

# IEEE COMMUNICATIONS MAGAZINE

January 2017, Vol. 55, No. 1

- Enabling Mobile and Wireless Technologies for Smart Cities
- Impact of Next-Generation Mobile Technologies on IoT-Cloud Convergence
- Network and Service Management
- Ad Hoc and Sensor Networks
- Next Generation 911



IEEE

IEEE ComSoc<sup>™</sup>  
IEEE Communications Society

A Publication of the IEEE Communications Society  
[www.comsoc.org](http://www.comsoc.org)

While the world benefits from what's new,  
IEEE can focus you on what's next.

Develop for tomorrow with  
today's most-cited research.

Over 3 million full-text technical documents  
can power your R&D and speed time to market.

- IEEE Journals and Conference Proceedings
- IEEE Standards
- IEEE-Wiley eBooks Library
- IEEE eLearning Library
- Plus content from select publishing partners

**IEEE Xplore® Digital Library**

Discover a smarter research experience.

Request a Free Trial  
[www.ieee.org/tryieeexplore](http://www.ieee.org/tryieeexplore)

Follow IEEE Xplore on  

 **IEEE**  
Advancing Technology  
for Humanity

#### 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 Subrahmanyam, 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, U. of Hawaii, Manoa (USA)

Ciprian Dobre, Univ. Politehnica of Bucharest (Romania)

Silvia Giordano, Univ. 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, University of Tokyo (Japan)

##### *Consumer Communications and Networking*

Ali Begen, Cisco (Canada)

Mario Kolberg, University of Sterling (UK)

Madjid Merabti, Liverpool John Moores U. (UK)

##### *Design & Implementation*

Vijay K. Gurbani, Bell Labs/Alcatel Lucent (USA)

Salvatore Loreto, Ericsson Research (Finland)

Ravi Subrahmanyam, Invisage (USA)

##### *Green Communications and Computing Networks*

Song Guo, University of Aizu (Japan)

John Thompson, Univ. of Edinburgh (UK)

RangaRao V. Prasad, Delft Univ. of Tech. (The Netherlands)

Jinsong Wu, Alcatel-Lucent (China)

Honggang Zhang, Zhejiang Univ. (China)

##### *Integrated Circuits for Communications*

Charles Chien, CreoNex Systems (USA)

Zhiwei Xu, SST Communication Inc. (USA)

##### *Network and Service Management*

George Pavlou, U. College London (UK)

Juergen Schoenwaelder, Jacobs University (Germany)

##### *Networking Testing and Analytics*

Ying-Dar Lin, National Chiao Tung University (Taiwan)

Erica Johnson, University of New Hampshire (USA)

Irena Atov, InClusive Technologies (USA)

##### *Optical Communications*

Admela Jukan, Tech. Univ. Braunschweig, Germany (USA)

Xiang Lu, Futurewei Technologies, Inc. (USA)

##### *Radio Communications*

Thomas Alexander, Ixia Inc. (USA)

Amitabh Mishra, Johns Hopkins Univ. (USA)

#### Columns

##### *Book Reviews*

Piotr Cholda, AGH U. of Sci. & Tech. (Poland)

##### *History of Communications*

Steve Weinsten (USA)

##### *Regulatory and Policy Issues*

J. Scott Marcus, WIK (Germany)

Jon M. Peha, Carnegie Mellon U. (USA)

##### *Technology Leaders' Forum*

Steve Weinsten (USA)

##### *Very Large Projects*

Ken Young, Telcordia Technologies (USA)

#### Publications Staff

Joseph Milizzo, Assistant Publisher

Susan Lange, Online Production Manager

Jennifer Porcello, Production Specialist

Catherine Kemelmacher, Associate Editor



IEEE

IEEE ComSoc  
IEEE Communications Society

# IEEE COMMUNICATIONS MAGAZINE

JANUARY 2017, vol. 55, no. 1

[www.comsoc.org/commag](http://www.comsoc.org/commag)

- 4 THE PRESIDENT'S PAGE
- 6 CONFERENCE REPORT/IEEE ONLINEGREENCOMM 2016
- 8 CONFERENCE PREVIEW/IEEE WCNC 2017
- 9 WORKSHOP REPORT/IEEE COMSOC WOMEN'S WORKSHOP ON COMMUNICATIONS AND SIGNAL PROCESSING
- 10 BOOK REVIEWS
- 12 CONFERENCE CALENDAR
- 13 GLOBAL COMMUNICATIONS NEWSLETTER

## IMPACT OF NEXT-GENERATION MOBILE TECHNOLOGIES ON IOT-CLOUD CONVERGENCE

GUEST EDITORS: M. SHAMIM HOSSAIN, CHANGSHENG XU, YING LI, AL-SAKIB KHAN PATHAN, JOSU BILBAO, WENJUN ZENG, AND ABDULMOTALEB EL SADDIK

- 18 GUEST EDITORIAL
  - 20 WHO MOVED MY DATA? PRIVACY PROTECTION IN SMARTPHONES  
Wenyun Dai, Meikang Qiu, Longfei Qiu, Longbin Chen, and Ana Wu
  - 26 SECURITY AND PRIVACY FOR CLOUD-BASED IOT: CHALLENGES, COUNTERMEASURES, AND FUTURE DIRECTIONS  
Jun Zhou, Zhenfu Cao, Xiaolei Dong, and Athanasios V. Vasilakos
  - 34 HIGH-EFFICIENCY URBAN TRAFFIC MANAGEMENT IN CONTEXT-AWARE COMPUTING AND 5G COMMUNICATION  
Jianqi Liu, Jiafu Wan, Dongyao Jia, Bi Zeng, Di Li, Ching-Hsien Hsu, and Haibo Chen
  - 41 BEYOND 5G VISION FOR IOLITE COMMUNITY  
Doruk Sahinel, Cem Akpolat, Manzoor A. Khan, Fikret Sivrikaya, and Sahin Albayrak
  - 48 FEDERATED INTERNET OF THINGS AND CLOUD COMPUTING PERSVASIVE PATIENT HEALTH MONITORING SYSTEM  
Jemal H. Abawajy and Mohammad Mehedi Hassan
  - 54 WEARABLE 2.0: ENABLING HUMAN-CLOUD INTEGRATION IN NEXT GENERATION HEALTHCARE SYSTEMS  
Min Chen, Yujun Ma, Yong Li, Di Wu, Yin Zhang, and Chan-Hyun Youn
  - 62 MILLIMETER-WAVE WIRELESS COMMUNICATIONS FOR IOT-CLOUD SUPPORTED AUTONOMOUS VEHICLES: OVERVIEW, DESIGN, AND CHALLENGES  
Linghe Kong, Muhammad Khurram Khan, Fan Wu, Guihai Chen, and Peng Zeng
  - 69 SMART HEALTH SOLUTION INTEGRATING IOT AND CLOUD: A CASE STUDY OF VOICE PATHOLOGY MONITORING  
Ghulam Muhammad, SK Md Mizanur Rahman, Abdulhameed Alelaiwi, and Atif Alamri
- ## ENABLING MOBILE AND WIRELESS TECHNOLOGIES FOR SMART CITIES
- GUEST EDITORS: EJAZ AHMED, MUHAMMAD IMRAN, MOHSEN GUIZANI, AMMAR RAYES, JAIME LLORET, GUANGJIE HAN, AND WAEEL GUIBENE
- 74 GUEST EDITORIAL
  - 76 LTE/LTE-A RANDOM ACCESS FOR MASSIVE MACHINE-TYPE COMMUNICATIONS IN SMART CITIES  
Md Shipon Ali, Ekram Hossain, and Dong In Kim
  - 84 EFFICIENT ENERGY MANAGEMENT FOR THE INTERNET OF THINGS IN SMART CITIES  
Waleed Ejaz, Muhammad Naeem, Adnan Shahid, Alagan Anpalagan, and Minho Jo
  - 92 GEO-CONQUESTING BASED ON GRAPH ANALYSIS FOR CROWDSOURCED METATRAILS FROM MOBILE SENSING  
Bo-Wei Chen, Wen Ji, and Seungmin Rho

#### 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 Garca Gomez  
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:** \$27 per year print subscription. \$16 per year digital subscription. Non-member print subscription: \$400. Single copy price is \$25.

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

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

**SUBMISSIONS:** The magazine welcomes tutorial or survey articles that span the breadth of communications. Submissions will normally be approximately 4500 words, with few mathematical formulas, accompanied by up to six figures and/or tables, with up to 10 carefully selected references. Electronic submissions are preferred, and should be 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>. For further information contact Tarek El-Bawab, Associate Editor-in-Chief ([telbawab@ieee.org](mailto:telbawab@ieee.org)). All submissions will be peer reviewed.



#### 98 EFFICIENT MEDIA STREAMING WITH COLLABORATIVE TERMINALS FOR THE SMART CITY ENVIRONMENT

Jordi Mongay Batalla, Piotr Krawiec, Constandinos X. Mavroumoustakis, George Mastorakis, Naveen Chilamkurti, Daniel Negru, Joachim Bruneau-Queyreix, and Eugen Borcoci

#### 105 NAMED-DATA-NETWORKING-BASED ITS FOR SMART CITIES

Safdar Hussain Bouk, Syed Hassan Ahmed, Dongkyun Kim, and Houbing Song

#### 112 ENABLING COMMUNICATION TECHNOLOGIES FOR SMART CITIES

Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Yasir Mehmood, Abdullah Gani, Salimah Mokhtar, and Sghaier Guizani

#### 122 SECURITY AND PRIVACY IN SMART CITY APPLICATIONS: CHALLENGES AND SOLUTIONS

Kuan Zhang, Jianbing Ni, Kan Yang, Xiaohui Liang, Ju Ren, and Xuemin (Sherman) Shen

### NEXT GENERATION 911: WHERE ARE WE? WHAT HAVE WE LEARNED? WHAT LIES AHEAD?

GUEST EDITORS: CAROL DAVIDS, VIJAY K. GURBANI, SALVATORE LORETO, AND RAVI SUBRAMANYAN

#### 130 GUEST EDITORIAL

#### 132 EUROPEAN NG112 CROSSROADS: TOWARD A NEW EMERGENCY COMMUNICATIONS FRAMEWORK

Fidel Liberal, Jose Oscar Fajardo, Cristina Lumbreras, and Wolfgang Kampichler

#### 139 EMYNOS: NEXT GENERATION EMERGENCY COMMUNICATION

Evangelos K. Markakis, Asimakis Lykourgiotis, Ilias Politis, Anastasios Dagiuklas, Yacine Rebahi, and Evangelos Pallis

#### 146 UTILIZING AN NG 9-1-1 TEST LAB TO VALIDATE STANDARDS COMPLIANCE

Walter R. Magnusson, Ping Wang, and Yangyong Zhang

#### 152 IMPLEMENTATION OF NG9-1-1 IN RURAL AMERICA—THE COUNTIES OF SOUTHERN ILLINOIS: EXPERIENCE AND OPPORTUNITIES

Barbara Kemp

#### 159 IN-VEHICLE EMERGENCY CALL SERVICES: eCALL AND BEYOND

Risto Orni and Ana Goulart

### AD HOC AND SENSOR NETWORKS

SERIES EDITORS: EDOARDO BIAGIONI, SILVIA GIORDANO, AND CIPRIAN DOBRE

#### 166 SERIES EDITORIAL

#### 168 A DECADE OF RESEARCH IN OPPORTUNISTIC NETWORKS: CHALLENGES, RELEVANCE, AND FUTURE DIRECTIONS

Sacha Trifunovic, Sylvia T. Kouyoumdjieva, Bernhard Distl, Ljubica Pajevic, Gunnar Karlsson, and Bernhard Plattner

#### 174 A SOCIAL-AWARE FRAMEWORK FOR EFFICIENT INFORMATION DISSEMINATION IN WIRELESS AD HOC NETWORKS

Yanru Zhang, Lingyang Song, Chunxiao Jiang, Nguyen H. Tran, Zaher Dawy, and Zhu Han

#### 180 AUTOMATIC VIDEO REMIXING SYSTEMS

Sujeet Mate and Igor D. D. Curcio

#### 188 THE LOVE-HATE RELATIONSHIP BETWEEN IEEE 802.15.4 AND RPL

Oana Iova, Fabrice Theoleyre, Thomas Watteyne, and Thomas Noel

### NETWORK AND SERVICE MANAGEMENT

SERIES EDITORS: GEORGE PAVLOU AND JURGEN SCHONWALDER

#### 196 SERIES EDITORIAL

#### 197 INCREASING DNS SECURITY AND STABILITY THROUGH A CONTROL PLANE FOR TOP-LEVEL DOMAIN OPERATORS

Cristian Hesselman, Giovane C. M. Moura, Ricardo de Oliveira Schmidt, and Cees Toet

#### 204 SERVICE PROVIDER DEVOPS

Wolfgang John, Guido Marchetto, Felician Nemeth, Pontus Skoldstrom, Rebecca Steinert, Catalin Meirosu, Ioanna Papafili, and Kostas Pentikousis

### ACCEPTED FROM OPEN CALL

#### 212 IEEE 802.15.7r1 REFERENCE CHANNEL MODELS FOR VISIBLE LIGHT COMMUNICATIONS

Murat Uysal, Farshad Miramirkhani, Omer Narmanlioglu, Tuncer Baykas, and Erdal Panayirci

#### 218 A MULTI-SERVICE ORIENTED MULTIPLE ACCESS SCHEME FOR M2M SUPPORT IN FUTURE LTE

Nassar Ksairi, Stefano Tomasin, and Merouane Debbah

**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](http://www.comsoc.org)**

## MID-TERM PROGRESS REPORT

As I write this President's Page, my first year as ComSoc President has just finished. My goal as President, as expressed in my January 2016 column, is to "make the Society the go-to place for communications information, standards, technology, and community." At this half-way point, it is my belief that we have made great strides in reaching this goal.

The first order of business was to create the organizational structure (see the chart below) needed to generate, and then provide, communications information to our current members and to the many potential members we need to grow our Society. To accomplish this task, two new positions were created: the Chief Marketing Officer (CMO) and the Chief Content Officer (CCO).

The CMO oversees the Society's marketing initiatives across multiple platforms and media driving ComSoc quality, engagement, and brand consistency. Among the responsibilities of the CMO are the adoption of emerging media techniques such as podcasts, infographics, slide-shares, boards, blogs, and videos, along with the use of social platforms such as Facebook, Twitter, Tumblr, and Instagram.

The CCO oversees the Society's online content, including training platforms, webinars, newsletters, etc. Working closely with the CMO, the CCO provides the information that is needed to bring our messages, technologies, activities, and the like, to our constituents and to a wider audience.

Equivalent to our Director of Technical Committees, we created a new position, Director of Industry Communities. Where the Director of Technical Committees oversees our 20+ committees, such as Computer Communications, Data Communications, etc., the Director of Industry Communities oversees the first three communities that we created: 5G, IoT, and SD. Currently these communities have more than 30 industry leaders participating.

Changes in our Board Committees were also made. The Ad Hoc Young Professionals Committee was elevated to a permanent Standing Committee. This committee is tasked to bring more

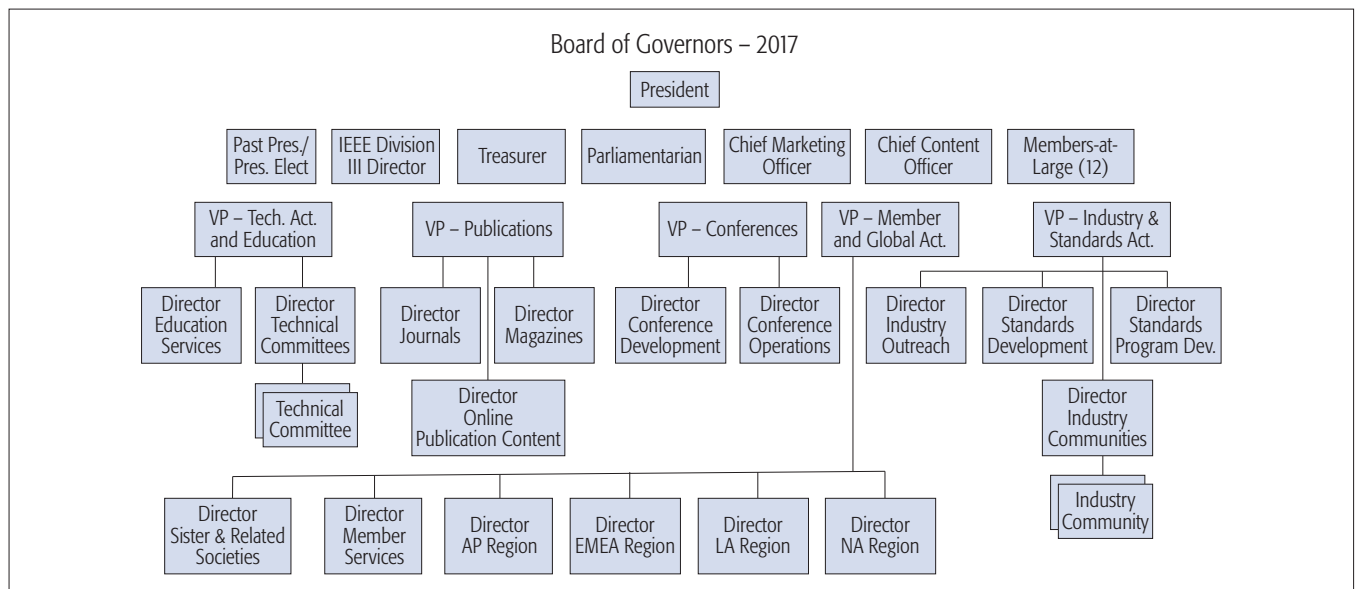


Harvey Freeman

professionals that are within 15 years of their degree into the ComSoc fold. An Ad Hoc Policy committee was formed to determine what Policy activities the IEEE Communications Society should participate in and how to pursue a Policy agenda that is appropriate and creates value for IEEE Communications Society membership.

In the Education and Training area, we held our 2nd Student Summer School in June where 40 students participated on-site in Torino, Italy, and 50-70 people took the courses via Internet streaming. Planning activities for the third session are underway (to be held in the U.S.). We held 26 online live courses with a total of 975 registrations as part of our ComSoc Training Program. The "Wireless for the Internet of Things" course was offered three times during 2016 and was our most popular course with 200 seats sold.

We have had great success in our Industry and Standards strategies. We have taken leadership positions in three new initiatives: 5G, IoT, and Fog Computing. In the 5G area, we started with a workshop in Princeton, NJ on 29-30 August 2016 with participation by more than 40 volunteers and IEEE staff representing 12 Societies/Operating Units. We established volunteer-led Working Groups to drive activity including: Web Portal, Content Development, Community Development, Standards, Education, Publications, Branding, Roadmap, Events and Industry Engagement. Six 5G summits were held across the world in the US, China, Germany, India, and Denmark, and drew from 100 to 400 in-person attendees and up to 1000 people watching Summits in the U.S. over IEEE TV. In 2017 we will continue to conduct summits on a global basis, launch an IEEE 5G Newsletter with original content focused on a broad audience, and refresh the portal that we launched in October on a regular basis with original content, including written Q&As, interviews, opinion pieces and podcasts. Lastly, we plan to establish and build a global, engaged IEEE 5G Community through a broad range of channels including Collabratec, Twitter, and LinkedIn.



The Internet of Things (IoT) initiative is sponsored by five major players: the Communications, Computer, Signal Processing, and Consumer Education Societies, and the Sensor Council. Some 15 other Societies are also participating. In 2016 we held a series of planning meetings and launched an IoT Directory and Registration Project. We completed a Draft Proposal for the IoT Magazine and selected an Editor-in-Chief. Planning is underway to hold our first IoT Vertical Summit in Anchorage, Alaska in September 2017.

Fog Computing is not as far along as the previous two. Initial funding is provided by ComSoc, with proposals being submitted to the IEEE Future Directions Committee for additional funding as a new Initiative. Papers were solicited for a feature topic of *Communications Magazine*, and so many good papers were received that the topic will be published in two parts. To collaborate with industry more, we are working with the OpenFog Consortium. We have attended all of their meetings and are actively influencing OpenFog directions.

We are putting the pieces together to create a one-stop Information and Communications Technology (ICT) Service System. The ComSoc Member Relationship Management Database/Tool was approved by our Board of Governors and vendors were selected for its implementation. The go-live date is expected in the first quarter. Our single sign on (SSO) work has been completed and we can now create a "member-wall" for web content, thereby making ComSoc membership more valuable. A team is now in place discussing products and services to be offered to members only.

Lastly, we have been working to establish enhanced value propositions for our constituents: academics/industry researchers, students, and industry practitioners and management. For academics, the issue thus far is not to retain academics but to give value for membership money, and possibly attract academics from adja-

cent disciplines who otherwise would not have joined the IEEE. Ways of accomplishing this include entrepreneurship, intermediate ComSoc membership/recognition status, stronger support of young academics, wider course portfolios, and an Xplore app with offline content.

Currently it seems that ComSoc does not provide value for students, or they do not understand the value we provide. The challenge then is to convince potential members that there is some exclusive content or service (i.e., value) that is obtainable only by ComSoc members. These may include special initiatives directed to students, exclusive content accessible only by ComSoc members, online forums, and interaction between students and world-class technical leaders.

Defining a concise value proposition for non-academic ComSoc members is a challenging task. The problems start with the definition of the target group. Basically, we are trying to define a value proposition for all ComSoc members not engaged with technology research, be it in academia or in industry. For the technical engineering professional, we could provide continuous professional education, create and manage online technical communities, create a magazine that provides content focused on the industry professional, and improve how ComSoc fosters professional networking. For executives, analysts, and business focused professionals, we could provide information about technology trends as well as awards and recognition.

In December I will report back to you what and how we accomplished the goals that I set forth in my first column in January 2016. In addition, I will provide our next President a guide to what still needs to be done to increase our membership, our value to our membership, and our standing in the international community.

# PROPEL YOUR NETWORK R&D TO A HIGHER ORBIT

**Technologies**

- 802.11 a/b/g/n/ac and e
- MANET
- WSN
- Cognitive Radio
- IOT
- VANETs
- LTE/LTE-A
- Military Radios
- Emulator for connecting real devices and more....

**Applications**

- Network R&D
- Military Communications
- Network Capacity Studies

**Used by**

- Universities
- Defence Organizations
- Network Equipment Manufacturers

**Model & Simulate**

**Data Center Network**

Write to us for a  
**FREE**  
Evaluation

**Graphical Analysis**

**Write your own code**

**Wireshark Packet Capture**

Wireshark is a registered trademark of Wireshark Foundation

**NETWORK SIMULATION/EMULATION SOFTWARE**  
With Open Protocol Source Code

**NetSim™**  
Model - Predict - Validate

Over 300+ customers across 15 countries  
[www.tetcos.com](http://www.tetcos.com) | [sales@tetcos.com](mailto:sales@tetcos.com) | + 91 76760 54321

## SIXTH ANNUAL IEEE ONLINEGREENCOMM 2016 CONFERENCE

BY JOSIP LORINCZ, FESB-UNIVERSITY OF SPLIT, CROATIA, MICHELA MEO, POLITECNICO DI TORINO, ITALY, EMAD ALSUSA, UNIVERSITY OF MANCHESTER, UK, MARCO RUFFINI, TRINITY COLLEGE, IRELAND, PAOLO MONTI, KTH, SWEDEN, ANDRES GARCIA SAAVEDRA, NEC EUROPE, GERMANY, CHIN KEONG HO, A\*STAR, SINGAPORE

From November 14–17, 2016, the IEEE Communications Society (ComSoc) hosted the annual IEEE Online Conference on Green Communications (IEEE OnlineGreenComm'16). This sixth IEEE OnlineGreenComm conference served as forum for presenting and discussing novel research and technological developments in the area of green communications, networking, and computing. By keeping the online concept of previous events, IEEE OnlineGreenComm'16 continued in spirit of "energy efficient discussion about energy-efficiency," through the conference virtual communication platform, which offers a unique and unsurpassed networking and interaction experience.

With the international webcasting of the complete conference program, IEEE OnlineGreenComm'16 attendees from industry and academia participated in conference events from the comfort of their own work and/or home environments. This ecological conferencing approach of IEEE OnlineGreenComm'16 contributes to the reduction of greenhouse gas emissions and costs by eliminating the need for conference travel, while providing the participants with time-flexible access to live or recorded OnlineGreenComm'16 events. In this way, the conference format of IEEE OnlineGreenComm'16 offers to presenters a truly worldwide audience, with the ability to answer questions in real time through interaction with session chairs and with the aid of webcast managers.

In the framework of the IEEE OnlineGreenComm'16 program, a number of different technical sessions related to green wireless communications, green optical communications, sustainability of communication systems, and smart cities were organized. The conference hosted four full days of keynotes, invited speeches, regular papers, panel and networking sessions, exploring topics ranging from energy harvesting of sensor ad hoc networks and green communications in fifth generation (5G) networks to energy-efficiency in wired optical networks and improvements in vehicle charging management.

IEEE OnlineGreenComm 2016 officially started on Monday, November 14, with conference chair Michela Meo of Politecnico di Torino in Italy welcoming all attendees to this year's conference. During the opening speech, Michela also presented an overview of the OnlineGreenComm'16 comprehensive program, which included nearly 25 presentations from global contributors representing the following regions: Europe/Middle East/Africa, Asia/Pacific, United States, and Canada. Michela's introduction emphasized the conference's benefits in terms of reduced travel and registration costs accompanied by experimenting with new media and interaction paradigms enabling all conference participants to be in line with a concept of "simply be green."

Following the opening remarks, the first keynote speaker, Louise Krug from British Telekom, UK, was introduced. In her inspiring presentation, Louise offered insights on the topic "High performance, high availability green networks—how far have we come, what remains problematical, and what solutions might be available?" Krug stated that the energy consumption of telecom networks such as the Internet have been identified as both significant and potentially fast growing. She looked back over the past six years at what has been achieved in terms of improving



telecom network energy efficiency, and looked at what remains a problem today. The last part of the keynote speech was dedicated to network function virtualization as a solution for improving network energy efficiency, based on the container approach, which

enables quick, robust, and adaptable upgrades of network resources in an energy-efficient manner.

The first day's program continued with a technical session on sustainability in communications, chaired by Michela Meo. This first technical session contained one regular paper entitled "Architecture Exploration of Multi-Source Energy Harvester for IoT Nodes," and one invited paper titled "An Assessment of the Impacts of Product Packaging on Overall Reliability and Environmental Performance: A Life Cycle Assessment Approach." The first day of the conference concluded with a green wireless session chaired by Mohammed Bahbahani from the University of Manchester, UK. In this session, two regular papers were presented on the topics "Spectral and Energy-Efficient Transmission over Frequency-Orthogonal Channels," and "Green OFDMA Resource Allocation in Cache-Enabled CRAN."

On the second day, two technical sessions chaired by Paolo Monti, KTH, Sweden, were organized. The first technical session on sustainability highlight research results of two regular papers titled "Lifetime Maximization of Connected Differentiated Target Coverage in Energy Harvesting Directional Sensor Networks," and "Grid Energy Consumption of Mixed-Traffic Cellular Networks with Renewable Energy Sources." The second technical session on green optical networks was devoted to the presentation of one regular paper titled "Asynchronous Delivery Oriented Efficient Resource Allocation in TWDM-PON Enabled Fronthaul," and one invited paper titled "Green Optical Transport Network Design for 5G Mobile Networks."

The extensive program on the second day concluded with an industry panel discussion on the topic "ICT Industry Experience and Perspectives on the Energy Efficiency Challenges for Today, 2020 and Beyond." Contributions to the panel discussion were given by three prominent panellists, moderated by Julio Montalvo from Telefónica, Spain. Julio opened the panel discussion with comments on the drivers for improving energy efficiency of the telecommunication sector, European Union (EU) targets in energy consumption reductions by 2030, and the role of information and communication technologies (ICT) as a key enabler for improving energy efficiency.

After his remarks, Julio introduced the first panellist, Erik Fernández from Telefónica, Spain, who spoke about the "global energy efficiency program" of mobile telecom operator Telefónica. During his presentation, Erik spoke about energy consumption trends and challenges in reducing telecom operator energy consumption, with implemented projects dedicated to overcoming these challenges. The second panellist, Pål Frenger from Ericsson AB, Sweden, discussed "5G energy performance," pointing to the low load dependence of energy consumption in the newest mobile networks, and the necessity of addressing low traffic cases in future networks through ultra-lean designs based on discontinuing transmission. The last panellist, Azzedine Gati from Orange Labs, France, spoke about "the challenge of energy efficiency in future ICT



networks." This presentation addressed the reasons why green is important for mobile operators, emphasizing the equipment vendor's commitment to increase energy-efficiency x2000 in the next ten years.

The third day of the conference featured the second keynote speech given by Thas Nirmalathas, Director of Melbourne Networked Society Institute, Australia, on the topic of "Sustainable Growth of Network Services—Accounting the Energy Consumption at a Service Level and Balancing the Use-Phase and Embodied Energy Consumption." In his extraordinary keynote speech, Thas offered a framework for analyzing and understanding energy consumption at the service level, and established the foundation for the development of a sustainable growth model for the Internet. To illustrate how network energy consumption can be linked to consumption at the service level, an approach based on modelling the energy consumption of over-the-top mobile wireless services was presented. In the last part of his presentation, a sustainable growth model balancing the use-phase and embodied energy consumption was discussed.

After this keynote speech, the third day of the conference proceeded with the second part of the technical session on green optical networks, chaired by Marco Ruffini from Trinity College, Ireland. In the framework of this technical session, one regular paper titled "Electricity Cost and Emissions Reduction in Optical Networks," and two invited papers titled "Light Trail Design for Energy-Efficient Traffic Grooming in Light-trail Optical WDM Networks" and "Energy Efficiency in Future PON," were presented. The content-rich third day of the conference was concluded with the first part of the smart cities technical session chaired by Josip Lorincz from FESB, University of Split, Croatia. This technical session highlighted two regular papers titled "Using IEC 61850 and IEEE WAVE Standards in Ad-Hoc Networks for Electric Vehicle Charging Management" and "Online QoS-Aware Charging Scheduling in Battery Swapping Stations Under Dynamic Energy Pricing."

On the conference's last day, the global audience was again provided with virtual access to two technical sessions and one panel discussion, followed by a virtual happy hour. The second part of the green wireless session was the first technical session of the last day. Within this session, chaired by Chin Keong Ho, from the Institute for Infocomm Research, Singapore, one regular paper titled "User Selection Scheme for Amplify-and-Forward Relaying with Zero Forcing," and two invited papers titled "The 5G Main Building Blocks" and "5G Energy Performance: Challenges and Solutions," were presented. The second part of the smart cities session was the last technical session presented at the conference. This session, chaired by Josip Lorincz, included two regular papers titled "Agent-Based Charging Scheduling of Electric Vehicles" and "Peak Load Reduction of Multiple Water Heaters: Respecting Consumer Comfort and Money Savings."

The last panel discussion on the fourth day was organized by the IEEE Young Professionals community, with the goal of presenting views of "IEEE Young Professionals on the Trends of Green Communications." The panel format included a five minute talk by each of the panellists followed by 30 minutes of discussion moderated by Rentao Gu, from the School of Information and Communication Engineering, Beijing University, China. During the panel discussion, the first panelist, Peerapol Tinnakornrisuphap from Qualcomm Research, USA, presented expertise in the area of "Smart Home and Smart Energy Management in Qualcomm Research." The second panelist, Xiodong Xu from Beijing University of Posts and Telecommuni-

cations, spoke on the topic "Challenges for 5G Green Cellular Networks."

The closing event of IEEE OnlineGreenComm'16 was organized as a virtual happy-hour and moderated by Lola Awoniyi-Oteri from Qualcomm Research, USA. This was the first time this type of event was offered, providing a completely new and exciting experience to the conference participants in terms of socializing and networking. The virtual happy-hour provided conference participants with an informal atmosphere in which to exchange experiences, ideas, and novel initiatives in the field of green communications. The pleasant ambience of the virtual happy-hour provided industry and academia participants of different generations and backgrounds with the opportunity to establish new research connections and collaborations.

Above all, IEEE OnlineGreenComm'16 was successfully organized in a manner keeping with the tradition of one of the most relevant global events for presenting the newest research results and ideas in the area of energy-efficient communications. The conference organizers handled the huge job of soliciting conference participation, promoting the conference agenda, and scheduling the review process. Gratitude goes to each person involved in organizing IEEE OnlineGreenComm'16. Acknowledgement also goes to all the reviewers who volunteered their time and professional expertise during the process of reviewing submitted papers, of which 31 percent were accepted for publication. In addition, special thanks go to the conference manager David Stankiewicz from IEEE for his great help in conference organization and coordination of all activities. Last but not least, the conference organizers express their gratitude to the keynote speakers, panelists, and contributing authors for sharing their expertise and state-of-the-art research results.

For more information on IEEE OnlineGreenComm 2016, please visit:

<http://onlinegreencomm2016.ieee-onlinegreencomm.org/>

Visitors to the conference website are encouraged to network with peers and colleagues, discuss IEEE OnlineGreenComm conference events and share their professional experiences through the conference's LinkedIn, Twitter, and Facebook pages.

All IEEE OnlineGreenComm 2016 content, including papers, recorded video presentations, and presenter slides, will be available online to registered conference participants on the following virtual platform link:

<https://vts.inxpo.com/Launch/Event.htm?ShowKey=34085>

New visitors can register for IEEE OnlineGreenComm'16 on-demand access for \$US 20 at the following link:

<http://www.cvent.com/d/tvq735>

On-demand access will provide full access to all conference materials over the conference virtual platform. Access to the conference virtual platform will be available until December 31, 2016. All accepted and presented papers from IEEE OnlineGreenComm'16 are also included in the IEEE Xplore database.

Finally, the IEEE OnlineGreenComm 2016 conference organizers invite all interested professionals to provide their contributions to the upcoming IEEE OnlineGreenComm conference, by proposing novel online conference formats or submitting their latest research results in this exciting and fast evolving field of green communications. All interested professionals seeking information about previous conference editions, ongoing conference updates, and detailed paper submission instructions, are invited to visit

<http://www.ieee-greencom.org>

Looking forward to meeting you at the next IEEE OnlineGreenComm conference!

## IEEE WIRELESS COMMUNICATIONS AND NETWORKING CONFERENCE 19-22 MARCH 2017, HYATT EMBARCADERO, SAN FRANCISCO, CA

We invite you to participate in IEEE WCNC 2017 which will be held in beautiful San Francisco, close to the heart of Silicon Valley. In addition to technical presentations on state-of-the-art wireless research, IEEE WCNC 2017 will feature a dazzling array of keynotes by luminaries in the field, “hot topic” panels, and industry talks, demos, and posters. A new innovation offered by IEEE WCNC 2017 is its “Startup City”, which will showcase the technologies of the most promising wireless startups. In addition, the conference will have a dynamic social program for its participants, including a ComSoc Young Professional Event and an event geared towards women in communications. The conference banquet will be held at McCormick and Kuleto’s in historic Ghirardelli Square with exquisite dining and stunning views of the bay.

The conference begins Sunday, 19 March with tutorials and workshops highlighting the latest in wireless communications. Tutorial topics include 5G, next-generation WiFi, massive MIMO, spectrum policy, millimeter wave systems, and molecular communications. The workshops will complement the tutorials and their foci range from 5G and the Tactile Internet to polar codes, M2M communications, and energy harvesting. That evening the conference will host a Welcome Reception at the conference venue, the Hyatt Regency.

Monday through Wednesday, 20–22 March, the technical



programs, panels, student program and industry program will highlight the future of wireless communications. This year, in addition to the original three technical tracks of WCNC, PHY and Fundamentals, MAC and Cross-Layer Design, and Wireless Networks, we have

added a new track: Emerging Technologies, Architectures and Services. The “hot topics” panels span the most timely, important, and controversial topics facing wireless researchers and the industry today. The panels are populated with leading experts in wireless communications from academia and industry that don’t necessarily share the same views, and we expect some sparks to fly during these discussions.

Student Program events include a mentoring session, student posters (with awards) and a recruiting session where students can meet representatives from the established companies and startups participating in the conference. The poster session is just before the recruiting event so students can discuss the details of their work with conference attendees from both industry and academia.

The industry program is multi-faceted and will provide an in-depth view into the most compelling wireless communications products on the market today. Industry posters, talks, and demos will be gleaned from open call submissions. Our exhibit floor will showcase the technologies of our 14 patron and exhibitor companies as well as 10-15 startups; these companies represent the best-known and most innovative companies in the wireless field. Our plenary speakers are leaders in academia and industry, including John Cioffi (ASSIA/Stanford), Erik Ehudden (Ericsson), Gerhard Fettweis (TU Dresden), Matt Grob (Qualcomm), Chih-Lin I (China Mobile), Ken Stewart (Ericsson), Marcus Weldon (Nokia Bell Labs), and Yongxing Zhou (Huawei). In addition to plenary talks, we will have invited industry forum talks where experts will discuss the latest wireless products and the future of the wireless world.

This IEEE WCNC is not like any in the past. It is a bold new conference that showcases the most advanced wireless research results as well as the best wireless technology. It includes many features completely new to IEEE WCNC and, in fact, to IEEE conferences in general. We invite you to participate in IEEE WCNC 2017 in the heart of San Francisco, 19–22 March at the Hyatt Regency, Embarcadero, San Francisco <http://wcnc2017.ieee-wcnc.org/>.



San Francisco Golden Gate Bridge. Photo Courtesy of San Francisco Travel Association.



Site of the Welcome Reception for WCNC '17, Photo Courtesy of the Hyatt Regency, Embarcadero, San Francisco



View from the Hyatt Regency, site of WCNC '17, overlooking the San Francisco Ferry Building and Bay Bridge. Photo Courtesy of the Hyatt Regency, Embarcadero, San Francisco.

## IEEE COMSOC THIRD WOMEN'S WORKSHOP ON COMMUNICATIONS AND SIGNAL PROCESSING: ALREADY A TRADITION

By OCTAVIA A. DOBRE, CHAIR OF THE WICE COMMITTEE (MEMORIAL UNIVERSITY, CANADA),

ANA GARCIA ARMADA (UNIVERSIDAD CARLOS III DE MADRID, SPAIN), AND JEAN ARMSTRONG (MONASH UNIVERSITY, AUSTRALIA)

The third edition of the IEEE ComSoc Women's Workshop on Communications and Signal Processing took place in Washington DC, on December 4, 2016, in conjunction with the IEEE GLOBECOM conference.

It was a fantastic event, which provided the opportunity for junior and senior researchers to discuss recent developments in the communications field, and fostered a warm environment for mentoring and networking between attendees. It hosted invited talks by senior researchers from industry, academia, and governmental agency. Dr. Peying Zhu (Huawei) presented new developments in the emerging 5G systems, Dr. Antonia Tulino (Nokia Bell Labs) talked about coding for caching in 5G networks, and Prof. Andrea Goldsmith (Stanford University) shared with us her vision on the future of wireless and what it will enable. Dr. Grace Wang (National Science Foundation) showed statistics and trends about the changing landscape in engineering research, while Monique Morrow (Cisco) revealed how to use technology to achieve gender neutrality. The speakers also shared inspiring stories about their career paths.

In addition, social components were included in the agenda, such as discussions focusing on the role of women in the IEEE and IEEE ComSoc, as well as panel discussions to address questions regarding career development in both industry and academia. The panel was lead by Prof. Katie Wilson (Santa Clara U.), and the invited panelists were Sheila Hemami (Draper Labs and IEEE VP Publications), Muriel Medard (MIT), Elza Erkip (NYU), and Antonia Tulino.

Activities for mentoring both junior and senior attendees took place, such as discussions on topics like The myth of doing it all; Work-life balance: is it possible?; How assertive should a woman engineer be?; When to say yes and how to say no?; How to harness networking and mentoring opportunities to accelerate your career?; Working together towards equality: dealing with biases in the workplace; Transitioning from engineering to management: is it for me?; Leadership gap: what would you do if you weren't afraid?; Publish or perish; Engineer your bliss: create the career you want; High impact communication.

In total, the workshop was attended by 54 participants, including both seniors and juniors from Australian, Canada, China, France, Germany, Ireland, Italy, India, New Zealand, Spain,

Switzerland, Sweden, Turkey, UK, and USA.

At this edition of the workshop, the IEEE ComSoc Women in Communications Engineering (WICE) Awards were presented: Prof. Elza Erkip (NYU) was the recipient of the Outstanding Achievement Award for significant contributions to cooperative communication and related fields; Prof. Shalinee Kishore (Lehigh U.) received the Outstanding Service Award for exceptional contributions and demonstrated leadership to the Women in Communications Engineering community; Prof. Urbashi Mitra (USC) was presented with the Mentorship Award for exceptional commitment and contribution to training and mentoring of Women in the Communications Engineering community.

The experiences that junior attendees had at the event were shared through feedback forms: "It was great to know that my worries are similar to others in my field. We have similar challenges and we learn from each other's experiences"; "It was a fantastic opportunity to network and talk to very interesting people"; "Excellent all around. Inspiring talks, outstanding technical program."

This fantastic event would not have been possible without the effort and generous support of different people. In particular, we would like to thank Dr. Harvey Freeman (IEEE ComSoc President) and Prof. Stefano Bregni (IEEE ComSoc Vice-President for Member Relations) for their support of WICE activities, as well as to Carol Cronin, Ting Qian, Bruce Worthman, and Susan Brook from IEEE ComSoc. We also thank the organizing committee for their effort and precious contribution, namely: Mari Carmen Aguayo-Torres (Universidad de Malaga), Jean Armstrong (Monash U.), Irena Atov (Microsoft), Irem Bor-Yaliniz (Carleton U.), Maite Brandt-Pearce (U. of Virginia), Sinem Coleri Ergen (Koc U.), Octavia A. Dobre (Memorial U.), Ana Garcia Armada (Universidad Carlos III de Madrid), Philippa Martin (U. of Canterbury), Shalinee Kishore (Lehigh U.), Urbashi Mitra (USC), Nada Philip (Kingston U.), Meryem Simsek (Technische Universität Dresden), Meixia Tao (Shanghai Jiao Tong U.), and Yahong Rosa Zheng (Missouri U. of Science and Technology).

The success of the first three editions of the workshop have paved the road for future editions, and we are looking forward to another great experience at the 2018 IEEE ComSoc Women's Workshop on Communications and Signal Processing!



Group photo: IEEE ComSoc Third Women's Workshop on Communications and Signal Processing (December 4, 2016).

## SOFTWARE DEFINED NETWORKING DESIGN AND DEPLOYMENT

By Patricia A. Morreale and James M. Anderson  
CRC Press, 2015, ISBN 978-1-4822-3863-1,  
hardcover, 172 pages

Reviewer: Piotr Borylo

The topic of Software Defined Networking (SDN) has been drawing attention from academia and industry for the last couple of years. The SDN concept enables us to introduce sophisticated mechanisms for computer networks, making it attractive for researchers. Simultaneously, the advantages of SDN allow network operators to reduce operational costs, improve resource utilization and, as a result, increase revenues.

