G technologies. He holds a master of science degree from Linkoping University of Technology, Sweden, and a Ph.D. in radio communication systems from the Royal Institute of Technology in Stockholm, Sweden. Since 2013 he has been an adjunct professor at the Royal Institute of Technology in Wireless Infrastructures.

Dominic A. Schupke (dominic.schupke@airbus.com) is with Airbus in Munich, Germany, driving research and innovations for wireless communications. Prior to that he was with NSN, Siemens, and the Institute of Communication Networks at Munich University of Technology (TUM). He received his Dipl.-Ing. degree from RWTH Aachen in 1998 and his Dr.-Ing. degree from TUM in 2004. He has over 15 years experience in the area of communication networks, especially their design and optimization. Since April 2009 he has taught the course 'Network Planning' at TUM. He is an author or co-author of more than 120 journal and conference papers. His research interests include network architectures and protocols, routing, recovery

methods, availability analysis, critical infrastructures, security, virtualization, network optimization, and network planning. He is Senior Member of IEEE, and member of Comsoc, VDE/ITG, and VDI.

MATS NILSON (matsnils@kth.se) is assisting center director for Wireless@kth. He received an M.S degree in electrical engineering (1980) from KTH. He has since then been deeply engaged in R&D for mobile communications with experiences as a developer and in management positions with mobile operators and SME type of radio equipment manufacturers. He has been engaged by KTH on a part time basis for more than 10 years.

SANDRA HOFMANN (sandra.s.hofmann@airbus.com) received her M.Sc. degree in electrical engineering from Technische Universitat Munchen, Germany in 2016. In the same year she joined Airbus, Munich, Germany, where she is pursuing a Ph.D. degree in the area of wireless communications. Her current focus is on providing high capacity communication for aerial vehicles using 5G.

CICEK CAVDAR (cavdar@kth.se) is a senior researcher in the Communication Systems Department at KTH Royal Institute of Technology. She has been leading a research group in the Radio Systems Lab composed of eight researchers focusing on the design and planning of intelligent network architectures, direct air to ground communications and IoT connectivity platforms. She finished her Ph.D. studies in computer science, University of California, Davis in 2008, and at the Istanbul Technical University, Turkey in 2009. After her Ph.D., she worked as an assistant professor in the Computer Engineering Department, Istanbul Technical University. She has been chairing several workshops on green mobile broadband technologies and green 5G mobile networks the last few years co-located with IEEE ICC and IEEE Globecom. She served as chair of the Green Communication Systems and Networks Symposium at ICC 2017 in Paris. At the Wireless@KTH Research Center, she has been leading EU EIT digital projects such as 5GrEEn: Towards Green 5G Mobile Networks and Seamless DA2GC in Europe. She is serving as the leader of the Swedish cluster for the EU Celtic Plus project SooGREEN (Service Oriented Optimization of Green Mobile Networks).

ACCEPTED FROM OPEN CALL

Self-Sustaining Caching Stations: Toward Cost-Effective 5G-Enabled Vehicular Networks

Shan Zhang, Ning Zhang, Xiaojie Fang, Peng Yang, and Xuemin (Sherman) Shen

The authors investigate cost-effective 5G-enabled vehicular networks to support emerging vehicular applications, such as autonomous driving, in-car infotainment and location-based road services. To this end, self-sustaining caching stations (SCSs) are introduced to liberate on-road base stations from the constraints of power lines and wired backhauls.

ABSTRACT

In this article, we investigate cost-effective 5G-enabled vehicular networks to support emerging vehicular applications, such as autonomous driving, in-car infotainment and location-based road services. To this end, self-sustaining caching stations (SCSs) are introduced to liberate on-road base stations from the constraints of power lines and wired backhauls. Specifically, the cache-enabled SCSs are powered by renewable energy and connected to core networks through wireless backhauls, which can realize "drop-and-play" deployment, green operation, and low-latency services. With SCSs integrated, a 5G-enabled heterogeneous vehicular networking architecture is further proposed, where SCSs are deployed along the roadside for traffic offloading while conventional MBSs provide ubiquitous coverage to vehicles. In addition, a hierarchical network management framework is designed to deal with high dynamics in vehicular traffic and renewable energy, where content caching, energy management and traffic steering are jointly investigated to optimize the service capability of SCSs with balanced power demand and supply in different time scales. Case studies are provided to illustrate SCS deployment and operation designs, and some open research issues are also discussed.

INTRODUCTION

Vehicular communication networks hold the promise to improve transportation efficiency and road safety, by enabling vehicles to share information and coordinate with each other. Several potential vehicular networking solutions have been proposed, such as the IEEE 802.11p standard and cellular-based techniques [1]. Compared with other candidates, cellular-based vehicular networking can benefit from the existing cellular network infrastructures to provide ubiquitous coverage and better quality of service (QoS) [2]. In fact, 80 percent of on-road wireless traffic is served by cellular networks [3]. Therefore, cellular-based vehicular networking has drawn extensive attention from both academia and industry. Specifically, the 3rd Generation Partnership Project (3GPP) is currently specifying LTE enhancements to support both vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communications, by integrating cellular and device-to-device interfaces [4]. The corresponding specification work will be finalized as a part of Release 14 in 2017, which can provide a full set of technological enablers from air interface to protocols. In addition, extensive LTE-vehicular trial testing is now ongoing in different places, like Germany and China. In particular, LTE-based vehicular networking has demonstrated its advantages in achieving significant message coverage gain, compared with IEEE 802.11 technologies in both high speed highway and congested urban scenarios [5].

Despite the favorable advantages, cellular networks still face tremendous challenges to meet the needs of future vehicular communications, and the most pressing one is network capacity enhancement. Currently, on-road wireless traffic accounts for 11 percent of cellular traffic [3], which is expected to dramatically increase due to the proliferation of connected vehicles and emerging applications such as autonomous driving, in-car infotainment, augmented reality, and location-based road services. Deploying on-road base stations is the most effective way to increase vehicular network capacity. However, conventional base stations require power lines and wired backhaul connections, making on-road deployment greatly challenging and costly. Furthermore, densification of on-road base stations may also lead to huge energy consumption and bring heavy burdens to backhauls, which can increase operational costs and degrade service performance.

In this article, we first introduce a new type of 5G-enabled on-road base station, namely self-sustaining caching stations (SCSs), to enhance vehicular network capacity in a cost-effective way. Specifically, SCSs have three features:

- Powered by renewable energy instead of the power grid.
- Connected to the core network via millimeter wave (mmWave) backhauls.
- Cache-enabled for efficient content delivery.

By leveraging these 5G technologies, SCSs can be deployed flexibly in a "drop-and-play" manner without wired connections, enable green network operation without additional on-grid energy consumption, and improve delay performance by relieving backhaul pressures. Then, we propose a cost-effective heterogeneous vehicular network architecture, where SCSs are deployed along the roadside to enhance network capacity while conventional macro base stations provide ubiquitous coverage and control signaling.

Digital Object Identifier:
10.1109/MCOM.2017.1700129

Shan Zhang and Xuemin (Sherman) Shen are with University of Waterloo; Ning Zhang is with Texas A&M University-Corpus Christi; Xiaojie Fang is with Harbin Institute of Technology; Peng Yang is with Huazhong University of Science and Technology.

0163-6804/17/$25.00 © 2017 IEEE

To harness the potential benefits of proposed network architecture, we further design a hierarchical management framework to deal with challenges such as intermittent renewable energy supply and highly dynamic traffic demand. In particular, cached contents are updated to maintain content hit rate considering vehicular mobility, while energy management and traffic steering are performed to balance power demand and supply in both large and small time scales. In addition, case studies are provided to illustrate the implementation of the proposed architecture in detail, including cache size optimization and sustainable traffic-energy management.

The remainder of this article is organized as follows. In the next section, the basics of SCSs are introduced, based on which a heterogeneous vehicular network architecture is proposed. Then, a hierarchical network management framework is designed, and case studies are provided. Finally, we discuss future research topics, followed by the conclusions.

Vehicular Network Architecture with SCSs

Cellular-Based Vehicular Networks

With existing infrastructures and state-of-the-art technical solutions, cellular-based vehicular networks hold the promise to provide ubiquitous coverage and support comprehensive QoS requirements in different scenarios. For example, the hidden terminal problems of the 802.11p standard can be totally avoided [6]. In addition, low latency and high reliability can be guaranteed even in dense traffic scenarios, with effective congestion control and resource management schemes.

In spite of the aforementioned advantages, cellular-based vehicular networks still face significant challenges. With the rapid development of information and communication technologies, massive advanced on-road technologies and applications are emerging, such as autonomous driving, argumented reality, infotainment services, and other location-based road services. As these data-hungry applications will result in a surge in wireless traffic, improving vehicular network capacity has become an urgent issue. To this end, on-road base stations need to be deployed. However, conventional base stations are connected in a wired manner due to the requirement of on-gird power supply and backhaul transmission, which can cause the following problems. First, conventional base stations rely on power lines and wired backhaul (e.g., optical fiber) to function, resulting in inflexible deployment, especially in areas with undeveloped power lines or fiber connections (such as highways and rural areas). Second, the huge energy consumption can cause high operational expenditure as well as environmental concerns. Furthermore, with the popularity of multimedia and localized services on the wheel, conventional base stations that only offer connectivity might fail to provide satisfactory QoS, due to time-consuming file fetching from remote servers.

Self-Sustaining Caching Stations (SCS)

Considering the characteristics of vehicular networking and emerging on-road applications, we leverage promising 5G technologies and propose to deploy SCSs in addition to existing cellular networks to enhance vehicular network capacity in a cost-effective way. Specifically, SCSs are equipped with energy harvesting techniques and content caching units, which are connected to the core network through mmWave wireless backhauls.

Equipped with solar panels or wind turbines, SCSs can harvest renewable energy to operate in a self-sustaining manner without the support of the power grid.[1] Exploiting renewable energy as a supplementary or alternative power source is an inevitable trend in the 5G era and beyond, as wireless network energy efficiency expects to be improved by 1000 times [8]. In addition, renewable energy harvesting can liberate network deployment from power lines. Wireless backhaul can be supported by mmWave wireless communication technologies. With large bandwidth unlicensed, mmWave bands can realize broadband wireless communication based on massive multiple input multiple output (MIMO) and beamforming technologies [9]. Therefore, SCSs, which combine both energy harvesting and mmWave backhaul techniques, can be deployed in a "drop-and-play" manner with no wired constraints.