This book is divided into nine chapters. Chapter 1 carefully describes virtualization. It is stated that valuable virtualization can be applied to the network infrastructure thanks to Software Defined Networking. Chapters 2 and 3 present the concept of SDN, including a brief history, architectural basics such as the separation of the control plane and the forwarding plane, router architecture, and finally the most important advantages, e.g., the availability of a global network view in the SDN controller. In Chapter 4 the properties of SDN are considered from the perspective of network operators, aimed mainly at increasing income. Chapter 5 discusses the application-oriented nature of the SDN concept as one of its most important advantages. To prove this claim, valuable and illustrative use cases are provided in the context of traditional networks. An especially interesting aspect is addressed in Chapter 6, where the SDN concept is presented as a driving factor for equipment vendors to reconsider their business models. Chapter 7 focuses on one of the most significant deployments of SDN made by Google in their wide area network interconnecting data centers. Valuable conclusions are drawn based on the deployment and challenges Google had to cope with in that case. Chapter 8 provides technical details about the OpenFlow Protocol. Operational principles of an OpenFlow switch, rules for packet and flow processing, management of flow tables, communication between switches and controllers are discussed. Finally, Chapter 9 considers how to transform legacy network infrastructures into SDN. This last chapter also includes illustrative examples of SDN applications, such as traffic-engineering in wide area networks. The content is complemented by description of the Frenetic and ElasticCon frameworks valuable in the context of network programming and SDN scalability.

The book provides basic information about the SDN concept and the related architectures as well as technical details about equipment, protocols, and frameworks. Therefore, the experienced as well as beginning network professionals will find it valuable. The book will be also useful for Ph.D. students who want to become familiar with the design of emerging networks. The strongest aspects of the book are, undoubtedly, its presentation of SDN's advantages from the business perspective, its description of steps taken by the industry to introduce SDN, and last but not least, its clear explanation of technical details regarding protocols and equipment. A minor drawback is related to the structure of the book: it seems that the introduction to virtualization could be briefer, while the length of Chapters 3 through 7 is not proportional to the remainder of the book. Simple consolidation of those chapters would probably improve the overall impression.

Summarizing, I recommend this book as a source of relevant and valuable information about emerging SDN topics. Architectural and conceptual basics are supplemented by technical details, making this work attractive for a wide spectrum of readers.

## FOUNDATIONS OF DECISION ANALYSIS

By Ronald A. Howard and Ali E. Abbas,  
Pearson Education, Inc., 2016, ISBN 978-0-13-233624-6, hardcover, 808 pages

Reviewer: Piotr Cholda

Decision making is not only an everyday life issue, but it has also practical and theoretical meaning in technology. This book deals with the both aspects. Thus, Howard and Abbas's work can be inspiring to a network specialist.

The book contains 40 short chapters, making the reading easy: one can become acquainted with the contents in small time chunks. The nature of the chapters is variable. Some of them introduce the reader into new aspects in a mild way, and others lead the reader through quite a complex modeling. Here is a short summary of the contents. Chapter 1 introduces the field of decision analysis, with the assumption that the reader may be a layman in the covered issues. The next chapter presents preliminaries on the relationship between the decision and probability theories. Chapter 3 deals with how values are involved in the decision processes. Next, the authors elaborate on how to effectively formulate and communicate decision problems. Then, the concept of the possibility description with

decision trees is introduced. Later, in Chapter 6 the uncertainty aspects are added; in consequence, the probability trees are covered. Additionally, more advanced material related to probabilistic modeling is given. Next, the description is extended with another important aspect of decision modelling, i.e., relevance. The following few chapters focus on the basics of decision taking: the related fundamental rules in Chapter 8; a description of preferences in Chapter 9; monetary utilities in Chapter 10; risk coverage in Chapter 11; and sensitivity analysis of the decision solutions in Chapter 12. Suggestions on how to gather data supporting the decision are discussed in Chapter 13. Then, some useful tools for communication are further discussed, i.e., decision diagrams in Chapter 14 and a representation of probability in Chapter 15. Cognitive aspects of decision making are dealt with in Chapter 16, while bounding the decisions is framed in Chapter 17. Then a group of chapters increasing awareness of complex modeling follows: methods to apply multiple sources while deciding in Chapter 18; general option theory in Chapter 19; and valuation of detectors suggesting which decision to take in Chapter 20.

Now, the flow comes back to topics enriching the decision theory. Thus, Chapter 21 describes how to represent influence factors; Chapters 22 and 23 focus on specific utility characteristics (those based on the logarithmic function and the linear risk tolerance, respectively); Chapter 24 discusses how moments of some measures can be applied; Chapter 25 introduces the notion of probabilistic dominance to enable a decision-maker to compare various distribution functions; Chapters 26 through 28 cover how to deal with different preferences; and Chapter 29 discusses how to create deals when the betting parties differ in the involved beliefs. Then, the authors again extend the knowledge of the reader with more advanced theoretical results: on dealing with the inherent imperfection of information in Chapter 30; an elaboration on auctions in Chapter 31; risk sharing in Chapter 32; and taking decisions in the face of potential fatal consequences in Chapters 33 and 34.

The book ends with practical aspects of decision making: how to make the continuous probability distribution discrete (a problem important from the computational viewpoint) in Chapter 35; support of decision making with simulations (Chapter 36); general decision processes (Chapter 37); an institutional view of the topics covered (Chapter 38); coordination methods for decision processes (Chapter 39); and ethical prob-

lems in Chapter 40.

Each chapter ends with a set of ingenious problems helpful in consolidating the awareness of the presented issues. An interesting idea is to add the so called “food for thought”, that is, one or more questions to be thought over by a reader. They also form a good basis for a more general discussion of the topics during exercise classes or seminars. Also, I appreciate that short appendices showing mathematical technicalities (e.g., on covariance) or more elaborate examples are located immediately after the

chapters, where they are necessary. The mathematical material is kept to a minimum. This makes the work most advantageous, especially to those who are just entering the field. However, it will be useful for decision theory experts mainly if they plan to use this position as the basis for teaching rather than research.

Some of the topics are presented in the context close to financial management (e.g., options or risk sharing), but generally the methods provided can be easily transferred to the technical field. While the book is not necessarily aimed

at networking engineers or researchers and presents the topic in the most general way, there are some examples that can be interesting for specialists working on current research topics in network management. In particular, I would like to indicate the problems where risk treatment is assumed. This problem is now quite broadly discussed in network security and reliability. Some of the covered aspects are also closely related to game theory (e.g., bidding problems for auctions), and can be inspiring for researchers in resource allocation.

#### CALL FOR PAPERS

##### *IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING*

Editor-in-Chief: E. Ayanoglu (UC Irvine)  
ayanoglu@uci.edu

#### BACKGROUND

The IEEE JSAC Series on Green Communications and Networking (JSAC-SGCN) is now *IEEE Transactions on Green Communications and Networking* (TGCN)! The three issues of JSAC-SGCN received a very large number of submissions, and the two issues published so far have substantially more papers than any other JSAC issues in recent history. As a result, IEEE has launched the new journal, *IEEE Transactions on Green Communications and Networking*. The journal will be published in an online-only format on a quarterly basis, with the first issue scheduled to be published in March 2017. The journal is co-sponsored by the IEEE Communications Society, the IEEE Signal Processing Society, and the IEEE Vehicular Technology Society.

The goal of this journal is to advance and promote significant technology advances in green communications and networks, including wireline, optical, and wireless communications and networks. Green communications and networking in this context means sustainable, energy-efficient, energy-aware, and environmentally aware communications and networking. The journal will promote innovations, new technologies, concepts, and principles toward a sustainable information and communications technology.

We invite submissions of high-quality manuscripts in the area of green communications and networking. We are seeking papers that have not been published and are not currently under review by any other journal. Topics of interest include but are not limited to:

- Green Wireline, Optical, and Wireless Communications and Networks
- Network and physical layer design, strategies, algorithms, protocols, and scheduling that consider environmental factors
- Energy-efficient and energy-aware heterogeneous networks, self-organized, and low-power sensor networks
- Energy efficiency in machine-to-machine communications, cooperative communications, and smart grid networks.
- Energy harvesting, storage, and recycling for network cross-layer optimization
- Environmentally-aware designs of communications and networking devices and systems
- Communications and networking for environmental protection monitoring

#### SUBMISSIONS

Prospective authors should prepare their submissions in accordance with the rules specified in the “Information for Authors” section of the TGCN guidelines ([www.comsoc.org/tgcn/information-for-authors](http://www.comsoc.org/tgcn/information-for-authors)). The paper should be compiled as a PDF file and uploaded to [mc.manuscriptcentral.com/tgcn](http://mc.manuscriptcentral.com/tgcn). The journal’s web site is [www.comsoc.org/tgcn](http://www.comsoc.org/tgcn).

#### IEEE Transactions on Green Communications and Networking Editorial Board

##### Energy Efficiency in Wireless Communications and Networking:

V. Leung (UBC, Area Editor), X. Ge (Huazhong U. S&T), S. Guo (Hong Kong Polytechnic), H. Ji (BUPT), V. Mancuso (IMDEA), A. Tulino (Bell Labs), R. Vaze (TIFR Mumbai), R. Yu (Carleton)

##### Optimization and Resource Allocation for Energy Efficiency:

Z. Niu (Tsinghua, Area Editor), K. Adachi (Univ. E-C), E. Bjornson (Linkoping), M. C. Gursoy (Syracuse), G. Miao (KTH), W. K. Ng (Univ. NSW), J. Thompson (Edinburgh), G. Yu (Zhejiang), Y. Zhang (Oslo)

##### Energy Harvesting and Wireless Energy Transfer:

S. Ulukus (Maryland, Area Editor), D. Gunduz (Imperial College), K. Huang (U. Hong Kong), H. Suraweera (Peradeniya), R. Zhang (National University of Singapore)

##### Energy-Efficient Networking and Protocols:

F. Granelli (Trento, Area Editor), R. Bruschi (Genoa), A. Cianfrani (Rome “La Sapienza”), V. Prasad (Delft UT), D. Rossi (Telecom ParisTech), C. Verikoukis (CTTC)

##### Devices for Energy Efficiency and Green Optical Communications:

J. Wu (U. Chile, Area Editor), L. Chiaraviglio (CNIT), D. Kilper (Arizona), H-S Kim (Michigan), P. Monti (KTH)

##### Energy Efficiency in Data Storage and Sensors:

M. Meo (Torino, Area Editor), C. Chaudet (Telecom ParisTech), B. Kantarci (Ottawa), S. Ren (UC Riverside), N. Tran (Kyung Hee U.)

UPDATED ON THE COMMUNICATIONS SOCIETY'S WEB SITE  
[www.comsoc.org/conferences](http://www.comsoc.org/conferences)

**2017**

**JANUARY**

*COMSNETS 2017 — Int'l. Conference on Communication Systems & Networks, 4–8 Jan.*  
 Bangalore, India  
<http://www.comsnets.org/>

**IEEE CCNC 2017 — IEEE Consumer Communications and Networking Conference, 8–11 Jan.**  
 Las Vegas, NV  
<http://ccnc2017.ieee-ccnc.org/>

*ICNC 2017 — Int'l. Conference on Computing, Networking and Communications, 26–29 Jan.*  
 Santa Clara, CA  
<http://www.conf-icnc.org/2017/>

**FEBRUARY**

*ICTACT 2017 — Int'l. Conference on Advanced Communication Technology, 19–22 Feb.*  
 Pyeongchang, Korea  
<http://www.icact.org/>

*WONS 2017 — Wireless On-Demand Network Systems and Services Conference, 21–24 Feb.*  
 Jackson Hole, WY  
<http://2017.wons-conference.org/>

**MARCH**

*NCC 2017 — Nat'l. Conference on Communications, 2–4 Mar.*  
 Madras, India  
<http://ncc2017.org/>

**IEEE DYSPAN 2017 — IEEE Dynamic Spread Spectrum Access Symposium, 6–9 Mar.**  
 Baltimore, MD  
<http://dyspan2017.ieee-dyspan.org/>

*ICIN 2017 — Conference on Innovations in Clouds, Internet and Networks, 7–9 Mar.*  
 Paris, France  
<http://www.icin-conference.org/>

*NETSYS 2017 — Int'l. Conference on Networked Systems, 13–17 Mar.*  
 Göttingen, Germany  
<http://netsys17.uni-goettingen.de/>

**IEEE WCNC 2017 — IEEE Wireless Communications and Networking Conference, 19–22 Mar.**  
 San Francisco, CA  
<http://wcnc2017.ieee-wcnc.org/>

**OFC 2017 — Optical Fiber Conference, 19–23 Mar.**  
 Los Angeles, CA  
<http://www.ofcconference.org/>

**IEEE CogSIMA 2017 — IEEE Conference on Cognitive and Computational Aspects of Situation Management, 27–31 Mar.**  
 Savannah, GA  
<http://cogsima2017.ieee-cogsima.org/>

*WD 2017 — Wireless Days 2017, 29–31 Mar.*  
 Porto, Portugal  
<http://www.wireless-days.com/>

**APRIL**

**IEEE ISPLC 2017 — IEEE Int'l. Symposium on Power Line Communications and its Applications, 3–5 Apr.**  
 Madrid, Spain  
<http://isplc2017.ieee-isplc.org/>

*WTS 2017 — Wireless Telecommunications Symposium, 26–28 Apr.*  
 Chicago, IL  
<http://www.cpp.edu/~wtsi/>

**MAY**

**IEEE INFOCOM 2017 — IEEE Int'l. Conference on Computer Communications, 1–4 May**  
 Atlanta, GA  
<http://infocom2017.ieee-infocom.org/>

*ICT 2017 — Int'l. Conference on Telecommunications, 3–5 May*  
 Limassol, Cyprus  
<http://ict-2017.org/>

**IFIP/IEEE IM 2017 — IFIP/IEEE Int'l. Symposium on Integrated Network Management, 8–12 May**  
 Lisbon, Portugal  
<http://im2017.ieee-im.org/>

*IEEE EIT 2017 — IEEE Int'l. Conference on Electro Information Technology, 14–17 May*  
 Lincoln, NE  
<http://engineering.unl.edu/eit2017/>

*ISNCC 2017 — Int'l. Symposium on Networks, Computers and Communications, 17–19 May*  
 Marrakesh, Morocco  
<http://www.isncc-conf.org/>

**IEEE ICC 2017 — 2017 IEEE Int'l. Conference on Communications, 21–25 May**  
 Paris, France  
<http://icc2017.ieee-icc.org/>

**IEEE BlackSeaCom 2017 — IEEE Int'l. Black Sea Conference on Communications and Networking, 5–9 June**  
 Istanbul, Turkey  
<http://blackseacom2017.ieee-blackseacom.org/>

*GIoTS 2017 — Global Internet of Things Summit, 6–9 June*  
 Geneva, Switzerland  
<http://iot.committees.comsoc.org/global-iot-summit-2017/>

**IEEE CTW 2017 — IEEE Communication Theory Workshop, 11–14 June**  
 Natatola Bay, Fiji  
<http://ctw2017.ieee-ctw.org/>

**IEEE LANMAN 2017 — IEEE Workshop on Local & Metropolitan Area Networks, 12–15 June**  
 Osaka, Japan  
<http://lanman2017.ieee-lanman.org/>

**IEEE SECON 2017 — IEEE Int'l. Conference on Sensing, Communication and Networking, 12–14 June**  
 San Diego, CA  
<http://secon2017.ieee-secon.org/>

–Communications Society portfolio events appear in bold colored print.

–Communications Society technically co-sponsored conferences appear in black italic print.

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



January 2017  
ISSN 2374-1082

MEMBERSHIP SERVICES

## Europe, Middle East and Africa Region Interview with Andrzej Jajszczyk, Director of the EMEA Region

By Stefano Bregni, Vice-President for Member and Global Activities, and Andrzej Jajszczyk, Director of the EMEA Region

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

In this issue, I interview Andrzej Jajszczyk, Director of the Europe, Middle East and Africa Region (EMEA). Andrzej is a professor at the AGH University of Science and Technology in Krakow, Poland, and President of the Krakow Branch of the Polish Academy of Sciences. He is the author or co-author of 12 books and over 300 research papers, as well as 19 patents in the areas of telecommunications switching, high-speed networking, network management, and reliability. He has held important positions in the IEEE Communications Society, such as Director of Magazines, Director of Europe, Africa, and Middle East Region, and Vice President-Technical Activities. He is an IEEE Fellow.

I had the honor to get to know Andrzej 20 years ago, when he was Editor-in-Chief of *IEEE Communications Magazine*. He was one of my best mentors in ComSoc when I first approached its Committees offering to serve as a volunteer. He was the founding editor of the *Global Communications Newsletter*, so it is a true honor and pleasure for me today to interview Andrzej and offer him this opportunity to outline his current activities and plans as Director of the Europe, Middle East and Africa Region.

**Bregni:** Andrzej, we might begin by outlining the main characteristics of the EMEA Region.

**Jajszczyk:** The EMEA Region covers a vast geographic area stretching from Cape Town, South Africa in the South, to Lisbon, Portugal, in the West, through Hammerfest, Norway, in the North, to Vladivostok, Russia in the East. The Region is served by 51 local chapters, together accounting for about one quarter of ComSoc members worldwide. The chapters provide a local connection for our society members. Their activities include talks organized within the Distinguished Lecturer Tour (DLT) or Distinguished Speaker Program (DSP) frameworks, social events, workshops, seminars, special events, etc.

**Bregni:** The Distinguished Lecturer Program (DLP) and the Distinguished Speaker Program (DSP) are particularly appreciated by our members. During my previous term as VP-Member

Relations, we increased the budget allocated to these programs, to allow more and better DLTs in all regions. What is your perception of these programs? How are they organized in the EMEA Region?

**Jajszczyk:** DLTs provide the means for ComSoc chapters to identify and arrange lectures by renowned experts on communications and networking related topics. ComSoc's DSPs enable current and past distinguished lecturers as well as ComSoc officers, IEEE Fellows, and prominent speakers to schedule lectures while traveling on business trips.

The DLTs and DSPs are coordinated by Charalabos (Harry) Skianis from Greece. In 2016, six DLTs were organized in the EMEA Region, and we have already approved three additional DLTs to be held in early 2017.

What is important, most of these tours were or are to be organized in areas with a rather low ComSoc membership, such as the Middle East or Africa. We do hope to encourage more professionals from these areas to join our Society. The four talks within the Distinguished Speaker Program held in 2016 involved such countries as Spain, Austria, Portugal, and Kuwait. To cope with the vast geographic area, and related travel costs, of the EMEA Region, we recently modified our Policies & Procedures for DLT/DSP.

**Bregni:** The EMEA Region Board assigns awards to recognize the contributions of its outstanding members. Please describe these awards in some detail?

**Jajszczyk:** An important facet of the EMEA activities, coordinated by Emilio Calvanese Strinati from France, is our Award Program, which includes the EMEA Young Researcher Award and the EMEA Region Distinguished Service Award.

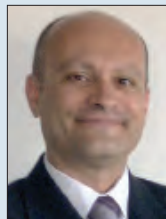
In 2016, the recipients of three EMEA Young Researcher Awards were: Sheng Yang from Centrale-Supelec, Gif Sur Yvette, France; Daniel Enrique Lucani Roetter from Aalborg University, Denmark; and Claudia Campolo from the University Mediterranea of Reggio Calabria, Italy. The EMEA Region Distinguished Service Award was given to Fambirai Takawira from the University of the Witwatersrand, Johannesburg, South Africa. Every year, we also select a particularly active chapter to receive the Chapter Achievement Award. In 2016, this award was given to the Romania Chapter and presented at Globecom 2016 in Washington, DC, USA.

**Bregni:** What are the major challenges that you currently see in the EMEA Region?

**Jajszczyk:** We have two major challenges: one is membership growth in our region, and the other is increasing industry-academia cooperation.

The coordination of the latter is the responsibility of Ali C. Begen from Turkey, supported by Peter Rost from Nokia Networks, Munich, Germany. The task is not easy in the current business environment, taking into account industry restructuring, growing competition, and changes in the overall markets. However, I do hope that we will be able propose working initiatives to strengthen cooperation between industry and academia within the EMEA Region.

(Continued on Newsletter page 4)



Stefano Bregni



Andrzej Jajszczyk

## Blaze of Glory: IEEE Communications Society and Vehicular Technology Society Malaysia Joint Chapter

### Winner of the 2016 IEEE ComSoc Chapter-of-the-Year Award

By F. Hashim, A. Sali, A. A. El-Saleh, W. Nurdiana, K. Anuar, K. Abdullah, M. Y. Alias, M. Roslee, N. Ramli, H. Mohamad, B. M. Ali, and M. Ismail, IEEE Malaysia Communications Society and Vehicular Technology Society Joint Chapter

#### 2016: THE YEAR IT IS

The year 2016 has been an exhilarating year for the ComSoc/VTS Malaysia Chapter. It is the year Kuala Lumpur, for the first time, hosted IEEE ICC. It is also the year the Joint Chapter won, for the second time in three years, the Chapter of the Year Award and the 2016 Chapter Achievement Award (Asia Pacific). Further, 2016 also marks 24 years of IEEE Malaysia Communication Society (ComSoc) Chapter and 10 years of the IEEE Malaysia ComSoc/VTS Joint Chapter. Over the past three years the Joint Chapter has received:

- 2014 and 2016 IEEE Communication Society Chapter of the Year Award.
- 2014 and 2016 IEEE Communication Society Chapter Achievement Award (Asia Pacific).
- 2013 and 2015 IEEE Vehicular Technology Society Chapter of the Year Award.
- IEEE Malaysia Section 2014 Chapter of the Year Award.

This is indeed a great honor for our Chapter, which started from a humble beginning in 1992 with about 60 members. The Chapter gradually grew until now we have more than 600 members, making our Chapter one of the largest chapters in the Malaysia Section. The recipe for success stems from the selfless commitment of successive groups of Executive Committees now led by Fazirulhisyam Hashim from the Universiti Putra Malaysia. The committee members are dynamic and dedicated people hailing from both academia and industry, making it an effective combination when conducting programs for the members.

#### ACTIVITIES

To highlight our 2015 activities, this Chapter organized 10 membership development programs at various universities and 68 technical, professional, and continuing education activities. Out of these, 11 are DLT/DSP talks given by six IEEE Distinguished Speakers from ComSoc and VTS. Apart from DLTs and DSPs, the Chapter organized and technically co-sponsored three



The Joint Chapter organized technical visits to TLDM, Media Prima Bhd, and Sapura Bhd.



The MICC 2015 Organizing Committee.



The ICC 2016 venue: Kuala Lumpur Convention Center (photo courtesy of KLCC).

conferences (TAFGEN, ICONSPACE and MICC), and a number of administrative and strategic meetings.

#### REACHING OUT TO INDUSTRY AND COMMUNITY

One of the highlights of 2015 is our engagement in community service. In early 2015 many areas in Malaysia, especially in the East Coast region, were inundated by a huge flood caused by exceptionally heavy rain. More than 200,000 people were affected, and many houses were totally wiped out, as were crops and livestock. The Chapter participated in the national Flood Humanitarian and Relief Mission to provide support to the affected areas by providing clothes, food, and cash for the affected people. We also provided support to rebuild the houses of the victims as well as motivating the children.

The Chapter realized the need to enhance its visibility in technical activities as well as in professional, community, educational activities, etc. In particular, we did the following:

- Initiated new collaboration with industry, such as Sapura and the Royal Navy, by means of technical visits and discussions to enhance our reach to members in industry. We also held some of our IEEE DLTs/DSPs and other technical talks at industry locations.

- Increased visibility of IEEE among university management, industry, local and global organizations through collaboration and joint activities.

- Conducted meetings and social gatherings with members (e.g., Eid Celebration and social gatherings during IEEE Day).

We also organized a number of collaborative activities with the IEEE Malaysia Sec-



## International Conference on Localization and GNSS (ICL-GNSS 2016), Barcelona, Spain, 28–30 June 2016

By G. Seco-Granados, J. Nurmi, J. A. Lopez-Salcedo and E.S. Lohan, Spain, Chairs and Steering Committee Members, ICL-GNSS 2016

The sixth edition of the International Conference on Localization and GNSS was organized in Barcelona, Spain on 28-30 June 2016 jointly by the Universitat Autònoma de Barcelona (UAB), Spain and Tampere University of Technology (TUT), with the support of the European Commission DG Joint Research Center (JRC). The conference was chaired by Prof. Gonzalo Seco-Granados and Prof. J. A. Lopez-Salcedo from UAB. Assoc. Prof. Elena-Simona Lohan from TUT was TPC Chair, and José A. Del Peral Rosado from UAB was the Publication Chair. The conference was technically co-sponsored by IEEE through the IEEE Spain Section and the Spanish chapters of the IEEE AESS and SPS/ComSoc. The event brought together more than 70 participants from academia and industry for three days to the conference center Casa Convalescència, which is a historic modernist building located next to the historic site of the Hospital de la Santa Creu i Sant Pau.

There were four high-caliber keynote speakers. Dr. Ignacio Fernández from the European Commission, Belgium, spoke about GNSS authentication features and key management, and the combination of authenticated and non-authenticated signals and sensors in GNSS and A-GNSS. Prof. Danijela Cabric from UCLA, United States, explained the concept of cognitive radio (CR) in the context of wireless positioning, and explored a set of CR cooperative localization algorithms for single and multiple primary users. Assoc. Prof. Klaus Witrisal from Graz University of Technology, Austria, concentrated on location-aware communications and how the location information can benefit the network. Dr. Javier de Salas, Broadcom, Spain, focused on the carrier phase positioning experiences in consumer global navigation satellite systems (GNSS) devices.



Attendees at the conference venue, Casa Convalescència, getting ready for the technical sessions.

The session topics were security aspects in GNSS, GNSS receivers, cooperative and hybrid positioning, positioning via 5G & UWB, indoor navigation, RSS-based positioning, design, prototyping and testing of positioning devices, ionospheric, tropospheric and scintillation effects in GNSS, and a MULTI-POS special session on multi-technology positioning, resulting in 43 oral presentations. There was one award given at the event: the best paper award given to Josef Kulmer, a Ph.D. student from Graz University of Technology, Austria, for the work entitled “Cooperative localization and tracking using multipath channel information.” It is worth remarking that this year’s edition of the conference was focused almost equally on positioning with GNSS and positioning with terrestrial technologies, attracting participants from both communities and facilitating interaction between them.

The talks and the social events were compelling and of high quality. On the first evening, there was the welcome reception at the Noble Halls of the Barcelona City Council, where Dr. Miquel Angel Essomba, the Commissioner for Education and Universities of the Barcelona City Council, gave a warm-hearted speech. The Gala dinner took place on the second evening, in the La Gavina restaurant, after an exciting guided visit at Parc Güell, one of the best known and representative Barcelona landmarks and an example of the architectural genius of Antoni Gaudi.

In 2017, the seventh edition of ICL-GNSS will be located in Nottingham, United Kingdom. For more information, see [www.icl-gnss.org](http://www.icl-gnss.org)



Group picture on the visit to Park Güell.



Welcome reception at the Saló de Cent, one of the noble halls of the Barcelona city council.

## MALAYSIA CHAPTER/Continued from page 2

tion and other chapters in Malaysia. Through these collaborations, we gave motivational talks and activities to IEEE student members to engage and participate in community service as well as sharing experiences with IEEE student branches in managing their branches, membership drives, and organizing various successful IEEE events.

In addition, Final Year Project Excellence Award in Communications were also organized at several universities around the country. Apart from evaluating the students' project work and awarding the winners, our ExCom used this activity to meet our members and deliver short talks to promote IEEE and ComSoc/VTS.

In 2015 we also organized our very first ComSoc/VTS Awards, giving prizes to the Best Paper published in selected IEEE journals, Best Social Activity and Outstanding Dissertation Awards. These awards are intended to encourage IEEE members to produce high quality activities and projects, and to obtain recognition from IEEE.

Aduwati Sali, one of our ExComm member since 2014, was consulted by local authorities as well as appearing in a national satellite TV broadcast channel to give her expert opinion on the disappearance of flight MH370 in March 2014, highlighting the role of satellite communications technology in tracing the whereabouts of the missing aircraft. Apart from broadcasting media, she was also asked to conduct training and talks on satellite communications, as well as to explain how the data log helped the search and rescue operations.

The year 2015 was nicely concluded by the Chapter with the organization of our flagship Malaysia International Conference on Communications (MICC 2015) on 23-25 November 2015 in Kuching, Sarawak. This conference was officially opened by Rt. Hon. Datuk Sylvester Entri Muran, Deputy Minister of Public Facilities, Sarawak. The event was well covered by the local media.

### IEEE MALAYSIA COMSOC/VTS JOINT CHAPTER 10 YEAR CELEBRATION

As 2016 marks the 10th year celebration of the Joint Chapter, various activities have been planned to commemorate the event. Some of the key activities are:

**Reaching Out to Flood-Prone Areas:** This is a community project to educate the local community in flood-prone areas. A one-day event is being planned comprising a talk on amateur radio and how this communication technology can help the local community contact the flood control center and rescue workers.

**Building Next Generation Workforce in Communication and Vehicular Technology:** This seminar is aimed at industry players and researchers as well as students, and is focused on how communications and vehicular technology have shaped the workforce in Malaysia.

**Defy Gravity—From Technical Expertise to Society:** This sharing session is a motivational talk for IEEE members and volunteers,

focused on how we can reach out to the society, using our technical expertise.

### 2016 IEEE INTERNATIONAL CONFERENCE ON COMMUNICATIONS (ICC 2016)

The Chapter and the international Organizing Committee (OC) organized an extraordinary IEEE International Conference on Communications (ICC 2016) on 23-27 May 2016 in Kuala Lumpur. The OC included the Executive Co-Chairs Hikmet Sari (Supelec, France) and Borhanuddin Mohd Ali (Universiti Putra Malaysia); the Advisory Executive Vice-Chair, Datuk Hod Parman (Malaysia); the Conference Operations Chair, Hafizal Mohamad (MIMOS Berhad, Malaysia); the Technical Program Co-Chairs, Stefano Bregni (Politecnico di Milano, Italy) and Nelson Fonseca (State University of Campinas, Brazil); the Industry Forums & Exhibition Chair, Khaled B. Letaief (Hong Kong University of Science and Technology); and the Exhibition Chair, Nordin Ramli (MIMOS Berhad, Malaysia).

The venue, the Kuala Lumpur Convention Centre (KLCC), is a modern five-star convention facility situated strategically at the foot of the iconic PETRONAS Twin Towers, overlooking a spacious, well landscaped public park.

The conference attracted 1,839 attendees from 58 countries. IEEE ICC 2016 featured interesting and thought-provoking keynote talks from John Cioffi (ASSIA Inc.), Seizo Onoe (NTT DOCOMO), Henning Schulzrinne (Columbia University and the FCC), Mischa Dohler (King's College London), and San-qi Li (Huawei Technologies).

Technical Symposia consisted of 967 peer-reviewed papers highlighting the latest research and business policies surrounding communications advancements worldwide. CTO Forums were also held featuring six CTOs from Axiata, Celcom, Telecom Malaysia, PLDT, Vodafone, M1 and Idea, discussing the fast evolution of technology and challenges for emerging countries.

There were 14 Industry Panels with representatives from various industry giants such as Samsung, Nokia, Intel, Microsoft, NEC, China Mobile, and NTT Docomo, addressing issues related to 5G, IoT, and shared spectrum. Attendees had the opportunity for professional development by attending one or more of the 19 tutorials and 14 workshops on the latest breakthroughs in information and communications technology. We are grateful for all the patrons and supporters of the conference, which include Axiata, Huawei, National Instruments, the Malaysian Communications and Multimedia Commission, Keysight Technologies, the Ministry of Communications and Multimedia Malaysia, the Malaysian Convention and Exposition Bureau, and Tenaga Nasional Berhad.

### MEMBERSHIP SERVICES/Continued from page 1

One of the issues that I would like to address as well is ComSoc's role in the current migration crisis in our region. I do believe that we have to find ways to help communication engineers from the Middle East and war affected countries of Africa who are currently trying to find refuge abroad, or are trying to survive in their countries. Frankly speaking, I do not have clear ideas on how to do that, so I would appreciate any input from our members.

**Bregni:** Besides those EMEA Board Members who you have mentioned already, would you like to acknowledge the contribution of others in particular?

**Jajszczuk:** I would like to mention here three other people who are playing very important roles within the EMEA Board. They are the past director, Hanna Bogucka, who serves as an advisor; Piotr Borylo, who is responsible for membership development and EMEA web visibility; and Carol Cronin, the staff representative whose role in running all our activities simply cannot be overestimated.

**Bregni:** Among the many activities that you are running, what will be the highlight next year?

**Jajszczuk:** At ICC'2017 in Paris, France, we will organize the EMEA Regional Chapters Chairs Congress. The Congress is intended to facilitate and encourage sharing, feedback, and networking among Chapter Chairs, staff, and ComSoc officers.

**GLOBAL COMMUNICATIONS NEWSLETTER**

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

**IEEE COMMUNICATIONS SOCIETY**  
STEFANO BREGNI, VICE-PRESIDENT FOR MEMBER AND GLOBAL ACTIVITIES  
CARLOS ANDRES LOZANO GARZON, DIRECTOR OF LA REGION  
SCOTT ATKINSON, DIRECTOR OF NA REGION  
ANDRZEJ JAJSCZYK, 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**  
SIMONA LOHAN (elena-simona.lohan@tut.fi)



www.comsoc.org/gcn  
ISSN 2374-1082

# IEEE Wireless Communications and Networking Conference

19-22 March 2017 // San Francisco, CA

## The Future of Wireless Technology is HERE at IEEE WCNC 2017

### EXCITING INDUSTRY PROGRAM

Panels, posters, demonstrations and an exhibition featuring the latest developments in broadband and wireless technologies, systems and services.

### ENGAGING TECHNICAL PROGRAM

Sessions covering the best in PHY, MAC, wireless networks and emerging technologies.

### INTERACTIVE STUDENT PROGRAM

Poster and demonstration session, graduate student mentorship session and an industry recruitment event.

### STARTUP CITY - NEW AT IEEE WCNC

- Exhibit area devoted to startups with a Startup Showcase
- Attendees and panel will select the "Hottest Wireless Startup"

### KEYNOTES



**John Cioffi**  
Assia & Stanford



**Erik Ekudden**  
Ericsson



**Gerhard Fettweis**  
TU Dresden



**Matt Grob**  
Qualcomm



**Chih-Lin I**  
China Mobile



**Kenneth Stewart**  
Intel



**Marcus Weldon**  
Nokia



**Yongxing Zhou**  
Huawei

### PATRONS & EXHIBITORS

#### Gold



#### Silver



#### Bronze



#### Publisher



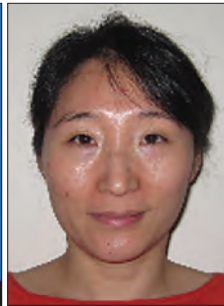
#### Exhibitors



## IMPACT OF NEXT-GENERATION MOBILE TECHNOLOGIES ON IoT-CLOUD CONVERGENCE



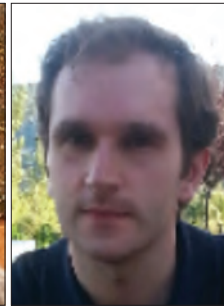
Changsheng Xu



Ying Li



Al-Sakib Khan Pathan



Josu Bilbao



Wenjun Zeng



Abdulmoteleb El Saddik

The Internet of Things (IoT) has brought the vision of a more connected world into reality with an emerging volume of data and numerous services that are delivered via heterogeneous access networks. Correspondingly, cloud computing has emerged to provide enormous storage, computing facilities, and data sharing opportunities. However, the convergence of IoT and cloud can provide new opportunities for both technologies. It can open a new horizon of ubiquitous sensing, interconnection of devices, service sharing, and provisioning to support better communication and collaboration among people and things in a more distributed and dynamic manner. Thus, such convergence can enable the development of innovative applications in various emerging areas to improve all aspects of life.

While researchers have been making advances in the study of IoT and cloud convergence, many issues still remain to be addressed with regard to the impact of next generation mobile and wireless networking technologies. This Feature Topic (FT) collates articles from a wide range of perspectives that stem from different industrial and research communities aiming to advance the IoT-cloud convergence paradigm. Following a rigorous review process, from quite a good number of submissions, only eight articles have been selected for publication in this January 2017 issue of *IEEE Communications Magazine*.

Cloud computing, with its powerful calculation and storage ability, will be the best partner of IoT devices. However, private data are leaked while enjoying the convenience and entertainment brought by various IoT devices. 93 percent of

iOS apps and 89 percent of Android apps face the issue of privacy leakage. The article “Who Moved My Data? Privacy Protection in Smartphones” summarizes this kind of data over-collection phenomenon, and proposes a way out, which is to use cloud service to provide fine-grained permission authorization. The challenging security and privacy issues are of paramount importance for next-generation mobile technologies on IoT-cloud convergence. The article “Security and Privacy for Cloud-Based IoT: Challenges, Countermeasures, and Future Directions” addresses the issues by proposing a new framework of efficient privacy preserving data aggregation without exploiting public key homomorphic encryption.

The article “High-Efficiency Urban Traffic Management in Context-Aware Computing and 5G Communication” proposes a four-tier architecture assisted by fifth generation (5G) networks, mobile edge computing (MEC), software defined networking (SDN), and remote cloud service, which enables closer connection among traffic units. The convergence of new generation mobile communication technologies and cloud computing technologies brings vehicles from mutual sensing to mutual cooperation and control, alleviate urban traffic collision, and improve urban traffic efficiency.

The article “Beyond-5G Vision for IOLITE Community” explains IoT-cloud integration over a smart kitchen example in order to show how cloud-IoT integration is going to impact the daily lives of people. In this article, a sketch of a potential solution with needed features is provided, such as flexibility in control via SDN and self-organizing networks

(SONs), elastic infrastructure and resources via virtualization, near-zero latency, and high throughput via MEC and SON.

The article “Federated Internet of Things and Cloud Computing Pervasive Patient Health Monitoring System” presents a remote patient health status monitoring framework that is flexible, energy-efficient, and scalable, and provides efficient and high-quality service. The proposed framework utilizes the benefits of modern technologies such as IoT, mobile, and cloud computing for pervasive monitoring. The article “Wearable 2.0: Enabling Human-Cloud Integration in Next Generation Healthcare System” comprehensively investigates the disadvantages of the existing healthcare systems and analyzes the trend of wearable computing. In Wearable 2.0, the user’s physiological data are unconsciously and sustainably collected, and personalized healthcare services are enabled based on big data analytics on clouds.

Vehicular communication systems are indispensable components to share road conditions in a wireless manner. As the next-generation wireless technology, millimeter-wave (mmWave) is advanced in its multi-gigabit transmittability and beamforming technique. Based on these features, the article “Millimeter-Wave Wireless Communications for IoT-Cloud Supported Autonomous Vehicles: Overview, Design, and Challenges” proposes a novel design of vehicular mmWave system combining the advantages of IoT and cloud computing. The article “Smart Health Solution Integrating IoT and Cloud: A Case Study of Voice Pathology Monitoring” proposes a framework of health monitoring integrating the IoT and the cloud. The proposed monitoring framework can be extended to other types of health monitoring using the IoT and the cloud.

In closing, the Guest Editors would like to thank all the authors who significantly contributed to this FT, and the reviewers for their efforts in respecting deadlines and their constructive reviews for this FT. We are also grateful to the Editor-in-Chief, Osman Gebizlioglu, for his support and the *IEEE Communications Magazine* publication staff as well who collaborated with us on every step. We hope that researchers and practitioners in this field will find the articles of this FT constructive and interesting, and hope this FT will inspire further research and development ideas for the next-generation mobile technologies on IoT-cloud convergence.

## BIOGRAPHIES

M. SHAMIM HOSSAIN [SM’09] (mshossain@ksu.edu.sa) is an associate professor at King Saud University, Riyadh, Saudi Arabia. He was the recipient of the 2016 *ACM Transactions on Multimedia Computing, Communications and Applications* (TOMM) Best Paper Award. He is on the Editorial Boards of *IEEE Access*, *Computers and Electrical Engineering* (Elsevier), and the *International Journal of Multimedia Tools and Applications*. Currently, he is serving as a lead Guest Editor for *IEEE Communications Magazine*, *IEEE Transactions on Cloud Computing*, and *IEEE Access*. (Photo not available at time of printing.)

CHANGSHENG XU [M’97, SM’99, F’14] is a professor at the National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, and executive director of the China-Singapore Institute of Digital Media. He is an Associate Editor of *IEEE Transactions on Multimedia*, *ACM Transactions on Multimedia Computing, Communications and Applications*, and the *ACM/Springer Multimedia Systems Journal*. He is on the Editorial Boards of the *Journal of Multimedia* and *ACM Transactions on Multimedia Computing, Communications and Applications*.

YING LI [SM’07] received her Ph.D. in electrical engineering from the University of Southern California in 2003. She is currently a principal data scientist at IBM. Her research interests include computer vision, business analytics, and cognitive computing. She has been actively involved in organizing conferences/workshops, serving on the Editorial Boards of journals and on the Program Committees of various IEEE and ACM conferences.

AL-SAKIB KHAN PATHAN [SM’14] received his Ph.D. in computer engineering in 2009 from Kyung Hee University, South Korea, and his B.Sc. in computer science and information technology from the Islamic University of Technology, Bangladesh, in 2003. He is currently an associate professor in the CSE Department at Southeast University, Bangladesh. He has served as a Chair and Committee Member of numerous international conferences and in editorial roles for several renowned journals.

JOSU BILBAO [M’05] obtained his Ph.D degree in computer science from the University of Navarra. He is currently a senior researcher at the Embedded Systems Group and responsible for the IoT research team at IK4-IKERLAN. He is responsible for M2M and IoT architectures for industrial products, and leads the communications and security roadmap definition for several entities. He has been involved in the organizing committees of several international conferences, and workshops.

WENJUN ZENG [F’12] is a principal research manager overseeing the Internet Media Group and the Media Computing Group at Microsoft Research Asia. He is an Associate Editor-in-Chief of *IEEE Multimedia Magazine*, was an Associate Editor of *IEEE Transactions on Circuits & Systems for Video Technology*, *IEEE Transactions on Information Forensics & Security*, and *IEEE Transactions on Multimedia*, and is/was on the Steering Committees of *IEEE Transactions on Mobile Computing* (current) and *IEEE Transactions on Multimedia* (2009–2012).

ABDULMOTALEB EL SADDIK [F’09] is Distinguished University Professor and University Research Chair in the School of Electrical Engineering and Computer Science at the University of Ottawa. His research focus is on multimodal interaction with digital information in smart cities. He is a Senior Associate Editor of *ACM Transactions on Multimedia Computing, Communications and Applications*, an Associate Editor of *IEEE Transactions on Multimedia*, and a Guest Editor for several IEEE transactions and journals.

# Who Moved My Data? Privacy Protection in Smartphones

Wenyun Dai, Meikang Qiu, Longfei Qiu, Longbin Chen, and Ana Wu

We introduce a mobile-cloud framework to provide fine-grained permission authorization service for IoT devices, and show its performance by experimental results. Nevertheless, there are still lots of issues to be solved for the perfect IoT-cloud architecture. We list and analyze some obstacles and trends in the field of IoT-cloud.