Content caching empowers SCSs to store popular content at the edge of networks, and thus reduce duplicate transmission from remote servers. As a matter of fact, the main on-road mobile applications are now generated by video streaming and map services, which are responsible for 80 percent and 15 percent of total traffic, respectively. The popularity of video content has been found to follow power-law distribution. Accordingly, caching popular video content in SCSs can effectively offload traffic from existing cellular systems. Moreover, emerging on-road applications are expected to be location-based with concentrated requests, which further makes a strong case for content caching. In addition to capacity enhancement, caching can also reduce transmission latency and relieve backhaul burdens, with content stored closer to end users. Furthermore, caching schemes can be devised with respect to specific objectives, such as mobility-aware caching. Specifically, content can be pre-fetched and stored in the next cells before the vehicles conduct handover, to realize smoother handover with high vehicle mobility.

By combing these 5G technologies, SCSs can offer the three-fold benefits of flexible deployment, green operation and enhanced QoS, paving the way to cost-effective vehicular networking.

5G-Enabled Heterogeneous Vehicular Network Architecture

With SCSs integrated, a heterogeneous vehicular network architecture is shown as Fig. 1. The conventional macro base stations (MBSs) and small cell base stations (SBSs) are connected with high speed wired backhauls and powered by the conventional power grid, which mainly provide network coverage and control for reliability. Meanwhile, the SCSs are densely deployed for capacity enhancement, and mainly provide high speed data access based on stored contents. Furthermore, V2V communications are also enabled through device-to-device (D2D) links. The control plane and user plane are separated (i.e., C/U plane separation) for reliable and flexible access. Specifically, vehicles maintain dual connectivities, one with MBSs for signaling and control informa-

Caching schemes can be devised with respect to specific objectives, such as mobility-aware caching. Specifically, the content can be pre-fetched and stored in the next cells before the vehicles conduct handover, to realize smoother handover with high vehicle mobility.

[1] The typical solar panel with 15 percent conversion efficiency can harvest 100 W of energy only by a 82 cm × 82 cm solar panel under rated sunlight radiation, which is sufficient to power a micro (/pico) base station with power demand of 80 W (/8 W). [7]

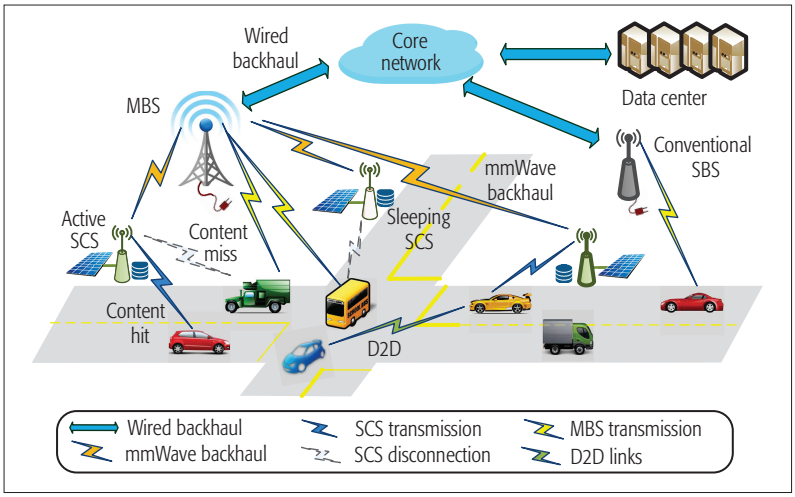


Figure 1. Vehicular network architecture with self-sustaining caching stations.

tion, and the other with SCSs for high data rate transmissions or with other vehicles for instant message exchange. As MBSs can provide ubiquitous signaling coverage with a large cell radius, such a separation architecture can better support vehicle mobility with less frequent handover.

The proposed architecture can support both safety-related and non-safety-related vehicular applications. For the safety-related use case, critical-event (such as collisions or emergency stops) warning messages can be exchanged locally via V2V communications with extremely low latency. For the non-safety-related use cases, MBSs and SCSs can enable better driving experiences, through services such as road condition broadcast, parking assistance and in-car infotainment. In fact, the non-safety-related applications can be data hungry, and account for more than 90 percent of on-vehicle traffic [3]. Accordingly, we mainly focus on V2I (vehicle-to-MBS and vehicle-to-SCS) communications, and investigate cost-effective solutions for the increasing vehicular traffic demand.

With SCSs offloading traffic from MBSs, the service process for vehicle users is as follows. The vehicle user can be directly served if its required content is stored at the associated SCS, which is called the content hit case as shown in Fig. 1. Otherwise, the vehicle user is served by the MBSs, which is called the content miss case. To serve the content miss users, the associated MBS needs to fetch content from remote data centers via wired backhauls, according to conventional cellular communication technologies.

With sufficient cache size and well-designed caching schemes, deploying SCSs can effectively reduce the traffic load of MBSs. Furthermore, the content hit vehicle users can enjoy better quality of experience (QoE) with lower end-to-end delay. As such, the proposed network architecture can provide high capacity for vehicular communications at lower cost.