## ABSTRACT

While enjoying the convenience brought by various kinds of IoT devices, our private data are leaked. Smartphones, as the typical pivot of IoT devices, use various kinds of apps, which collect our private data. In fact, private data leakage, as a potential hazard, is caused by the current design trend in industry, which is bigger and bigger. However, few people notice the side-effect, while we seem never bored by pursuing “smart” devices. In this article, we introduce the behaviors of data collection, and illustrate the motivations and reasons behind them. Thankfully, cloud computing with sufficient resources and exquisite services is a perfect way out. We introduce a mobile-cloud framework to provide fine-grained permission authorization service for IoT devices, and show its performance by experimental results. Nevertheless, there are still lots of issues to be solved for the perfect IoT-cloud architecture. We list and analyze some obstacles and trends in the field of IoT-cloud.

## INTRODUCTION

In the past decade, the Internet of Things (IoT) has subtly influenced our daily life. All kinds of physical objects, including groceries, vehicles, buildings, and others, are connected and combine into a network with the help of all kinds of electronics, such as sensors, mobile devices, and other wearable equipment. Everything is becoming smart and convenient for users. Meanwhile, due to the limited resources of IoT devices, cloud servers with sufficient resources can be used to take charge of data processing and storage [1]. We take Fig. 1 as an example to introduce three main aspects of the influence of IoT and the cloud. For home and lifestyle, IoT techniques can be used to implement electronic doors, automatic lighting and temperature control, and indoor video monitoring. We do not need to carry a lot of keys for our houses, offices, cabinets, and cars. In the field of transportation, every component in a vehicle is connected to the IoT to offer advanced driving assistance, navigation, and even automatic driving function. As a result, we do not need to check the abnormal conditions of oil or tire pressure of our cars, and we do not need to worry about getting lost when driving to some unfamiliar places. In the field of healthcare, vari-

ous kinds of wearable equipment monitor users’ physiological data, connected to the IoT, and offer diverse health management suggestions [2]. With the help of IoT techniques, we do not need to go to hospitals every month for physical examinations, which saves a huge amount of time and improves accuracy.

The most important characteristic and indispensable part of IoT is devices, which are connected to the IoT and communicate with each other to build an intelligent modern ecological system. These devices play extremely vital roles, especially smartphones. According to a new Pew Research Center report, “Technology Device Ownership: 2015,” 68 percent of adults in the United States own a smartphone. With the rapid development of IoT techniques, smartphones are not only communication equipment, but also health assistants, work secretaries, entertainment mates, and electronic IDs. Various kinds of smartphones are entering our daily lives with go-anywhere apps, which provide a wide array of enterprise, social, financial, and recreational services. The streamlining of marketing, installation, and updating creates low barriers for mobile app developers to bring their products to the market, and even lower barriers for users to obtain and use apps. To enjoy the vivid functions and services offered by apps, users have to permit the authority for accessing local data to apps [3]. However, these data may include our account numbers, email addresses, home addresses, photos, and other private information. It seems really convenient to store data in smartphones and further use them anywhere and anytime, but this kind of behavior brings about serious potential privacy hazard.

We define a term, data over-collection, to describe the most frequently occurring and most serious privacy leakage behavior in smartphones. Apps collect users’ data more than needed for the original function while within the permission scope, including tracking location, accessing photos, accessing address book, accessing calendar, tracking International Mobile Station Equipment Identity (IMEI) and Unique Device Identifier (UDID), and more. According to a report from Appthority [4], “App Reputation Report,” 93 percent of iOS apps exhibit at least one kind of data over-collection behavior, and 89 percent of Android apps have the same problem.

This work is partially supported by NSF CNS 1457506.

Digital Object Identifier:  
10.1109/MCOM.2017.1600349CM

Meikang Qiu, Wenyun Dai, Longbin Chen, and Ana Wu are with Pace University; Longfei Qiu is with the Nanjing Foreign Language School. Meikang Qiu is the corresponding author.

## DATA OVER-COLLECTION BEHAVIORS

Current mobile phone operating systems, such as iOS, Android, and Windows Phone, only provide coarse-grained permissions for regulating whether an app can access the data stored in smartphones. Meanwhile, few users actually notice and understand the permission agreement information shown during installation. Even knowing that an app may access their private information, few users choose to stop installing or to uninstall an app when it asks for permission authority. In fact, it is not users' responsibility to clearly know the permissions and cautiously manage the authority of apps. In this section, we choose some common data over-collection behaviors, analyze their inherent causes, and introduce the potential risks.

### LOCATION

Users' location data can be used in various kinds of apps, including navigation, photo organization, social networking service (SNS), restaurant recommendation, weather, and travel. Normally, users are warned once an app intends to obtain location information, but they usually grant permission in order to use the functions or service offered by the app. The privileged apps always keep obtaining users' information data to offer location-based function or service accurately and quickly. However, these apps over-collect users' location data. From the report of Appthority, 50 percent of the top iOS free apps and 24 percent of the top iOS paid apps track users' locations.

The iOS system offers a system service about location, named Frequent Locations, which is used to record the places users frequently visit. It is easy to disable this service, clear the record history, and stop it from running any time after initialization. However, users' location information is still collected by the operating system, and this information is just invisible and unavailable to users. Furthermore, this service meticulously records the amount of visits, the date, the time, and duration of staying. To offer such detailed information, this service must keep obtaining all location information in the form of geographic location and time, since then the most frequent records are sorted by frequency.

Much worse than iOS, from the report of Appthority, 82 percent of the top Android free apps and 49 percent of the top Android paid apps track user location. Due to the open developing and marketing environment of Android, it is extremely hard to restrict or prevent app developers to hide some data leakage codes into their Android apps and to put them into the market. W. Enck *et al.* [5] studied 1100 Android apps, and found that half of these apps exposed users' location information to third-party advertisement servers without requiring implicit or explicit user consent.

### PHOTOS

Compared to traditional cell phones, one remarkable function of current smartphones is taking photos. Smartphone users not only take photos for memories, but also for convenience. For example, taking photos of some slides instead of writing them down in notes is extremely convenient

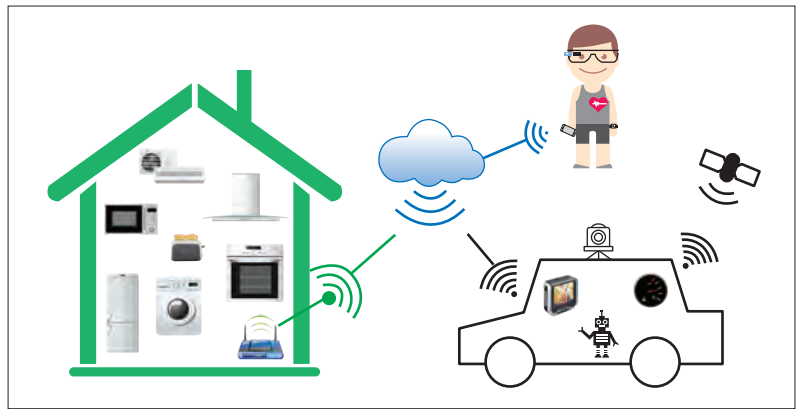


Figure 1. How IoT and the cloud influence our lives.

and time saving. Meanwhile, with the increasing popularity of SNS, smartphone users form the habit of posting photos to show what they are doing via SNS apps. At restaurants, we can always see this universal situation. The first thing to do after waiters or waitresses serve the dishes is to take a photo, but not enjoy them at once. Besides SNS apps, there are some other kinds of apps that obtain users' photos as well, such as cloud storage, wallpaper, customized albums, and photo decorating. As a result, it is very easy for these apps to get permission to access albums and cameras from users. Users seemed to not really care about the accessing permission of their photos until the celebrity photo leakage scandal happened in 2014. However, most responsibility was shirked to the cloud storage server. In fact, the origin culprit should be photo data over-collection behaviors of apps.

By analyzing users' general purposes, we can see that they use these apps to deal with just several or parts of photos, not all of them. For example, I just want to post one photo via Facebook, and only want to authorize permission to access this one photo to my Facebook app. However, current smartphone operating systems, including iOS and Android, just offer coarse-grained permission authorization. The coarse-grained permission authorization only allows two modes of data accessing, which are all and none, and the users' authorization are always one-time operations. Once a user authorizes the permission of an app to access one photo, this app will hold this permission to the whole album forever. iOS gives users a way of escape, allowing users to manually disable the permission of an app to the album in Setting/Privacy/Photos. However, Android users have no way to disable the permission of some app other than uninstalling it.

With current advanced techniques, most photos taken by smartphones are embedded with extra information about the location, time, device type, and more. As a result, the data over-collection behavior toward photos leaks not only users' photos, but also other private information. Users' photos reveal their daily lives. The exposure of photos not only infringes on users' rights such as profiles, but also may damage users' reputations. Even worse, some photos have fabulous value. Third party organizations can mine these photos for further commercially valuable information. This is the same as stealing assets from users. For

Apple hires about 100 employees to manually review iOS apps being published. However, they are only concerned with certain aspects, mainly focusing on the user interface (UI) and functions, such as crashes and bugs, broken links, advertisements, placeholder content, incomplete information, inaccurate descriptions, and repeated submission.

example, one product designer is working on an innovative product, and he/she takes a photo of his/her design draft for recording. It will be a direct asset loss if the photo of a product design draft is obtained by some third party organizations and sold to someone else in the same domain as him/her.

#### ADDRESS BOOK

An address book is a traditional function provided by mobile phones, and this function is improved by adding more relevant information about contacts at the platform of smartphones. For convenience, smartphone users create new contacts when they make new friends, update existing contacts with email addresses, extra phone numbers, addresses, face portraits, and remarks. These advanced functions of address books really offer lots of benefits to users. They do not need to remember contact information about their friends or cooperators, which not only saves great time and efforts, but can also guarantee that there are no errors.

The address book, usually a pre-loaded system app, provides a uniform interface for the apps running on its operating system to access users' address books. This uniform interface works as the bridge to connect apps and address books. Through this uniform interface, users can easily know which of their friends are using the same app and invite their friends to use this app. On the other hand, app developers are very willing to use this uniform interface, which can help to popularize their apps. As a result, apps can easily get permission to access address books from users.

Similar to the operation on users' photos, mobile operating systems only provide coarse-grained permission authorization. It seems more reasonable for apps to access the whole address book, because users may want to check the status of all their friends about using some app. However, once an app gets permission to access a user's address book, it can keep obtaining contact data until the user manually disables the permission. For example, after we grant permission to access our contact data to a Facebook app, it always sends us a notification to invite more friends who do not have a Facebook account. This behavior shows that the Facebook app keeps obtaining our address book data. Facebook apps are not alone; there are various kinds of SNS apps, game apps, and commercial apps that have the same behavior of contact data over-collection. From the report of Appthority, 26 percent of the top iOS free apps and 8 percent of the top iOS paid apps access users' address books, and 30 percent of the top Android free apps and 14 percent of the top Android paid apps access users' address books. In fact, it is unnecessary to allow apps to access users' contact data all the time, and normally "share with friends" is just a one-time operation. The permission authorization should be flexible to suit this kind of one-time operation.

#### IMEI/UDID

The IMEI and UDID make up the unique ID of one mobile phone, and cannot be deleted after manufacture. Similar to the idea of web cookies, the IMEI and UDID can be used to "remember" devices. From the report of Appthority, 88 per-

cent of the top Android free apps and 65 percent of the top Android paid apps access IMEI/UDID, and 57 percent of the top iOS free apps and 28 percent of the top iOS paid apps access IMEI/UDID. This information is innocent itself, but combined with other kinds of information, the hazard becomes a juggernaut. IMEI/UDID works as the primary key in a relational database. It is the identification of all kinds of data, and can be used to integrate these data for one specific smartphone. In other words, all over-collected data can be labeled in the form of smartphones, which makes data more valuable for mining.

As we all know, currently we do data mining research work based on anonymous data, even if they are recorded from real data. With the help of IMEI/UDID, users' behaviors can be correlated among multiple apps and matched to one unique device. Data over-collected by different apps can be integrated just via this ID, even though these data have nothing in common. Furthermore, IMEI/UDID can be used to build complete profiles with users' real data, including names, locations, accounts, and so on. For example, users' locations, names, and account data are collected by different apps and bought by one third-party company. This company can use the IMEI/UDID to create detailed profiles for in-depth views of users, which is undisputed privacy infringement.

After discussing the data over-collection behaviors, we briefly analyze the motivations behind these behaviors.

**Loose Development Limitations.** With the rapid popularity of smartphones, the number of mobile app developers keeps increasing. However, there are no established development limitations for app developers. In other words, app developers are permitted to implement any kinds of functions in their apps. Meanwhile, some app developers are not so familiar with app development, and they apply some open source libraries to achieve some functions of their apps. In the libraries, there are lots of code blocks implementing the obtaining of data functions. Without strict development limitations, it is hard to avoid the user using libraries with hidden data over-collection behaviors.

**Incomplete Censor Mechanisms.** Apple hires about 100 employees to manually review iOS apps being published. However, they are only concerned with certain aspects, mainly focusing on the user interface (UI) and functions, such as crashes and bugs, broken links, advertisements, placeholder content, incomplete information, inaccurate descriptions, and repeated submission. In fact, the reviewers from App Store only check basic UIs and functions. For example, we submitted an iOS app recording users' operation functions to App Store for review. The result sent back showed that the reviewer just checked the home screen and some basic functions of our app, and most of our detailed functions were not checked. Worse, for Android, until April 2015, Google Play did not have manual censors for Android apps, which indicates that all Android apps could be on the market. Furthermore, data over-collection behaviors are much more complicated to detect than malware, because they happen with users' permission, and it is almost impossible to determine the exact amount of data needed for functionality.



**Third-Party Companies.** Third-party companies or organizations are willing to buy users' data for commercial purposes. Third-party companies can be any kind of companies, even research organizations. As we know, the amount of customers is the most important parameter for the market share of one product. As a result, attracting potential customers to their products is the main reason for companies to obtain users' data from over-collection behaviors. Meanwhile, in this big data era, data are treasures. By analyzing users' behaviors, one company can accurately and quickly get the big picture of market trends.

From the view of smartphones, the main reason for data over-collection is the defects of current mobile operating systems, including coarse-grained permission authorization, one-time permission authorization, and no different levels of privacy. Coarse-grained permission only provides two kinds of permission authorization: none or all. Once one app gets permission to access some kind of data from a user, it can obtain all of the same kind of data. One-time permission authorization indicates that once users authorize permission to an app, it can keep this permission and access to data. Furthermore, the permission authorization operations only occur once, whether accepted or rejected. For example, once you accept the first request of a Facebook app to access your album, a Facebook app can access your album without your permission again forever until you manually deny this permission. In addition, there are no different levels of users' data based on how private they are. In other words, current mobile operating systems treat users' data equally without discrimination. This kind of strategy is convenient for management and operation, but fails to protect users' private information.

### CLOUD-BASED SOLUTION

It is impossible to force app developers not to share users' data with advertisement networks and other third party organizations, and it is unreasonable to expect that all smartphone users can understand permissions clearly and protect their privacy carefully. In fact, the data over-collection behaviors of apps are created by us. We have improved the mobile phone from traditional communication equipment to advanced smartphones with various kinds of apps. As Albert Einstein said, "The significant problems we face cannot be solved at the same level of thinking we were at when we created them." To solve the mobile data privacy problem, we have to change our pattern of thinking, which makes everything bigger and bigger. We have to eradicate it in advance, but not deal with it in the aftermath.

Meanwhile, current IoT devices have limited resources due to portability. These devices cannot undertake the due obligations of increasing requirements, including amount of storage, performance of calculation, availability, and other aspects. To solve the privacy issue of mobile data and to break through the resource limitations of IoT devices, a cloud-based solution is the best method. Thanks to cloud computing, we can offload a huge part, or even all, of the storage and calculation burdens to cloud servers at very low cost [6]. Furthermore, after offloading the mobile data to cloud servers, we can use cloud comput-

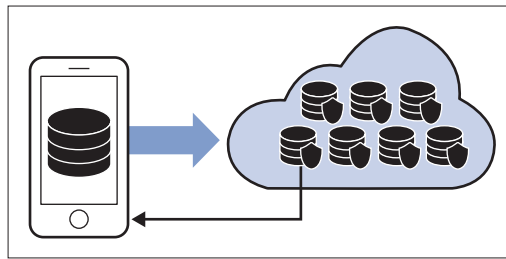


Figure 2. Mobile-cloud framework for data privacy protection.

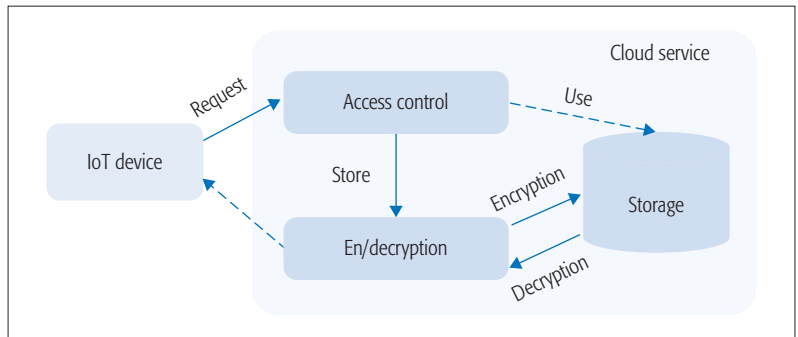


Figure 3. Working mechanism of the mobile-cloud framework.

ing techniques to provide fine-grained permission authorization. In addition, users are freed from complicated management of their data and permissions. Compared to users, cloud service providers are much more professional in assigning permissions to apps based on different privacy conditions.

Several months ago, we proposed a mobile-cloud framework for data privacy protection in smart cities [7]. In this framework, as shown in Fig. 2, we first separate mobile data into privacy levels from 1 to 3. The higher the value is, the more private the piece of data is. We assign privacy levels to users' data as follows. Location data: 3; photo data: 3 or 2; audio and video data: 1. Then we migrate location, photo, audio, and video data from smartphones to remote servers. Every time one app requests access permission for some kind of data, the access control service will determine whether to accept this request. Meanwhile, we design a grading system to evaluate our approach to data privacy. In this grading system, we use a formula to calculate the risk of apps using the total amount of data, the amount of over-collected data, and the privacy level of the data.

The detailed working mechanism of this framework is shown in Fig. 3. IoT devices send the request to access data to the cloud. The access control service receives the requests and validates the authorization. The en/decryption service encrypts and decrypts data before data is stored in storage and sent back to IoT devices.

If we allow the access control service to authorize apps to access one specific type of data with the same privacy level, our approach can reduce over 2.5 times the privacy grade than original coarse-grained permission authorization, as shown in Fig. 4a. If we set the access control service to only allow apps to access the specific pieces of data they need, our approach can reduce over 35 times the privacy grade than the original

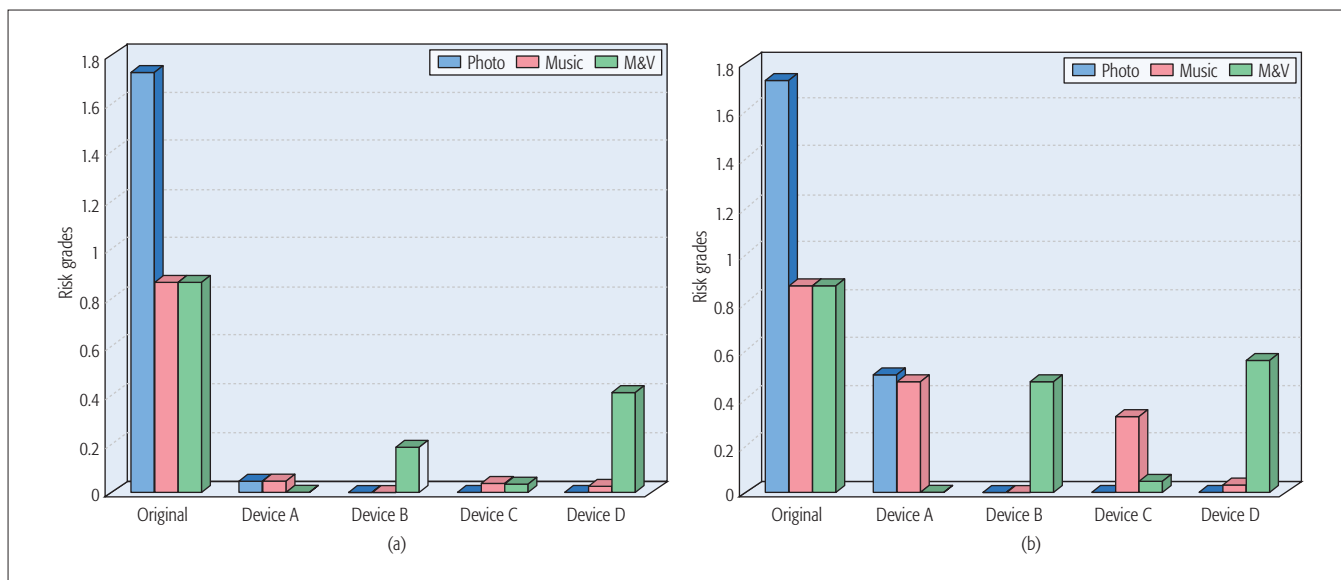


Figure 4. Risk grades of four devices in two experimental environments: a) normal default privacy level; b) extremely high default privacy level.

coarse-grained permission authorization strategy, as shown in Fig. 4b.

### OPEN RESEARCH ISSUES

Using cloud techniques to solve the mobile data privacy issue has not yet been fully studied. Some issues are even impossible to solve in current conditions. More research and effort are needed to completely achieve mobile data privacy protection. We list and analyze some thought-provoking and important research issues as follows.

#### PRIVACY CLASSIFICATION

To provide fine-grained permission authorization, the first thing is to classify users' data based on their privacy. After separating data into different levels, we can use cloud storage service to store users' data in different servers or virtual machines. Without privacy classification, we have to separate all data into pieces and further check their detailed privacy parameters to determine the access control decision, which is really time consuming and tedious. As a result, privacy classification is not only a prerequisite of cloud-based fine-grained permission authorization, but also can make the access control strategy smooth and efficient. Once one app requests access permission to high-privacy-level data, it can be automatically denied.

However, it is not easy to achieve the privacy classification of users' data. In the first step, we can separate data by types. For example, account information is much more private than photos; thus, they can be classified to different levels. Privacy classification by data types can be implemented without considering users' specific situations, because the types always have common characteristics and uses [8]. Nevertheless, it is extremely hard to classify one kind of data into different privacy levels, because these data are user determined. For example, there are two photos in one user's album. One is a photo of a normal file that may be downloaded online, while the other one is a print-screen of a confidential

document. It is obvious that these two photos should be assigned different privacy levels, but it is complicated for computers to tell the difference between these two photos. Machine learning and pattern recognition techniques seem to be suitable to solve this issue, but these methods have three deficiencies. The first one is that they must access the data in advance, which is also a hidden danger to users' privacy. The second one is that access control operations are dynamic, since users may keep creating, deleting, and updating their data. However, current machine learning and pattern recognition methods are not powerful enough to support such flexibility [9]. The third one is that the rules of classification are closely related to users. Different users may have different attitudes toward the same piece of data. Thus, using these two kinds of methods to achieve classification still needs extra user operations, which is not very practical.

#### IOT NETWORKING

Without networking, the Internet of Things (IoT) is useless to some extent. IoT devices, including sensors, smartphones, and other gadgets, need cloud servers to store and analyze data due to their limited resources. However, current networking techniques are far from ideal. Besides traditional problems, such as bandwidth and stability, the security issue is the main enemy in the privacy protection field [10]. Hyper-connectivity threatens users' privacy in increasing ways. Hackers are given more opportunities and targets to attack, since almost everything is connected to the IoT. Using cloud resources is a wise method to offer fine-grained permission authorization by cutting off the ways of data over-collection behaviors of apps. However, we need to ensure the security of IoT networks to avoid attending to one while neglecting the other. Meanwhile, this issue is closely related to the advanced networking techniques, such as software-defined networking [11], fifth generation, and Internet 2, which make IoT networking research field complicated but profound [11].

Currently, there is some research focusing on operations offloading from mobile devices to cloud servers. Restricted by the communication cost, it is impractical to offload everything [12]. Data offloading or migration seems much easier than operations, but in fact, it is much more complex. In mobile devices there are lots of internal or embedded data that are hidden and invisible unless rooted [13]. For example, IMEI/UDID is one kind of embedded data of a device, which is similar to the medium access control address of the computer network card.

Why do we need to offload mobile data to clouds? There are two reasons. The first one is that mobile or IoT devices will be totally released from the burdens of processing miscellaneous tasks that should not be in their charge. The second one is that this kind of framework is more suitable for both software developers and users [14]. Developers do not need to develop various versions of their products for different platforms, and they do not need to upload every update to the Internet. Users never need to update their apps or software, and they can get rid of tedious data management. With the rapid development of service computing technology, the traditional client-server model has been replaced by the browser-server model, and this tide certainly will happen in the field of IoT. Mobile apps or software will disappear and be replaced by various kinds of services. The only thing needing to be installed in a device is a browser. All data are stored and processed in cloud servers, which can provide fine-grained access control and high-quality service.

## CONCLUSION

IoT technologies bring lots of convenience to our daily life. The smartphone is the pivot, and can be used to control various IoT devices. However, data over-collection behaviors are ubiquitous, due to the deficiencies of current mobile operating systems. They only provide coarse-grained permission authorizations and general privacy management. Cloud computing with sufficient resources and fine-grained access control service can be used to solve the data privacy issue. However, there are many technical challenges to implementing the IoT-cloud framework practically. After introducing a basic mobile-cloud framework we designed before, we analyze some thought-provoking research issues to protect users' mobile data.

## REFERENCES

- [1] W. Dai, H. Chen, and W. Wang, "RaHeC: A Mechanism of Resource Management for Heterogeneous Clouds," *Proc. IEEE 17th Int'l. Conf. High Performance Computing and Commun.*, 2015, pp. 40–45.

- [2] M. S. Hossain and G. Muhammad, "Cloud-Assisted Industrial Internet of Things (IIoT)-Enabled Framework for Health Monitoring," *Computer Networks*, vol. 101, 2016, pp. 192–202.
- [3] Y. Li *et al.*, "Intercrossed Access Control for Secure Financial Services on Multimedia Big Data in Cloud Systems," *ACM Trans. Multimedia Computing Commun. and Applications*, vol. 12, no. 4s, p. 67, 2016.
- [4] Appthority, App reputation report, 2014; <https://www.appthority.com/app-reputationreport/report/AppReputationReportSummer14.pdf>
- [5] W. Enck *et al.*, "A Study of Android Application Security," *Proc. 20th USENIX Conf. Security*, 2011.
- [6] W. Dai *et al.*, "RMORM: A Framework of Multi-Objective Optimization Resource Management in Clouds," *Proc. IEEE 9th World Congress on Services*, 2013, pp. 488–94.
- [7] Y. Li *et al.*, "Privacy Protection for Preventing Data Over-Collection in Smart City," *IEEE Trans. Computers*, vol. 65, no. 5, pp. 1339–1350, 2016.
- [8] G. Wu *et al.*, "A Decentralized Approach for Mining Event Correlations in Distributed System Monitoring," *J. Parallel and Distrib. Computing*, vol. 73, no. 3, 2013, pp. 330–40.
- [9] M. Qiu *et al.*, "Phase-Change Memory Optimization for Green Cloud with Genetic Algorithm," *IEEE Trans. Computers*, vol. 64, no. 12, 2015, pp. 3528–40.
- [10] M. Qiu *et al.*, "Informer Homed Routing Fault Tolerance Mechanism for Wireless Sensor Networks," *J. Sys. Architecture*, vol. 59, no. 4, 2013, pp. 260–70.
- [11] T. Chen *et al.*, "Software Defined Mobile Networks: Concept, Survey, and Research Directions," *IEEE Commun. Mag.*, vol. 53, no. 11, Nov. 2015, pp. 126–33.
- [12] W. Li *et al.*, "Mechanisms and Challenges on Mobility-Augmented Service Provisioning for Mobile Cloud Computing," *IEEE Commun. Mag.*, vol. 53, no. 3, Mar. 2015, pp. 89–97.
- [13] K. Gai *et al.*, "Dynamic Energy-Aware Cloudlet-Based Mobile Cloud Computing Model for Green Computing," *J. Network and Computer Applications*, vol. 59, 2016, pp. 46–54.
- [14] K. Kumar and Y.-H. Lu, "Cloud Computing for Mobile Users: Can Offloading Computation Save Energy?" *Computer*, no. 4, 2010, pp. 51–56.

## BIOGRAPHIES

WENYUN DAI is a Ph.D. student of computer science at Pace University. He received his B.E. and M.E. degrees from Xiamen University and Shanghai Jiao Tong University, respectively. His research interests include cloud computing, mobile computing, and file systems.

MEIKANG QIU received his B.E. and M.E. degrees from Shanghai Jiao Tong University, and his Ph.D. degree in computer science from the University of Texas at Dallas. Currently, he is an adjunct professor at Columbia University and an associate professor of computer science at Pace University. He is an ACM Senior Member. He has published 330 peer-reviewed journal and conference papers.

LONGFEI QIU is with Nanjing Foreign Language School, China. His research interests include cyber security, big data analytics, mobile app design, and cloud computing. He has won the First Prize at the National Olympics in Informatics 2014 of China, and the Best Student Paper Award of IEEE SmartCloud 2016 in New York.

LONGBIN CHEN is currently a Ph.D. student of computer science at Pace University. He received his B.E. degree from Xiamen University. His research interests include mobile and cloud computing.

ANA WU is a mobile developer at Obie HR LLC. She received her B.E. and M.A. degrees from Xiamen University and Peking University, respectively. Her research interests include mobile computing and distributed computing.

Cloud computing with sufficient resources and fine-grained access control service can be used to solve the data privacy issue. However, there are many technical challenges to implement the IoT-cloud framework practically.

# Security and Privacy for Cloud-Based IoT: Challenges, Countermeasures, and Future Directions

Jun Zhou, Zhenfu Cao, Xiaolei Dong, and Athanasios V. Vasilakos

The authors introduce the architecture and unique security and privacy requirements for the next generation mobile technologies on cloud-based IoT, identify the inappropriateness of most existing work, and address the challenging issues of secure packet forwarding and efficient privacy preserving authentication by proposing new efficient privacy preserving data aggregation without public key homomorphic encryption.

## ABSTRACT

The Internet of Things is increasingly becoming a ubiquitous computing service, requiring huge volumes of data storage and processing. Unfortunately, due to the unique characteristics of resource constraints, self-organization, and short-range communication in IoT, it always resorts to the cloud for outsourced storage and computation, which has brought about a series of new challenging security and privacy threats. In this article, we introduce the architecture and unique security and privacy requirements for the next generation mobile technologies on cloud-based IoT, identify the inappropriateness of most existing work, and address the challenging issues of secure packet forwarding and efficient privacy preserving authentication by proposing new efficient privacy preserving data aggregation without public key homomorphic encryption. Finally, several interesting open problems are suggested with promising ideas to trigger more research efforts in this emerging area.

## INTRODUCTION

The Internet of Things (IoT) is composed of physical objects embedded with electronics, software, and sensors, which allows objects to be sensed and controlled remotely across the existing network infrastructure, facilitates direct integration between the physical world and computer communication networks, and significantly contributes to enhanced efficiency, accuracy, and economic benefits [1, 2]. Therefore, IoT has been widely applied in various applications such as environment monitoring, energy management, medical healthcare systems, building automation, and transportation. Unfortunately, due to the resource constraints of IoT devices, they always delegate highly complex computation to the energy abundant cloud for considerably enhanced efficiency. However, both the inputs, outputs, and function of the underlying computation may be closely related to the privacy of IoT users, which cannot be exposed to collusion between malicious cloud servers and malicious IoT users. Therefore, how to design new efficient privacy-preserving solutions for next generation mobile technologies with IoT-cloud convergence is a crucial issue of great concern.

## MOTIVATION

According to the functionality, cloud-based IoT can be categorized into static and mobile, the latter of which is more challenging in protocol design. Therefore, in this article, we mainly focus on the security and privacy issues and corresponding countermeasures in mobile cloud-based IoT. The fast development of next generation mobile technologies such as fifth generation (5G) on IoT-cloud convergence has cast light on types of security and privacy issues unaddressed for years.

The characteristics of resource-constrained short-range communication and mobility result in the unique features of packet forwarding in cloud-based IoT. Specifically, it lacks the end-to-end continuous connectivity between mobile IoT users (IoT users, nodes, and devices are used interchangeably in the rest of this article), and message delivery needs to be fulfilled by cooperation among a social group of IoT users directed toward the destination. However, selfish nodes would not be willing to participate in this energy-consuming task due to their limited resources unless they can obtain maximized gain from it. It is obvious that the more packets one IoT user transmits, the more benefit it will obtain. However, it would also be more likely to be selected as the compromise target by the adversary from the side channel attack through analyzing the packet flow around each IoT node and lose all earned utility. The reason is that if such an IoT node were compromised, the adversary would obtain more packets from one single attack, which we name target-oriented compromise. Therefore, it is required to design a secure incentive mechanism to stimulate collaboration for packet forwarding in cloud-based IoT. In addition, malicious IoT users could collude to illegally increase their utility by detouring the packet transmission routing. Until now, how to prevent a layer-adding attack in collaborative packet forwarding in IoT is still an open problem requiring convincing solutions.

On the other hand, cloud-based IoT has also provided a convincing platform to guarantee distributed location-based service (LBS) by periodically collecting and broadcasting certain kinds of passing service information such as all the restaurants satisfying the user's query conditions in the neighborhood and the traffic conditions for spe-

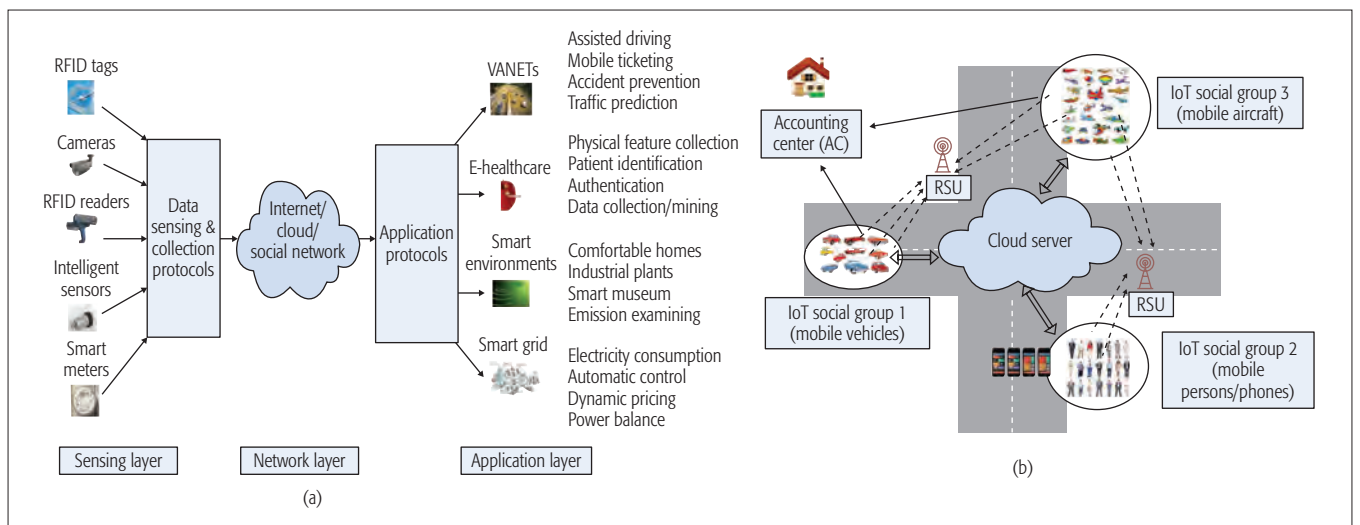


Figure 1. Network architecture of cloud-based IoT.

cific road sections during required time periods. Unfortunately, it actually faces various sophisticated attacks such as eavesdropping, modification, and repudiation. A malicious IoT user intends to forge its identity, manipulate the transmitted messages, and even to escape from crimes due to the lack of efficient tracking mechanisms. This false information would lead to other users' inconvenience and even disasters, which should be prevented from dissemination by designing effective authentication mechanisms in IoTs. Moreover, it is likely that mobile IoT users geographically in the neighborhood of each other would collect the information about the same event to generate considerably redundant packets. On the other hand, IoT users' conditional identity privacy and location privacy closely related to their private living habits should be well protected. Last but not least, it is reported that although appropriately powered, each resource-constrained IoT device is generally required to verify about 1000–5000 messages per second, the high computational complexity of which is still intolerable [5]. Therefore, it is required to propose privacy-preserving lightweight authentication with message content filtering to avoid duplicate packet transmission and reduce both the computational and communication cost.

Although several works have studied the security issues in IoT [3–5, 7, 9, 13] w.r.t. secure packet forwarding and lightweight message authentication, the challenging problems of the target-oriented compromise attack, the layer-adding collusion attack, and the privacy preserving message content filtering from generation still cannot be well addressed. Moreover, most of the existing work [3–5, 7, 9, 11, 13] considered the IoT and cloud computing system as independent entities. The researchers in IoT rarely took the cloud as an underlying primitive to execute the outsourced storage or outsourced computation for resource-constrained IoT users; while the studies on secure outsourced computation widely exploited public key homomorphic encryption, which is so computationally intensive that it cannot adapt well to the efficiency requirement for IoT users. In this article, we discuss the unique security and privacy challenges brought

by the new architecture of IoT–cloud convergence and propose new convincing solutions to the above-mentioned challenging issues in cloud-based IoT.

### CONTRIBUTIONS

In this article, we first give an overview and the network architecture of cloud-based IoT. Then we identify the unique security and privacy requirements in cloud-based IoT, propose a new method of efficient privacy preserving data aggregation without exploiting public key homomorphic encryption, and further exploit it to address secure packet forwarding by designing outsourced aggregated transmission evidence to resist layer-adding attack and efficient authentication by devising outsourced privacy-preserving message filtering in cloud-based IoT. Finally, we suggest some future research directions with promising ideas.

## NETWORK ARCHITECTURE OF CLOUD-BASED IoT

In this section, the network architecture of cloud-based IoT is presented. It is assumed that mobile IoT users are generally categorized into social groups since the ones at specific times and locations are always moving in the same pattern such as the direction and velocity [2,10]. IoT devices allow both mutual communication with each other and with the cloud. Therefore, whenever two IoT users are moving into transmission range of each other, they can exchange packet bundles. In addition, IoT users in each group passing specific locations would generate duplicate/redundant bundles reporting similar events at an overwhelmingly high probability. Finally, the resource-constrained IoT devices such as sensors, RFID tags, cameras, and smart meters would outsource the computations of high complexity to the cloud for efficiency optimization [12]. Figure 1a demonstrates the sensing layer, network layer, and application layer in cloud-based IoT, while Fig. 1b shows the network model of mobile cloud-based IoT with the following unique characteristics.

**Resource Constraints:** IoT devices are always

Items	Internet of Things	Traditional networks
Node energy	Constrained	Abundant
Node mobility	High mobility	Static
Architecture	Self-organized	Hierarchical
Communication range	Short	Long
Routing	Intermittent and dynamically constituted	Continuous end-to-end connection
Packet delivery mode	Cooperative, DTN type, and need incentive mechanism to stimulate	Guaranteed delivery

**Table 1.** Characteristic comparison between cloud-based IoT and traditional networks.

resource-constrained and comply with the store-carry-and-forward method of packet forwarding only when their storage is available. Computationally intensive tasks are intolerable by IoT nodes and must be outsourced to the cloud, both the storage and computational resources of which are assumed to be abundant. Therefore, the resource-constrained property requires lightweight protocol design for efficiency and practicability, especially on the IoT users' ends.

**Mobility:** Moving IoT users (i.e., vehicles and mobile electricity consumers) are dynamically categorized into multiple social groups according to their directions, velocities, and accelerations, and assumed to be uniformly distributed. All IoT nodes in each group are in communication range of each other, broadcast their collected content bundles on demand, and share a dynamically updated group key negotiated by all of them. The group leader is located at the group center, invulnerable to certain threats, and periodically updated due to dynamic group formulation.

**Self-Organization:** Mobile IoT users frequently collect and broadcast packet bundles within communication range of each other. The cloud intervenes only when computations of high complexity need to be delegated from resource-constrained IoT devices, but does not frequently participate in the distributed content bundle generation and authentication.

**Short-Range Communication:** Due to both the mobility and short-range communication, there is no guaranteed connection (routing) between the source and destination in mobile cloud-based IoT. All IoT users constitute a delay-tolerant network (DTN). Packet transmission is fulfilled through cooperation among IoT users, and the accounting center (AC) is responsible for charging and rewarding. Table 1 demonstrates the characteristic comparison between cloud-based IoT and traditional computer networks.

## SECURITY AND PRIVACY REQUIREMENTS FOR CLOUD-BASED IOT

We mainly focus on the security threats for cloud-based IoT, especially in the aspects of secure packet forwarding with outsourced aggregated transmission evidence generation and efficient privacy-preserving authentication with outsourced message filtering. Besides the traditional data confidentiality and unforgeability, the unique security

and privacy requirements in cloud-based IoT are presented:

**Identity Privacy:** Conditional identity privacy refers to the fact that the mobile IoT user's real identity should be well protected from the public; on the other hand, when some dispute occurs in emergency cases, it can also be effectively traced by the authority. The technique of pseudonyms has been widely adopted to achieve this target, but the periodically updated pseudonyms and certificates lead to intolerable computational cost for resource-constrained IoT nodes. More seriously, it cannot resist the physically dynamic tracing attack we identified for location privacy.

**Location Privacy:** Location privacy seems especially critical in IoTs, since the frequently exposed location privacy would disclose the living habit of the IoT user. The widely adopted technique is to hide its location through pseudonyms. However, since the location information is not directly protected, it cannot resist the physically dynamic tracing attack. Specifically, a set of malicious IoT users in collusion can be dispatched to the positions where the target IoT user occasionally visited, to physically record sets of real identities of passing nodes during specific time periods by observation or traffic monitoring video, and further identify the target IoT user's real identity. If the adversary knows that the target node with pseudonym  $pi_d$  occasionally visits  $n$  locations  $Loc_1, Loc_2, \dots, Loc_n$ ,  $n$  sets of nodes' real identities passing by these  $n$  locations  $Veh_1, Veh_2, \dots, Veh_n$  can be observed. The intersection would definitely reveal the target node's real identity and its private activities in other regions.

**Node Compromise Attack:** Node compromise attack means the adversary extracts from the resource-constrained IoT devices all the private information including the secret key used to encrypt the packets, the private key to generate signatures, and so on, and then reprograms or replaces the IoT devices with malicious ones under the control of the adversary. The target-oriented compromise attack means an adversary with global monitoring ability would select the IoT node holding more packets as the compromise target by watching the traffic flow around all nodes in IoT. Therefore, from one single compromise, it is likely that the adversary obtains more packets for recovering the original message or impeding its successful delivery by interruption.

**Layer Removing/Adding Attack:** The layer removing attack occurs when a group of selfish IoT users remove all the forwarding layers between them to maximize their rewarded credits by reducing the number of intermediate transmitters sharing the reward. On the contrary, the layer adding attack means colluding IoT users maliciously detour the packet forwarding path between them for increased credits by increasing the total obtainable utility.