Hierarchical Network Management Framework

Management Challenges

Network Heterogeneity: MBSs and SCSs exhibit distinct features with respect to coverage, user capacity, content access, and so on. MBSs guarantee ubiquitous coverage with a large cell radius (e.g., several kilometers), and hence the associated users can enjoy less handover when moving at a high speed. However, the large coverage radius may also bring massive connections to each MBS. As a result, MBSs can only provide limited radio resources to each vehicle user at low transmission rates. On the contrary, each SCS covers a relatively smaller area and serves fewer vehicle users at high transmission rates. Also, SCS users can get files without backhaul transmissions, which further reduces end-to-end delay. Nevertheless, SCSs mainly target popular file transmission, and their small coverage radius may cause frequent handover issues. In addition to the heterogeneity of network infrastructures, vehicular services are at a wide-range requiring heterogeneous QoS requirements. For example, the safety and control messages are delay-sensitive but occupy limited radio resources, whereas non-safety-related applications such as social networks on the wheel and map downloading require large bandwidth but can endure longer delay. The heterogeneity of network resources and traffic demand need be taken into consideration for user association and resource management.

Highly Dynamic Traffic Demand: On-road wireless traffic is highly dynamic in both time and spatial domains due to the variations of vehicle intensity. For example, traffic volume during rush hour can be 90 times the volume late at night, while traffic volume in one direction can be four times the volume in the opposite direction at the same road segment [10]. Such traffic non-uniformity can pose great challenges to network management. Bursty traffic during rush hour may lead to service outage due to limited network resources, whereas network resources cannot be fully utilized during off-peak periods. Also, the spatial traffic imbalance may also lead to congestion in some cells while resources are underutilized in other cells, degrading both service quality and network efficiency. In spite of traffic volume variation, the popularity distribution of different content also varies with time, and thus SCSs need to update their content cache to maintain high content hit rates.

Intermittent Energy Supply: Unlike the conventional power grid, renewable energy arrives randomly in an intermittent manner, which is likely to be a mismatch with traffic demand. For example, solar powered SCSs cannot provide adequate service after sunset, when vehicular networks may still be heavily loaded. In contrast, on-road traffic can be very light at noon, when solar energy can be harvested at peak rates. The unbalanced power demand and supply can cause energy outages as well as battery overflow, which degrades system reliability and also leads to energy waste. Accordingly, energy sustainability is critical to the proposed network paradigm, which requires intelligent network management to minimize the probability of energy outages and overflows.

In addition to the above mentioned challenges, there are also other issues that need to be addressed, such as vehicle mobility, and time-varying mmWave backhaul capacity. To summarize, the network should fully utilize heterogeneous network resources to provide reliable on-demand service, so as to minimize operational cost while

meeting different QoS requirements of on-road mobile applications.

Hierarchical Network Management

To address the above mentioned challenges, we propose a hierarchical network management framework, as shown in Fig. 2. The proposed framework mainly includes three components: energy management, content caching, and traffic steering. Furthermore, network management is conducted in both large (e.g., minutes or hours) and small time scales (e.g., seconds) with different strategies.

Energy management mainly deals with the randomness of renewable energy supply. Specifically, we propose dynamic SCS sleeping and radio frequency (RF) power control to reshape renewable energy supply by manipulating the process of charging and discharging. Notice that the power consumption of an SCS consists of two parts: constant power, which is irrelevant with traffic load, and RF power, which scales with traffic demand by adjusting the transmit power level or the number of utilized subcarriers. RF power control can reduce the RF power consumed by wireless transmission, while dynamic SCS sleeping can further reduce the constant part by completely deactivating the SCS. Although dynamic SCS sleeping is more effective for power saving, frequent switching may cause additional cost. Thus, SCS sleeping can be performed in a large time scale, and then each active SCS further adjusts the RF power in a small time scale. Hierarchical energy management can reshape renewable energy supply in the time domain to match the power demand at SCSs. For example, the SCSs with insufficient energy can switch to sleep mode, while active SCSs with oversupplied energy can enlarge transmit power to offload more vehicular traffic. In this way, SCSs can achieve energy-sustainable operation with balanced power demand and supply.

Content caching schemes are critical for system performance, due to the limited storage capacity and constrained mmWave backhauls. Specifically, we consider two design objectives, i.e., content hit rate and mobility support. Content hit rate determines the maximal amount of traffic offloaded from MSBSs to SCSs, which reflects the service capability of SCSs. Meanwhile, mobility-aware caching can be implemented to realize seamless handover, where content can be pro-actively fetched and stored at candidate cells based on handover prediction [11]. To realize these two objectives, the cache can be divided into two parts, one for popular content to guarantee content hit rate,[2] and the other for mobility-aware caching. Notice that mobility-aware caching requires frequent content fetching at the same time scale of vehicle handover, whereas the content popularity distribution may vary at a relatively slow pace. As the capacity of mmWave backhaul is constrained and varies dynamically with channel conditions, mobility-aware caching can be conducted in a timely manner on a small time scale, whereas popular content can be updated on a large time scale opportunistically based on channel status. Furthermore, each RSU should update content based on their own locations, since on-road mobile traffic requests can show location-based popularity.

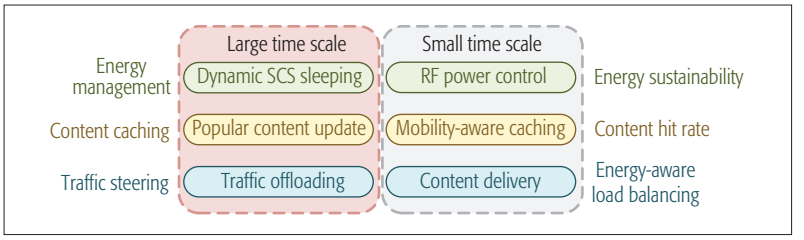


Figure 2. Hierarchical vehicular network management framework.

Traffic steering further reshapes traffic distribution to match the renewable energy supply, i.e., energy-aware load balancing. To this end, traffic offloading and content delivery are performed on different time scales, corresponding to energy management operations. In the large time scale, traffic offloading optimizes the amount of traffic served by each active SCS based on their renewable energy supply. For instance, SCSs with lower battery life can serve fewer vehicle users, and vice versa. In the small time scale, energy-aware content delivery optimizes transmission scheduling based on the SCS transmit power, to further improve QoS performance. For example, the delivery of best effort content can be delayed when the transmit power is reduced, while SCSs can pro-actively push popular content to vehicle users before requests when renewable energy is oversupplied. In essence, traffic offloading tunes the traffic load of each SCS (i.e., spatial traffic reshaping) while content delivery further adjusts traffic load at each time slot (i.e., temporal traffic reshaping). As such, traffic demand can be balanced with respect to renewable energy supply status.

Notice that these three operations jointly affect the performance of SCSs. For each SCS, content delivery control should be conducted based on the available battery life, stored content, and offloaded traffic status, as shown in Fig. 3. Therefore, the joint optimization of caching, energy and traffic management can help improve system performance, at the price of higher operational complexity.

Case Studies

Under the proposed management framework, many implementation problems still need to be addressed, such as caching design, intelligent energy and traffic management. In this section, we introduce two specific design examples for caching size optimization and sustainable traffic-energy management. Numerical results will be presented to offer insight into practical network deployment and operations.

We consider a two-way highway scenario where SCSs are deployed regularly with a coverage radius of 500 m. The file library consists of 1000 files whose popularity distribution follows a Zipf function with exponent γ_f. The headway among neighboring vehicles follows exponential distribution of parameter λ_v. In fact, λ_v reflects the vehicle density, and a larger λ_v characterizes denser vehicle scenarios. Assume all vehicles are greedy sources with average data rate requirement of 10 Mb/s, and each SCS can simultaneously serve 10 vehicle users at most due to radio resource limitations.

[2] Storing the most popular content can maximize content hit rate if SCSs do not cooperate with each other [12].

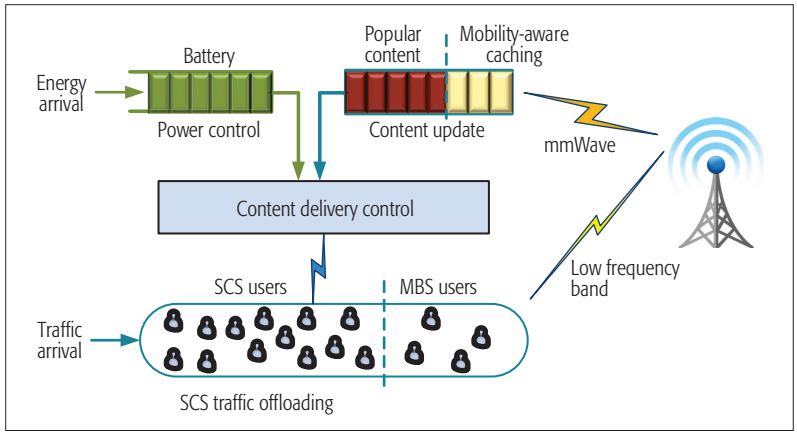


Figure 3. SCS management.

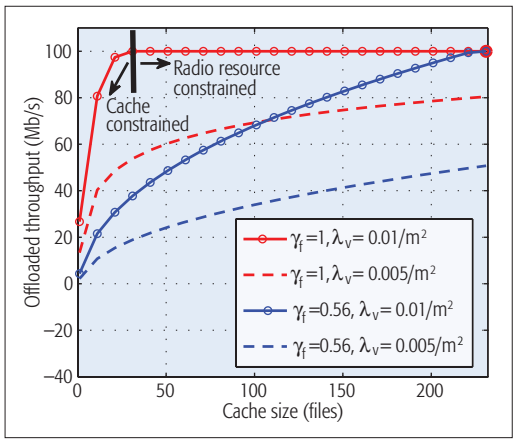


Figure 4. Service capability of SCSs with different cache sizes.