**Forward and Backward Security:** Due to the mobility and dynamic social group formulation in IoT, it is necessary to achieve forward and backward security. The former means that newly joined IoT users can only decipher the encrypted messages received after but not before they join the cluster; while the latter means that revoked IoT users can only decipher the encrypted messages before but not after leaving.

**Semi-Trusted and/or Malicious Cloud Security:** For the convergence of the cloud with IoT, the security and privacy requirements for the cloud should be especially considered. The semi-trusted model means that the cloud would faithfully comply with the protocol specification, but try its best to extract secret information from the interactions with IoT users; while the malicious model means that the adversary can arbitrarily destroy the protocol execution. Therefore, for outsourced computation, the following three security targets should be achieved:

- Input privacy: The data owner's individual inputs should be well protected even from collusion between the cloud and authorized data receivers.
- Output privacy: The computation result should only be successfully deciphered by authorized data receivers.
- Function privacy: The underlying function must be well protected from even the collusion of the cloud and malicious IoT users.

Table 2 demonstrates the main security and privacy threats in cloud-based IoT with the corresponding countermeasures.

## SECURE PACKET FORWARDING IN CLOUD-BASED IOT

We mainly focus on secure packet forwarding in cloud-based IoT DTNs, especially the techniques to address the kinds of attacks w.r.t. the bundle delivery, such as fairness for obtaining interest from transmitting packets, free riding attack, layer removing/adding attack, and node compromise attack identified earlier.

A secure credit-based incentive scheme, SMART, was presented to stimulate packet forwarding collaboration among DTN nodes [9]. Different from SMART, Lu *et al.* devised a secure and practical incentive protocol Pi addressing the fairness of charging and rewarding for packet delivery cooperation by adding some incentive on each bundle. Unfortunately, the existing work [9, 13] merely considered the outsider threats, leaving the target-oriented node compromise attack untouched. Consideration of the incentive schemes for multiple-copy algorithms is required to resist a single node compromise leading to the original message failing to be recovered. More significantly, the problem of layer adding collusion attack cannot be well addressed by either solution [9, 13].

To tackle the issue of compromise, by designing a modified model of population dynamics, a new threshold credit-based incentive mechanism, TCBI [10], is proposed in cloud-based IoT DTNs to efficiently resist node compromise attacks, stimulate packet transmission cooperation, optimize IoT users' utility, and achieve fairness among IoT users.

To resist layer adding attack, an outsourced aggregated transmission evidence generation algorithm is proposed by devising a new technique of secure outsourced data aggregation without public key homomorphic encryption. The sketch of the construction can be described as follows. First, each roadside unit (RSU)  $L$  encodes a randomness  $R$  using the one-way trapdoor permutation  $f$ , which is further adopted as the sym-

Security threats	Countermeasure
Identity privacy	Pseudonym [4, 5, 9], group signature [5], connection anonymization [7, 13]
Location privacy	Pseudonym [4, 5, 9], one-way trapdoor permutation [6, 10]
Node compromise attack	Secret sharing [8, 10, 14], game theory [7], population dynamic model [10]
Layer removing/adding attack	Packet transmitting witness [9, 10, 13], aggregated transmission evidence [10]
Forward and backward security	Cryptographic one-way hash chain [4, 5]
Semi-trusted/malicious cloud security	(Fully) homomorphic encryption [11], zero knowledge proof [15]

**Table 2.** A taxonomy of main security threats in cloud-based IoT.

metric key to encrypt the individual velocities of all mobile IoT users passing by using an appropriately selected symmetric homomorphic mapping (SHM) instead of traditional public key homomorphic encryption. Then the cloud computes the aggregated velocity in the ciphertext domain and transmits the encrypted result to an AC, which can successfully decipher it as the aggregated packet transmission evidence by using secret key  $sk_f$ . Each mobile IoT node's velocity privacy is well protected by a blinding factor  $r_l$  added to SHM, and the layer adding attack is well resisted by the packet transmission evidence. It is noted that, different from the existing work exploiting public key homomorphic encryption to realize secure outsourced data aggregation, any public key encryption can be utilized only once in TCBI to achieve privacy preservation for  $n$  inputs on the resource-constrained IoT users' end, where both the computational and communication costs are dramatically saved. For simulation, ElGamal encryption is adopted to implement the one-way trapdoor permutation in the proposed TCBI. Figures 2a and 2b demonstrate the advantages of the proposed TCBI in computational cost and communication cost over the existing work (i.e., SMART [9]) exploiting Paillier's homomorphic encryption [11] as the primitive.

## PRIVACY-PRESERVING AUTHENTICATION IN CLOUD-BASED IOT

For privacy-preserving authentication, we mainly focus on two aspects, the identity/location privacy protection and lightweight authentication solutions in cloud-based IoT.

Conditional identity privacy is traditionally achieved by a group signature [5]; however, the public key infrastructure (PKI) leads to the verification algorithm being inefficient and intolerable for the resource-constrained IoT devices due to the additional verification cost for the sender's public key certificate. To improve the efficiency, the pseudonym technique was proposed: Each IoT user is initialized with an anonymous public and secret key pair  $(PK_i, SK_i)$  in the registration phase, where the associated anonymous certificate is  $Cert_i$  w.r.t. its pseudonym  $psm_i$ . The registration authority privately keeps a tuple composed of the IoT user's real identity and its pseudonym, and reveals this relationship when some disrup-

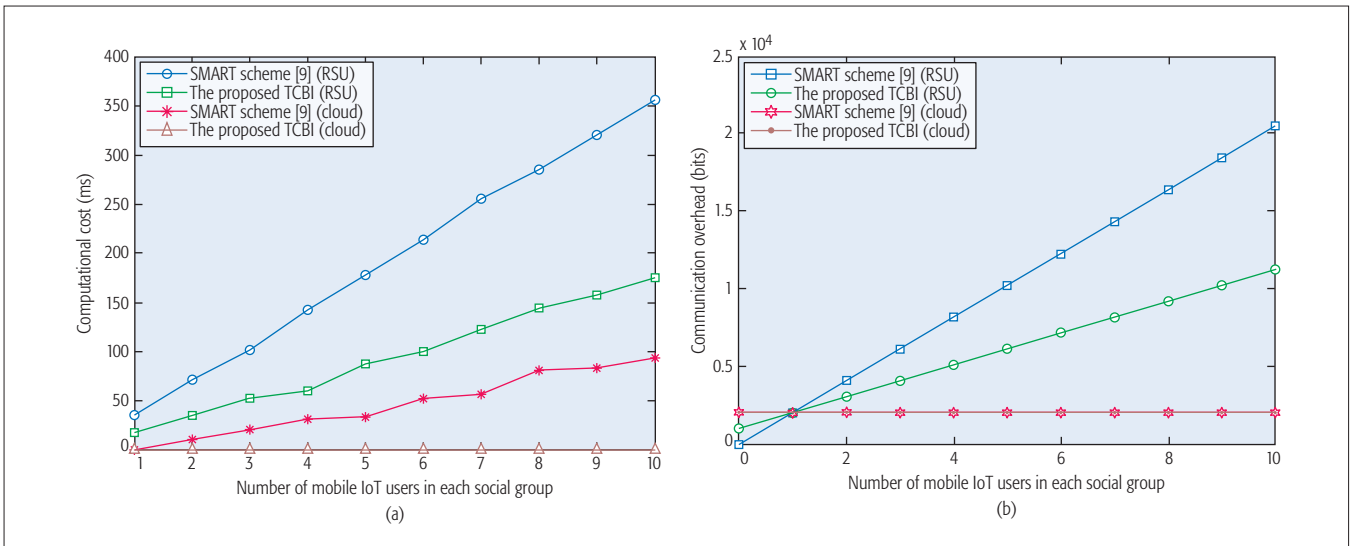


Figure 2. Efficiency comparison between SMART [9] and TCBI: a) computational cost; b) communication cost.

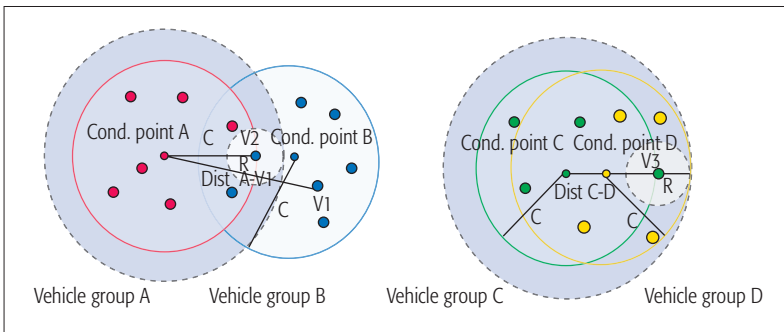


Figure 3. LBS content filtering mechanism with dynamic social group formulation in cloud-based IoT.

tion occurs. Each pair of keys has a short lifetime, which can be updated periodically to protect the IoT user's real identity and driving route from exposure. However, the frequent pseudonym updating would lead to intolerably high complexity on the resource-constrained IoT user's end.

X. Lin *et al.* proposed a timed efficient and secure communication (TSVC) scheme with privacy preservation in vehicular IoT [4]. By utilizing the techniques of hash chain and message authentication code, it aimed to minimize both the signature generation and verification overhead on the vehicle's side without compromising the underlying security and privacy requirements. However, the release delay of the private authentication key led to the construction only being appropriate for regular traffic events predefined in each time period. To overcome these shortcomings, Lin *et al.* proposed an efficient cooperative message authentication, also in vehicular IoT [5]. It minimized redundant authentication efforts from the receiver's aspect in that each message is verified by a single vehicle user, which reports afterward the verification result in its neighborhood. Unfortunately, the duplicate/redundant messages collected by vehicles passing the same road sections have not been reduced and still occupy a great deal of redundant bandwidth. More significantly, the intervention of an online trusted authority (TA) for token generation incurred considerable overhead.

Sen suggested privacy preserving authentication to verify the authenticity of the messages disseminated by IoT users by exploiting the technique of secure multiparty computation [7]. Roman *et al.* proposed a key management system for sensor networks in the context of IoT [8] to achieve both the forward and backward secrecy while IoT users join and/or are revoked from their current communication group. Recently, an efficient privacy-preserving relay filtering scheme, PReFilter, was proposed for DTNs in vehicular IoT communications [13]. It avoided junk packet delivery through setting and distributing an interest policy by message receivers for their friends, but still did not delete the redundant packets from the source. More seriously, all the constructions presented above cannot resist the physically dynamic tracking attack we identified in the multiple-pseudonym technique.

To address the challenging security issue, an efficient privacy-preserving authentication scheme SAVE for location-based service (LBS) in cloud-based IoT is proposed. Different from the existing work [5] which saved the verification cost from the receiver's view, an efficient privacy-preserving LBS bundle filtering mechanism with dynamic social group formulation is novelly designed from the sender's aspect to simultaneously prevent duplicate LBS contents from aggregation.

Let  $K$ ,  $C$ ,  $R$ , and  $Dist_{x,y}$  be the number of independent IoT social groups, the maximum communication range between mobile IoT users, the IoT user's LBS content sensing range, and the distance between  $x$  and  $y$ , respectively, where  $x$ ,  $y$  refer to either the IoT user's location or the condensing point (group center) position. Figure 3 shows the dynamic social group formulation in cloud-based IoT. There are four social groups, A, B, C, and D, denoted by red, blue, green, and yellow circles, respectively, with their own IoT users and condensing points at the group centers. From the left of Fig. 3, it is observed that IoT user  $V_2$  belonging to social group B is located at the edge of social group A with  $Dist_{V_2,CP_A} = C$ . Additionally, the LBS content sensing domain of  $V_2$  represented by the dashed circle with center



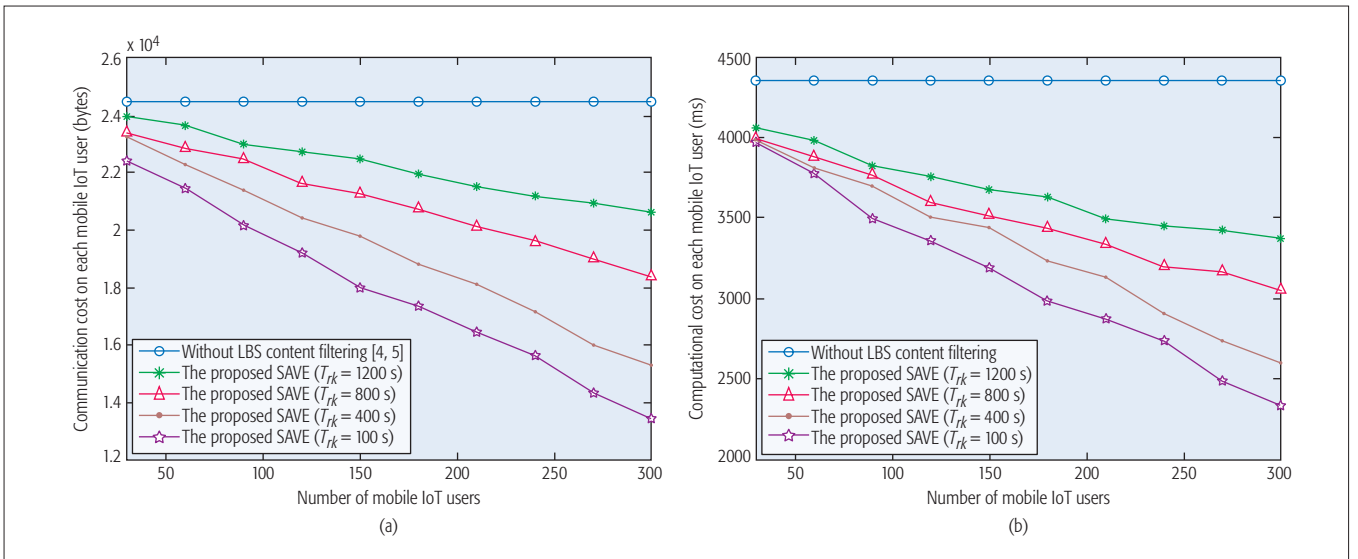


Figure 4. Efficiency of LBS content filtering mechanism in SAVE: a) communication cost; b) computational cost.

$V_2$  and radius  $R$  is just included in the LBS content sensing domain of IoT users belonging to group  $A$  represented by the dashed circle with center  $A$  and radius  $C + R$ . In other words,  $Dist_{V_2, CPA} \leq C$  means the LBS content collected by  $V_2$  is invalid to social group  $A$ , since it overlaps the LBS content collected by mobile IoT users in group  $A$ , becomes redundant, and should be efficiently filtered at the sender  $V_2$ 's aspect and prevented from further aggregation and dissemination in social group  $A$ .

From the right of Fig. 3, we focus on all the IoT users belonging to group  $C$  and located at the edge of group  $C$  (i.e., IoT user  $V_3$  in Fig. 3). The LBS content sensing domain of IoT users in group  $C$  represented by the dashed circle with center  $C$  and radius  $C + R$  just includes the domain of social group  $D$  when  $Dist_{CP_C, CP_D} = C - (C - R) = R$ . Therefore,  $Dist_{CP_C, CP_D} \leq R$  means social groups  $C$  and  $D$  are required to be merged into a new one, since the LBS contents collected by them are extremely redundant, and all the LBS content collected by group  $D$  should be prevented from aggregation in group  $C$  to save both computational and communication costs on resource-constrained IoT users.

On the other hand, the physically dynamic tracing attack can be prevented to achieve two levels of location privacy. Location privacy level I means that each IoT user's private location can only be obtained for the LBS system at the network initialization phase and in the scenario of disputes, but not for other unauthorized IoT users. It can be achieved by the new technique of efficient privacy-preserving data aggregation presented earlier. Only the LBS system holding secret key  $sk_f$  can successfully decipher the authentic distances and allocate each IoT user to the corresponding group where the distance between the IoT user and the selected group center is the shortest.

Location privacy level II must be achieved in the efficient privacy-preserving LBS bundle filtering phase. It means that only the distance comparison result for deciding duplicate/redundant LBS contents, but not each IoT user's private

locations (driving route), can be obtained by unauthorized IoT users. Note that the distance comparison result between  $Dist_{pid_i, CP_l}$  and  $C$  is the metric to decide whether the newly arrived message carried by IoT user  $pid_i$  is redundant to social group  $C$  with the condensing point  $CP_l$ . The privacy-preserving distance comparison can be realized by exploiting the technique of zero knowledge proof [15], where the real distance  $Dist_{pid_i, CP_l}$  implying the position of IoT user  $pid_i$ , would not be disclosed. To further guarantee data confidentiality, forward/backward security, and reduced communication overhead, the technique of self-healing group key distribution can be exploited to prevent the key establishment material from retransmission due to the packet loss from the mobility and short-range communication in cloud-based IoT. Figures 4a and 4b demonstrate the advantage of communication cost and computational cost on each IoT user in the proposed SAVE compared to the existing work [4, 5], reducing the authentication overhead from the receiver's aspect.

## CONCLUSIONS AND OPEN RESEARCH ISSUES

In this section, we conclude this article by identifying a series of challenging open research issues with convincing solution ideas.

1. The first problem is fine-grained ciphertext access control in cloud-based IoT. It is well known that LBS allows each mobile IoT user to obtain timely and useful responses from the server according to her/his query interest. Unfortunately, due to the "pay-per-use" manner of the LBS cloud server, only an IoT user entering the regions in which her/his corresponding LBS has been registered to the local server can successfully decipher the encrypted query responses. It is obviously observed that this problem can also be extended into the multiple dimension scenario and possesses wide applications in outer space security. Designing lightweight attribute-based encryption (ABE) [14] provides a convincing solution to this issue.

2. Besides data confidentiality, location privacy and query privacy for cloud-based IoT users in

Besides data confidentiality, location privacy and query privacy for cloud-based IoT users in LBS should also be well protected, since the moving route exposure would reveal IoT users' living habits and the query privacy would disclose their private favorites.

LBS should also be well protected, since the moving route exposure would reveal IoT users' living habits, and the query privacy would disclose their private favorites. To address the challenging open problem, designing policy-hidden ABE exploiting the technique of a noninteractive proof system for bilinear groups [15] would give us a promising solution.

3. Our proposed efficient privacy preserving technique of one-way trapdoor permutation was only exploited for secure data aggregation from one single user. It is required to extend our proposed efficient privacy preserving technique to thwart the security and privacy threats in other types of cloud-based IoTs. For example, in smart grid IoT, it is also required to protect each user's real-time power usage from exposure while judging the peak/off-peak status by outsourced computing of the total power consumption of all power consumers in a specific region and comparing it to a predefined threshold. Therefore, how to extend our proposed new efficient privacy-preserving technique to achieve secure data aggregation from multiple users in other kinds of cloud-based IoT to meet their unique security requirements also considerably appeals to both the academia and the industry.

4. For the next generation mobile technologies such as 5G on IoT-cloud convergence, dramatically increasing batches of data are required to be processed with privacy preservation. Another interesting open research issue is privacy-preserving outsourced data mining in cloud-based IoT. For example in vehicular IoT, it is required for each vehicle user to monitor the real-time traffic conditions in its neighborhood, which can be exploited to infer the traffic status afterward (i.e., exploiting an appropriate curve fitting algorithm on the collected traffic data days before to forecast the traffic status during the same time period days after, or using the data hours before to infer the condition hours later in the same day) and recommend to corresponding vehicular users the most unobstructed route from source to destination. However, it is also required to protect the user's identity privacy and location privacy, guarantee the correctness of an outsourced mining result, and ensure that the result can only be accessed by authorized entities. Therefore, how to design verifiable outsourced data mining in the ciphertext domain becomes a challenging open problem.

5. For the security and privacy of cloud-based IoT w.r.t. big data, public key fully homomorphic encryption (FHE) undoubtedly suggests an alternative to generalized secure outsourced computation supporting both addition and multiplication operations in the ciphertext domain (i.e., not limited to secure data aggregation); however, to the best of our knowledge, despite great efforts on designing lightweight FHE, the huge volume of computational complexity still significantly impedes its wide application on resource-constrained users in cloud-based IoT. Fortunately, to construct a new generalized framework of lightweight secure outsourced computation by extending our proposed efficient privacy preserving data aggregation without public key homomorphic encryption would definitely contribute to the blooming of cloud-based IoT.

## ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grants 61373154, 61371083, 61632012, 61672239, and 61602180, in part by the NSFC-Zhejiang Joint Fund for the Integration of Industrialization and Informatization under Grant U1509219, and in part by the Natural Science Foundation of Shanghai under Grant 16ZR1409200.

## REFERENCES

- [1] Z. Sheng et al., "A Survey on the IETF Protocol Suite for the Internet of Things: Standards, Challenges, and Opportunities," *IEEE Wireless Commun.*, vol. 20, no. 6, 2013, pp. 91–98.
- [2] X. Li et al., "Smart Community: An Internet of Things Application," *IEEE Commun. Mag.*, vol. 49, no. 11, 2011, pp. 68–75.
- [3] R. Roman, P. Najera, and J. Lopez, "Securing the Internet of Things," *Computer*, vol. 44, no. 9, 2011, pp. 51–58.
- [4] X. Lin et al., "TSVC: Timed Efficient and Secure Vehicular Communications with Privacy Preserving," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, 2008, pp. 4987–98.
- [5] X. Lin and X. Li, "Achieving Efficient Cooperative Message Authentication in Vehicular Ad Hoc Networks," *IEEE Trans. Vehic. Tech.*, vol. 62, no. 7, 2013, pp. 3339–48.
- [6] J. Zhou et al., "4S: A Secure and Privacy-Preserving Key Management Scheme for Cloud-Assisted Wireless Body Area Network in m-Healthcare Social Networks," *Info. Sciences*, vol. 314, 2015, pp. 255–76.
- [7] J. Sen, "Privacy Preservation Technologies in Internet of Things," *Proc. Int'l. Conf. Emerging Trends in Mathematics, Technology, and Management*, 2011.
- [8] R. Roman et al., "Key Management Systems for Sensor Networks in the Context of the Internet of Things," *Computer & Electrical Engineering*, vol. 37, no. 2, 2011, pp. 147–59.
- [9] H. Zhu et al., "SMART: A Secure Multilayer Credit-Based Incentive Scheme for Delay-Tolerant Networks," *IEEE Trans. Vehic. Tech.*, vol. 58, no. 8, Oct. 2009, pp. 4628–39.
- [10] J. Zhou et al., "Secure and Privacy Preserving Protocol for Cloud-Based Vehicular DTNs," *IEEE Trans. Info. Forensics and Security*, vol. 10, no. 6, 2015, pp. 1299–314.
- [11] P. Paillier, "Public Key Cryptosystems Based on Composite Degree Residuosity Classes," *Eurocrypt '99*, pp. 223–38.
- [12] Y. Saleem, F. Salim, and M. H. Rehmani, "Resource Management in Mobile Sink Based Wireless Sensor Networks through Cloud Computing," in *Resource Management in Mobile Computing Environments*, Springer-Verlag, vol. 3, 2014, pp. 439–59.
- [13] R. Lu et al., "Pi: A Practical Incentive Protocol for Delay Tolerant Networks," *IEEE Trans. Wireless Commun.*, vol. 9, no. 4, Apr. 2010, pp. 1483–92.
- [14] J. Zhou et al., "TR-MABE: White-Box Traceable and Revocable Multi-Authority Attribute-Based Encryption and Its Applications to Multi-Level Privacy-Preserving e-Healthcare Cloud Computing Systems," *IEEE INFOCOM 2015*.
- [15] J. Groth and A. Sahai, "Efficient Noninteractive Proof Systems for Bilinear Groups," *Advances in Cryptology@EUROCRYPT 2008*, Springer Berlin, 2008., pp. 415–32

## BIOGRAPHIES

JUN ZHOU (jzhou@sei.ecnu.edu.cn) received his Ph.D. degree in computer science from Shanghai Jiao Tong University (SJTU) and joined East China Normal University in 2015. His research interests mainly include fine-grained ciphertext access control and secure outsourced computation.

ZHENFU CAO [SM'10] (zfciao@sei.ecnu.edu.cn) received his B.Sc. degree in computer science and technology and his Ph.D. degree in mathematics from Harbin Institute of Technology, China, in 1983 and 1999, respectively. His research interests mainly include number theory, cryptography, and information security. Since 1981, he has had more than 400 academic papers published in Journals and conferences. He was promoted to associate professor in 1987, became a professor in 1991, and is currently a Distinguished Professor at East China Normal University. He also serves as a member of the expert panel of the National Nature Science Fund of China. He is actively involved in the academic community, serving as Committee/Co-Chair and Program Committee member for several international conferences: IEEE GLOBECOM (since 2008), IEEE ICC (since 2008), and others. He is the Associate Editor of *Computers and Security* (Elsevier) and *Security and Communication*

---

*Networks* (Wiley), an Editorial Board member of *Fundamenta Informaticae* (IOS) and *Peer-to-Peer Networking and Applications* (Springer-Verlag), and a Guest Editor of *Wireless Communications and Mobile Computing* (Wiley), *IEEE Transactions on Parallel and Distributed Systems*, and others. He has received a number of awards, including the Youth Research Fund Award of the Chinese Academy of Science in 1986, the Ying-Tung Fok Young Teacher Award in 1989, the National Outstanding Youth Fund of China in 2002, and the Special Allowance by the State Council in 2005, and was a corecipient of the 2007 IEEE International Conference on Communications-Computer and Communications Security Symposium Best Paper Award in 2007. He is also the leader of the Asia 3 Foresight Program (61161140320) and the key project (61033014) of the National Natural Science Foundation of China.

XIAOLEI DONG (dongxiaolei@sei.ecnu.edu.cn) is a Distinguished Professor at East China Normal University. After her graduation with a doctorate degree from Harbin Institute of Technology, she pursued her postdoctoral study at in SJTU from September 2001 to July 2003. Then, in August 2003, she joined the Department of Computer Science and Engineering of SJTU. Her primary research interests include number theory, cryptography, and trusted computing. Since 1998, she has published

more than 80 academic papers. As the first author, she has two textbooks published by Science Press and China Machine Press respectively. Her "Number Theory and Modern Cryptographic Algorithms" project won the first prize of the China University Science and Technology Award in 2002. Her "New Theory of Cryptography and Some Basic Problems" project won the second prize of the Shanghai Nature Science Award in 2007. Her "Formal Security Theory of Complex Cryptographic System and Applications" won the second prize of the Ministry of Education Natural Science Progress Award in 2008. Currently, she hosts a number of research projects supported by the National Basic Research Program of China (973 Program), the special funds on information security of the National Development and Reform Commission and National Natural Science Foundation of China, and more. She is an Associate Editor of *Security and Communication Networks* (Wiley).

ATHANASIOS V. VASILAKOS (th.vasilakos@gmail.com) is currently a visiting professor at the National Technical University of Athens, Greece. He has served or is serving as an Editor for many technical journals, such as *IEEE TNSM*, *IEEE TSMC-PART B*, *IEEE TITB*, *ACM TAAS*, and *IEEE JSAC* Special Issues in May 2009, and January and March 2011. He is Chairman of the Council of Computing of the European Alliances for Innovation.

# High-Efficiency Urban Traffic Management in Context-Aware Computing and 5G Communication

Jianqi Liu, Jiafu Wan, Dongyao Jia, Bi Zeng, Di Li, Ching-Hsien Hsu, and Haibo Chen

With the increasing number of vehicle and traffic jams, urban traffic management is becoming a serious issue. The authors propose a novel four-tier architecture for urban traffic management with the convergence of VANETs, 5G networks, software-defined networks, and mobile-edge computing technologies.

## ABSTRACT

With the increasing number of vehicle and traffic jams, urban traffic management is becoming a serious issue. In this article, we propose novel four-tier architecture for urban traffic management with the convergence of VANETs, 5G networks, software-defined networks, and mobile edge computing technologies. The proposed architecture provides better communication and more rapid responsive speed in a more distributed and dynamic manner. The practical case of rapid accident rescue can significantly shorten the rescue time. Key technologies with respect to vehicle localization, data pre-fetching, traffic lights control, and traffic prediction are also discussed. Obviously, the novel architecture shows noteworthy potential for alleviating traffic congestion and improving the efficiency of urban traffic management.

## INTRODUCTION

With the growth of urbanization, the problem of urban traffic congestion has become a serious concern. In 2014, urban commuters in the United States collectively lost 6.9 billion hours and 3.1 billion gallons of fuel to traffic delay, and the excess fuel and lost productivity cost US\$160 billion. Similarly, as nearly a third of the world's 50 most congested cities are in China, the traffic problem is worse than in the United States. Researchers have added extra infrastructure to reduce the congestion such as building dedicated lanes for bus rapid transit. However, the effect is limited because the construction speed of extra roads is far slower than the increasing speed of new vehicles. Therefore, the new urban traffic management solution is expected to explore strategies to use emerging technologies to mitigate urban traffic congestion [1]. In order to achieve high-efficiency urban traffic management, there are at least three key issues that need to be addressed.

**Perception of real-time traffic conditions:** Numerous high-resolution roadside sensors and onboard sensors need to be deployed to sense all real-time traffic conditions including vehicle speed, direction, location, road throughput, weather conditions, temperature/humidity and so on.

**Low-latency communication and massive data storage:** Sensors intermittently produce huge amounts of raw data, and easily touch the petabyte order of magnitude in size. In view of the difference in data types, dimensionality, and huge volume, the bandwidth of communication networks, storage capability, and data processing speed need to expand than ever.

**Traffic prediction and real-time responsiveness:** Massive traffic data helps city planners to monitor traffic density, throughput, and events in real time, and the traffic control systems should have real-time responsive ability to respond to traffic events, and make immediate decisions based on traffic prediction algorithms to guide traffic flow. Designing a context-aware traffic light control system to decrease the waiting time at intersections, a rapid accident rescue system to improve emergency responsiveness, and a novel data analysis and traffic prediction system based on massive traffic data to optimize the efficiency of the existing roads are considered as key elements.

Unfortunately, the existing data collection system, vehicular ad hoc networks (VANETs), and the traditional traffic flow prediction model [2] are far from sufficient to solve the above-mentioned issues. For example, VANETs support a variety of services and achieve success in a certain aspect such as a vehicular collision avoidance system using vehicle-to-vehicle or vehicle-to-infrastructure communication [3]. However, its inherent defects such as unbalanced traffic flow and low bandwidth impact the deployment of urban traffic management [4]. Therefore, exploring a high-efficiency urban traffic management system is becoming extremely urgent.

The rapid development of emerging technologies such as fifth generation (5G), software defined networking (SDN) [5], and mobile edge computing (MEC) [6] is expected to boost the advancement of urban traffic management. Introducing the 5G and SDN technologies into the vehicular network, the new SDN-based heterogeneous vehicular network shall offer high-bandwidth communication service with flexibility and programmability[7]; thus, the environmental sensing will become more agile. Meanwhile, MEC places the computing resource at the edge of the

mobile vehicular network, and performs critical missions with real-time or near-real-time response speeds. The main contributions of this article are to propose a novel architecture combined with 5G wireless network, SDN, and MEC technologies, use a paradigm of road accident rescue to validate the high efficiency of the proposed architecture, and discuss the key technologies and potential solutions. In short, the novel urban traffic management architecture would be more highly efficient than ever before with the features of intelligent sense, low-latency communication, and real-time response.

The article is organized as follows. We propose a novel four-tier architecture. Afterward, rapid accident rescue, a practical case, is used to validate the high efficiency of the proposed architecture, and several key technologies toward the architecture are discussed. Finally, the conclusion is made.

## ARCHITECTURE OF URBAN TRAFFIC MANAGEMENT

Nowadays, tremendous traffic congestion makes the commuter extremely stressed. However, various emerging technologies provide a potential opportunity to improve traffic congestion and exhaust emission by monitoring traffic conditions. Based on these technologies, it is necessary to construct a new architecture to improve emergency responsiveness, balance the traffic flow, and save fuel and time in the transportation of citizens.

When artificial intelligence combines with big data, a new data-driven computing model, such as deep learning, is explored. Apart from data collection and communication, the new architecture needs to carefully take into consideration the storage, access, and analysis technologies. The four-tier architecture for urban traffic management is shown in Fig. 1, including the environment sensing layer, communication layer, MEC server layer, and remote core cloud server (RCCS) layer.

### ENVIRONMENT SENSING LAYER

Similar to IoT applications, the data collection layer is the foundation and plays a vital role [8]. Traffic data is mainly derived from the roadside infrastructure and onboard sensors. As usual, the roadside infrastructure such as the inductive loop is responsible for counting the number of vehicles, identifying license plate numbers, and so on. Such data from the detector is classified as passive data. But now, the vehicles that act as a big intelligent sensor employ the onboard sensors to sense vehicular status including engine speed, velocity, direction, and surrounding environment information, including location, lane, temperature, and humidity. This type of data is classified as active data. In the environment sensing layer, the precision and integrity is better than in the traditional way since high-resolution sensors are deployed massively, and the vehicles report their status and environmental data actively. In particular, vehicle location, which is important traffic data, can also be estimated by new wireless localization technology in urban environments where GPS coverage is not available.

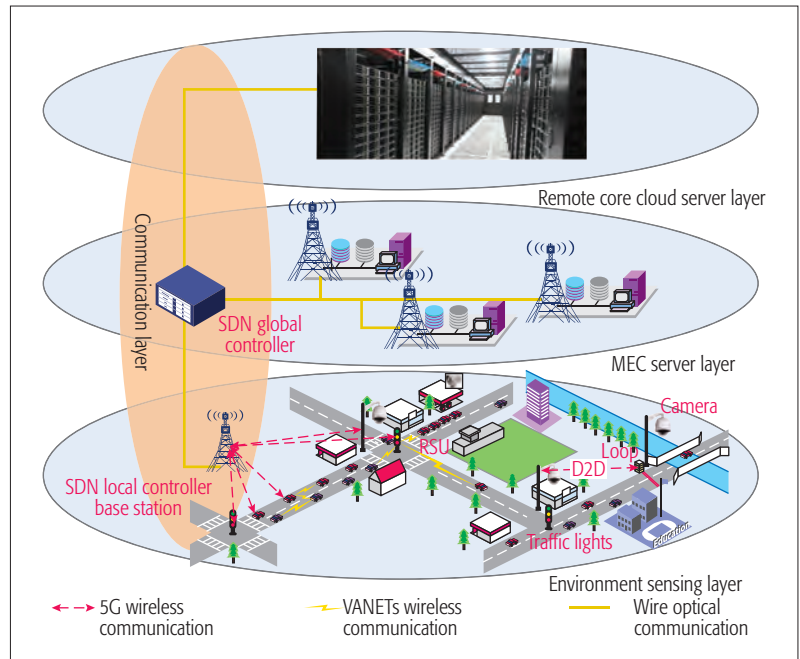


Figure 1. The four-tier architecture.

### COMMUNICATION LAYER

VANETs have been viewed as one of the most enabling technologies for “connected cars,” but huge amounts of traffic data bring some challenges to VANETs, such as unbalanced traffic flow [9]. In order to break the bottleneck of communication, we consider two emerging network paradigms: 5G and SDN. The 5G network employs multiple-input multiple-output (MIMO) and cognitive radio technologies to achieve 1.2 Gb/s communication speed in a mobile environment from a vehicle traveling at over 100 km/h, while device-to-device (D2D) technology can offer positioning service, in addition to more flexible and direct information exchange. SDN architecture, a new revolutionary idea in networking, decouples the network control (control plane) and the forwarding functions (data plane), enabling network control to become programmable. The underlying infrastructure can be abstracted from applications and network services. SDN-based networks provide flexibility, scalability, programmability, and global knowledge of the network. The SDN model operates with OpenFlow protocol, as shown in Fig. 2, the components of which are as follows.

**SDN global controller:** The central global controller maintains all the network behaviors of all the SDN-based heterogeneous wireless and wire networks. It belongs to the control plane, and communicates with the data plane using a data-controller plane interface (D-CPI) and with the application plane using an application-controller plane interface (A-CPI).

**5G base station (BS):** The BS plays three types of roles:

- A router as global mobile communication device acts as a resource (router).
- MEC could act as a server; the portable virtual machine (VM) is deployed on the BS.
- The SDN local controller controls the network elements, roadside units (RSUs), and vehicles in a certain local area.

The MEC application platform includes infrastructure-as-a-service and a set of middleware. Applications are deployed on an independent VM. In order to facilitate the urban traffic management, the proposed architecture offers four basic service components.

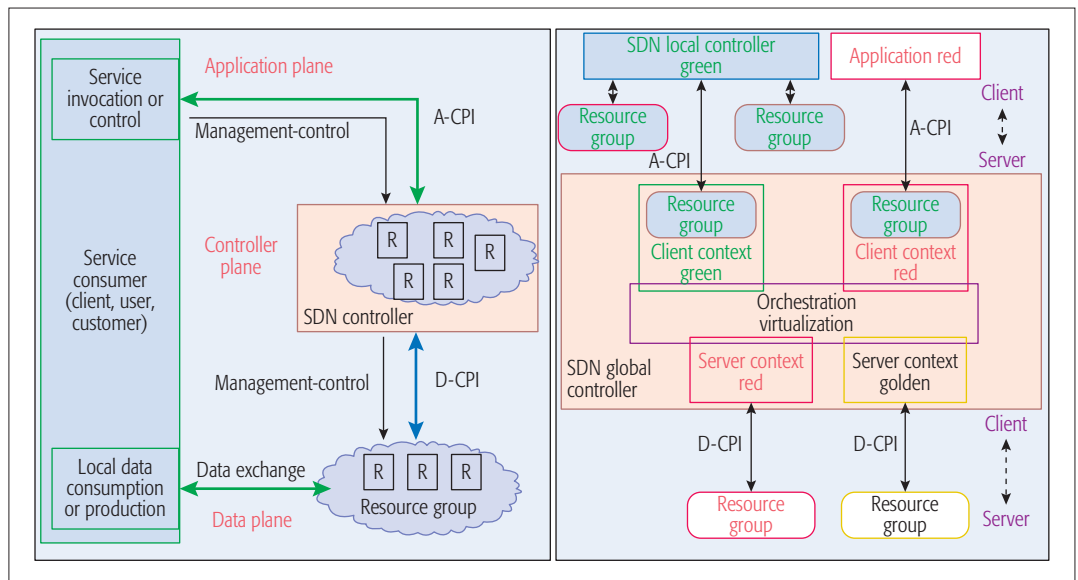


Figure 2. SDN-based network model.

**SDN RSUs:** RSUs acting as resources (switches) are responsible for forwarding the data and are controlled by the SDN local controller (i.e., BS). They belong to the data plane, and provide service using D-CPI.

**SDN wireless node:** The vehicles, equipped with onboard RF modules, act as the resource (transceiver). Their role is to provide client-to-request communication services.

The SDN-based heterogeneous network offers various advantages. First, due to the separation of the control and data planes, we can design a forwarding policy to make more informed routing decisions that can resolve the unbalanced traffic flow issue, and offer more flexible path selection strategy with the network's programmability. Second, the channel or frequency selection becomes more flexible. Vehicles are often equipped with several wireless modules to support different modes of communication; the SDN-based communication layer offers a selection policy according to the cognitive radio and channel allocation policy, which enables low-latency and high-bandwidth communication. The information exchange becomes easier and more agile in the communication layer.

#### MOBILE EDGE COMPUTING SERVER LAYER

MEC, initially introduced by IBM, aims at optimizing the existing mobile infrastructure service, and minimizing the mean delay of general traffic flow in the LTE downlink [10]. Due to its characteristics of low latency, on-premises presence, and location awareness, MEC is introduced into the proposed architecture to improve the responsiveness for traffic light control and accident rescue. The MEC server, deployed on the 5G BS, is close to the end user at the edge of the mobile vehicular network, and each application runs on it. Besides, in order to facilitate VM migration, data pre-fetching, and synchronization, a uniform framework and management infrastructure is necessary, which is illustrated in Fig. 3.

The framework includes the MEC hosting infrastructure, MEC application platform, and applications layer. The MEC hosting infrastructure

consists of a hardware virtualization layer and a set of hardware resources, especially vehicular communication devices. The MEC application platform includes infrastructure as a service and a set of middleware (service management, communication components, cognitive radio network information, and traffic offload function). Applications are deployed on an independent VM. In order to facilitate urban traffic management, the proposed architecture offers four basic service components.

**Vehicle localization service:** This service is deployed only on an MEC server, and aims to support reliable location service in a complex urban environment. As the satellite signal is blocked by skyscrapers, this service employs direction of arrival (DOA)/time of arrival (TOA) or vision to realize GPS-free localization based on 5G antenna or D2D communication.

**Traffic data process service:** This service is deployed on the MEC server and the RCCS simultaneously. The raw traffic data in certain areas are converged into the MEC server, but these data contain some invalid and coarse data that needs to be cleansed. This service in the MEC server aims at filtering the unnecessary data and storing it temporarily. In the RCCS, this service aims at data pre-fetching, data synchronizing, and data storage.

**Traffic prediction service:** The traffic prediction service is deployed on the MEC server and the RCCS simultaneously, and aims to alleviate the traffic congestion and offer personalized services such as dynamic route planning. In the RCCS, based on the massive traffic data, the service employs a deep learning algorithm to learn the generic features for predicting the traffic flow in the short term. In the MEC server, the service is responsible for dissemination of messages to the drivers according to the prediction result.

**Traffic lights control service:** The traffic lights control service is deployed only on the MEC server, and aims to control the traffic lights in real time. Due to its proximity to the source of traffic data, the service is particularly useful to capture key information for local traffic flow analysis, and

can realize dynamic management according to the traffic flow, which can increase the throughput at the intersection.

### REMOTE CORE CLOUD SERVER LAYER

The RCCS, the same as Google Hadoop, provides on-demand network access to a shared pool of resources including the processing power, storage, applications, and services [11]. The framework of the RCCS is the same as that of the MEC server, and the VM (i.e., service) can migrate freely among the RCCS and the MEC servers. In comparison to the MEC server, the difference here is that the RCCS has more storage resource and more high-performance computing power, which may make up for the MEC server's drawback of limited resources. The MEC server focuses on critical missions with real-time responsiveness; in contrast, the RCCS focuses on big data storage and analysis. Therefore, all of the traffic data eventually converges into the RCCS and is permanently stored, while the traffic prediction service employs a deep learning algorithm to predict traffic flow in the short term based on the massive traffic data.

### CASE STUDY: RAPID ROAD ACCIDENT RESCUE

The *golden hour* philosophy indicates that casualties have a much poorer chance of survival if they are not delivered to the hospital and receive definitive care within one hour, including the time taken for call-out, traveling to the accident spot, extrication, and transport to the hospital. The tight collaboration of on-site personnel, medical workers, firefighters, and traffic management agencies would reduce the entrapment time and consequently mortality rates through better organization and a methodical approach to extrication [12]. In the urban environment, the crucial obstacle to rescue is that of severe traffic congestion. In practice, once a road accident blocks the traffic in an urban area, an ambulance is not able to rapidly get to the emergency spot due to lack of special rescue lanes. The accident causes traffic congestion, and the congestion decreases the rescue speed and results in further congestion. This phenomenon looks like a deadlock, and the existing rescue mechanism offers no solution.

The proposed architecture has a potential opportunity to change this impasse, as shown in Fig. 4. With the assistance of the high-bandwidth and low-latency SDN-based heterogeneous network and rapid-responsive MEC server, the novel rescue system would be improved by three significant measures.