Cache Size Design

Figure 4 illustrates the amount of traffic that can be offloaded to each SCS under different traffic density λ_v and content popularity distributions γ_f,[3] which first increase but then level off with the increase of cache size. The reason is that the amount of offloaded traffic is also constrained by available radio resources.[4] Accordingly, system performance can be divided into two regions, i.e., a cache constrained region and a radio resource constrained region, as shown in Fig. 4. In the cache constrained region, content hit rate is low and fewer vehicle users can be offloaded to SCSs, which corresponds to the non-saturated case with under-utilized radio resources. As the cache size increases, more users can be offloaded to SCSs with higher content hit rate. Accordingly, the traffic of SCSs becomes saturated, and the throughput of SCSs no longer increases due to the limitation of available radio resource.

The obtained results reveal the Pareto optimality of cache size and SCS density, and offer insight into practical network deployment. For example, the optimal cache size should be larger than 31 files when the SCS coverage is 500 m, vehicle density is 0.01/m, and the popularity parameter γ_f = 1. Furthermore, the cost-optimal combination of cache size and SCS density for the given network capacity can also be found, given the cost functions of cache size and SCSs.

Sustainable Traffic-Energy Management

To reveal the importance of sustainable traffic-energy management, we study the service capability of the SCS under different traffic-energy management schemes. The greedy scheme is adopted as a baseline, where the SCSs always stay active and work at maximal transmit power. With sustainable traffic-energy management, an SCS goes into sleep mode if the available energy is insufficient to support its constant power consumption, otherwise it remains active and adjusts transmit power and offloaded traffic amount based on the instant energy arrival rate. Redundant energy is saved in the battery for future use, and the battery capacity is considered to be large enough without overflow.

Figure 5a illustrates the normalized traffic and energy profiles. Specifically, the two peaks of the traffic profile correspond to the on-road rush hours in the morning and afternoon. Meanwhile, solar energy harvesting is considered, and the daily energy arrival rate is modeled as a sine function with a peak at noon. Under the considered traffic and energy profiles, the normalized offloaded traffic (i.e., the percentage of vehicle users offloaded to the SCSs) is shown as Fig. 5b, where the peak energy arrival rate equals the maximal power consumption of each SCS, and the highest traffic density corresponds to SCS capacity. As shown in Fig. 5b, the sustainable traffic-energy management method outperforms the greedy scheme. Specifically, sustainable traffic-energy management can increase SCS capacity by nearly 1.7 times compared with the greedy scheme, realizing cost-effective management. In fact, the greedy scheme can minimize the probability of battery overflow, which performs well with sufficient energy supply. The sustainable traffic-energy management scheme further reduces the probability of battery outage through dynamic SCS sleeping, which can better utilize energy with higher efficiency.

Open Research Issues

As the study of cost-effective vehicular networking is still in its infancy, there are many research issues that remain unsolved.

Caching Scheme Design: Under the proposed management framework, efficient caching schemes should be designed to maximize content hit rate while minimizing handover cost, by determining cache size splitting, popular content updates, and mobility-aware caching. For the given cache splitting, the problems of popular content updates and mobility-based caching can both be modeled as a Markov decision process (MDP), and dynamic programing or machine learning provide powerful solutions. Then, optimal cache splitting can be further explored based on the designed schemes of popular content updates and mobility-aware caching schemes. Notice that there exists a trade-off between content hit rate and handover delay with different cache size splitting ratios. Accordingly, Pareto optimality can serve as the design criteria.

Sustainable Traffic and Energy Management: As demonstrated in the case study, conventional greedy traffic offloading and energy management schemes are insufficient, due to the randomness of renewable energy and highly dynamic vehicular

[3] γ_f = 0.56 comes from real data measurement of Youtube video streaming [13]. γ_f = 1 can describe location-based services (e.g., map downloading) whose requests may present higher similarity.

[4] Notice that each SCS can simultaneously serve 10 users at most, each with a data rate of 10 Mb/s.

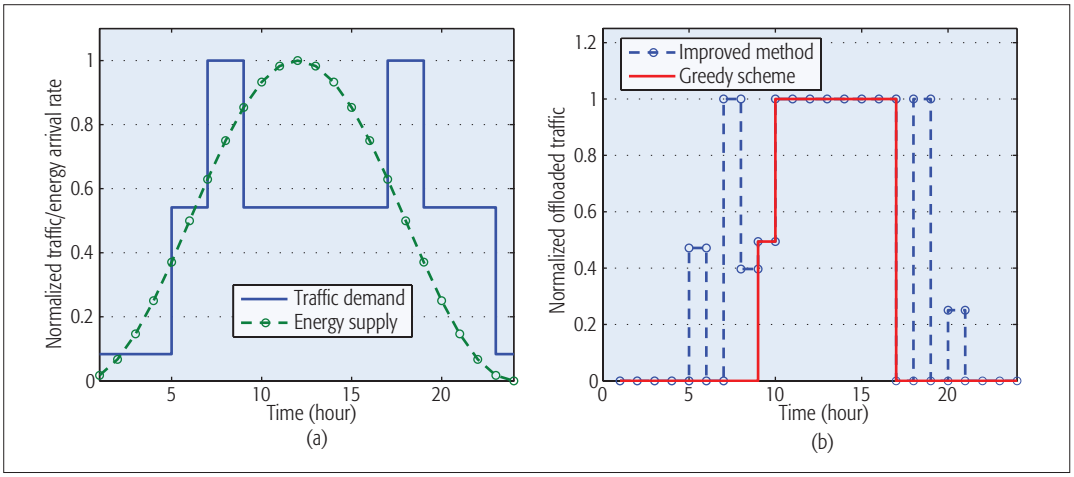


Figure 5. System performance with sustainable traffic-energy management: a) daily traffic and renewable energy profiles; b) normalized offloaded traffic.