**Remote video diagnosis and initial assessment:** In the process of traditional rescue, when the emergency medical responders receive the aid request, they immediately depart for the incident scene in the shortest time. Due to the limited network bandwidth, there is no initial assessment of the casualties and no advice to on-site personnel for avoiding the risk of another collision and removal of the injured from the vehicle in the right way. In the proposed architecture, the damage-related information collected by sensors along with the vehicle location would be forwarded to the rescue center through the heterogeneous vehicular network. The real-time video diagnosis can be conveniently guaranteed with high bandwidth, and the emergency medical responders

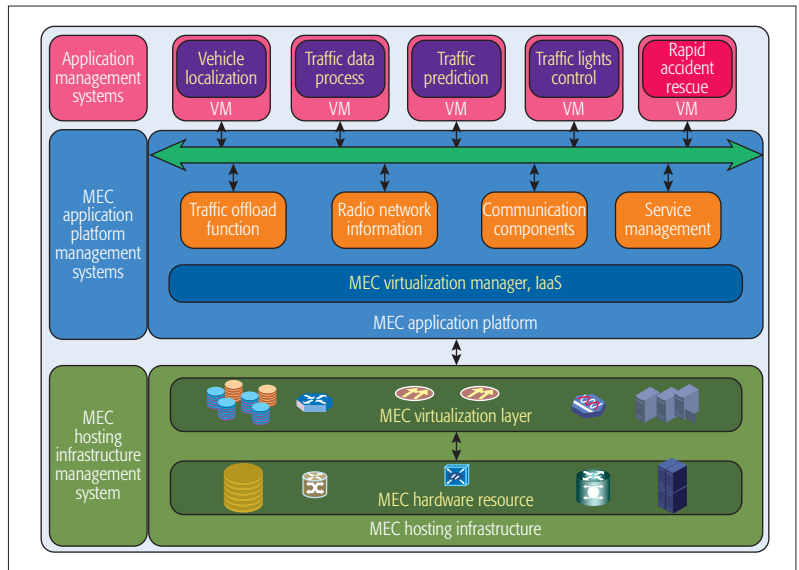


Figure 3. The portable MEC server framework.

can acquire the first-hand information required to evaluate the extent of the injuries [13], and thus prepare the suitable doctors and devices to further treatment.

**Early-warning and no-go area:** The main reason for the accident-related traffic congestion is that drivers are not aware of the accident time and location. The vehicular stream unknowingly enters the accident area, which results in traffic congestion. With the assistance of vehicle localization service, the rescue system can obtain the location of every vehicle. The system can set a virtual early-warning and no-go area on the electronic map, and then disseminate warning messages to the drivers. Once a vehicle enters into the early-warning area, the driver would be alerted by a message to keep away from the accident spot. Meanwhile, based on the traffic lights control service, the system can adjust the “green time” of traffic lights at the edge of the no-go area and force the vehicles to bypass the accident spot.

**Green rescue lane for emergency vehicles:** Rapid removal of the injured to the hospital will improve their chances of survival. In order to minimize the entrapment time from the hospital to the accident spot, the proposed rescue system calls for the traffic lights control service to coordinate all traffic lights in this certain area, and clear a green rescue lane for the ambulance and other emergency vehicles.

In order to simplify the verification of the proposed system, we select 10 accidents randomly from all traffic accidents in a month, and use the traffic data to simulate the practical traffic flow. Because the “ambulance priority” law is not observed well in some developing countries, we make two assumptions:

- 60 percent of drivers or pedestrians keep clear of an ambulance without an electronic police (E-police) monitor.
- 95 percent of drivers or pedestrians keep clear of an ambulance with an E-police monitor.

Figure 5 shows the preliminary validation of the proposed system in normal times and in rush hour. We first compare three solutions in normal

With the widespread installation of on-board and roadside sensors, a huge amount of traffic data is being collected. The objective of traffic prediction is to utilize the massive historical and real-time data to mine the trend of traffic flows in near future time and guide route planning for drivers.



Figure 4. Rapid road accident rescue system.

times without congestion. The novel solution can save 2.8 min on average without an E-police monitor, and 5.5 min with an E-police monitor at the first-arrival time, and saves 4.9 min and 6.8 min, respectively, on average in traffic recovery time. It seems that the improved effect is not obvious. Second, we compare them during rush hour. The novel solution saves on average 9.9 min without an E-police monitor and 12.5 min with an E-police monitor in the first-arrival time of ambulance prospect, and saves on average 15.5 and 19.5 min, respectively, in traffic recovery time. The rapid road accident rescue system can decrease the time consumed and increase the chance of survival.

### DISCUSSION AND FUTURE WORK

The rapid road accident rescue has achieved great success with the help of the proposed architecture, but there are still some key technologies that need to be addressed. We focus on four aspects (vehicle localization, data pre-fetching strategy, traffic prediction, and traffic lights control) to state the challenges and possible solutions.

#### VEHICLE LOCALIZATION IN THE URBAN ENVIRONMENT

With the growth of urbanization, dense skyscrapers have impacted GPS-based localization, and therefore high-reliability GPS-free localization needs to be addressed. The 5G ultra-dense networks are expected to operate under the coverage area of multiple BSs simultaneously. This technology brings a new opportunity to provide localization services for vehicles. First, the DOA estimation method based on the 5G antenna array is feasible. Second, the received signal strength can be utilized to estimate the distance between vehicles and BSs. Third, the millimeter waves of 5G have higher bandwidth and more severe propagation losses, which enable TOA

estimation with high accuracy. Finally, as the MEC server is placed at the edge of the vehicular network with near-real-time response speed, the vision-based localization method could be used. Vehicle-centric localization is gradually replaced by network-centric localization due to the higher accuracy and better robustness.

#### PREDICTABLE DATA PRE-FETCHING

Mobile vehicles are typically subjected to network fluctuations and intermittent downtimes, which lead to an unpleasant experience. Hence, in order to compensate for link disconnections, it is important to pre-fetch and buffer the data locally. In order to further enhance the response speed, a two-tier predictable pre-fetching strategy is proposed as the vehicular traversal route can be predictable. First, the MEC server pre-fetches all of the traffic data from the previous MEC server and RCCS simultaneously through high-speed optical networks. Second, the vehicles pre-fetch the data from the MEC sever. The location-based predictable pre-fetching strategy can efficiently decrease the latency in data transmission, enhance the responsiveness of urban traffic management systems, and improve quality of user experience.

#### TRAFFIC PREDICTION

With the widespread installation of onboard and roadside sensors, a huge amount of traffic data is collected. The objective of traffic prediction is to utilize the massive historical and real-time data to mine the trend of traffic flow in the near future and guide route planning for drivers. In the past, some prediction models were developed with a small amount of traffic data collected in a specific small area, and the accuracy of traffic prediction was dependent on the traffic flow features embedded in the collected spatiotempo-



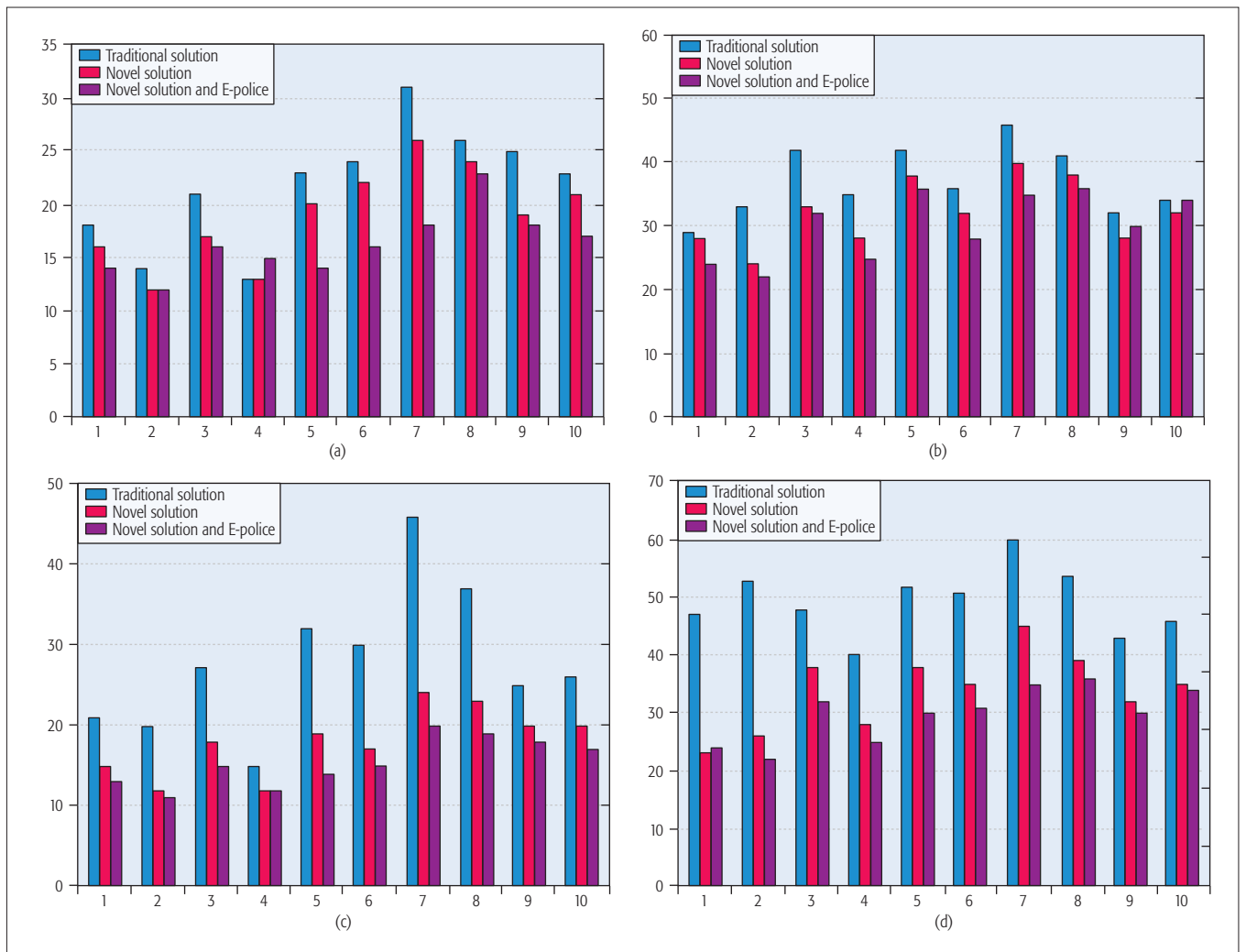


Figure 5. The comparison of consuming time in rescue action (units: minute): a) first arrival time in normal times; b) traffic recovery time in normal times; c) first arrival time in rush hour; d) traffic recovery time in rush hour.

ral traffic data. Practically, these models have not been successful. Recently, big data is increasingly being used to analyze global problems and find solutions using massive data; Yisheng Lv *et al.* have proposed a novel prediction model based on a deep learning algorithm to alleviate traffic congestion. This model has the ability to improve traffic conditions and reduce travel delays. The proposed architecture is more suitable for the deep-learning-based prediction model. First, the active and passive traffic data is more comprehensive than ever. These massive data can help the models to discover latent traffic flow features such as the nonlinear spatial and temporal correlations from the traffic data. Second, the low latency makes it easier for the model to utilize the real-time traffic data. Third, the cloud server is divided into an MEC cloud server and an RCCS in the proposed architecture. As real-time responsive tasks are turned over to the MEC server, the RCCS can be dedicated to deep learning calculation, and the prediction performance is better.

#### INTELLIGENT TRAFFIC LIGHTS CONTROL

The traffic lights controller is designed to ensure that the traffic flow moves as smoothly and safely as possible, while the pedestrians

are protected when they cross the intersection. Some researchers have long used historical data to design signal plans that optimize the “green time” to improve traffic flow with sophisticated systems using different plans for different times of day. As accommodating practical dynamic traffic conditions is difficult, the existing traffic lights control strategy is unable to adapt real-time response to the change in traffic flow.

Fortunately, the SDN-based heterogeneous network and the MEC technologies shall reform the method of data collection. First, the vehicle actively reports its status and environmental information, including trajectory and route. Second, high-definition cameras and high-performance sensors are deployed to take high-quality video or raw traffic data, while the delay in data transmission and processing can be constrained within near real time. In addition, pedestrians and bicycles can be identified and quickly counted by camera. Based on the new traffic data collection system, the traffic light control service can make the traffic flow more smooth and efficient. At heavy intersections, this can extend the “green time” and offer preference to the longest line of vehicles.

The proposed architecture may decrease traffic congestion, and improve the ability to manage urban traffic.

The paradigm of the rapid road-accident rescue application has validated the feasibility and high efficiency of the proposed framework.

## CONCLUSIONS

Urban traffic congestion has become a difficult problem, and has motivated many researchers to find effective solutions. Fortunately, many emerging technologies have the potential to alleviate traffic congestion. First, the SDN-based heterogeneous network with the convergence of VANETs, 5G wireless networks, and SDN technologies can provide high-bandwidth and low-latency communication services along with programmability. Second, the MEC cloud server offers a real-time or near-real-time response speed for critical missions. Finally, the RCCS can run deep-learning-based prediction algorithms with high-performance computing power and ultra-large storage space. The proposed architecture may decrease traffic congestion and improve the ability to manage urban traffic. The paradigm of the rapid road accident rescue application has validated the feasibility and high efficiency of the proposed framework.

## ACKNOWLEDGMENT

The corresponding author is Dr. Jiafu Wan. This project was supported in part by grants from the National Natural Science Foundation of China (Nos. 61262013, 11104089, and 51575194), the Guangdong Province Special Project of Industry-University-Institute Cooperation (No. 2014B090904080), the Guangdong Natural Science Foundation (Nos. 2016A030313734, and 2016A030313735), and the Fundamental Research Funds for the Central Universities (No. 2015ZZ079). This work was also supported in part by optiTruck project (No. H2020/713788).

## REFERENCES

- [1] J. Wan *et al.*, "Context-Aware Vehicular Cyber-Physical Systems with Cloud Support: Architecture, Challenges, and Solutions," *IEEE Commun. Mag.*, vol. 52, no. 8, Aug. 2014, pp. 106–13.
- [2] J. Wan *et al.*, "Mobile Crowd Sensing for Traffic Prediction in Internet of Vehicles," *Sensors*, vol. 16, no. 1, 2016, Article ID 88.
- [3] K. Zheng *et al.*, "Reliable and Efficient Autonomous Driving: The Need for Heterogeneous Vehicular Networks," *IEEE Commun. Mag.*, vol. 53, no. 12, Dec. 2015, pp. 72–79.
- [4] W. Li and H. Song, "ART: An Attack-Resistant Trust Management Scheme for Securing Vehicular Ad Hoc Networks," *IEEE Trans. Intelligent Transportation Systems*, vol. 17, no. 4, 2016, pp. 960–69.
- [5] L. Hu *et al.*, "Software defined Healthcare Networks," *IEEE Wireless Commun.*, vol. 22, no. 6, Dec. 2015, pp. 67–75.
- [6] S. Nunna *et al.*, "Enabling Real-Time Context-Aware Collaboration through 5G and Mobile Edge Computing," *Proc. 12th IEEE Int'l. Conf. Info. Technology-New Generations*, 2015, pp. 601–05.
- [7] I. Ku *et al.*, "Towards Software-Defined VANET: Architecture and Services," *Proc. 13th Annual IEEE Mediterranean Ad Hoc Networking Wksp.*, 2014, pp. 103–10.
- [8] M.S. Hossain and G. Muhammad, "Cloud-Assisted Industrial Internet of Things (IIoT) – Enabled Framework for Health Monitoring," *Computer Networks*, vol. 101, 2016, pp. 192–202.
- [9] N. B. Truong, G. M. Lee and Y. Ghamri-Doudane, "Software Defined Networking-Based Vehicular Adhoc Network with Fog Computing," *Proc. IFIP/IEEE Int'l. Symp. Integrated Network Management*, 2015, pp. 1202–07.
- [10] J.O. Fajardo, I. Taboada and F. Liberal, "Radio-Aware Service-Level Scheduling to Minimize Downlink Traffic Delay through Mobile Edge Computing," *Mobile Networks and Management*, Springer, 2015, pp. 121–34.
- [11] R. Buyya *et al.*, "Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing as the 5th Utility," *Future Generation Computer Systems*, vol. 25, no. 6, 2009, pp. 599–616.
- [12] Y. Zhang *et al.*, "iDoctor: Personalized and Professionalized Medical Recommendations Based on Hybrid Matrix Factorization," *Future Generation Computer Systems*, vol. 66, 2017, pp. 30–35.
- [13] M. S. Hossain and G. Muhammad, "Cloud-Assisted Speech and Face Recognition Framework for Health Monitoring," *Mobile Network App.*, vol. 20, no. 3, 2015, pp. 391–99.

## BIOGRAPHIES

JIANQI LIU [M] (liujianqi@ieee.org) is an associate professor with Guangdong Mechanical & Electrical College, China. He received his Ph.D. and M.S. degrees from the Guangdong University of Technology (GDUT), China. His current research interests are IoV, WSNs, and CPS.

JIAFU WAN [M] (jiafu\_wan@ieee.org) is a professor in the School of Mechanical and Automotive Engineering at South China University of Technology (SCUT). Thus far, he has authored/co-authored more than 70 journal papers (with 60+ indexed by ISI SCIE) and 30 international conference papers. His research interests include cyber-physical systems, Industry 4.0, smart factory, industrial big data, industrial robots, and the Internet of Vehicles. He is a member of ACM.

DONGYAO JIA (jiady@163.com) received his Ph.D. degree in computer science from City University of Hong Kong in 2014. He is currently a research fellow at the Institute for Transport Studies (ITS), University of Leeds, United Kingdom. He worked as a senior engineer in the telecom industry in China from 2003 to 2011. His research interests include vehicular CPS, cooperative driving, and IoT.

BI ZENG (zb9215@gdut.edu.cn) is a professor with the School of Computers, GDUT. She received her Ph.D. and M.S. degrees from GDUT, and she is a member of CCF, multi-valued logic and fuzzy logic committee, China. Her current research interests include embedded systems, robot control techniques, and WSNs.

DI LI (itdili@scut.edu.cn) is currently a professor with the School of Mechanical and Automotive Engineering at SCUT. She has directed over 50 research projects, including ones with the National Natural Science Foundation of China. Thus far, she has authored/co-authored over 180 scientific papers. Her research interests include embedded systems, computer vision, and cyber-physical systems.

CHING-HSIEN (ROBERT) HSU [SM] (chh@chu.edu.tw) is a professor in the Department of Computer Science and Information Engineering, at Chung Hua University, Taiwan. His research includes high-performance computing, cloud computing, parallel and distributed systems, big data analytics, and ubiquitous/pervasive computing and intelligence. He has published 200 papers in refereed journals, conference proceedings, and book chapters in these areas.

HAIBO CHEN (H.Chen@its.leeds.ac.uk) received his B.E. degree in mechanical engineering from Central South University, China, and his Ph.D. in mechatronics from the University of Dundee, United Kingdom. He is a Principal Research Fellow in ITS at the University of Leeds. Currently, he is leading several EU projects at Leeds including Viajeo+, FOT-Net Data, optiTruck, and AUTOPILOT. His research interests include urban mobility solutions, automatic road transport support systems, and automated driving.

# Beyond 5G Vision for IOLITE Community

Doruk Sahinel, Cem Akpolat, Manzoor A. Khan, Fikret Sivrikaya, and Sahin Albayrak

## ABSTRACT

Connecting a massive number of sensors and actuators with energy and transmission constraints is only possible by providing a reliable connection despite the increase in data traffic due to the Internet of Things, and by guaranteeing a maximum end-to-end delay for applications with real-time constraints. Next generation network architectures need to satisfy these two requirements while connecting IoT sources producing data at massive scales to cloud resources that provide the capability to process and store this data. For this reason, realization of IoT in next generation cellular networks faces the problem of delivering cloud services over the network to things that are placed anywhere. In this study, we explain how the technologies envisioned for next generation networks can respond to the challenge of realizing IoT over a use case prepared for the IoT smart home platform called IOLITE. We start by discussing capacity requirements and focus on network densification as a promising solution method. The challenges of network densification can be overcome by integrating the technological enablers such as SDN, C-RAN, SON, and mobile edge computing. For this reason, we provide a thorough survey on the state of the art in integrating these enablers for a flexible network architecture at all network segments. Finally, we discuss how the needs of the IOLITE community use case scenario can be satisfied by implementing a service-centric abstraction layer on top of a flexible infrastructure for beyond 5G IoT applications.

## INTRODUCTION

Internet of Things (IoT) concepts offer lots of services that can be used in various domains such as smart city, smart grid, and smart home. Even though IoT concepts have long existed in the literature, it is time to give more attention to these concepts as we are on the verge of a transformation by IoT in our lifestyles from the routine experience within the world we know into an environment where smart interaction with millions of devices is made possible by novel platforms.

Smart home is an IoT application concept that is going to provide massive data ranging from home sensors such as smart meters, temperature sensors, and light controllers to security cameras and multimedia services. Currently, the sensory data from most of these devices are processed locally and do not pose communication-related challenges. However, we believe that the read-

ers are no longer strangers to terms including virtual reality, augmented reality, tactile Internet, real-time pattern recognition, semantic recommendations, and many more. One common fact among all these applications is remote processing, that is, the sensory data must be processed in a computation-rich remote location (e.g., virtualized servers in remote clouds), and the outcome must be made available to the user. Here, the main challenge lies not in remote processing but in timely availability of the results to the users. This requires a communication network that guarantees near-real-time transportation of the data over the network stretch between the sensors and their remote computation servers [1]. In order to follow the crux of this article in a better way, we discuss an IoT initiative of DAI-Labor, IOLITE.<sup>1</sup>

The basic idea of IOLITE is to realize the vision of a smart home. This platform enables the interconnection of various sensors in a future home. It implements the philosophy of one platform connecting all devices by extensive development of a universal platform that supports a broad range of protocols. IOLITE enables the design of applications for a variety of use cases and allows these applications to manage all devices in the smart home environment. The big picture of IOLITE's vision is depicted in Fig. 1. In the following, we detail smart kitchen application of IOLITE to explain the network requirements of IoT use cases.

The current IOLITE version does not create a great challenge to networks. However, let us take IOLITE to a new level – the IOLITE community, which is also DAI-Labor's vision for its evolution. For instance, IOLITE-enabled apartments in a society will generate the demand for similar video contents in similar time slots, and this will most likely stress out the network infrastructure. In addition, an increasing number of smart home devices have to be able to share data with each other and their users within the community. The IoT traffic demands of the projected IOLITE community motivates implementation of novel capacity enhancement techniques inside the region where the community lives. In the next section, we discuss network densification to meet the capacity demands in a region. We also briefly discuss interference, mobility, and energy efficiency challenges of network densification.

In order to make the best use of capacity enhancement, we believe that the future network paradigm should enable concepts like demand attentiveness (i.e., estimating the content demands on temporal and spatial domains) and a soft-

The authors explain how the technologies envisioned for next generation networks can respond to the challenge of realizing IoT over a use case prepared for the IoT smart home platform called IOLITE. They start by discussing capacity requirements and focus on network densification as a promising solution method. The challenges of network densification can be overcome by integrating technological enablers such as SDN, C-RAN, SON, and MEC.

<sup>1</sup> <http://www.iolite.de>, accessed Sept. 13, 2016.

The use of millimeter wave frequency bands in cellular networks also requires careful dense small cell planning as non-line-of-sight attenuation, rain attenuation and oxygen absorption allow transmissions only when radio node coverage radius is below 200 meters.

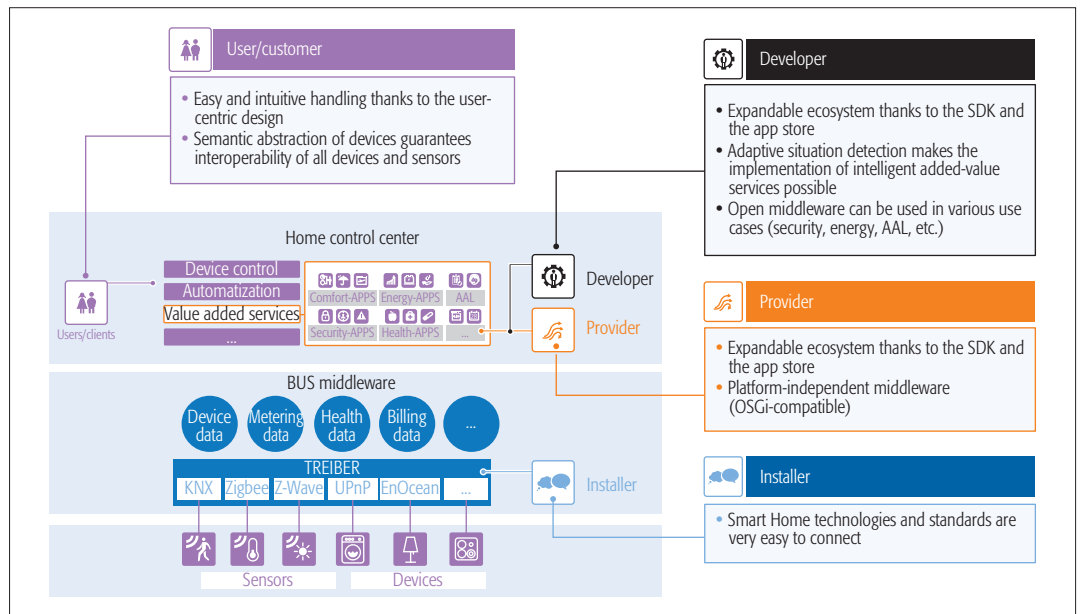


Figure 1. IOLITE smart home environment.

ware-defined networking (SDN)-based architecture for dynamic topology reconfiguration. Furthermore, bringing not only capacity resources, but also storage and processing resources to the living environment of the community can meet local content sharing challenges. Mobile edge computing (MEC) is a promising technology to provide local storage and processing at the edge of the network with low latency, and local network management can provide efficient local content sharing with SON functionalities. The motivation behind using these technological enablers is explained in more detail later, while the “IOLITE smart kitchen” application, a content sharing example for the IOLITE community, is presented after that.

The timeline for the fifth generation (5G), as provided in [2], introduces the term “beyond 5G” for the network development period after 2020. The technological enablers of next generation wireless networks such as SDN, self-organizing networking (SON), the cloud radio access network (C-RAN), and MEC are also likely to be reshaped in order to enable an IoT framework with ubiquitous communication between massive sensor deployments. Information-centric networking (ICN) is likely to introduce a prominent architectural change as beyond 5G envisions building a content-specific topology instead of an IP-based topology. We present the idea of dynamically forming an overlay network to move the content to the edge in an ICN-based environment for beyond 5G networks. The article is finalized with a discussion and concluding remarks.

### CAPACITY CHALLENGES OF NEXT GENERATION NETWORKS

The massive growth of the number of connected devices, the amount of data produced with IoT, and the diversity of IoT use cases bring a prominent challenge to next generation wireless networks. In these networks, the capacity of a coverage area must significantly increase in order to handle the massive amount of data produced by IoT services. The capacity can be increased by

utilizing more spectrum bands, by deploying new technologies to increase the spectral efficiency (bits per second per Hertz rate), by increasing the number of access nodes in the coverage area [3], or by using device-to-device communication techniques [4]. Utilizing more spectrum bands for cellular networks is possible either via deploying unused spectrum bands such as millimeter wave or by exploiting the spectrum bands underutilized by other technologies, such as TV bands. New modulation and multiplexing techniques can increase the spectral efficiency in spectrum bands. However, Rysavy [3] argues that network capacity enhancement with increasing spectral efficiency will be limited as spectral efficiency is close to theoretical bounds, and forming dense small cell deployments can enhance capacity well beyond other techniques. The use of millimeter-wave frequency bands in cellular networks also requires careful dense small cell planning as non-line-of-sight attenuation, rain attenuation, and oxygen absorption allow transmissions only when radio node coverage radius is below 200 m [5].

The network densification definition given by Bhusan *et al.* [4] includes both spectral densification by use of new spectrum bands and spatial densification with dense cellular node deployments. This network densification definition covers all the above-mentioned capacity enhancement methods, which are greatly required for realizing the massive data traffic of IoT. For this reason, the network densification concept cannot be left out of next generation 5G network architectures and the challenge of cloud-IoT integration.

Despite its necessity for 5G, network densification also creates new challenges for network management. The increasing number of small cells is likely to lead to more interference, as the distance between the channels using the same frequency decreases. As cell coverage areas get smaller, the number of cell edges, the areas that are more vulnerable to interference, is also going to increase. Furthermore, irregularly positioned small cells (e.g., privately owned small cells) or device-to-device communications using the same frequency


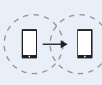

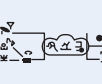
ND challenges	SDN	SON	C-RAN	MEC
 <p>Spectral efficiency</p>	<ul style="list-style-type: none"> <li>+ Dynamic frequency resource allocation</li> <li>+ Centralized transmission power control</li> </ul>	<ul style="list-style-type: none"> <li>+ Adjust transmission power to mitigate interference by listening to local environment</li> <li>+ Plug and play self-configuration decreases signaling load</li> </ul>	<ul style="list-style-type: none"> <li>+ Implementation of CoMP and enhanced inter-cell interference coordination (eICIC) for interference mitigation with coordinated RRHs</li> <li>+ Reduce signaling load of RRHs by support of BBU pool</li> </ul>	<ul style="list-style-type: none"> <li>+ Reduce signaling overhead between at backhaul and core by keeping signaling traffic at edge network</li> </ul>
 <p>Mobility management</p>	<ul style="list-style-type: none"> <li>+ Coordination and optimization of handover signaling</li> <li>+ Storage of mobility data and learning methods for prediction</li> </ul>	<ul style="list-style-type: none"> <li>+ Use self-optimization functions such as MRO and MLB for adaptive decision making based on user information, propagation paths, etc.</li> </ul>	<ul style="list-style-type: none"> <li>+ Centralized and cooperative mobility management solutions with BBU pool to prevent frequent handovers between RRHs</li> </ul>	<ul style="list-style-type: none"> <li>+ Predict mobility behavior of users based on local mobility patterns with location-awareness</li> </ul>
 <p>Energy efficiency</p>	<ul style="list-style-type: none"> <li>+ Minimization of active cells with topology manager</li> <li>+ Moving energy consuming features (e.g. spectrum sensing) from small cells to data centers</li> </ul>	<ul style="list-style-type: none"> <li>+ Small cells can switch themselves on and off based on traffic information exchange with neighbors</li> <li>+ Transmit power adjustment with SON</li> </ul>	<ul style="list-style-type: none"> <li>+ Circuit power consumption reduced compared to the traditional RAN</li> </ul>	<ul style="list-style-type: none"> <li>+ Decrease data traffic between at backhaul and core to save energy</li> </ul>
 <p>Cloud-IoT integration</p>	<ul style="list-style-type: none"> <li>+ Centralized frequency resource utilization</li> <li>+ Considering QoS/QoE of IoT-Cloud communication</li> </ul>	<ul style="list-style-type: none"> <li>+ Respond to varying network conditions dynamically to keep a reliable communication channel between IoT and Cloud</li> </ul>	<ul style="list-style-type: none"> <li>+ A single BBU pool for signal processing for many small cells to reduce processing delays and provide low latency</li> </ul>	<ul style="list-style-type: none"> <li>+ Decrease latency by processing and storing data locally</li> <li>+ Extend processing capability for IoT applications in cooperation with cloud</li> </ul>

Figure 2. Network densification challenges and solutions provided by 5G enablers.

bands as dense cellular networks can be a source of interference. Handover requirements of dense networks are also more challenging as the number of mobile users to manage will substantially increase, and the smaller coverage area of small cells are likely to cause more handovers. Expansion of network nodes is going to increase energy consumption in cellular networks; therefore, energy efficiency has to be taken into consideration. Based on these facts, it can be concluded that integrating the solution methods offered by different technological enablers such as SDN, SON, C-RAN, and MEC plays a central role in reaching the capacity and performance objectives of dense networks. Figure 2 provides a list of network densification challenges and the solution methods offered by 5G technological enablers, and in the next section we explain how these enablers can play their roles in creating a flexible wireless architecture.

## TECHNOLOGICAL ENABLERS OF NEXT GENERATION NETWORKS

### SOFTWARE DEFINED NETWORKING

Next generation cellular networks need to deal with a massive number of small cells and a massive amount of data traffic. In addition, expansion of backhaul traffic and hard-to-reach locations of small cells make dense network backhauling a complicated problem. Efficient management of such issues cannot easily be taken care of with traditional network architectures, as coupling of data and control planes in current network architectures makes it almost impossible to reconfigure the network to enforce new policies. Lack of timely reconfiguration in current networks is highly likely to cause quality of service (QoS) problems for IoT applications when frequent changes occur in network topology.

The SDN concept addresses the ossification problem of current networks by decoupling the

data plane from the control plane [7], where data plane becomes simple forwarding hardware and the control plane has decision making ability. SDN has a layered architecture, with a controller layer placed between the application layer above and the device layer below. The application layer on top involves application logic. The SDN controller resides in the controller layer and communicates with programmable services of the application layer via the northbound interface, whereas the communication with the SDN-enabled switches in the device layer is handled by the southbound interface. SDN architecture has programmable centralized controllers, meaning that the network rules and instructions can be reconfigured to respond to increasing connectivity demands and fluctuations in network topology. The programmability of SDN provides the much needed agility to deploy new protocols or services on demand. However, a very commonly known challenge is SDN's scalability due to its characteristic of a logically centralized controller. Despite the great amount of research effort on SDN, this challenge is still not addressed thoroughly, especially in the control plane.

### FLEXIBLE NETWORK ARCHITECTURE WITH SDN AND C-RAN

SDN has the potential to revolutionize fronthaul, backhaul, and core network designs of next generation wireless architecture, and can play an enabler role in RAN architecture for 5G. Technological advances such as C-RAN make an easier and energy-efficient cellular RAN design possible for massive small cell deployments [8]. The deployment of small cells with C-RAN architecture reduces signaling when many small cells are supported by a single baseband unit (BBU) pool. In this section, we provide some proposals from the literature that can play an enabler role in obtaining a flexible network architecture with SDNized fronthaul, backhaul, and core parts.

Heterogeneous C-RAN (H-CRAN), an archi-

Next generation cellular networks need to deal with a massive number of small cells and a massive amount of data traffic. In addition, expansion of backhaul traffic and hard-to-reach locations of small cells make dense network backhauling a complicated problem.



(NFV) and MEC integration, is proposed with the title of software defined mobile edge computing (SD-MEC) in [13]. A four-layer architecture is designed for SD-MEC, which involves a device layer, a network layer, a control layer, and an application layer. The device layer enables communication between a massive number of devices through various technologies such as LTE and WiFi. The network layer has an SDN gateway that is essential for the interoperability between different communication protocols and heterogeneous networks. The control layer enables the use of network applications such as network orchestration and computation, topology management, computation, and scheduling algorithms. Finally, the application layer is centralized, and it feeds the information provided by the control layer to create business layer applications for IoT.

### SMART KITCHEN APPLICATION FOR AN IOLITE COMMUNITY

Before presenting our approach built on 5G technological enablers introduced earlier, we first present an exemplary beyond 5G IoT application. One of the main differences for beyond 5G applications is that pushing only simple content may not be enough for some use cases. Pushing the service together with computation and using cloud computing to fit the content to the user's profile might be necessary.

The IOLITE smart kitchen assistant helps users during cooking by retrieving information from IoT devices in the kitchen with device drivers. For instance, the smart kitchen assistant can check the availability of ingredients in the refrigerator to suggest recipes. A future option for a smart kitchen assistant is to recommend and stream the optimal video recipe based on stored eating habits of family members. The framework also aims to enable sharing of user recipes.

Our example beyond 5G scenario for the IOLITE smart kitchen assistant is visualized in Fig. 4. In this scenario, user 1 creates a recipe with his/her own oven and kitchen appliances, and pushes this content to the IOLITE storage cloud. If the second user demanding this recipe owns a different oven or different kitchen appliances, the recipe must first be processed in the cloud and adjusted according to the user's equipment and then be presented to the second user. The food missing in the kitchen for this recipe might also be ordered on demand by using an online shopping application. As seen in this example, not only the video containing the recipe, but also the cloud service resources to process the data according to the user's profile (containing the second user's kitchen data in this use case) should be thought of as part of the user's demand.

Now, imagine that it is dinner time for an IOLITE community. Almost every user is looking for a video recipe being produced by a person in the community. Together with high quality of experience (QoE) and real-time video recipe streaming, cloud services are required to tailor the recipe based on every user's equipment profile; for example, each oven has to be adjusted to the optimal heating level for different recipes, and ventilation in the kitchen should be adjusted. The need for physical layer and cloud resources in the

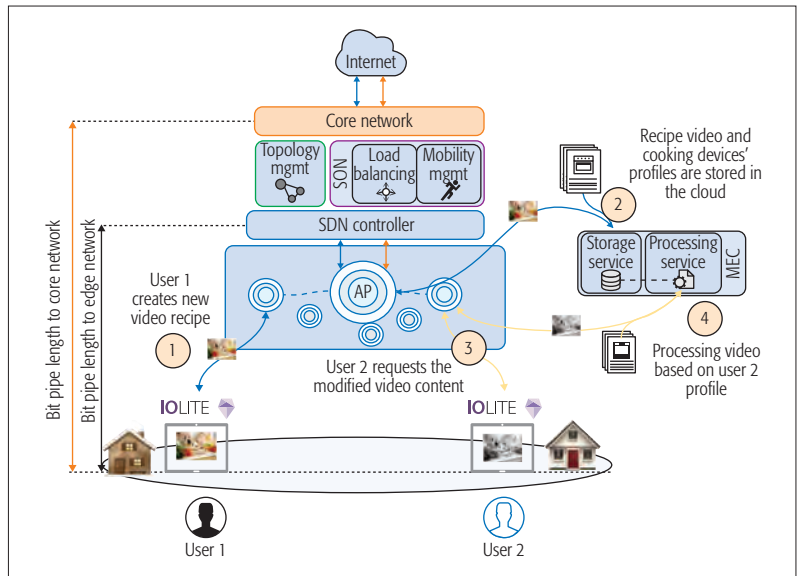


Figure 4. IOLITE smart kitchen dinner time use case.

local community is obviously going to increase. On the network side, this dynamic demand pattern should be recognized, and extra radio and cloud resources should be made available for the community. In the next section, we provide a two-layered approach that responds to this dynamic demand pattern. Figure 5 gives a list of the challenges for the IOLITE smart kitchen use case and explains how one or more technological enablers deal with those challenges.

### FLESTIC APPROACH FOR SCALABLE IoT COMMUNICATION

Technological enablers of 5G have the potential to overcome many of the envisioned challenges for cloud-IoT integration in the near future. However, the real-time and reliable service perspective of beyond 5G applications [2] forces the research community to come up with innovative solutions. In this section, we present a network slicing approach that can realize the smart kitchen application envisioned for IOLITE community.

Flexibility and elasticity are core elements of 5G, driven by the convergence of next generation networks. In our approach, SDNized RAN, backhaul, and core segments of the flexible infrastructure layer brings in the flexible management of the whole network by dynamically adapting the routing strategy and allowing function placement for varying networking tasks. For example, dinner time in an IOLITE community may require a different latency in the data plane than other times, and different routing strategies may be needed at the RAN and backhaul when one or more users share real-time video recipes among their neighbors. Elasticity brings dynamic resource allocation required when a dynamic change occurs on demand (e.g., at dinner time). Resource allocation involves both turning on inactive small cells and enabling device-to-device communication to provide more radio resources in the last mile and cloud service resources to process data according to the user's profile. We name our high-level approach *flectic*, a short name capturing *flexibility* and *elasticity*, to reflect the importance of these two characteristics.





Beyond 5G challenges	IOLITE community smart kitchen scenario requirements	Approaches provided by enablers (MEC, SDN, SON, C-RAN)
<b>Scalability</b> 	At dinner time, most of the users are actively using IOLITE community platform to reach video recipes. More and more devices in the kitchen need communication to prepare food for dinner.	Dynamic resources can be provisioned by MEC (storage and computing) and C-RAN (bandwidth), whereas their dynamic management can be operated via SDN and SON.
<b>Low latency</b> 	Sharing recipe video content and in the meantime processing computation power to minimize the latency; thus, the end user can efficiently watch and edit the video content.	MEC deployment can ensure low latency by providing service at the edge and reducing the bit pipe. SDN can dynamically configure topology for shorter end-to-end paths and allow function placement for low latency.
<b>High throughput</b> 	Smart kitchen video recipes is a multimedia service that needs high throughput and an acceptable QoS for all end users. This high throughput should be provided to every user within the community.	Dense small cell deployments with C-RAN provides extra radio resources that can be switched on and off. SON functionalities and learning with random matrix games adjust transmission power to mitigate interference.
<b>Reliability</b> 	A stable network communication able to withstand varying network conditions is an intrinsic requirement for IOLITE community apps.	Varying network conditions can be managed through SDN and SON, and the required network resources could be solved in cooperation with MEC and C-RAN.

Figure 5. IOLITE smart kitchen dinner time requirements and solutions provided by beyond 5G technological enablers.

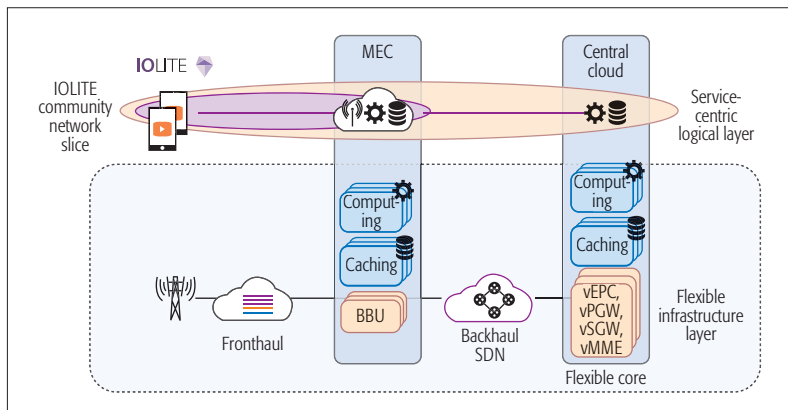


Figure 6. A two-layered flestic approach to network slice formation.

Based on our understanding of beyond 5G, we discuss a layered flestic approach in Fig. 6, where the layers correspond to the abstraction levels on the network control. At the bottom, the *flexible infrastructure layer* is the provisioning of SDNized, programmable mobile network infrastructure enabled by the aforementioned network virtualization technologies (C-RAN, SDN, and MEC). This layer allows creation and customization of a network slice for an IOLITE community during dinner time by forming an end-to-end service-specific virtual network connecting users' devices and cloud-based resources. The *service-centric logical layer* gives a different view of the virtualized infrastructure from the service perspective. Having known the temporal and spatial demands, the service providers react to them by orchestrating virtualized network stretches, bringing the contents nearer, and creating a service-centric topology on top of the physical topology. This potentially reduces the bit pipe lengths between service providers and consumers, which in turn ensures much lower delays.

The communication within an IOLITE com-

munity at dinner time demands a network slice that provides high throughput and low latency. We now present how more resources are made available to an IOLITE community, and how the service on demand is shifted to the last mile. It should be mentioned that technological enablers do not work only in our flestic approach. Some of the challenges are solved by bringing many enablers together.

For a beyond 5G IOLITE community, radio resources are made available by small cells, while storage and processing resources are provided by MEC. The envisioned flexible infrastructure layer for an IOLITE community region consists of an H-CRAN cluster with a macrocell BS and many small cell BSs. To provide extra radio resources, inactive small cells are turned on dynamically and in a self-x (e.g., self-configuration, self-deployment, self-organization) fashion. The bit pipe is shortened by bringing the access point closer to the user or by directing the data traffic between users via device-to-device communication. Self-organization of device layer parameters (transmission power, sub-carrier spacing, etc.) deals with the trade-off between energy efficiency and spectral efficiency. In our approach, self-organization for device layer resource utilization (e.g., power adjustment) is to be achieved by distributed learning based on random matrix games, where each small cell acts as a gamer making their own decisions [14]. Random games are chosen for learning, as the pay-offs not only depend on the action of the gamer but also on the stochastic nature of beyond 5G networks. Moreover, integration of MEC to this architecture enables applications running at the edge with low latency, and at the same time reduces backhaul and core traffic of the network.

Pushing the on-demand service close to the last mile requires SDN control over the content layer instead of the switch-based topology of current control systems. This brings us to merging ICN with SDN. The ICN paradigm offers a new



distribution method in networks by decoupling information and location with named data objects (NDOs) that correspond to the role of IP in the traditional network [15]. When ICN and SDN concepts are merged, SDN controllers should detect ICN requests and reroute the traffic for ICN request-response pairs instead of finding IP servers. Integration of ICN, SDN, and H-CRAN architecture for dense networks is proposed in [15], where the logically centralized BBU pool of H-CRAN obtains the network-wide view to decide on request-response pairs. The integration of SDN and H-CRAN in the ICN domain creates a form of smart traffic, in which closer communication between the demanding user and the content provider can be established with offloading, and the traffic can be cached closer to demand so that the user does not have to go through the content servers.

## DISCUSSION AND CONCLUDING REMARKS

Realization of IoT on a large scale depends on its integration with cloud computing. Limited energy, storage, and computing capabilities mean that massive data produced by devices are only of use when the resources of cloud technologies are deployed together with IoT. In this article, we highlight the limitations of current networks and the required changes for next generation networks over a smart kitchen application.

Enhancing the capacity of next generation wireless networks is essential but not enough to meet cloud-IoT requirements. Smart network functions are needed to complement the radio technologies to generate a flexible network architecture. For this reason, resource virtualization techniques and dynamic management enablers have to be integrated to form a flexible network. In this article, we have given examples from the literature explaining how technological enablers can play their role in establishing flexible networks.

The IOLITE smart kitchen application is chosen as a beyond 5G example. In order to realize this application in the dinner time use case, we discuss a high-level approach that forms a network slice with an infrastructure layer consisting of dynamically reconfigurable radio and cloud resources and a logical topology managed by an ICN-based SDN controller. We do not offer a concrete solution but a hint on the road map toward the beyond 5G phase.

## ACKNOWLEDGMENTS

The work of the authors at GT-ARC is funded in part by the German Federal Ministry of Education and Research within the ISCO Project. The work of the authors at DAI-Labor is funded in part through the Huawei Innovation Research Program (HIRP).

## REFERENCES

- [1] S. Y. Lien *et al.*, "Ultra-Low-Latency Ubiquitous Connections in Heterogeneous Cloud Radio Access Networks," *IEEE Wireless Commun.*, vol. 22, no. 3, June 2015, pp. 22–31.
- [2] R. L. Aguiar, "White Paper for Research Beyond 5g," <http://networkd2020.eu/wp-content/uploads/2015/11/B5G-Vision-for-Researchv-1.0-for-public-consultation.pdf>, 2015; accessed Sept. 13, 2016.
- [3] P. Rysavy, "Challenges and Considerations in Defining Spectrum Efficiency," *Proc. IEEE*, vol. 102, Mar. 2014, pp. 386–92.

- [4] N. Bhusan *et al.*, "Network Densification: The Dominant Theme for Wireless Evolution into 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 82–89.
- [5] S. Rangan, T. Rappaport, and E. Erkip, "Millimeter-Wave Cellular Wireless Networks: Potentials and Challenges," *Proc. IEEE*, vol. 102, Mar. 2014, pp. 366–85.
- [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] B. Nunes *et al.*, "A Survey of Software-Defined Networking: Past, Present, and Future of Programmable Networks," *IEEE Commun. Surveys & Tutorials*, vol. 16, 3rd qtr. 2014, pp. 1617–34.
- [8] A. Checko *et al.*, "Cloud RAN for Mobile Networks Technology Overview," *IEEE Commun. Surveys & Tutorials*, vol. 17, 1st qtr. 2015, pp. 405–26.
- [9] M. Arslan, K. Sundaresan, and S. Rangarajan, "Software-Defined Networking in Cellular Radio Access Networks: Potential and Challenges," *IEEE Commun. Mag.*, vol. 53, no. 1, Jan. 2015, pp. 150–56.
- [10] D. Bojic *et al.*, "Advanced Wireless and Optical Technologies for Small-Cell Mobile Backhaul with Dynamic Software-Defined Management," *IEEE Commun. Mag.*, vol. 51, no. 9, Sept. 2013, pp. 86–93.
- [11] K. Pentikousis, Y. Wang, and W. Hu, "Mobileflow: Toward Software-Defined Mobile Networks," *IEEE Commun. Mag.*, vol. 51, no. 7, July 2013, pp. 44–53.
- [12] O. Aliu *et al.*, "A Survey of Self Organisation in Future Cellular Networks," *IEEE Commun. Surveys & Tutorials*, vol. 15, 1st qtr. 2013, pp. 336–61.
- [13] O. Salman *et al.*, "Edge Computing Enabling the Internet of Things," *2015 IEEE 2nd World Forum on Internet of Things*, Dec 2015, pp. 603–08.
- [14] M. A. Khan and H. Tembine, "Random Matrix Games in Wireless Networks," *2012 IEEE Global High Tech Congress on Electronics*, Nov. 2012, pp. 81–86.
- [15] C. Yang *et al.*, "When ICN Meets C-RAN for Hetnets: an SDN Approach," *IEEE Commun. Mag.*, vol. 53, no. 11, Nov. 2015, pp. 118–25.

## BIOGRAPHIES

DORUK SAHINEL ([doruk.sahinel@gt-arc.com](mailto:doruk.sahinel@gt-arc.com)) received his Master's degree in information and communications systems from Hamburg University of Technology, Germany, and his Bachelor's degree in electrical and electronics engineering from Middle East Technical University, Turkey. He is currently working as a researcher at the German-Turkish Advanced Research Center for ICT (GT-ARC). His main areas of research include mobility management in next-generation networks, software defined networking, and network virtualization.

CEM AKPOLAT ([cem.akpolat@gt-arc.com](mailto:cem.akpolat@gt-arc.com)) received his Master's degree in computer science from the Technical University of Berlin (TU Berlin), Germany, and his Bachelor's degree in computer science from Galatasaray University, Turkey. He is currently working as a researcher at GT-ARC. His main areas of research include IoT, software defined networking, and network virtualization.

MANZOOR A. KHAN ([manzoor-ahmed.khan@dai-labor.de](mailto:manzoor-ahmed.khan@dai-labor.de)) received his Ph.D. degree in computer science from TU-Berlin. He is currently the co-director of the competence center Network and Mobility at DAI-Labor, TU Berlin. His research focuses mainly on cloud computing, SDN, mobile networks, learning in agent based autonomic networking, user-centric networking, and QoE.

FIKRET SIVRIKAYA ([fikret.sivrikaya@gt-arc.com](mailto:fikret.sivrikaya@gt-arc.com)) received his Ph.D. degree in computer science from Rensselaer Polytechnic Institute, New York, and his Bachelor's degree in computer engineering from Bogazici University, Istanbul, Turkey. He is currently the scientific coordinator at GT-ARC and co-director of the Network and Mobility group at DAI-Labor, TU Berlin. His research interests include future mobile networks, user-centric networking, wireless communications, vehicular networks, multihop ad hoc networks, distributed algorithms, and optimization.

SAHIN ALBAYRAK ([sahin.albayrak@dai-labor.de](mailto:sahin.albayrak@dai-labor.de)) is a professor and head of the chair of Agent Technologies in Business Applications and Telecommunication (AOT) at TU Berlin. He is the founder of DAI-Labor, Deutsche Telekom Innovation Laboratories, and the founding director of the Connected Living Association as well as GT-ARC. His research interests cover next generation telecommunication services and infrastructures, service-centric architectures, autonomous systems, agent-based architectures, smart home and smart cities, preventive health, and cybersecurity.

The integration of SDN and H-CRAN in the ICN domain creates a form of smart traffic, in which closer communication between the demanding user and the content provider can be established with offloading, and the traffic can be cached closer to demand so that the user does not have to go through the content servers.

# Federated Internet of Things and Cloud Computing Pervasive Patient Health Monitoring System

Jemal H. Abawajy and Mohammad Mehedi Hassan

The authors present a pervasive patient health monitoring (PPHM) system infrastructure. PPHM is based on integrated cloud computing and Internet of Things technologies. In order to demonstrate the suitability of the proposed PPHM infrastructure, a case study for real-time monitoring of a patient suffering from congestive heart failure using ECG is presented.

## ABSTRACT

The exponentially growing healthcare costs coupled with the increasing interest of patients in receiving care in the comfort of their own homes have prompted a serious need to revolutionize healthcare systems. This has prompted active research in the development of solutions that enable healthcare providers to remotely monitor and evaluate the health of patients in the comfort of their residences. However, existing works lack flexibility, scalability, and energy efficiency. This article presents a pervasive patient health monitoring (PPHM) system infrastructure. PPHM is based on integrated cloud computing and Internet of Things technologies. In order to demonstrate the suitability of the proposed PPHM infrastructure, a case study for real-time monitoring of a patient suffering from congestive heart failure using ECG is presented. Experimental evaluation of the proposed PPHM infrastructure shows that PPHM is a flexible, scalable, and energy-efficient remote patient health monitoring system.

## INTRODUCTION

Healthcare costs in many countries are increasing at an unsustainable rate. In the United States, for instance, healthcare spending is expected to be \$4.8 trillion in 2021, which is close to 20 percent of gross domestic product [1]. Factors accounting for the increasing healthcare spending include chronic diseases, waste, and inefficiencies such as over-treatment, and redundant, inappropriate, or unnecessary tests and procedures. In addition, advances in medicine over the last decades have significantly increased the average life expectancy while simultaneously decreasing the rate of mortality substantially. As a result, the number of elderly people has been rising constantly, which is placing a strain on the healthcare services. The need to bring healthcare costs into a sustainable range is an urgent issue that needs to be addressed [2].

One possible way to address the challenges facing the healthcare industry is by caring for patients in their environments such as their residences. A lot of patient categories such as those with chronic disease who need only therapeutic supervision, elderly patients, and patients with congenital heart defects do not need to use a hos-

pital bed as they can be cared for in their homes [2–4]. The challenge, however, is how healthcare professionals can accurately, reliably, and securely monitor the health status of their patients without physically visiting them at their residences. The system must be able to facilitate patient mobility, while at the same time improve their safety and increase their autonomy.

This study addresses this challenge by augmenting existing healthcare systems with inexpensive but flexible and scalable pervasive technologies that enable long-term remote patient health status monitoring. Recent advances in the Internet of Things (IoT) [12] and cloud computing (CC) [13] have made it practically possible to transform the healthcare sector. As the healthcare system increasingly values efficiency and outcomes, the adoption and diffusion of IoT and cloud can play a significant role in arresting the spiraling healthcare costs without impacting the quality of care provided to patients [4]. Although the integration of IoT and CC would be a great innovation in contemporary medical applications [7], remote patient health status monitoring systems that integrate IoT and CC have received less attention [4]. Therefore, despite all of the possibilities that IoT and CC technologies offer, there are some significant obstacles that need to be overcome before their full potential can be realized [9].

In this article, we propose a remote pervasive patient health monitoring (PPHM) framework. The proposed framework leverages the combined strong synergy of IoT, CC, and wireless technologies for efficient and high-quality remote patient health status monitoring. The article makes the following contributions:

- A flexible, energy-efficient, and scalable remote patient health status monitoring framework
- A health data clustering and classification mechanism to enable good patient care
- A case study where the capabilities of the PPHM framework are exploited for patients with heart disease
- Performance analysis of the PPHM framework to show its effectiveness

The rest of the article is organized as follows. First, we provide related work. We then present the proposed cloud and IoT integrated remote health status monitoring framework. Next, we

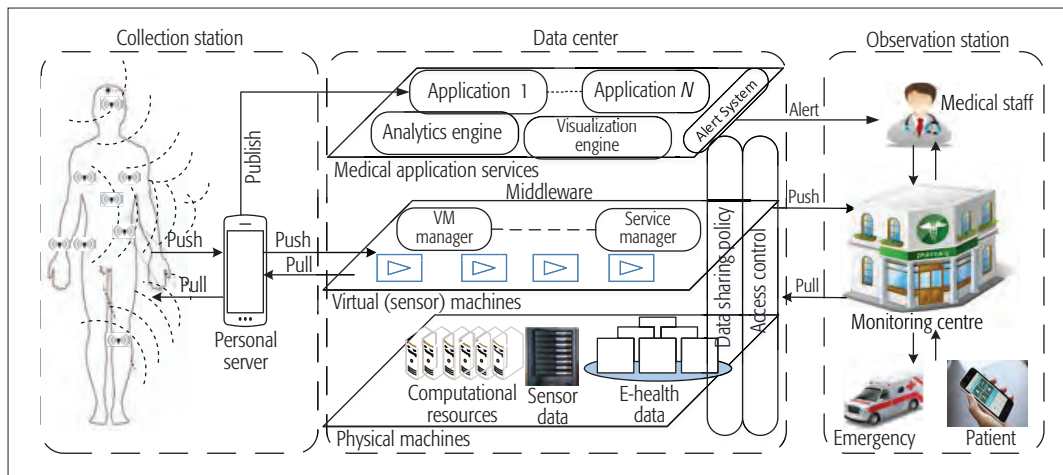


Figure 1. Internet of Things and cloud-based architecture for remote healthcare monitoring.

present an ECG process analysis using the proposed monitoring system. We report performance analysis of the PPHM framework. Finally, the conclusion is outlined.

## RELATED WORK

The question addressed in this article is how to remotely monitor and evaluate the health of patients in the comfort of their own homes. Integrating IoT and CC for patient health and activity monitoring has been an active research area lately. A complex framework that encompasses several health ecosystems, where data from the sensors is watermarked for security purposes and transmitted to the cloud for feature extraction and classification is discussed in [2]. One-class support vector machine classification is used in the framework to classify an ECG as abnormal or not. A privacy-preserving data collection and secure transmission framework is presented in [6]. BodyCloud [7] is a three-tier integrated software as a service (SaaS)-based cloud and body sensor networks (BSNs) architecture that enables the development and deployment of cloud-assisted BSN applications. A mobile healthcare system for wheelchairs that exploits BodyCloud components is discussed in [10]. The framework in [5] integrates TCP/IP and Zigbee for interoperability in the coordinator devices. The framework discussed in [8] is designed to perform diagnosis of chronic illnesses such as diabetes. Patient data are collected through body sensors and stored in the cloud for subsequent analysis and classification. This client-server model framework does not consider energy consumption. In the architecture proposed in [9], patient data is transferred through the home gateway to the cloud, where it is processed and then made available to healthcare professionals or patients. How this is done is not really explained.

Our work is motivated by these previous works and complements them in many ways. As in [7], our PPHM framework is three-tiered with push-pull communication between the three tiers. Thus, in our model, an authorized healthcare professional can request and obtain the real-time data collected by a particular sensor in an IoT subsystem. This capability is generally absent from these works. As in [2], our framework integrates data analytics based on our prior work

[11, 13]. Unlike [2], we use data clustering and classification mechanisms to improve classification accuracy. We also consider optimization of the communication and energy consumption at all levels of the system. Unlike the previous studies, we assume that the cloud is used by many competing applications, and proper service provisioning is used to allocate cloud resources to the competing applications.

## REMOTE HEALTH STATUS MONITORING FRAMEWORK

This section describes the general three-tier architecture of the proposed PPHM framework shown in Fig. 1. In the following subsections, we explain the major components of the framework.

### OBSERVATION STATION

The observation station consists of an IoT subsystem that is tasked with remote physiological and activity monitoring of patients. The core monitoring infrastructure of the IoT subsystem is the wireless BSNs. This subsystem contains a set of  $n$  BSNs,  $B = \{b_1, \dots, b_n\}$ . Each  $b_i \in B$  represents a patient and is defined as  $b_i = \langle S, P \rangle$ , where  $P$  is a personal server and  $S = \{s_1, \dots, s_m\}$  is a set of  $m$  energy-constrained lightweight wireless sensor nodes. Each sensor  $s_i \in S$  has enough capability to collect patient data, aggregate it, perform basic processing, and transmit it to a personal server for further processing. These sensors can be implantable, worn or attached, to everyday objects such as clothes unobtrusively to gather specific physiological parameters such as a patient's blood sugar levels, blood glucose, capnography (i.e.,  $\text{CO}_2$  level and breathing), and pulse oximetry and ECG continuously or on demand. Continuous monitoring is performed when intensive monitoring is needed for patients. In this case, sensors continuously collect vital data and send it to the personal server. The on-demand monitoring occurs when a request from any authorized person within the system, such as a patient, doctor, or nurse, is generated.

The personal server provides a link between the IoT subsystem and the cloud infrastructure. The personal server is a dedicated per-patient machine (e.g., a tablet or smartphone) with built-in features such as a GPS module, Bluetooth radio

Continuous monitoring is performed when an intensive monitoring is needed for patients. In this case, sensors continuously collect vital data and send it to the personal server. The on-demand monitoring occurs when a request from any authorized person within the system such as patients, doctors or nurses is generated.

The cloud also hosts the middleware system, virtual sensors, and application services that allow medical staff to analyze and visualize patients' data as well as to identify and raise alerts when events requiring urgent intervention are observed.

module, and SQLite database. We assume that the personal server can compatibly interact with various local networks such as WiFi and LTE [4]. Each sensor within a given BSN is wirelessly connected via a single hop to a dedicated personal server. We assume that the default communication between the sensor nodes to a personal server is via Bluetooth. The personal server receives a stream of sensor data from the sensors. It performs basic data analysis and aggregation, generating alarm signals, making the data available to the entities subscribed to be notified (e.g., patient), or pushing the data (along with the location of the patient) out to the cloud for further analysis and sharing by healthcare professionals. In order to manage bandwidth and energy consumption, a fuzzy-based data fusion technique that distinguishes and aggregates only the true values of the sensed data [14] is used. This method decreases the processing and transmission of the sensed data as well as removes redundant data, thus minimizing energy depletion while prolonging the network lifetime. In addition to transferring data from the sensors to the cloud, the personal server can possibly receive a request for specific data from cloud applications or an end user.

#### DATA CENTER SUBSYSTEM

The cloud relieves the IoT subsystem by performing heavy functions that require storing, processing, and analyzing the collected patient health data from the IoT subsystem. Cloud storage offers benefits of scalability and accessibility on demand at any time from any place. The healthcare provider data center hosts the cloud subsystem, which delivers storage resources and provides computational capability for analyzing and processing of the collected data. The cloud also hosts the middleware system, virtual sensors, and application services that allow medical staff to analyze and visualize patients' data as well as to identify and raise alerts when events requiring urgent intervention are observed. The major components of the cloud subsystem are described below.

**Patient Data Storage:** The cloud storage resources are used for long-term storage of patients' medical information (e-Health) and the data from the IoT subsystem (sensory data). E-Health contains the conventional clinical data (e.g., clinic observation and lab test results) while the sensory data contains longitudinal patient data provided by BSNs. Based on the access control configuration, healthcare practitioners or emergency centers can access the stored information without visiting the patient. The physicians, having access to the sensory data along with the e-Health data supported by decision support systems, can improve the quality of patient health in remote locations by making better and quicker prognoses, intervention, and treatment recommendations.

**Health Data Sharing Policy (HDSP):** One of the aims of the healthcare service providers for collecting clinical data from patients is to share them with authorized healthcare professionals. As data security and privacy are important issues in healthcare systems [2], we use an access control mechanism (e.g., signature or certificate) that ensures only legitimate end users can access the

data in the cloud. We also use policy to control the sharing of data. HDSP governs how the patient data is shared among the authorized entities and used to verify the identity of the user with access authority. For instance, the policy can define that access to the sensor reading in the sensor data storage and the corresponding analysis results can only be accessed by the doctors in the neurology department. HDSP also ensures that patient unique identities and associated profiles should be anonymized before the data is shared with other entities such as a research center. In the proposed framework, the data monitoring unit is responsible for setting up the HDSP taking into account regulatory compliance requirements and the need for sharing to provide the best possible care for the patient.

**Cloud Middleware:** The middleware consists of a virtual machine (VM) manager and a service scheduler, among others. The VM manager is responsible for managing the virtual sensors, which are virtualized counterparts of physical sensors in BSNs, collecting sensor data from personal servers, and storing those data in the "sensor data" store. As compared to the standard cloud workloads such as non-real-time data for scientific computation and storage, the workload from the IoT subsystem is characterized by high inter-arrival rates and highly variant runtimes but with low parallelism. Thus, it becomes important to have cloud resource management and scheduling that can be adapted to handle such different workloads. Thus, service scheduling is necessary to properly schedule many real-time and non-real-time service requests to improve resource usage efficiency. Also, the scheduler performs dynamic load balancing and adaptive resource management in an energy-efficient manner.

**Medical Application Services:** The cloud hosts various services that process clinical data collected from the IoT subsystem for clinical observation and intervention, and to dispatch ambulances or notify family members of patients. The analytics engine (AE) extracts features from the collected data and classifies the data to assist healthcare professionals to facilitate good patient care. For the healthcare professionals to use the results from the AE to reach accurate and appropriate responses and actions, the output from AE will be used by the visualization engine to make the data accessible to the healthcare professionals in a readily digestible format. The alert system raises alert signals when events requiring urgent interventions are observed. The alarm function generates alerts if the value of the sensed physiological parameters exceeds a predetermined threshold value. For example, an alarm signal is generated when abnormalities such as arrhythmia or hypotension are detected. This capability enables patient health problems to be detected without visiting a doctor, notifying healthcare providers if a check-up is needed, and generating emergency alerts to ambulances.

#### OBSERVATION STATION

The observation station is where data-driven clinical observation and intervention take place. At this tier, entities such as healthcare professionals (e.g., doctors), emergency response services, medical research centers, and patients have

presence. The monitoring center involves the participation of many healthcare actors, including doctors, patients, and nursing staff, in clinical observation, patient diagnosis, and intervention processes. Thus, all access requests for patient data are managed by the monitoring center. Any authorized user wanting to access the sensor data can do so by issuing a data request to the cloud through the monitoring center. If the requested data is available in the sensor data storage, the data will be returned to the user. Therefore, the healthcare professionals must have appropriate authentication and authorization credentials to access the data.

The framework also allows authorized users or applications to pull any missing or extra data on an on-demand basis from the personal server. The personal server will retrieve data from either its memory or a sensor node and send it to the end user or the application. Entities at this level can subscribe to the data service to be informed automatically when specific data or patterns are observed. For example, patients can subscribe to receive data for the purpose of self-health monitoring. This can happen, for example, after data analytics and health indicators, when the system provides medical advice to the user. In this case, the data is automatically published to the subscribers immediately when it becomes available. The patient can use data such as blood sugar levels to take appropriate actions in case of anomaly detection. Such knowledge-based decisions may lead to reduction in the number of visits to doctors, tests, and hospitalizations. It can also inform caregivers and emergency centers through SMS. The advantage of this service model is less network traffic and power consumption.

### CASE STUDY: CONGESTIVE HEART FAILURE

In this case study, we consider a patient suffering from congestive heart failure (CHF) requiring care on a regular basis at her home. CHF develops when the heart's blood pumping ability weakens due to factors such as coronary heart disease, hypertension, and arrhythmia [11]. The cardiac activity of the patient is monitored via ECG, which is a non-invasive diagnostic method for monitoring and detecting a range of heart diseases.

Figure 2 illustrates the proposed framework for remote patient monitoring. In the example, a physician initiates the monitoring and the execution of the data analysis processes. The physician can define the start and end times of the monitoring period. The monitoring center goes through the setup process, which includes confirmation that the requesting agent is an authorized individual, the setup of the personal server that is capable of collecting, aggregating, and sending patient data through the Internet based on patient location (e.g., home, hospital, or outdoors), the registration of the personal server, the initiating doctor, thresholds to be checked for alert initiation, and the exchange of encryption keys between the cloud and the personal server. In addition, the type of monitoring service (continuous or on request) needs to be selected by the physician.

After the setup step is completed, the IoT subsystem starts gathering key physiological parameters and forwarding the data to the personal server, where the data is aggregated and

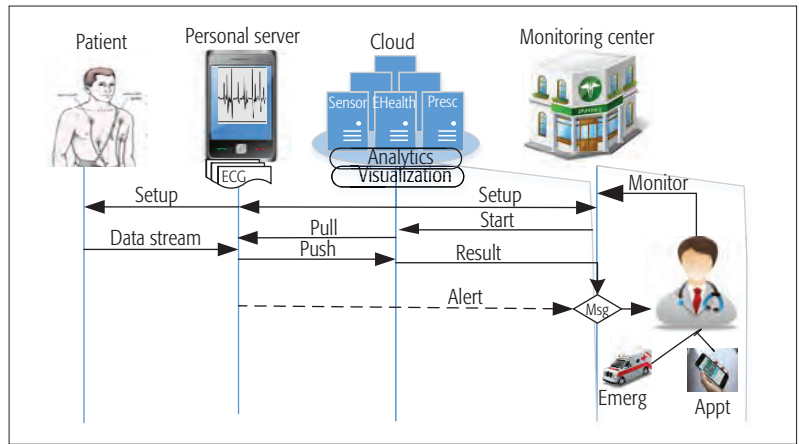


Figure 2. The PPHM framework monitoring process.

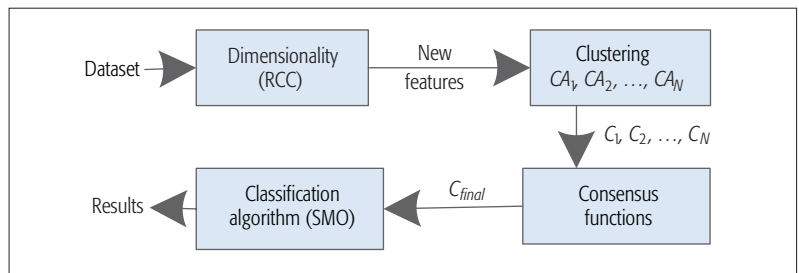


Figure 3. ECG data classification.

relayed to the sensor data storage linked to the patient e-Health records in the cloud system. In conventional settings, the physicians examine the ECG recordings visually for important features. As manual inspection of ECG heartbeats can lead to inaccurate decisions [11], automatic classification of the ECG signals is important for clinical diagnosis of various heart diseases. However, the ECG dataset is highly dimensional, large, and noisy in nature. To address this problem, we use an approach that combines feature reduction, consensus clustering, and classification algorithms for ECG data profiling. Figure 3 shows the components of the multistage system model.

The ECG dataset is processed by the dimensionality reduction algorithm, which is the rank correlation coefficient (RCC) algorithm to obtain fewer features that effectively capture the behavior of the ECG signals. The output from RCC is fed into a set of unsupervised clustering algorithms (i.e., Cobweb, Expectation Maximization, Farthest First, and Simple  $K$ -Means) algorithms. This step generates a set of  $n$  independent clusters  $C = \{C_1, C_2, \dots, C_n\}$ . We used the hybrid bipartite graph formulation (HBGF) consensus function to combine the  $C$  clusters and produce a final consensus cluster ( $C_{final}$ ). HBGF is based on a bipartite graph, and the  $C_{final}$  is determined by the way HBGF partitions all elements of the data set. Finally, we used sequential minimal optimization (SMO) with a polynomial kernel supervised classification algorithm to classify the dataset.

### PERFORMANCE EVALUATION

In this section, we evaluate the proposed framework using an emulator-based approach [7, 15] on real ECG signals from the BIDMC Congestive Heart Failure Database (CHFD). The CHFD

dataset contains ECG recordings from 15 subjects with severe congestive heart failure. The individual recordings are each approximately 20 h in duration. They contain two main ECG signals, each sampled at 250 samples/s with 12-bit resolution over a range of  $\pm 10$  mV. As in [15], we use the ECG Sensor Emulator, implemented in Matlab, to generate an ECG data stream by converting each ECG sample from the CHFD dataset to a series of pairs of 16-bit frames and transfer them to the personal server over Bluetooth.

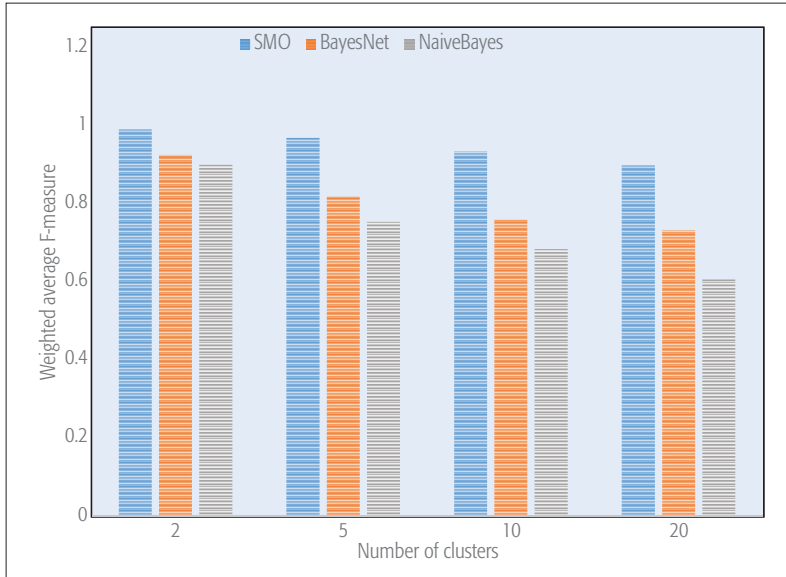


Figure 4. Classification accuracy.

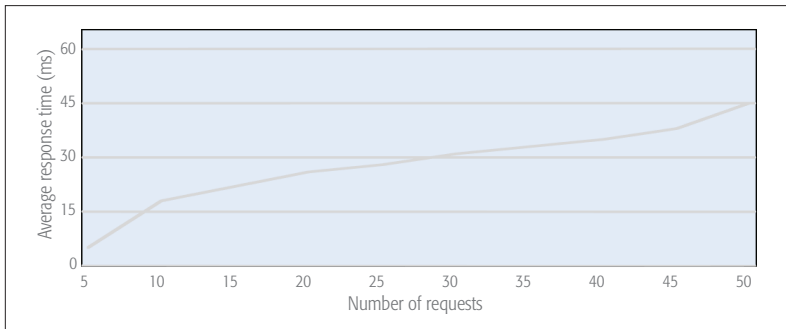


Figure 5. System scalability.

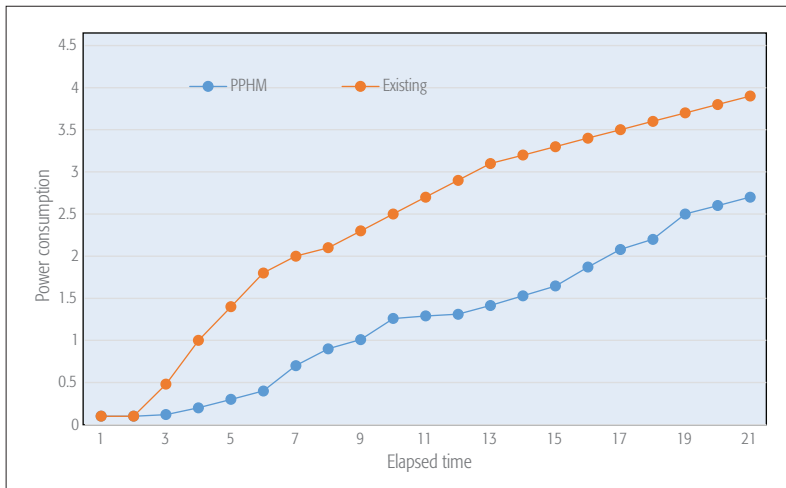


Figure 6. Commutative power consumption.

## ECG CLASSIFICATION

We studied the effectiveness of the proposed classification scheme using the weighted average F-measure. We used 10-fold cross validation and compared the SMO-based classification algorithms with the Bayes Network Learning (BayesNet) and Classical Naive Bayes (NaiveBayes) algorithms.

Figure 4 shows the performance (weighted average F-measure) of the three classification algorithms as a function of the number of clusters after their training on the initial consensus clustering data. In [2], the accuracy obtained was 87.7 percent with MIT-BIH database and 90.4 percent with a private database. In our case, we achieved 89.7 percent with 20 clusters and 98.9 percent with 2 clusters. The results establish that our classification algorithms achieve high accuracy with the SMO-based classifier achieving the best results. The result also demonstrates that the SMO-based classifier scales up much better as the number of clusters increases. These algorithms can be used in practical implementations for profiling of highly dimensional, noisy, and large ECG datasets.

## SCALABILITY ANALYSIS

To study the scalability of the system, we emulated a set of clients that concurrently transmit sensor data stream as in [7]. We model the request inter-arrival time as a Poisson process, while the service demand is randomly selected between 1 to 5 ms. We repeated the experiment 1000 times and took the average result. Figure 5 shows the average response time as the number of simultaneous requests vary. As the number of requests increase, we can see that the response time increases linearly.

## ENERGY CONSUMPTION

In order to study the energy consumption effectiveness of the proposed PPHM framework, we model energy consumption for sensing, computation, and transmission of the messages for a period of time and check the level of the energy usage. We send a  $b$ -bit message over a distance  $d$  as  $((E_{elec} + b) + (\epsilon_{amp} + b + d^2))$  and receive this message as  $(E_{elec} + b)$ . The  $E_{elec} = 50$  nJ/bit is the energy dissipated to run the transmitter or receiver circuitry, and  $\epsilon_{amp} = 0.1$  nJ/bit is the transmit amplifier. The initial energies of each sensor node is fixed at 1.0 J.

Figure 6 shows the cumulative power consumption as a function of the elapsed time. As the existing framework does not deploy any optimizations, it dissipates energy faster than our framework. In contrast, we deploy optimization techniques such as the fuzzy-based data fusion method to manage bandwidth and energy consumption. This method is able to decrease the transmission and the processing of the sensed data as well as remove redundant data, thus minimizing energy consumption while increasing the network lifetime.

## CONCLUSIONS

In the conventional hospital-centric healthcare system, patients are often tethered to several monitors. In this article, we develop an inexpensive but flexible and scalable remote health status monitoring system that integrates the capabilities of the IoT and cloud technologies for remote monitoring

of a patient's health status. Through experimental analysis, we have shown that the proposed framework is scalable and energy-efficient with very high classification accuracy. We believe that the proposed work can address the healthcare spending challenges by substantially reducing inefficiency and waste as well as enabling patients to stay in their own homes and get the same or better care. We are currently implementing the proposed algorithm and testing it in a real-life environment. We are also extending the proposed work to include the privacy and security aspects.

#### ACKNOWLEDGMENT

We extend our appreciation to Maliha Omar and the anonymous reviewers. We also appreciate the Deanship of Scientific Research at King Saud University for its funding of this research through the research group project no. RGP-281.

#### REFERENCES

[1] www.cms.gov, "Centers for Medicare and Medicaid Services, National Health Expenditures Projections 2011–2021"; <https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/nationalhealthexpenddata/downloads/proj2012.pdf>, accessed: 04/10/2016.

[2] M. S. Hossain and G. Muhammad, "Cloud-Assisted Industrial Internet of Things (IIoT) – Enabled Framework for Health Monitoring," *Computer Networks*, vol. 101, June 2016, pp. 192–202.

[3] B. Townsend and J. Abawajy, "Security Considerations for Wireless Carrier Agonistic Bio-Monitoring Systems," *Security and Privacy in Commun. Networks*, vol. 164, Jan. 2016, pp. 725–37.

[4] S. Ghanavati, J. Abawajy, and D. Izadi, "An Alternative Sensor Cloud Architecture for Remote Patient HealthCare Monitoring and Analysis," *2016 IEEE Int'l. Joint Conf. Neural Networks*, 24–29 July 2016, Vancouver, Canada.

[5] M. M. Hassan et al., "A Multimedia Healthcare Data Sharing Approach through Cloud-Based Body Area Network," *Future Generation Computer Systems*, vol. 66, Jan. 2017, pp. 48–58.

[6] K. Zhang et al., "Security and Privacy for Mobile Healthcare Networks: From a Quality of Protection Perspective," *IEEE Wireless Commun. Mag.*, vol. 22, no. 4, Aug. 2015, pp. 104–12.

[7] G. Fortino, D. Parisi, and V. Pirrone, "BodyCloud: A SaaS Approach for Community Body Sensor Networks," *Future Generation Computer Systems*, vol. 35, June 2014, pp. 62–79.

[8] P. D. Kaur and I. Chana, "Cloud Based Intelligent System for Delivering Health Care as a Service," *Computer Methods and Programs in Biomedicine*, vol. 113, no. 1, Jan. 2014, pp. 346–59.

[9] S. Luo, and B. Ren, "The Monitoring and Managing Application of Cloud Computing Based on Internet of Things," *Computer Methods and Programs in Biomedicine*, vol. 130, July 2016, pp. 154–61.

[10] L. Yang et al., "People-Centric Service for mHealth of Wheelchair Users in Smart Cities," *Internet of Things Based on Smart Objects*, Apr. 2014, pp. 163–79.

[11] J. H. Abawajy, A. V. Kelarev, and M. Chowdhury, "Multistage Approach for Clustering and Classification of ECG Data," *Computer Methods and Programs in Biomedicine*, vol. 112, no. 3, Dec. 2013, pp. 720–30.

[12] B. R. Ray, M. U. Chowdhury, and J. H. Abawajy, "Secure Object Tracking Protocol for the Internet of Things," *IEEE Internet of Things J.*, vol. 3, no. 4, May, 2016, pp. 544–53.

[13] M. Chowdhury et al., "A Clustering-Based Multi-Layer Distributed Ensemble for Neurological Diagnostics in Cloud Services," *IEEE Trans. Cloud Computing*, online, DOI: 10.1109/TCC.2016.2567389, 2016.

[14] D. Izadi et al., "A Data Fusion Method in Wireless Sensor Networks," *Sensors*, vol. 15, no.2, Jan. 2015, pp. 2964–79.

[15] A. Secerbegovic et al., "The Research mHealth Platform for ECG Monitoring," *Proc. 11th IEEE Int'l. Conf. Telecommun.*, 15–17 June 2011, Graz, Austria, pp. 103–08.

#### BIOGRAPHIES

JEMAL H. ABAWAJY [SM'11] (jemaal@deakin.edu.au) is a full professor at Deakin University, Australia, where he leads the distributed computing and security research group. He has served over 300 conferences in various capacities and on the Editorial Boards of several journals. He has published more than 300 refereed articles and given more than 50 invited/keynote presentations.

MOHAMMAD MEHEDI HASSAN [M'12] (mmhassan@ksu.edu.sa) is an assistant professor in the College of Computer and Information Sciences (CCIS), King Saud University (KSU), Riyadh, Kingdom of Saudi Arabia. He has published over 100 + research papers in journals and conferences of international repute. He received the Excellence in Research Award from CCIS, KSU in 2014 and 2015. His research areas of interest are cloud federation, multimedia cloud, sensor cloud, the Internet of Things, big data, and mobile cloud.

We believe that the proposed work can address the healthcare spending challenges by substantially reducing inefficiency and waste as well as enabling patients to stay in their own homes and get the same or better care. We are currently implementing the proposed algorithm and testing it in a real-life environment.

# Wearable 2.0: Enabling Human-Cloud Integration in Next Generation Healthcare Systems

Min Chen, Yujun Ma, Yong Li, Di Wu, Yin Zhang, and Chan-Hyun Youn

The authors propose a Wearable 2.0 healthcare system to improve QoE and QoS of next generation healthcare systems. In the proposed system, washable smart clothing, which consists of sensors, electrodes, and wires, is the critical component to collect users' physiological data and receive the analysis results of users' health and emotional status provided by cloud-based machine intelligence.

## ABSTRACT

With the rapid development of the Internet of Things, cloud computing, and big data, more comprehensive and powerful applications become available. Meanwhile, people pay more attention to higher QoE and QoS in a "terminal-cloud" integrated system. Specifically, both advanced terminal technologies (e.g., smart clothing) and advanced cloud technologies (e.g., big data analytics and cognitive computing in clouds) are expected to provide people with more reliable and intelligent services. Therefore, in this article we propose a Wearable 2.0 healthcare system to improve QoE and QoS of the next generation healthcare system. In the proposed system, washable smart clothing, which consists of sensors, electrodes, and wires, is the critical component to collect users' physiological data and receive the analysis results of users' health and emotional status provided by cloud-based machine intelligence.

## INTRODUCTION

Due to the ever growing number of elderly people coupled with limited resources in terms of medical facilities and personnel in many countries, the burden that conventional healthcare systems carry is becoming heavy. On the other hand, traditional human face-to-face communications are mostly replaced by networking in social and cyber spaces, which causes various unhealthy living habits, such as insufficient physical exercise, unhealthy diet, irregular sleeping, and more frequent "burning the midnight oil." All these factors are usually the keys to triggering chronic diseases, including cardiovascular disease, hyperlipemia, diabetes, tumor, obesity, and chronic respiratory disease. As reported in "The Leading Causes of Death and Disability in the United States,"<sup>1</sup> 50 percent of people in the United States suffer from one or more kinds of chronic diseases at different levels, while 80 percent of medical funds in the United States are used to treat chronic disease. In 2015, the United States spent \$2.7 trillion on chronic disease treatment, which accounts for 18 percent of the U.S. gross domestic product.

Therefore, it is a great challenge to design a cost-effective healthcare system for handling chronic diseases, especially considering the large population of elderly people and empty nesters,

most of whom suffer from one or more chronic diseases. To address this issue, we should lower the operating cost and improve their scalability of healthcare systems, which are expected to provide various basic healthcare services [1], such as physiological monitoring, early warning via abnormal vital signs, and online patient consultations [2]. Fortunately, with the assistance of cloud computing and big data, various advanced services become possible by the use of big data analysis for chronic disease detection and intelligent health monitoring [3]. However, it is difficult to solve the following undesirable issues in existing healthcare systems.

**Physiological Data Collection:** Without considering users' mobility, traditional physiological data collection is uncomfortable for users because they must wear multiple sensors or related devices. On the other hand, if only simple devices (e.g., smart bracelet and smart watch) are carried, the collected data is inaccurate. Thus, how to accurately collect sufficient physiological data in a comfortable way is still an unsolved problem..