Under the proposed management framework, efficient caching schemes should be designed to maximize content hit rate while minimizing handover cost, by determining caching size splitting, popular content update, and mobility-aware caching.

traffic. Sustainable traffic and energy management is desired to balance power demand and supply at each SCS, through the cooperation among neighboring SCSs and cellular networks. An optimization problem can be formulated to maximize the service capability of SCSs, subject to energy casualty and QoS requirements of all users. The decision variables include the work mode, offloaded traffic amount, transmit power, and content delivery scheduling of each SCS. However, this problem can be extremely complex due to the multi-dimensional coupled optimization variables. In this case, the hierarchical management framework can be exploited for problem decoupling. Specifically, we can deal with the work mode and offloaded traffic amount in the large time scale, while adjusting the transmit power and scheduling content delivery in the small time scale. Then, low-complexity management schemes can be proposed for practical implementations.

Cost-Effective SCS Deployment: The introduction of SCSs also poses new design issues for network deployment, as discussed in the case study of cache size optimization. In fact, the service capability of SCSs can increase with denser deployments of SCSs, larger cache size, or higher battery capacity. Accordingly, cost-effective SCS deployment should jointly optimize these system parameters to minimize long-term network cost while meeting vehicular traffic demand. Specifically, the trade-off among those system parameters should be carefully studied to obtain the cost-optimal combination. Stochastic geometry can be adopted for such large-scale system performance analysis, which can provide favorable closed-form results with reasonable approximations [14].

Conclusions

We have introduced a new type of on-road base station, called SCS, to exploit renewable energy harvesting, mmWave backhaul, and content caching techniques to achieve flexible, sustainable, and cost-effective vehicular networking. With promising 5G technologies, SCSs can enable "drop-and-play" deployment, green operation, and low-latency content delivery, paving the way to cost-effective vehicular networking. Furthermore, a heterogeneous vehicular network architecture has been proposed to provide high capacity and better QoS to vehicle users, by efficiently exploring the specific advantages of SCSs and MBSs. In addition, a hierarchical management framework has been designed, where energy management, content caching, and traffic steering are performed in both large and small time scales to deal with the dynamics of energy supply and traffic demand. Case studies on cache size optimization and sustainable traffic-energy management have been conducted to provide insights into the practical design of 5G-enabled vehicular networks. Important research topics on SCSs have also been discussed.

Acknowledgment

This work is sponsored in part by the Nature Science Foundation of China No. 91638204, and the Natural Sciences and Engineering Research Council of Canada.

References

[1] K. Abboud, H. A. Omar, and W. Zhuang, "Interworking of DSRC and Cellular Network Technologies for V2X Communications: A Survey," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, Dec. 2016, pp. 9457–70.

[2] S. Sun *et al.*, "Support for Vehicle-to-Everything Services Based on LTE," *IEEE Wireless Commun.*, vol. 23, no. 3, June 2016, pp. 4–8.

[3] F. Malandrino, C. Chiasserini, and S. Kirkpatrick, "The Price of Fog: A Data-Driven Study on Caching Architectures in Vehicular Networks," *Proc. First Int'l. Workshop on Internet of Vehicles and Vehicles of Internet*, Paderborn, Germany: ACM, July 2016, pp. 37–42.

[4] H. Seo *et al.*, "LTE Evolution for Vehicle-to-Everything Services," *IEEE Commun. Mag.*, vol. 54, no. 6, June 2016, pp. 22–28.

[5] S. Sorrentino, "LTE for Intelligent Transport Systems," Ericsson, Tech. Rep., June 2016, accessed Sept. 13, 2016; available: https://www.ericsson.com/research-blog/lte/lte-intelligent-transport-systems/

[6] H. A. Omar, W. Zhuang, and L. Li, "VeMAC: A TDMA-Based-MAC Protocol for Reliable Broadcast in VANETs," *IEEE Trans. Mobile Comput.*, vol. 12, no. 9, June 2013, pp. 1724–36.

[7] G. Piro *et al.*, "HetNets Powered by Renewable Energy Sources: Sustainable Next-Generation Cellular Networks," *IEEE Internet Comput.*, vol. 17, no. 1, Jan. 2013, pp. 32–39.

[8] J. G. Andrews *et al.*, "What will 5G Be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, May 2014, pp. 1065–82.

[9] J. Qiao, Y. He, and X. Shen, "Proactive Caching for Mobile Video Streaming in Millimeter Wave 5G Networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, Oct. 2016, pp. 7187–98.

[10] F. Bai and B. Krishnamachari, "Spatio-Temporal Variations of Vehicle Traffic in VANETs: Facts and Implications," *Proc. Sixth ACM Int'l. Workshop on VehiculAr InterNETworking*, Beijing, China, Sep. 2009, pp. 43–52.