**Negative Psychological Effects:** Besides uncomfotability, users might feel that they have health problems when they wear body sensors, which further causes stress and other negative emotions. More seriously, the negative psychological effects will result in some mental illness, especially when patients feel lonely or depressed. Hence, we need to rethink the design for an innovative healthcare system in terms of physiological data collection in a more comfortable, energy-efficient, and sustainable way [4–6]:

**Sustainable Big Physiological Data Collection:** Nowadays, wearable techniques are widely accepted in the market, including smart watches, smart bracelets, wearable sleep aid devices, sport monitoring and promotion, and so on. Although users may not have observable discomfort while wearing such devices in their daily life, the collected physiological data are usually simple and cannot be called "big data," even for long-term monitoring. Thus, these insufficient data have limited reference value for chronic disease diagnosis. If medical-level data are needed, more complicated medical devices should be used. In this case, a user's normal life will be disturbed. For example,

<sup>1</sup> <http://www.cdc.gov/chronicdisease/overview>



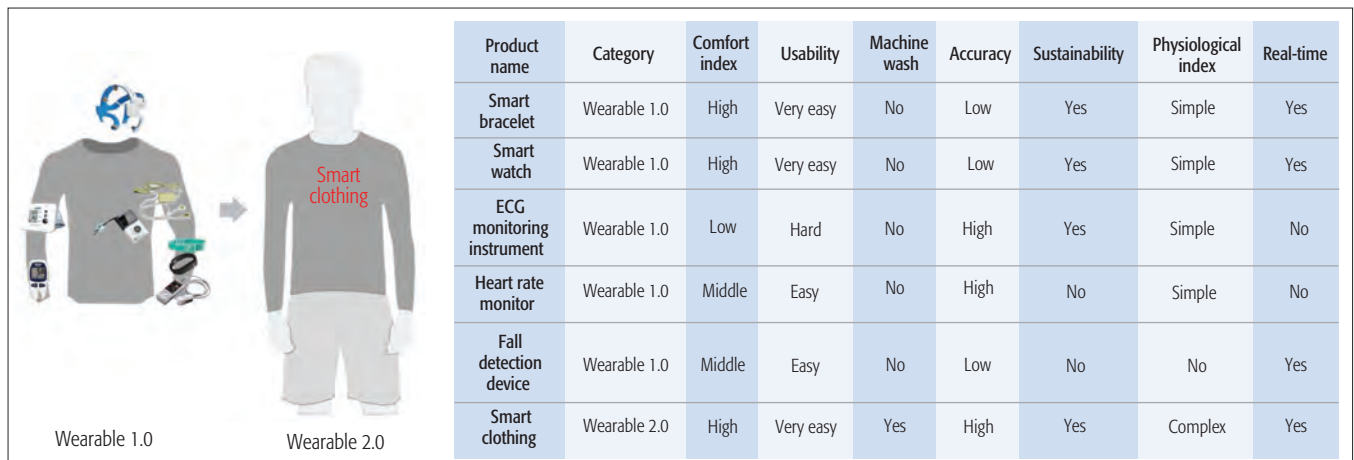


Figure 1. Motivation for proposing Wearable 2.0: a) comparison of a conventional wearable system and smart clothing; b) features of Wearable 1.0 and Wearable 2.0.

a portable electrocardiogram (ECG) monitoring device can be used for collecting detailed ECG curves, but long-term wearing is implausible, especially considering a user's mobility.

**Anti-Wireless for Body Area Networking:** In traditional body area networks (BANs), wireless technology is encouraged to replace wired cables [7]. However, there are various factors, such as user mobility and surrounding environment, that create wireless interference, bringing challenges to achieve high stability and accuracy during physiological data collection. On the other hand, the wireless bandwidth in BANs is limited to support medical-level data transmissions, while the energy consumption of wireless networking is also a main concern. In the next generation BANs, wireless should be avoided as much as possible for green communications.

To address the above challenges, this article investigates innovative wearable devices, especially washable smart clothing. When wearing these innovative wearable devices, users may not experience any different feeling compared to a normal T-shirt, as shown in Fig. 1a. The traditional wearable system is considered Wearable 1.0, where there are problems of insufficient data collection when carrying small or "uncomfortable wearing when professional devices are adopted [8, 9]. With continuous improvement of wearable techniques, Wearable 2.0 is proposed to incorporate multiple connected devices and cloud services that, together, offer more meaningful enhancements to users' lifestyles [10]. In Fig. 1b, we compare Wearable 1.0 with Wearable 2.0 in terms of comfort, usability, accuracy, washability, and support for real-time monitoring. Obviously, Wearable 2.0 is a good solution to the challenging issues in terms of sustainable data collection for health big data. Therefore, in this article we propose a Wearable 2.0-based healthcare services to improve quality of experience (QoE) and quality of service (QoS) in the next generation healthcare system.

In the remaining part of this article, we present the design and implementation issues of Wearable 2.0. The architecture of Wearable 2.0 is proposed. It describes a testbed of Wearable 2.0. Based on this testbed, applications of emotional care are discussed. We then conclude the article.

## DESIGN ISSUES OF WEARABLE 2.0

In order to integrate body sensors and cables with textile material perfectly, various issues need to be considered in Wearable 2.0 design, including details of the sensors, availability of the system, and the user experience. We classify the functional components of the smart clothing representing Wearable 2.0 into the following categories.

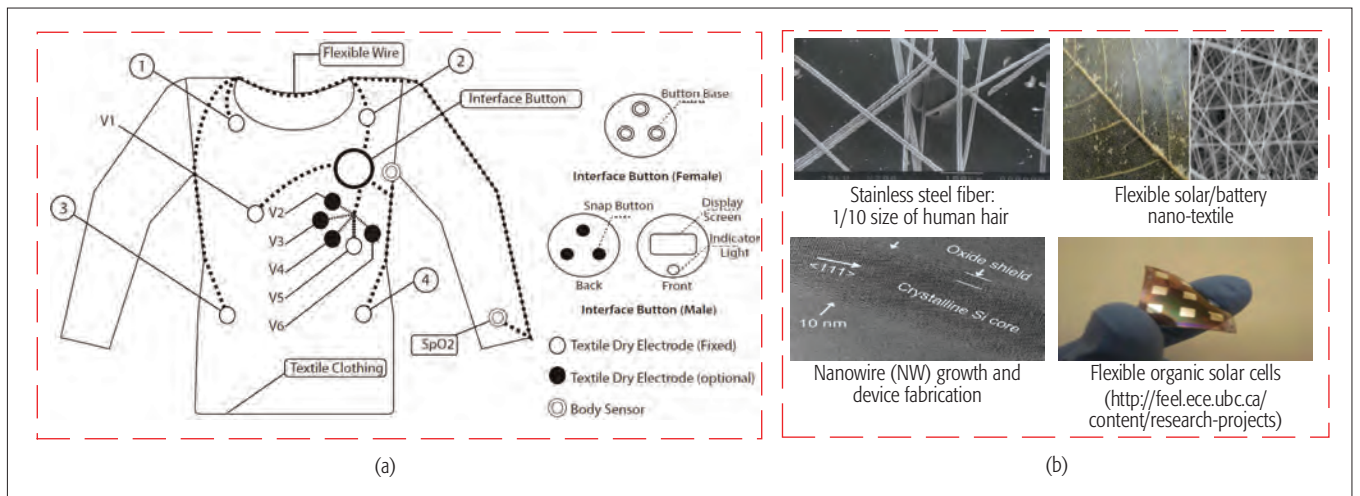
**Sensors Integration:** In our design, the pulse sensor is installed around the wrist; the body temperature sensor is put in the underarm seam; a set of ECG sensors are mounted on the chest, shoulders, and ribs; the myocardial sensor is embedded in the left part of the chest; and the SpO<sub>2</sub> sensor is deployed on the triceps of the left (or right) arm. As shown in Fig. 2b, more bio-sensors can be selected without the consideration of deployment cost.

**Electrical-Cable-Based Networking:** Generally, elastic textile cloth material should be selected for smart clothing. Furthermore, in order to guarantee comfort, flexible electrical cables should be embedded along the stitching of the cloth, as shown in Fig. 2a. In this way, the smart clothing is also more durable after it is washed repeatedly.

**Digital Modules:** Strictly speaking, digital modules do not belong to the components of smart clothing, since they are not washable. However, these add-on devices are closely attached to the smart clothing.

For example, the power module should be miniaturized or flexible, but it does not need to be an independent module, so the power module can be combined with the signal collection terminal (i.e., Smart Box), and provide power supply to both the Smart Box and the sensors in smart clothing. The common design for the Smart Box is expected to possess the capabilities of signal collection, short-term data storage, and local processing.

In our Smart Box, the sensory data is pre-processed through the digital signal processor (DSP) sub-module, compressed by the storage sub-module, and then transmitted to mobile devices (smartphone, portable computers, etc.) via the wireless communication sub-module. However, it is a challenging problem for the Smart Box to obtain an optimal trade-off among signal sample rate, complexity of signal processing, and computing cost.



**Figure 2.** Key design issues for Wearable 2.0: a) design of the conductive network based on the washable power cable; b) raw materials for smart clothing.

To address the above three challenging issues, Wearable 2.0 design exhibits the following features:

- All the sensors are wrapped and protected by textile cloth. Textile electrodes are covered under clothing, that is, they are invisible from the front side, as shown in Fig. 1a. These sensors are powered by miniature battery to collect physiological data.
- The sensors do not have to touch human skin tightly or continuously, since tight body contact causes discomfort for users. Opportunistic contacts during a user's movements will be enough to collect adequate data if the user feels comfortable and wears the smart clothing for a long time. Due to this feature, smart clothing can be personalized for a user as normal clothing.
- Digital modules are not merged with smart clothing directly. Instead, there is an "interface button" to interconnect smart clothing with the Smart Box, as shown in Fig. 2a.
- Before washing the smart clothing, the Smart Box should be disassembled through the interface button, since it is not waterproof. When the smart clothing is dried, the Smart Box can be installed again.

The interface button is a bridge to connect smart clothing with an outside communication device (i.e., the important interface for human-cloud integration). Thus, it is critical to decide where it is located. The interface button can be disguised as a shoulder strap or chest card, as shown in Fig. 2a. The selection of its position should consider comfort, aesthetics, and convenience for the user. In Fig. 2a, three buttons are designed: two of them are used for power supply, and the other one may be used to connect a data cable. If only two buttons are adopted, the cables will work as both electrical wires and data cables. In the proposed smart clothing, four basic physiological data are collected: ECG, oxygen saturation, body temperature, and heartbeat rate.

In Fig. 2a, 10 fixed electrodes are placed for ECG data collection. Electrodes 1 and 2 are close to the shoulders, while electrodes 3 and 4 are located in the lower front part of the clothing. Specifically, electrode V1 corresponds to the position of the fourth rib gap of the body right parasternal, while

electrode V5 corresponds to the position of the body left anterior axillary line. Another four electrodes (i.e., V2, V3, V4, and V6) are optional. Based on the specific requirements of sensory data accuracy, the user's comfort and wearing duration, those optional electrodes can be flexibly snapped on or taken off. When all 10 electrodes are used, standard medical-level ECG signals can be obtained. Practically, the various combinations of fixed and optional electrodes lead to different trade-offs between data accuracy, comfort, and power consumption.

Moreover, the energy cost of the Smart Box is also critical for Wearable 2.0. We recommend some recent radio technologies, such as Bluetooth 4.0+, which is convenient to connect with a smartphone in an energy-efficient fashion. Meanwhile, IEEE Low Power Wi-Fi (IEEE 802.11 ah) is also a good selection, since it has the features of long distance, low power, and low data rate, which is suitable for Internet of Things (IoT) applications.

It is obvious that the Wearable 2.0 healthcare system is complicated, which involves various research areas and technologies, including material science (Fig. 2b), costume design, electronic engineering, embedded systems, wireless communications, mobile networks, cloud computing, big data, and so on.

Thus, application-driven design is recommended for the implementation of the proposed system. It needs to decide which kinds of body signals should be measured, and then sensor types and corresponding data rates can be determined, as shown in the following scenarios.

**Patients with Cardiovascular Diseases:** Five physiological data are essential, including ECG, heart rate, inspiration, body temperature, and SpO<sub>2</sub>. Among these five kinds of data, ECG and heart rate have higher data priority in terms of accuracy and timeliness. According to the emergency level of a patient's illness, a specific sample rate can be decided.

**Long-Stay Patients Lying in Bed:** Due to lack of movement, indoor environments are more sensitive to long-term bed-ridden patients. Thus, the surveillance system can be deployed to monitor environmental parameters, such as temperature, humidity, noise, air quality (e.g., PM 2.5, volatile chemicals), and electromagnetic radiation. By jointly analyzing

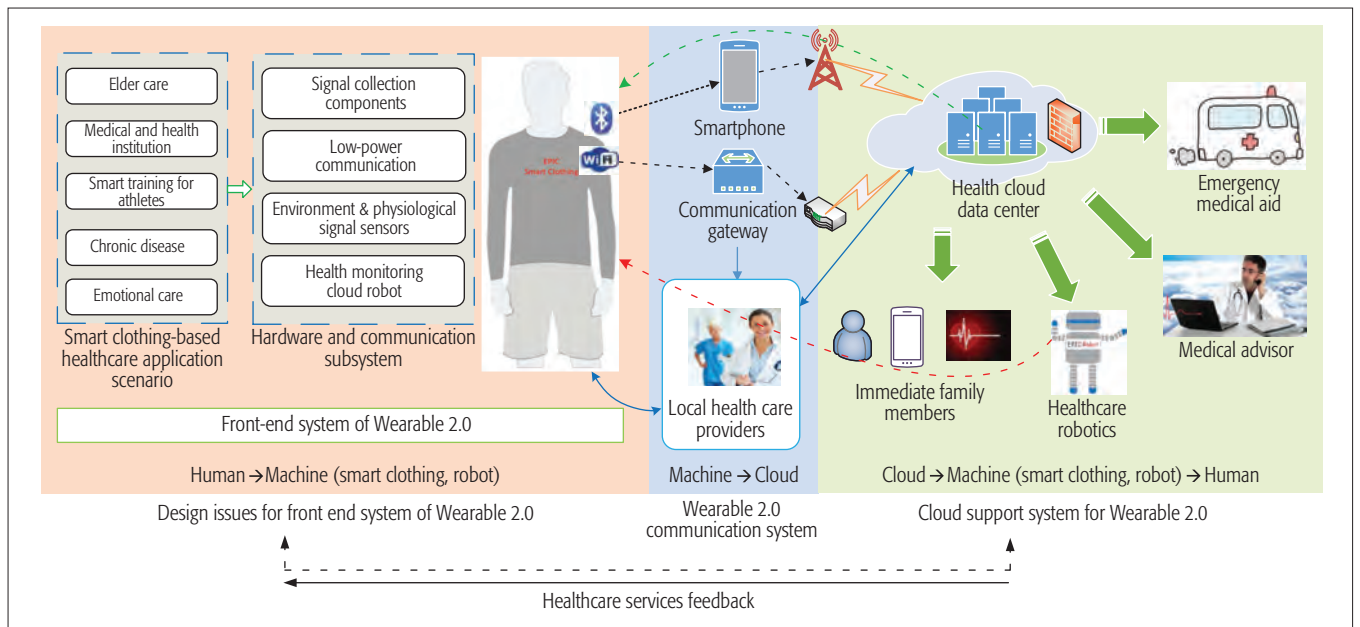


Figure 3. Design for a smart clothing-based healthcare system.

smart clothing-based physiological and environmental data, personalized diagnosis can be provided by medical experts for a healthier life. Therefore, the software system of smart clothing includes two parts: mobile applications for users and the applications for doctors or medical personnel.

## ARCHITECTURE OF THE WEARABLE 2.0 HEALTHCARE SYSTEM

In this section, the smart-clothing-centric sustainable health monitoring system is introduced. In Fig. 3, the whole Wearable 2.0 healthcare system includes the front-end system, communication system, and cloud supporting system.

### THE FRONT-END SYSTEM OF WEARABLE 2.0

The front-end system of Wearable 2.0 includes various sensors, and works as a long-term data source that plays an important role in collecting health big data. Moreover, the front-end system also works as a user interface. To achieve high user experience, healthcare robots can be implemented in the proposed front-end system. In particular, a mobile robot with a human-like shell can provide more friendly and personalized healthcare services. For example, when a heart attack occurs, and the user loses verbal capability, a healthcare robot can send a video recording and pictures to the remote medical center or immediate family members. Furthermore, when emotion-aware services are required, a humanoid robot with walking capability plays an important role in affective interaction. Thus, the integration of smart clothing and a humanoid robot is beneficial to increase the interoperability of the system in various complicated situations.

With the support of a mobile cloud system, healthcare big data can be stored over a long period, and big data analytics in the cloud can greatly enhance the intelligence and cognitive capability of a humanoid robot. Thus, real-time affective human-computer interaction is available with the support of the humanoid robot through a certain understanding of human emotion and

the user's intent. Furthermore, the robot also can work as a mobile sink to collect environmental data. In a word, the smart clothing supports high mobility, while the robot provides efficient data sensing and health monitoring.

### THE WEARABLE 2.0 WIRELESS COMMUNICATION SYSTEM

In the Wearable 2.0-based healthcare system, a wireless communication system for smart clothing is critical to achieve human-cloud integration.

Seamless design for interconnecting smart clothing and other devices should focus on energy efficiency and user experience:

**Normal Users:** Because of self-care ability and mobile capability, smartphones carried by users serve as personal gateways to forward the sensory physiological data to the cloud. As the major communication bridge between smart clothing and the cloud, the smartphone also stores, processes, and visualizes the health data locally. The mobile healthcare applications also enable users to understand their health status.

**Special Users:** There are still some users who seldom use smartphones or have difficulty using smartphones (e.g., elderly, children, disabled people, and patients with Alzheimer diseases). For those users, access points should be deployed in the areas where frequent activities happen. Therefore, Wi-Fi with low power consumption is the most suitable for smart clothing to connect with the cloud in this scenario.

### BACK-END SUPPORTING CLOUDS FOR THE WEARABLE 2.0 SYSTEM

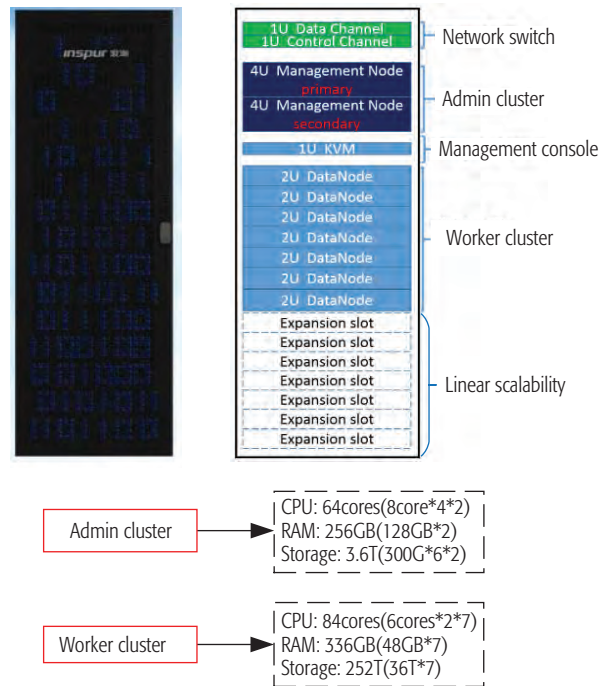
The back-end supporting cloud is the brain of the whole system to provide intelligence for cognitive healthcare applications. The health data from the lower layer are stored in the mobile local cloud (e.g., ad hoc cloudlet, edge cloud, and mobile cloud at the network edge) for further physiological data fusion [11]. The upper layer applications are available to get computing resources support, as shown in Fig. 4. In order to provide elastic services based on the real-time healthcare dataset and



(a)



(b)



(c)

Figure 4. Supporting cloud system Wearable 2.0: mobile health cloud platform: a) control panel for system administrator of mobile clouds; b) interface for mobile cloud user; c) supporting hardware system (i.e., Inspur SDA30000 data center) for mobile clouds.

available hardware devices, software defined networking can be used to enable flexible network control by separately personalizing the data collection, data analysis, and action feedback [12]. Furthermore, the cloud also provides support to the applications for the end-user and third-party medical care institution. In summary, smart clothing software is the foundation of data collection; mobile applications are the bridges of human-cloud

integration; and the cloud is the center of various healthcare services and cognitive applications.

## TESTBED FOR THE WEARABLE 2.0 HEALTHCARE SYSTEM

In this section, we briefly introduce the Wearable 2.0 healthcare system developed by the Embedded and Pervasive Computing (EPIC) lab at



**Figure 5.** Wearable 2.0 testbed: a) ECG dry electrodes and signal acquisition module in the EPIC Wearable 2.0 testbed; b) three modes for body signal visualization: mobile phone, PC, and cloud.

Various software is deployed on the smart clothing, smartphone, and mobile cloud platform, which involves the embedded system development, mobile application software development, and big-data-based cloud software development.

Huazhong University of Science and Technology. The EPIC Wearable 2.0 system mainly consists of smart clothing, a smartphone, big data cloud, and a humanoid robot. Although there is various software providing different services independently, all the software modules work cooperatively and form a comprehensive software ecosystem. Furthermore, the kernel software is deployed at our data center with Inspur SDA30000, while the basic architecture of the mobile health cloud is based on Openstack. On the other hand, mobile applications are developed based on Android 5.0, which has the following main functions:

- Connecting to smart clothing for gathering the physiological data, and setting the parameters and transmitting sensory data
- Providing personalized services, such as healthcare data visualization and early health alert

Specifically, when ECG monitoring is implemented in EPIC smart clothing, two electrodes are selected in order to decrease the cost and complexity. The ECG electrodes are made of textile, as shown in the circles marked 1 in Fig. 5. Then flexible wire (marked 2) is used to connect two snap fasteners (marked 3). ECG data is collected and transmitted through a black box. Finally, the ECG module transmits an ECG signal via wireless to the smartphone, computer, or cloud.

In the cloud, real-time detection and analysis are available to process the user's health data through the established health indicator threshold and data model based on the user's health (e.g., the user's ECG model). The analysis results of the user's health status are provided immediately to the user or health service providers so as to provide timely health care. In order to improve the accuracy of health monitoring, other relat-

ed data should be transmitted to the cloud for deep analysis, such as the user's location, indoor environment condition, social network data, facial expressions, and voice records. Based on the big data processing platform and machine learning algorithms, a cognitive healthcare system can be developed in which the smart clothing is an important component for accessing cloud computing and big data technology. Various software is deployed on the smart clothing, smartphone, and mobile cloud platform, which involves the embedded system development, mobile application software development, and big-data-based cloud software development.

An ECG signal strongly reflects human emotion, so it is usually used for emotion detection and analysis. Therefore, the user's emotional models based on an ECG historical dataset are established in the cloud for supporting realtime user emotion prediction, and its availability is verified through testbed. Moreover, 10 volunteers (including four men and six women, aged from 23 to 30 years old, average age is 25.2 years old) are recruited for the experiment and evaluation, and their ECG signals are tagged with five emotional states (based on the original emotion proposed by Krech *et al.* [13]): normal, happy, angry, fear and sadness. All the ECG data are divided into training set (70 percent) and testing set (30 percent) for feature extraction separately.

Specifically, support vector machine is used to analyze the training set and establish the classification model, which is evaluated through testing set. Finally, in the emotion prediction result of the 10 users, only the accuracies of User2 and User7 are relative lower (80.17 percent and 81.23 percent respectively), and the average accuracy has reached 87.13 percent. Because there are different

The Wearable 2.0 healthcare system is based on smart clothing integrating various physiological sensors. Therefore, the system can collect various important human physiological indicators, and has the potential to achieve high user experience, QoS and QoE.

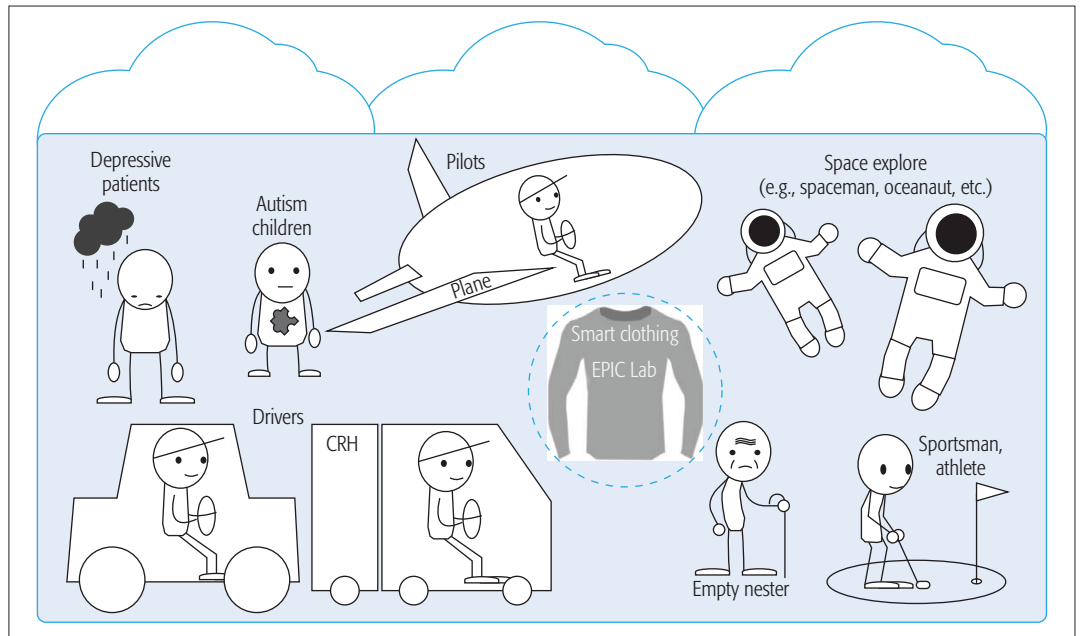


Figure 6. Wearable 2.0 based emotion care and health monitoring for special groups of population.

numbers of the samples for different emotions in the collected sample data, which may affect the model outputs accuracy during the training, and cause the different accuracy in emotion recognition. Specifically, the emotional detection accuracy of normal and happy reach 90 percent+, the accuracy of angry is 88.83 percent, the accuracy of fear and sadness are 85.65 percent and 81.28 percent.

### WEARABLE 2.0 HEALTHCARE APPLICATIONS

The Wearable 2.0 healthcare system is based on smart clothing integrating various physiological sensors. Therefore, the system can collect various important human physiological indicators, and has the potential to achieve high user experience, QoS and QoE. As shown in Fig. 6, the representative Wearable 2.0 healthcare applications include chronic disease monitoring, elderly people care, medical and health institution, smart training for athlete, emotion care, etc.

**Chronic Disease Monitoring:** Chronic disease monitoring is the main application of Wearable 2.0 healthcare system. Chronic disease patients wear smart clothing in their daily life for non-invasive physiological data collection. By processing and analyzing of the physiological data, cloud obtains realtime health status of the user and predicts the disease trend or health condition through learning from a large number of historical data with the support of big data technology. Based on the diagnosed health status, the system provides users with personalized healthcare services in multiple ways. For example, if heart attack is detected, the system will immediately notify the medical first aid agency or the user's family. Besides, if there is a health care robot in home, the system will control the robot to deliver emergency medicine to the patient immediately.

**Auxiliary Athlete Training:** For the sports project with less intensity but higher skill, some new challenges are brought. For instance, the accuracy of action has a decisive effect on the performance in the golf race. Thus, the system needs to deploy the sensors which can detect the movement of

the athletes through three-axes acceleration apparatus, gyroscope, etc.

**Emotion Care:** Emotion care is especially helpful for the empty nesters living alone, the long-distance truck drivers and patients suffering from mental diseases. Based on the physiological data related to the user's emotion, the system is available to provide emotion care. When detecting the user with a bad mood, the system supports emotional feedback, such as voice reminder, tuning suitable music, or playing selected video contents etc. If there is an interactive emotion-care robot, it will carry out a more accurate emotional interaction with the user, after receiving the user's emotion state and emotion interaction commands from the cloud. Traditional emotion detection methods are usually based on the data source from a single space, such as facial expression video, body signals, or posts on social networks. To overcome this shortcoming, the emotion detection of Wearable 2.0 can be more accurate by utilizing multi-dimensional data from Cyber-Physical-Social spaces. With the deployment of dedicated mobile terminal applications, it is convenient to integrate the user's social network data, location information, mobile phone call records, and so on. The physiological data from the health cloud platform can greatly improve the accuracy of emotional care. The concrete implementation method is to store and train the emotion model based on physiological data in the cloud and establish a unique emotion recognition model for each user. Specifically, according to the emotion recognition model, the user's emotional state is predicted by the trained model, while the related data are collected from the mobile terminal. When any negative emotions are detected, the relevant equipment with various resources are allocated to interact with users. For example, if sadness is detected, appropriate music is played to ease the grief of the user, and the system even sends a command to the indoor robot for effective interaction with users through a series of combinations of actions.

**The Applications of Virtual Reality/Augmented Reality Based on Smart Clothing:** The applications of virtual reality (VR)/augmented reality (AR) based on wearable technologies have shown great potential for the game industry or smart factory [14]. Specifically, because smart clothing may be closely integrated with the human body and collect more accurate physiological signals, the VR/AR applications (film and television, home computer games/video games, medical surgery, etc.) can implement more natural human-computer interaction with the help of smart clothing [15].

## CONCLUSION

In this article, we comprehensively investigate the disadvantages of the existing healthcare system and the trend of wearable computing. Then a Wearable 2.0 healthcare system is proposed based on smart clothing to improve QoE and QoS of the next generation healthcare system. In the proposed system, the user's physiological data is unconsciously collected, and personalized healthcare services are big data analytics on clouds. Furthermore, this article presents system architecture, functional components, and the design details of smart clothing based on a Wearable 2.0 healthcare system. Finally, a testbed with various compelling scenarios are presented to verify the feasibility of the proposed architecture.

## ACKNOWLEDGMENT

This research was supported by the Cross-Ministry Giga KOREA Project (GK16P0100, Development of Tele-Experience Service SW Platform Based on Giga Media) and the ITRC support program (IITP-2016-H8501-16-1015) supervised by the IITP (Institute for Information & Communications Technology Promotion), which are funded by the Ministry of Science, ICT and Future Planning, Korea.

Prof. Min Chen's work was supported by the National Natural Science Foundation of China (Grant No. 61572220).

## REFERENCES

- [1] M. S. Hossain and G. Muhammad, "Cloud-Assisted Industrial Internet of Things (IIoT)-Enabled Framework for Health Monitoring," *Computer Networks*, vol. 101, 2016, pp. 192–202.
- [2] S.-H. Seo, J.-W. Jang, and S.-W. Jang, "Design and Implementation of a Smart Clothing System Coping with Emergency Status," *Int'l. Info. Inst.*, Tokyo, vol. 19, no. 1, 2016, p. 175.
- [3] K. Zheng et al., "Big Data-Driven Optimization for Mobile Networks Toward 5G," *IEEE Network*, vol. 30, no. 1, Jan. 2016, pp. 44–51.
- [4] M. Chen et al., "Smart Clothing: Connecting Human with Clouds and Big Data For Sustainable Health Monitoring," *Mobile Networks and Applications*, 2016, pp. 1–21.
- [5] L. Hu et al., "Software Defined Healthcare Networks," *IEEE Wireless Commun.*, vol. 22, no. 6, Dec. 2015, pp. 67–75.
- [6] E. Strazdienė et al., "New Tendencies of Wearable Electronics Application in Smart Clothing," *Elektronika ir Elektrotechnika*, vol. 73, no. 1, 2015, pp. 21–24.
- [7] G. Fortino et al., "Cloud-Assisted Body Area Networks: State-of-the-Art and Future Challenges," *Wireless Networks*, vol. 20, no. 7, 2014, pp. 1925–38.
- [8] J. V. D. Brand et al., "Flexible and Stretchable Electronics for Wearable Health Devices," *Solid-State Electronics*, vol. 113, 2015, pp. 116–20.
- [9] K. Takei et al., "Toward Flexible and Wearable Human-Interactive Health-Monitoring Devices," *Advanced Healthcare Materials*, vol. 4, no. 4, 2015, pp. 487–500.
- [10] F. G. Retail and Technology, "The Wearables Report 2016: Reviewing a Fast-Changing Market," <https://www.fbcigroup.com/?q=publication/wearables-report-2016-fbic-global-retail-and-technology-june-21-2016>, 2016.

- [11] G. Fortino et al., "A Framework for Collaborative Computing and Multi-Sensor Data Fusion in Body Sensor Networks," *Information Fusion*, vol. 22, 2015, pp. 50–70.
- [12] X. Ge et al., "User Mobility Evaluation for 5G Small Cell Networks Based on Individual Mobility Model," *IEEE JSAC*, vol. 34, no. 3, 2016, pp. 528–41.
- [13] D. Krech, R. S. Crutchfield, and N. Livson, *Elements of Psychology*, Knopf, 1974.
- [14] P. Baudisch, "Virtual Reality in Your Living Room: Technical Perspective," *Commun. ACM*, vol. 58, no. 6, 2015, p. 92.
- [15] G. S. Ruthenbeck and K. J. Reynolds, "Virtual Reality for Medical Training: the State-of-the-Art," *J. Simulation*, vol. 9, no. 1, 2015, pp. 16–26.

## BIOGRAPHIES

MIN CHEN [SM'09] (minchen@ieee.org) has been a full professor in the School of Computer Science and Technology at Huazhong University of Science and Technology (HUST) since February 2012. He is Chair of the IEEE Computer Society Special Technical Community (STC) on Big Data. He was an assistant professor in the School of Computer Science and Engineering at Seoul National University (SNU). He worked as a postdoctoral fellow in the Department of Electrical and Computer Engineering at the University of British Columbia for three years. His Google Scholar Citations reached 7700+ with an h-index of 43. His top paper was cited 860+ times.

YUJUN MA (yujun.hust@gmail.com) received his B.Sc. degree in computer applications from Nanyang Institute of Technology (NIT), China, in 2003 and his Ph.D. degree in computer science and technology from HUST in 2016. Currently, he is an Associate Professor in Computer Network Center at NIT since 2016. His research interests include cloud computing, big data and the Internet of Things.

YONG LI (liyong07@tsinghua.edu.cn) received his B.S. degree in electronics and information engineering from HUST in 2007 and his Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2012. During July 2012 to August 2013, he was a visiting research associate with Telekom Innovation Laboratories and the Hong Kong University of Science and Technology, respectively. During December 2013 to March 2014, he was a visiting scientist at the University of Miami. He is currently a faculty member of the Department of Electronic Engineering, Tsinghua University. His research interests are in the areas of networking and communications.

DI WU (wudi27@mail.sysu.edu.cn) received his B.S. degree from the University of Science and Technology of China, Hefei, in 2000, his M.S. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2003, and his Ph.D. degree in computer science and engineering from the Chinese University of Hong Kong in 2007. He was a postdoctoral researcher with the Department of Computer Science and Engineering, Polytechnic Institute of New York University, Brooklyn, from 2007 to 2009, advised by Prof. K. W. Ross. He is currently a professor and the Assistant Dean of the School of Data and Computer Science with Sun Yat-sen University, Guangzhou, China.

YIN ZHANG [SM'16] (yin.zhang.cn@ieee.org) is an assistant professor of the School of Information and Safety Engineering, Zhongnan University of Economics and Law (ZUEL), China. He is an Excellent Young Scholar at ZUEL. He is Vice-Chair of the IEEE Computer Society Big Data STC. He was a postdoctoral fellow in the School of Computer Science and Technology at HUST. He serves as an Editor or Associate Editor for *IEEE Access*, *the IEEE Sensors Journal*, and other publications. He has been a Guest Editor for *Mobile Networks and Applications*, *Sensors*, *Multimedia Tools and Applications*, *the Journal of Medical Systems*, *the New Review of Hypermedia and Multimedia*, and so on. He also served as TPC Co-Chair of CloudComp 2015 and Local Chair of TRIDENTCOM 2014. He has published more than 50 prestigious conference and journal papers. His research interests include intelligent service computing, big data and social networks, and other areas.

CHAN-HYUN YOUN (chyou@kaist.ac.kr) is a professor at the School of Electrical Engineering of the Korea Advanced Institute of Science and Technology (KAIST), Daejeon. He is an associate vice-president of the office of planning and budgets and is also a director of the Grid Middleware Research Center in KAIST. His research areas are cloud computing, high-performance computing, and vehicle-cloud systems and applications. He was an Editor-in-Chief of *KIPS*, an Editor of the *Journal of Healthcare Engineering*, and served as head of the Korea branch (computer section) of IEICE.

Because smart clothing may be closely integrated with the human body and collect more accurate physiological signals, the VR/AR applications (film and television, home computer games/video games, medical surgery, etc.) can implement more natural human-computer interaction with the help of smart clothing.

# Millimeter-Wave Wireless Communications for IoT-Cloud Supported Autonomous Vehicles: Overview, Design, and Challenges

Linghe Kong, Muhammad Khurram Khan, Fan Wu, Guihai Chen, and Peng Zeng

The authors explore the capability of millimeter-wave communications for autonomous vehicles. As the next-generation wireless technology, mmWave is advanced in its multi-gigabit transmittability and beamforming technique. Based on these features, the authors propose the novel design of a vehicular mmWave system combining the advantages of the Internet of Things and cloud computing.

## ABSTRACT

Autonomous vehicles are a rising technology in the near future to provide a safe and efficient transportation experience. Vehicular communication systems are indispensable components in autonomous vehicles to share road conditions in a wireless manner. With the exponential increase of traffic data, conventional wireless technologies preliminarily show their incompetence because of limited bandwidth. This article explores the capability of millimeter-wave communications for autonomous vehicles. As the next-generation wireless technology, mmWave is advanced in its multi-gigabit transmittability and beamforming technique. Based on these features, we propose the novel design of a vehicular mmWave system combining the advantages of the Internet of Things and cloud computing. This mmWave system supports vehicles sharing multi-gigabit data about the surrounding environment and recognizing objects via the cloud in real time. Therefore, autonomous vehicles are able to determine the optimal driving strategy instantaneously.

## INTRODUCTION

An automotive revolution is taking place in autonomous vehicles. Without human intervention, autonomous vehicles [1] have the potential to remove more than 85 percent of traffic accidents caused by human errors. Moreover, drivers could get rid of boring tasks and enjoy their travels. Due to these benefits, “57 percent of consumers, globally, trust driverless cars — even more so in emerging markets” was reported in Cisco’s survey. To meet market demands, automobile manufacturers have contributed great efforts. Several semi-autonomous technologies have been applied in practice such as Cadillac’s super cruise and Benz’s park assist. Furthermore, 53 Google driverless cars are self-driving in California and Texas for field tests.

The technology of autonomous vehicles is a typical convergence of the Internet of things (IoT) [2] and cloud computing [3]. From the macro aspect, navigation relies on GPS, map service, and road conditions, which is provided by cloud computing. From the micro aspect, an autonomous vehicle determines its real-time moving strategy

depending on the dynamic surroundings. Many in-vehicle sensors make driverless cars go, but few are more important than the light detection and ranging (LiDAR) devices mounted on the roofs of vehicles. The LiDAR device scans more than 70 m in all directions, generating a precise three-dimensional map of a car’s surroundings.

However, LiDAR, even with other sensors, is inadequate to ensure safe and efficient self-driving. In February 2016, a Google driverless car was at fault in a crash. Before that, LiDAR-based Google cars were involved in 17 minor accidents in six-year 2-million-mile tests. The possible problems are first, LiDAR is constrained by line of sight and cannot see through a large obstacle such as a truck ahead. Second, LiDAR performs poorly in bad weather. Third, LiDAR may incorrectly recognize some harmless objects (e.g., plastic bags) as obstacles. Fourth, LiDAR cannot discern human signs.

Vehicular communication systems [4] are a feasible solution to compensate for the drawbacks of LiDAR/sensors. Through two wireless modes, vehicle-to-infrastructure (V2I) and vehicle-to-vehicle (V2V) [5], autonomous vehicles can acquire more traffic data to optimize their driving strategy. Existing works attempt to integrate commercial WiFi, Bluetooth, ZigBee, WiMax, and fourth generation (4G) into vehicles. In addition, the U.S. Department of Transportation has committed to use IEEE 802.11p-based dedicated short-range communications (DSRC) [6] on new light-duty vehicles beginning in 2017. Nevertheless, these conventional wireless communications have limited bandwidth. For example, the maximal bit rate of DSRC is 27 Mb/s. On the contrary, the traffic data are ever growing, such as LiDAR’s 3D imaging and a camera’s high-definition (HD) video.

The next-generation wireless technology, millimeter-wave (mmWave) [7], shows its potential to solve this dilemma. The much anticipated mmWave particularly works at 3–300 GHz [8], in which the available channel bandwidth is up to several gigahertz. Hence, mmWave can achieve multi-gigabit transmittability [9] for big data delivery. Moreover, mmWave exploits smart antenna arrays to realize the beamforming technique [10]. As a result, the constructive directional signal can track [11] and transmit to high-speed targets over



	DSRC	WiFi	Bluetooth	ZigBee	WiMax and 4G	
<b>Spectrum</b>	5.9 GHz	2.4/5.8 GHz	2.4 GHz	868 MHz/915 MHz/2.4 GHz	2–6 GHz	1880–2650 MHz
<b>Standard</b>	802.11p	802.11a/b/g/n	802.15.1	802.15.4	802.16e	LTE
<b>Bandwidth</b>	10 Mb/s	20, 40 MHz	1 MHz	2 MHz	1.75–20 MHz	20 MHz
<b>Bit rate</b>	3–27 Mb/s	6–600 Mb/s	1–24 Mb/s	250 kb/s	Peak upload: 56 Mb/s, Peak download: 128 Mb/s	Upstream: 75 Mb/s, Downstream: 300 Mb/s
<b>Modulation</b>	OFDM	MIMO, OFDM	FHSS, GFSK, $\pi/4$ -DPSK, 8-DPSK	DSSS, O-QPSK	OFDMA, MIMO	OFDMA, MIMO
<b>Tx range</b>	< 300 m	< 100 m	< 100 m	< 100 m	< 10 km	< 2 km
<b>Cost</b>	Cheap	Cheap	Cheap	Cheap	Expensive	Expensive

**Table 1.** Comparison of vehicular communication systems.

a long distance. These features exactly fit the demand of autonomous vehicles.

To fully exploit mmWave, we propose a novel vehicular mmWave system for autonomous vehicles. This system consists of both V2V and V2I mmWave communications. The V2V part enables the real-time exchange of sensory data (e.g., LiDAR data and HD video) among vehicles, helping to cover the blind areas and share vision in bad weather. The V2I part leverages the roadside infrastructure and cloud computing to feed back recognized objects and signs. Combining the multi-modal data, autonomous vehicles are able to immediately determine the optimal driving strategy. In this article, we introduce the framework design of a vehicular mmWave system and discuss the key design problems. Prototype and performance evaluation are conducted to demonstrate the feasibility of vehicular mmWave. The open issues and future directions are summarized at the end.

The proposed system is a general framework. It is easy to add customized components according to the demands of autonomous vehicles. We believe vehicular mmWave has wider implications and prospects for intelligent transportation applications than explored in this article.

## RECENT ADVANCES IN AUTONOMOUS VEHICLES

An autonomous vehicle [1] (driverless car, self-driving car, robotic car) is a vehicle that is capable of sensing its environment and navigating without human input. To realize self-driving, autonomous vehicles first detect their surroundings by vehicle-mounted sensors. Then advanced control systems interpret sensory information to identify appropriate navigation paths as well as obstacles. To increase the sensing accuracy, an autonomous vehicle is equipped with at least two independent systems: the sensor system as the main part and the communication system as the assistant.