[11] R. Wang *et al.*, "Mobility-Aware Caching for Content-Centric Wireless Networks: Modeling and Methodology," *IEEE Commun. Mag.*, vol. 54, no. 8, Aug. 2016, pp. 77–83.
[12] J. Gong *et al.*, "Joint Optimization of Content Caching and Push in Renewable Energy Powered Small Cells," *Proc. IEEE ICC'16*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.
[13] P. Gill *et al.*, "YouTube Traffic Characterization: A View from the Edge," *Proc. 7th ACM SIGCOMM Conf. Internet Measurement*, San Diego, USA, Oct. 2007, pp. 15–28.
[14] S. Zhang *et al.*, "How Many Small Cells Can Be Turned Off via Vertical Offloading under a Separation Architecture?" *IEEE Trans. Wireless Commun.*, vol. 14, no. 10, Oct. 2015, pp. 5440–53.

Biographies

Shan Zhang [M] (s372zhan@uwaterloo.ca) received her Ph.D. degree from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2016. She is currently a postdoctoral fellow in the Department of Electronical and Computer Engineering, University of Waterloo, Ontario, Canada. Her research interests include resource and traffic management for green communication, intelligent vehicular networking, and software defined networking. She received the Best Paper Award at the Asia-Pacific Conference on Communication in 2013.

Ning Zhang [M] (zhangningbupt@gmail.com) received the Ph.D. degree from the University of Waterloo in 2015. He is now an assistant professor in the Department of Computing Science at Texas A&M University-Corpus Christi. Before that, he was a postdoctoral research fellow at the BBCR Lab, University of Waterloo. He was the co-recipient of the Best Paper Award at IEEE GLOBECOM 2014 and IEEE WCSP 2015. His current research interests include next generation wireless networks, software defined networking, vehicular networks, and physical layer security.

Xiaojie Fang [s] (fangxiaojie@hit.edu.cn) received his B.Sc. and M.Sc. degrees from the Department of Electronics and Information Engineering, Harbin Institute of Technology in 2010 and 2012, respectively. From 2015 to 2016 he was a visiting scholar in the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. He is currently pursuing the Ph.D. degree in the Department of Electronics and Information Technology, Harbin Institute of Technology. His current research interests include physical layer security, coding and modulation theory.

Peng Yang [S] (yangpeng@hust.edu.cn) received his B.E. degree from the Department of Electronics and Information Engineering, Huazhong University of Science and Technology (HUST), Wuhan, China, in 2013. Currently, he is pursuing his Ph.D. degree in the School of Electronic Information and Communications, HUST. Since September 2015, he has been a visiting Ph.D. student in the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His current research interests include next generation wireless networking, software defined networking and fog computing.

Xuemin (Sherman) Shen [F] (xshen@bbcr.uwaterloo.ca) is a university professor, Department of Electrical and Computer Engineering, University of Waterloo, Canada. He is also the associate chair for graduate studies. His research focuses on resource management, wireless network security, social networks, smart grid, and vehicular ad hoc and sensor networks. He was an elected member of the IEEE ComSoc Board of Governor, and the chair of the Distinguished Lecturers Selection Committee. He has served as the Technical Program Committee chair/co-chair for IEEE Globecom'16, Infocom'14, IEEE VTC'10 Fall, and Globecom'07. He received the Excellent Graduate Supervision Award in 2006, and the Outstanding Performance Award in 2004, 2007, 2010, and 2014 from the University of Waterloo. He is a registered professional engineer of Ontario, Canada, an IEEE Fellow, an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering Fellow, and a Royal Society of Canada Fellow. He was a Distinguished Lecturer of the IEEE Vehicular Technology Society and the IEEE Communications Society.

IEEE Collabratec™

Bright Minds. Bright Ideas.

Introducing IEEE Collabratec™

The premier networking and collaboration site for technology professionals around the world.

IEEE Collabratec is a new, integrated online community where IEEE members, researchers, authors, and technology professionals with similar fields of interest can **network** and **collaborate**, as well as **create** and manage content.

Featuring a suite of powerful online networking and collaboration tools, IEEE Collabratec allows you to connect according to geographic location, technical interests, or career pursuits.

You can also create and share a professional identity that showcases key accomplishments and participate in groups focused around mutual interests, actively learning from and contributing to knowledgeable communities. All in one place!

Network.
Collaborate.
Create.

Learn about IEEE Collabratec at
ieeecollabratec.org

IEEE

Fraunhofer
Heinrich Hertz Institute

FRAUNHOFER INSTITUTE FOR TELECOMMUNICATIONS, HEINRICH HERTZ INSTITUTE, HHI

INDOOR Li-Fi

Make your Light Smarter

Fraunhofer HHI presents the next generation Gigabit Visible Light Communication (VLC) modules for wireless Internet access via light. Outstanding features are the smaller form factor, lower energy consumption, enhanced coverage and multi-user access. A standard Ethernet interface allows easy network integration. The new modules are immediately available for industrial prototyping and field tests.

Facts

- Use of standard high-power LEDs
- No interference with existing Wi-Fi networks
- Multi-user access possible
- Peak data rate 1 Gbps
- Small form factor

also available in other colors

Photonic Networks and Systems

Fraunhofer Heinrich Hertz Institute
Einsteinufer 37 | 10587 Berlin
Germany

products-pn@hhi.fraunhofer.de
www.hhi.fraunhofer.de/vlc