### VEHICULAR SENSOR SYSTEMS

A vehicular sensor system is usually composed of LiDAR, radar, GPS, odometry, and computer vision [1]. In these sensors, LiDAR is considered as the “eyes” of recent driverless cars. The commercial LiDAR employing 64 laser diodes to produce 2.8 million data points per second with

a 360° horizontal field of view and a 26.8° vertical field of view. By virtue of LiDAR, vehicles can detect obstacles and build 3D surroundings for safe navigation in dynamic environments. The effectiveness of LiDAR has been demonstrated in practice. But the other sensors are still indispensable, and play important roles in special applications such as side cameras for lane-keeping, infrared sensors for night detection, and sonar for distance measurement.

Nevertheless, the vehicular sensor system alone is not sufficient for a vehicle’s automatic cruise. First, blind areas exist in LiDAR as well as other sensors because of the line-of-sight constraint (e.g., a vehicle cannot see through the vehicle ahead, causing overtake difficulty and potential risk). Second, sensors perform poorly in bad weather. The sensing range of LiDAR is largely reduced in heavy rain or snow. Third, it is not easy to identify whether a small object is harmless or not; for example, a wrong estimation of a plastic bag or a small mound may lead to needless veering, decreasing driving efficiency. Fourth, LiDAR is able to detect a human but is not accurate enough to recognize human gestures; for example, it is difficult for sensors to distinguish the police gestures of “Go” and “Stop.”

### VEHICULAR COMMUNICATION SYSTEMS

In order to compensate for the drawbacks of sensor systems, a communication system is applied in vehicles [4]. Through wireless data sharing, a communication system is advanced in breaking the line-of-sight constraint and acquiring more data on surroundings, such as blind area information, even in bad weather. With more data, the vehicle can further optimize the driving strategy.

In the literature, scientists and engineers have attempted to implement various wireless standards into vehicular communication systems [5, 12]. A comparison of these standards is provided in Table 1.

Both DSRC and WiFi belong to the IEEE 802.11 family, the most common wireless protocol stack. The 802.11p-based DSRC [6] is specially designed for vehicular communications, which is close to 802.11a. The major difference is that the channel bandwidth of 11p is half that of 11a, so 11p’s bit rate is half as much and the transmission range is three times longer than 11a. In addition, with multiple-input multiple-output (MIMO), the bit rate of 802.11n is up to 600 Mb/s.

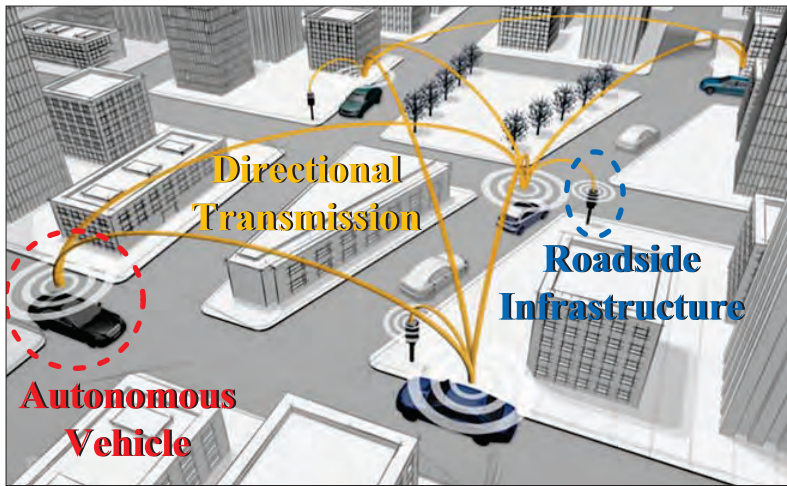


Figure 1. The vehicular mmWave system enables multi-gigabit transmission for V2V and V2I communication modes.

The advantage of Bluetooth and ZigBee is low power, where the power consumption of Bluetooth 4.0 is as low as 0.5 mW. However, these low-power standards adopt simple modulation techniques, leading to a low bit rate, where the bit rate of Bluetooth Low Energy (BLE) mode is 1 Mb/s and that of ZigBee is only 250 kb/s.

WiMax and 4G LTE are base station driven long-distance transmission technologies widely adopted in cellular networks. They can provide megabit wireless access service even in a high-speed mobile environment. However, the base stations are expensive, and vehicular users need to pay for the data traffic during their driving.

Although there are numerous standards for vehicular communication systems, none of them is qualified for autonomous vehicles. The common problem of existing standards is that their transmittability is limited at the megabit level. In contrast, LiDAR and HD cameras in autonomous vehicles generate huge amounts of data every second. Sharing these data among multiple vehicles, especially in a crowded scenario, urgently requires gigabit wireless transmission.

## MILLIMETER-WAVE WIRELESS COMMUNICATIONS

The next-generation mobile technology, mmWave [7], is envisioned to offer multi-gigabit wireless service for emerging applications [13]. Before applying mmWave to autonomous vehicles, we first introduce the promises and propagation characteristics of mmWave.

The first promise of mmWave is bandwidth. Take 60 GHz as an example. The unlicensed 60 GHz band provides 7 GHz bandwidth for mobile applications and is supported by IEEE 802.11ad, targeting indoor multi-gigabit wireless networks. Benefiting from the wide bandwidth, the bit rate of 802.11ad is up to 6.76 Gb/s [9]. TP-Link announced the world's first 802.11ad router in January 2016; the peak bit rate achieved is 7 Gb/s. The other key parameters in 802.11ad are listed in Table 2. If we transplant such multi-gigabit transmittability into autonomous vehicles, sensory data including LiDAR's 3D images and cameras' HD videos can be shared among all neighboring vehicles in real time.

Parameters	Values
Spectrum	57–64 GHz
Number of channels	4
Bandwidth	2.16 GHz
Bit rate	693 Mb/s–6.76 Gb/s
Modulation	OFDM
Tx range	< 10 m (omni-antenna)
Cost	Cheap

Table 2. Key parameters of IEEE 802.11ad for 60 GHz mmWave communications.

Besides the bandwidth, another promise is short wavelength. Since the wavelength of mmWave is at the millimeter level, it is possible to pack a large number of antennas into small space (e.g., a 100-element 60 GHz array can be integrated into 1 in<sup>2</sup>). Thus, the beamforming technique is handily applied in mmWave. Beamforming [10] is a signal processing technique to generate directional signal transmission by smart antenna array. Although the transmission range of mmWave is only 10 m in omnidirectional broadcast mode, beamforming can concentrate power in one direction and offer a transmission range that exceeds 130 m for 385 Mb/s and 79 m for 2 Gb/s. Beamforming is significantly helpful for autonomous vehicles. On one hand, the directional transmission assists the localization in the high-speed mobile environment. On the other hand, beamforming realizes concurrent transmissions by space-division multiple access (SDMA) and reduces the interference.

Moreover, mmWave has significantly different propagation characteristics compared to the 2.4/5.9 GHz band, where WiFi, Bluetooth, ZigBee, and DSRC operate:

**Propagation:** In free space, the signal strength is mainly lost due to oxygen absorption, where the loss of 60 GHz mmWave is about 16 dB/km [14]. Although it is difficult to realize a long-range (kilometer-level) link, mmWave has little effect within a short range because beamforming enhances the spatial reuse. For example, the loss due to oxygen absorption and heavy rain at 50 mm/hour is 36 dB/km, which works out to a modest 3.6 dB for a transmission range of 100 m.

**Penetration:** While 2.4/5.9 GHz signals penetrate through some objects, mmWave signals are easily blocked by most solid materials. Even a human body will introduce 20–50 dB of loss. In addition, since the transmission power is limited to 40 dBm by the Federal Communications Commission (FCC), mmWave does not have adequate power to burn through obstacles [8]. Therefore, it is challenging to guarantee robust mmWave connectivity in dynamic and obstacle-rich transportation environments.

**Doppler:** The Doppler effect depends on frequency and mobility. If the mmWave frequency is 3–60 GHz with mobility speed within 3–350 km/h, the Doppler shift will range from 10 Hz to 20 kHz. Due to the concentrated beam, there is a non-zero bias in the Doppler spectrum, which is largely compensated by automatic frequency control (AFC) [7] at the receiver side. As a result, the Doppler effect of mmWave can be well solved in vehicular communication systems.

## DESIGN OF VEHICULAR MMWAVE SYSTEMS

To satisfy the big data delivery in autonomous vehicles, we propose a novel vehicular mmWave system. The ideal vehicular mmWave system operates as shown in Fig. 1, where any vehicle directionally connects with other vehicles and roadside infrastructure. The proposed system has three principal members:

1. Every *autonomous vehicle* is equipped with an mmWave radio, LiDAR, a camera, and the other usual sensors.
2. The *roadside infrastructure* consists of an HD camera and an mmWave radio. In addition, the infrastructure has a wired connection with the cloud.
3. *Cloud computing* has strong computational capability for data analyzing and path planning.

The framework of the vehicular mmWave system is shown in Fig. 2. This framework provides services based on V2V and V2I communication modes.

**V2V mmWave communication:** With mmWave radios, vehicles are able to share real-time sensory data within transmission ranges, forming an IoT application. Thus, the blind area and bad weather problems are effectively addressed. In detail, when a vehicle observes a blind area in its sensing range, it asks for LiDAR or camera data from neighboring vehicles to compensate. In addition, although the LiDAR's sensing range is sharply reduced in bad weather, mmWave's transmission range has almost no influence. Leveraging the shared sensory data, a vehicle can reconstruct the 3D road conditions by multi-source multi-modal data analysis [15].

**V2I mmWave communication:** In this mode, the roadside infrastructure works as a relay to forward data between vehicles and the cloud. Therefore, the recognition problems can be tackled well. For example, when a vehicle senses but cannot identify an object or a human gesture, it transmits HD video to the cloud. Benefiting from big data and strong computation capability, the cloud is able to accomplish the recognition instantaneously and feed the result back.

Above all, the vehicular mmWave system is helpful to autonomous vehicles for safe and efficient driving. The main contribution is that mmWave changes the self-driving strategy from purely local control to collaborative control. However, the proposed system cannot work with only the abstract framework. Next, we discuss four key design problems and their potential solutions for this system.

### DATA PRIORITY

The objective of data priority is to determine which sensory data can be transmitted in advance when wireless collision occurs. We classify the communication needs into three priorities.

**Priority I:** Emergent data. Safety is the first criterion in autonomous vehicles. When a vehicle detects or estimates any dangerous surroundings such as a car crash, its highest priority is to immediately transmit these data to neighboring vehicles and infrastructures.

**Priority II:** Application-driven request. The communication needs are triggered by vehicular

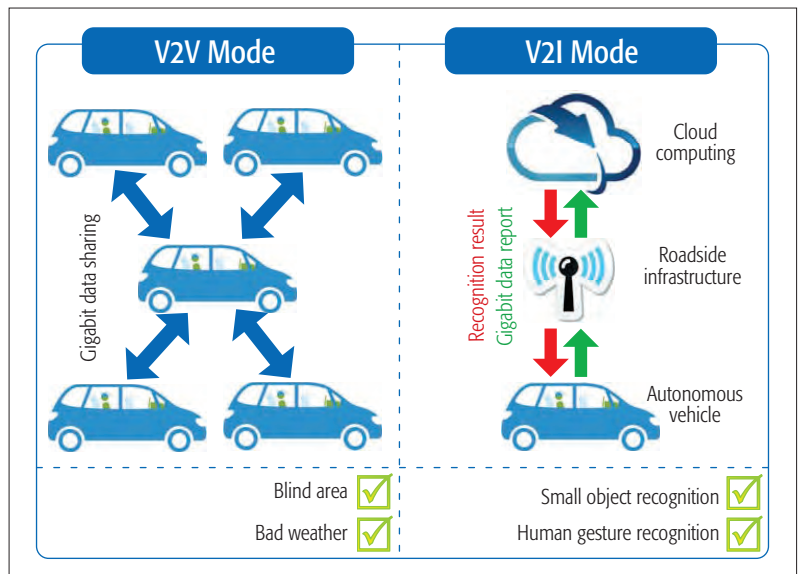


Figure 2. The framework of vehicular mmWave system.

applications. For example, when an autonomous vehicle plans to overtake a truck ahead and cannot sense the road conditions in front of the truck, this vehicle sends a request to the truck. Then the truck responds its LiDAR's and HD camera's data. The priority of the application-driven request is second only to the emergent data.

**Priority III:** Routine broadcast. When the mmWave channel is not occupied, routine messages are broadcast to all single-hop neighbors including vehicles and infrastructures. Routine messages can include GPS information, movement information, mmWave's channel state, and abstracted sensory data. Moreover, a vehicle transmits data every time it traverses an intersection.

In mmWave's medium access control (MAC) layer, we set data with the highest priority having the shortest backoff range, which can be sent first after collision. Similarly, data with the lowest priority has the longest backoff range.

### DEPLOYMENT PLAN

The deployment plan determines how many infrastructures need to be deployed along roadsides and their optimal locations. The deployment problem is studied from two dimensions.

From the space dimension, the deployment in the X-Y plane is planned by big data analysis. First, using the map information and the transmission range of directional mmWave, the lowest number of infrastructures can be calculated to satisfy the full coverage of all roads. Second, limited by the size of an antenna array, one infrastructure can serve only a finite number of vehicles simultaneously. The redundant coefficient is derived according to historical road conditions; for example, high redundancy is set for roads with frequent congestion or accidents. Leveraging the above two steps, the total number of infrastructures and their rough distribution are obtained. However, it can be proved that it is NP-hard to find their optimal locations. We adopt the combinatorial optimization method to reach a sub-optimal result. In the Z-axis, the height of an infrastructure's antennas follows the rule that a height which is too low too

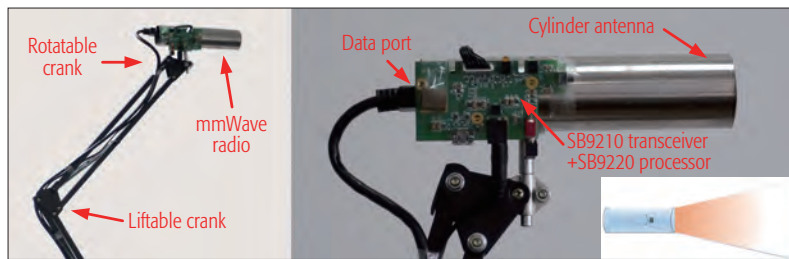


Figure 3. The prototype of a vehicular mmWave system, the crank arm, the cylinder antenna, and the beam model.

low will be blocked frequently, and a height that is too high will increase the path loss in long-distance transmission.

The view from the time dimension is also interesting and practical. We can imagine that the deployment of roadside infrastructures in an urban environment is a pretty long process. Similarly, the promotion of autonomous vehicles cannot be accomplished overnight. The report in *IEEE Spectrum* forecasted that the first vehicle with V2V and V2I communications would come into the market in 2018, and half of new cars will be autonomous by 2032. Hence, even if the optimal locations for deployment have been obtained, the deployment sequence should be further studied to keep pace with the market. We attempt to use economics theory to formulate the relationship between the numbers of infrastructures and vehicles with mmWave communications. Moreover, simulations based on the relationship are conducted to guide and amend the deployment plan.

#### BEAM CONTROL

Directional transmission is required in mmWave to overcome the path loss. To realize it, the directional antenna and beamforming are two candidate methods.

The major advantage of a *directional antenna* is its mature technology, as it is easy to implement at the current time using off-the-shelf devices. In [11, 13], a single horn antenna is adopted to concentrate the signal into a  $7^\circ$  beam, and a motor supports the rotation of this antenna. However, the drawbacks include the fact that one antenna provides only one wireless link, and the rotation motor introduces additional delay.

**Beamforming** can generate multiple links simultaneously by antenna array, and its direction change is fast enough to catch up with the vehicle's speed. Nevertheless, smart array devices are rare in the market. It is envisioned that beamforming will be a core technique in vehicular mmWave systems.

Using antenna array, the beam formation can be realized in the digital or analog domain. Digital beamforming is carried out by multiplying a particular coefficient to the modulated baseband signal. The strengths of digital beamforming include a higher degree of freedom and better transmission performance. Nevertheless, its drawback is the high complexity including the separate fast Fourier transform (FFT)/inverse FFT (IFFT) blocks, digital-to-analog converters (DACs), and analog-to-digital converters (ADCs) for every link. On the contrary, analog beamforming is a simple and effective method that generates high beamforming gains by controlling phase shifters and variable

gain amplifiers. However, analog beamforming requires a large number of antennas, and it is less flexible than the digital method.

The trade-off between flexibility and simplicity motivates us to propose a hybrid structure. In this structure, simple analog beamforming is used to quickly track high-speed vehicles, while flexible digital beamforming provides multiple beams if one infrastructure needs to connect multiple vehicles simultaneously.

#### HANDOVER STRATEGY

In conventional cellular networks, handover occurs when an established wireless link is redirected from the current cell to another. Compared to cellular networks, handover in vehicular mmWave communications is more complicated, which might be conducted as follows:

- When the vehicle is driving away from the coverage area of one mmWave radio and entering another radio's, the wireless link is transferred in order to avoid link termination. Even in this case, the handover operation is nontrivial because the vehicle has not only the V2I communication mode but also the V2V mode, which increases the destination diversity for handover.
- When one vehicle's wireless link is blocked by an object, such as a tree, a human, or other vehicles, this link has to be transferred to another mmWave radio quickly. Such a case never happens in cellular networks due to strong penetration capability. However, it is common in vehicular mmWave systems because of the directional transmission, poor penetration, and high speed.

Besides inheriting the state-of-the-art handover solution, we propose to add a prediction strategy to improve the handover performance. With the assistance of cloud computing, a vehicle can predict relatively precisely the movements of surrounding objects based on sensory data. Then, with the road map and the infrastructure locations, this vehicle schedules its handover beforehand with the objective of the optimal link selection constrained by bypassing the potential obstacles and minimizing the back-and-forth case.

#### PROTOTYPE AND EVALUATION

**Prototype:** To demonstrate the feasibility of a vehicular mmWave system, we build a prototype of 3D mmWave radio, shown in Fig. 3. This radio is supported by liftable and rotatable cranks, so its height and direction could be arbitrarily adjusted in 3D space, which can partially bypass the obstacle of line-of-sight communication. The radio frontend consists of a data port to exchange data with a computer, an SB9220 processor to operate the network control, an SB9210 transceiver to provide 4 Gb/s bit rate transmission in 60 GHz band, and a customized cylinder and metal waveguide as the antenna to form the signal into a beam. Then the beam can be considered as the cone model with the angle  $\alpha$ . We conduct outdoor testing of a pair of such radios by HD video transmission. The angle  $\alpha$  is nearly  $9^\circ$ , and the communication range is about 20 m without obvious lag.

**Performance evaluation:** Simulations are further conducted to evaluate the performance of vehicular mmWave systems. Our simulation is in

a 100 m × 15 m (3 lanes each direction) road segment. Six infrastructures are deployed at locations {(0,0), (20,15), (40,0), (60,15), (80,0), (100,15)}. The number of autonomous vehicles varies from 20 to 100. All infrastructures and vehicles are equipped with mmWave systems. According to our prototype, the mmWave radio is liftable (height range 2–3 m) and rotatable, communication range is 20 m, and  $\alpha = 9^\circ$ . The ratio of Priority I, II, III data is set 0.1:0–3:0.6. If two senders transmit data to one receiver, the sender with higher priority wins the link. We assume the antenna adjustment and the handover are quick enough without time delay.

Figure 4 illustrates the simulation result on the average number of effective wireless links, where one link is defined by two (end-to-end) or more (broadcast) connected radios. The comparison is between *mmWave* and *DSRC*; recall that DSRC is the 802.11p-based dedicated communication for vehicles. In Fig. 4, mmWave always performs better than DSRC, which demonstrates the efficient channel utilization by mmWave. If mmWave's multi-gigabit rate is further considered, its throughput in the whole network is much more than DSRC's, which has a maximal rate of 27 Mb/s. With the increase of density, the trends of two curves cannot maintain a linear increase because radios in each other's interference range cannot build new links. However, benefiting from the directional transmission, mmWave's trend slope is also better than DSRC's.

## SUMMARY AND DISCUSSION

Both academia and industry have contributed considerable efforts on autonomous vehicles. Several projects, such as Google's driverless car, have been carried out to develop related standards, technologies, and applications. Millimeter-wave spectrum can potentially provide the ability of multi-gigabit transmission, which is the most effective and straightforward solution to support the communication for autonomous vehicles in the next few decades and beyond.

In this article, we design an IoT-cloud supported vehicular mmWave system to fully exploit the advantages of mmWave and vehicles. On one hand, this system enables sensory data sharing among vehicles to tackle the blind area and bad weather problems. On the other hand, toward the accurate recognition of human gestures and small objects, this system leverages cloud computing via V2I communication of HD video.

Using mmWave in autonomous vehicles is a new concept. Several open issues are worth being deeply studied in the future. First, it is very important to build a systematical theory for vehicular mmWave systems. The theoretical derivation of data redundancy, trajectory prediction, and throughput could guide the design, strategy determination, and parameter setting. Second, the security and privacy mechanism is also an open issue. Traffic data sharing may expose one's location and trajectory, resulting in privacy leakage. It is desired to design a privacy preservation component for mmWave communications. Third, the proposed system still lacks an incentive mechanism. Such a mechanism is helpful to encourage more users to participate in data sharing for more accurate self-driving optimization.

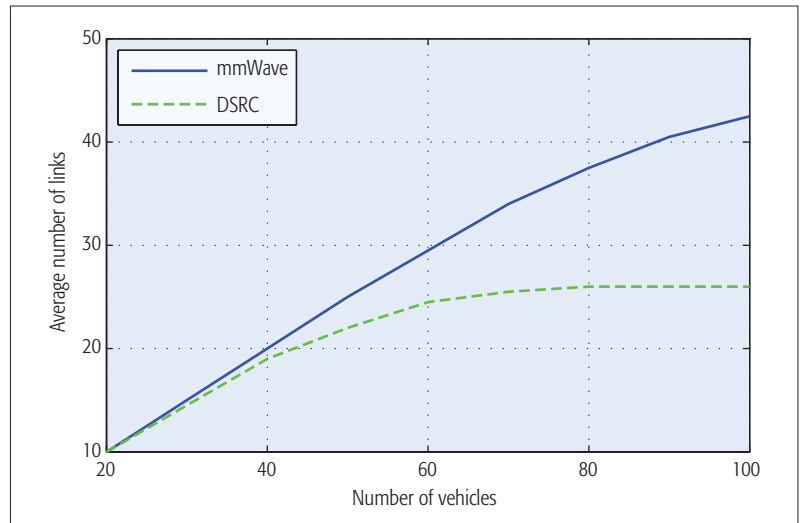


Figure 4. Average number of wireless links in simulation.

The framework of a vehicular mmWave system also produces several promising research directions. One valuable direction is to apply mmWave communications for other emerging applications, not only self-driving but also other mobile applications, such as entertainment and social networks, which demand big data transmission in high-speed environments. Moreover, a hybrid communication system is another practical direction. According to the above analysis, different wireless technologies possess different advantages. For example, Bluetooth is low-power and 4G covers a long transmission range. A communication system that consists of multiple wireless technologies can deal with complex requirements. Last but not least, since mmWave ensures adequate traffic data, a multi-modal data-based self-driving strategy can be studied to enhance the performance of self-driving.

## ACKNOWLEDGMENT

This research was partly supported by the State Key Development Program for Basic Research of China (973 project 2014CB340303) and NSFC grants 61672349, 61303202, 61472252, 61422208, and 61133006. The authors extend their sincere appreciation to the Deanship of Scientific Research at King Saud University for its funding of this Prolific Research Group (PRG-1436-16).

## REFERENCES

- [1] J. Baber *et al.*, "Cooperative Autonomous Driving: Intelligent Vehicles Sharing City Roads," *IEEE Robotics & Automation Mag.*, vol. 12, no. 1, 2005, pp. 44–49.
- [2] M. Hossain and G. Muhammad, "Cloud-Assisted Industrial Internet of Things (IIoT)-Enabled Framework for Health Monitoring," *Computer Networks*, 2016.
- [3] F. Liu *et al.*, "Gearing Resource-Poor Mobile Devices with Powerful Clouds: Architectures, Challenges, and Applications," *IEEE Wireless Commun.*, vol. 20, no. 3, June 2013, pp. 14–22.
- [4] K. Dar *et al.*, "Wireless Communication Technologies for ITS Applications," *IEEE Commun. Mag.*, vol. 48, no. 5, May 2010, pp. 156–62.
- [5] N. Kumar *et al.*, "Critical Applications in Vehicular Ad Hoc/Sensor Networks," *Telecommun. Systems*, 2014.
- [6] J. B. Kenney, "Dedicated Short-Range Communications (DSRC) Standards in the United States," *Proc. IEEE*, vol. 99, no. 7, 2011, pp. 1162–82.
- [7] Z. Pi and F. Khan, "An Introduction to Millimeter-Wave Mobile Broadband Systems," *IEEE Commun. Mag.*, vol. 49, no. 6, June 2011, pp. 101–07.

- [8] Y. Zhu *et al.*, "Demystifying 60GHz Outdoor Picocells," *ACM MobiCom*, 2014, pp. 5–16.
- [9] A. Ghosh *et al.*, "Millimeter-Wave Enhanced Local Area Systems: A High-Data-Rate Approach for Future Wireless Networks," *IEEE JSAC*, vol. 32, no. 6, 2014, pp. 1152–63.
- [10] W. Roh *et al.*, "Millimeter-Wave Beamforming as an Enabling Technology for 5G Cellular Communications: Theoretical Feasibility and Prototype Results," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 106–13.
- [11] T. Wei and X. Zhang, "mtrack: High-Precision Passive Tracking Using Millimeter Wave Radios," *ACM MobiCom*, 2015, pp. 117–29.
- [12] M. Saini *et al.*, "How Close Are We to Realizing a Pragmatic VANET Solution? A Meta-Survey," *ACM Computing Surveys*, vol. 48, no. 2, 2015, p. 29.
- [13] D. Halperin *et al.*, "Augmenting Data Center Networks with Multi-Gigabit Wireless Links," *ACM SIGCOMM Comp. Commun. Review*, vol. 41, 2011, pp. 38–49.
- [14] S. Geng *et al.*, "Millimeter-Wave Propagation Channel Characterization for Short-Range Wireless Communications," *IEEE Trans. Vehic. Tech.*, vol. 58, no. 1, 2009, pp. 3–13.
- [15] S. Qian *et al.*, "Multi-Modal Event Topic Model for Social Event Analysis," *IEEE Trans. Multimedia*, vol. 18, no. 2, 2016, pp. 233–46.

### BIOGRAPHIES

LINGHE KONG (linghe.kong@cs.sjtu.edu.cn) is currently an associate professor with the Department of Computer Science and Engineering at Shanghai Jiao Tong University, P. R. China. Before that, he was a postdoctoral researcher at Columbia University, McGill University, and Singapore University of Technology and Design. He received his Ph.D. degree from Shanghai Jiao Tong University in 2012, his Master's degree from Telecom SudParis in 2007, and his B. E. degree from Xidian University in 2005. His research interests include wireless communication, sensor networks, mobile computing, Internet of things, and smart energy systems.

MUHAMMAD KHURRAM KHAN [SM] (mkhurram@ksu.edu.sa) is currently working as a full professor at the Center of Excellence in Information Assurance (CoEIA), King Saud University,

Kingdom of Saudi Arabia. He is the Editor-in-Chief of the well reputed journal *Telecommunication Systems*. He is also on the Editorial Boards of several journals published by IEEE, Elsevier, Springer, Wiley, and others. He is an author of 275 research publications and an inventor of 10 U.S./PCT patents. His research areas of interest are cybersecurity, digital authentication, biometrics, multimedia security, and technological innovation management. He is a Fellow of the IET, BCS, and FTRA, a member of the IEEE Technical Committee on Security & Privacy, and a member of the IEEE Cybersecurity community.

FAN WU (fwu@cs.sjtu.edu.cn) is an associate professor in the Department of Computer Science and Engineering at Shanghai Jiao Tong University. He received his B.S. in computer science from Nanjing University in 2004, and his Ph.D. in computer science and engineering from the State University of New York at Buffalo in 2009. He visited the University of Illinois at Urbana-Champaign as a postdoctoral research associate. His research interests include wireless networking and mobile computing, algorithmic network economics, and privacy preservation.

GUIHAI CHEN (gchen@cs.sjtu.edu.cn) earned his B.S. degree from Nanjing University in 1984, his M.E. degree from Southeast University in 1987, and his Ph.D. degree from the University of Hong Kong in 1997. He is a Distinguished Professor of Shanghai Jiao Tong University. He had been invited as a visiting professor by many universities including the Kyushu Institute of Technology, Japan, in 1998, the University of Queensland, Australia, in 2000, and Wayne State University, Michigan, from September 2001 to August 2003. He has a wide range of research interests with focus on sensor networks, peer-to-peer computing, high-performance computer architecture, and combinatorics.

PENG ZENG (zp@sia.cn) received his B.S. degree in computer science from Shandong University in 1998 and his Ph.D. degree in mechatronic engineering from Shenyang Institute of Automation (SIA), Chinese Academy of Sciences (CAS), in 2005. Currently, he is a professor and Ph.D. supervisor at SIA, CAS. His research interests include industrial communication, smart grids, demand response, and wireless sensor networks. He is a member of IEC TC65 WG16 and a member of the SP100 Standard Committee, ISA.

# Smart Health Solution Integrating IoT and Cloud: A Case Study of Voice Pathology Monitoring

Ghulam Muhammad, SK Md Mizanur Rahman, Abdulhameed Alelaiwi, and Atif Alamri

## ABSTRACT

The integration of the IoT and cloud technology is very important to have a better solution for an uninterrupted, secured, seamless, and ubiquitous framework. The complementary nature of the IoT and the cloud in terms of storage, processing, accessibility, security, service sharing, and components makes the convergence suitable for many applications. The advancement of mobile technologies adds a degree of flexibility to this solution. The health industry is one of the venues that can benefit from IoT-Cloud technology, because of the scarcity of specialized doctors and the physical movement restrictions of patients, among other factors. In this article, as a case study, we discuss the feasibility of and propose a solution for voice pathology monitoring of people using IoT-cloud. More specifically, a voice pathology detection system is proposed inside the monitoring framework using a local binary pattern on a Mel-spectrum representation of the voice signal, and an extreme learning machine classifier to detect the pathology. The proposed monitoring framework can achieve high accuracy of detection, and it is easy to use.

## INTRODUCTION

The Internet of Things (IoT) refers to some intelligent and uniquely addressable objects (things) that are interconnected to a worldwide network. In general, the IoT can be a set of real-world small devices that are distributed, and have limited processing power and storage. On the other hand, cloud computing can have huge storage and processing power, where security can also be trusted. Therefore, a convergence of the IoT and the cloud can render a wide application in daily and social life; it is expected to raise the number of such applications in the near future [1].

There are some complementary attributes between cloud computing and IoT. These attributes include:

- Centralized vs. pervasive nature
- Virtual vs. real-world things
- Omnipresent vs. limited computation
- Huge vs. no or restricted storage

Table 1 shows the complementary nature of cloud computing and IoT.

The health industry is a major industry that

has no end of demands. It provides very crucial services to humans, and generates huge revenues. The competition between companies in this industry is to provide service that is sophisticated, always available, accurate, and low cost [2]. Hence, efficient use of the IoT and the cloud is one of the most researched topics in this important industry.

There are several categories of IoT devices related to healthcare. These categories include:

- Smart wristbands, such as FitBit, Healbe, Misfit, and Nabu Razer
- Wearable devices, such as portable insulin syringes
- Internal devices, such as implanted hearing aids
- Stationary devices, such as ECG machines and stroboscopes

Smart wristbands can communicate via Bluetooth to nearby personal smartphones; wearable devices can communicate with smartphones using a dedicated wireless protocol, while the internal or stationary devices can communicate wirelessly.

There are different types of sensors that can be used in the medical domain for simplicity and ubiquitous nature. For example, HealthPatch MD is a biosensor attached to the chest that can measure and track a person's heart rate, body temperature, respiratory rate, and body movement, in addition to fall detection. Zio XT Patch can detect an abnormal heartbeat rate over a certain period. Figure 1 shows a framework of the integration of the IoT and the cloud. A hosting device, most preferably a smart device, captures data from different IoT through a local area network (LAN) interface (e.g., using Bluetooth). The device then sends the heterogeneous data to the cloud using a wide area network (WAN) interface. The data transfer can be realized by using WiFi or fourth/fifth generation (4G/5G) technology. One of the concerns is secured transmission, which can be a task of the service provider [3].

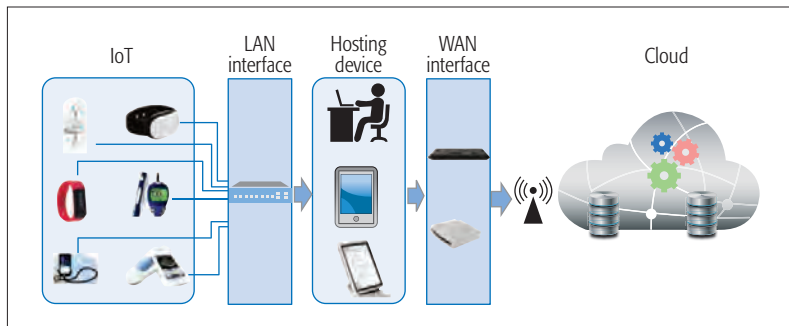
## RESEARCH CONTRIBUTION

In this article, we discuss the feasibility of voice pathology monitoring using the integration of the IoT and the cloud. We propose a framework (Fig. 2) where voice signals are captured through the IoT and are sent to a hosting device, such as a smartphone. Other signals about health and the

The authors discuss the feasibility of and propose a solution for voice pathology monitoring of people using IoT-cloud. More specifically, a voice pathology detection system is proposed inside the monitoring framework using a local binary pattern on a Mel-spectrum representation of the voice signal, and an extreme learning machine classifier to detect the pathology. The proposed monitoring framework can achieve high accuracy of detection, and it is easy to use.

Attributes	Cloud computing	IoT
Storage	Huge or unlimited	None or limited
Accessibility	Any time	Restricted
Processing	Huge power	Limited power
Things	Virtual	Real-world
Distance	Omnipresent, anywhere in the world	Limited
Components	Centralized	Heterogeneous
Big data	Managed	Cradled
Security	More secure	Less secure

**Table 1.** Comparative attributes of cloud computing and IoT.



**Figure 1.** Integration of IoT and cloud.

ambient conditions are also captured using different IoT sensors. These signals can be used to have extra information that can complement the decision making by a doctor. The hosting device sends the signals to the cloud, where the signals are processed after verifying the authenticity of the user. The medical doctor can get the processed data to analyze and make a decision, which is then fed back to the patients. For voice pathology detection, we use a local binary pattern (LBP) as a feature of the signals and an extreme learning machine (ELM) as a classifier. The LBP is a widely used texture descriptor in image processing applications, and the ELM is an efficient and less computationally expensive classifier. To the best of our knowledge, this is the first time that the combination of the LBP and ELM is used in voice pathology monitoring.

### VOICE PATHOLOGY MONITORING

Voice pathology can be defined as an abnormal development or growth of mass or tissue in the vocal folds resulting in reduced voice quality. Abnormal growths include polyps, cysts, nodules, sulci, and so on in the vocal folds. The voice of a patient with voice pathology sounds hoarse, strained, or breathy, and there is voice breaking or loss of pitch in the voice. These symptoms depend on the severity of the pathology [4]. Professionals in teaching and music, or where there is an excessive use of vocal cords, suffer from this voice pathology more than other professionals. There are some voice disorders, such as spasmodic dysphonia, caused by the involuntary move-

ment of the muscles; these are not limited to any particular profession. There are more than 7.5 million people suffering from voice pathology in the United States.

There are several methods to diagnose voice pathology. These methods can be broadly classified as subjective and objective. In subjective evaluation, there are two techniques: invasive and non-invasive. In the invasive technique, an expert doctor may use a laryngoscope or stroboscope to see the condition of the vocal folds; these devices are invasive and may cause discomfort to patients. They are also expensive. In the non-invasive technique, expert doctors listen to the voice and rate it using the consensus auditory perceptual evaluation of voice (CAPE-V) or grade, roughness, breathiness, asthenia, and strain (GRBAS). The problem of subjective evaluation is that it depends on the experience of the doctor, the severity of the pathology, the rating scale used, and the speaking style. Based on these limitations, researchers have been trying to develop an automated way to detect and classify voice pathology. The advantage of objective evaluation is that it is non-invasive, low-cost, and independent of expertise, and can be used offshore. Despite the advantage of objective evaluation, we firmly believe that this is only an assistive tool to the professional medical doctor; the final decision should be made by the doctor. In this aspect, software-defined healthcare networks are also used at present [5].

Investigating a sustained vowel /a/ is a popular choice for voice pathology detection, because it has clearly separated formants and is easy to pronounce by a patient; however, in practical life, we cannot limit a patient to pronouncing a specific phone for a certain period of time (sustained). There should be a mechanism that can detect voice pathology from daily conversational speech or continuous speech. Continuous speech has some interesting parts, such as voice onset and offset and voice break, which can be crucial in determining the voice pathology.

There is almost no previous work of voice pathology monitoring using IoT-cloud; however, this work is important due to several aspects. It has consulted with clinics and found that most patients, after having their treatment, do not go to the clinic again for follow-up. The follow-up is very crucial to check the progress of the vocal folds' abnormal growth and the voice quality. If the progress is minimal, the doctors advise them to have extra treatment. People do not go physically to the clinic for follow-up mainly because of finding time to go there. An IoT-cloud-based framework can solve this problem. The voice data can be obtained from the patient's home through the IoT, processed in the cloud, and sent to the clinic for evaluation by the doctors. The doctors evaluate the data, and send the patient necessary advice.

### THE PROPOSED SYSTEM

In the proposed system, the IoT related to capturing voice, body temperature, electrocardiogram, and ambient humidity is used. We exclude devices such as laryngoscope and stroboscope because they are difficult for a patient to operate. The data captured by the IoT are sent by



Bluetooth technology to the patient's smartphone using a developed app. For authentication purposes, a simple but robust watermark is embedded into the signals. The watermark is a personalized identification of the patient, which is created by the patient himself. Watermark embedding is a very important step in the proposed system, because it protects the ownership of the personal data. For watermarking, we use the algorithm proposed in [6] because of its robustness against different attacks. In this watermarking scheme, a discrete wavelet transform (DWT) and a singular value decomposition (SVD) based algorithm is utilized. The patient ID is used as the watermark, and it is embedded using SVD in the detailed coefficients subband at level 2 of the voice signal decomposition using DWT. The watermarked signals are transmitted to the cloud through the Internet.

The cloud has the following main components: authentication manager, data manager, feature extraction server, classification server, and storage [7]. The authentication manager first checks whether the patient is enrolled in the system or not; if enrolled, it extracts his/her identity. It also manages the request access by the doctors by verifying their identity. The data manager controls the flow of data to and from the servers [8]. In the feature extraction server, features are extracted from the signal, while in the classification server, the features are processed using a classifier to classify the signal. The decision is sent to the data manager, and the features and models of the classifier are stored in the storage.

For the feature extraction stage, the signal is first divided into overlapping frames of 30 ms. The frames are Hamming windowed. Any bias is removed by mean subtraction. The time-domain frames are converted into frequency-domain representation using the Fourier transform. A bank of band-pass filters, whose center frequencies are spaced on an auditory perceptual scale (e.g., Mel scale) is applied to the frequency-domain signal. After this stage, we get a spectro-temporal representation of the signal [9]. The LBP is applied to this representation. The LBP is a powerful texture descriptor that takes into account the relative contribution in a neighborhood compared to the center point in the neighborhood [10]. There are many variants of the LBP, but in our proposed system, we use the simplest one, which is a rectangular mask of  $3 \times 3$  dimensions. The LBP finds the dominant spectro-temporal information in the signal. This information is helpful for the classifier, because it has been observed that the dominant frequencies in the case of a voice having pathology are in the high-frequency regions due to its noisy characteristics.

In the classification server, the signal is classified as normal or pathological using the ELM classifier. The ELM is an efficient and fast learning algorithm that has recently been used in many applications, including optical character recognition, 3D shape classification, ECG classification, and traffic sign classification [11]. There are several interesting characteristics of the ELM:

- Learning is not required to be iterative-tuned.
- Layers can be learned one at a time.
- Output weights can be determined analytically.

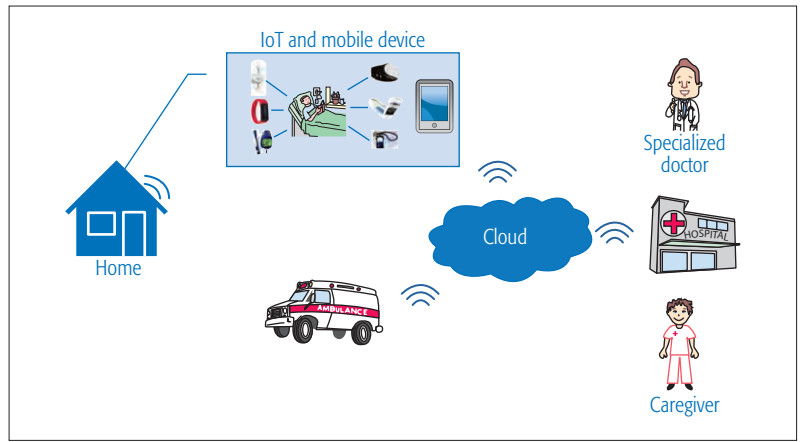


Figure 2. Health monitoring framework using IoT and cloud.

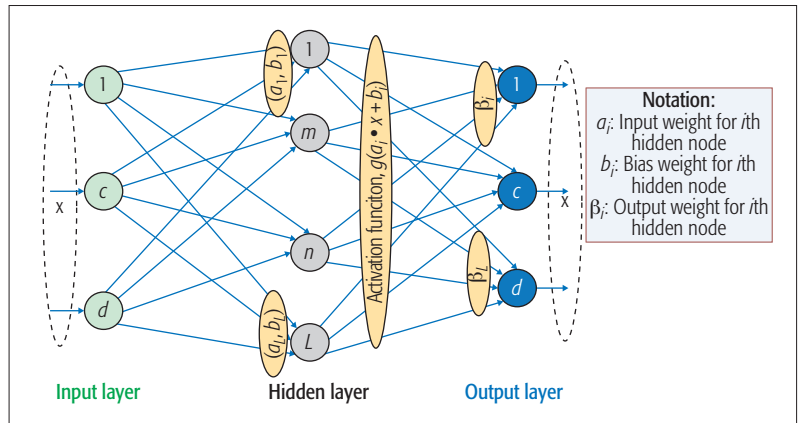


Figure 3. A simplified architecture of the ELM.

- The whole network can be solved in a few steps with less computational complexity.
- A uniform solution can be obtained for N-class and regression problems.
- Feature mapping can be realized on either a known space or an unknown space.

The ELM is based on a single-layer feed-forward network, where the  $L$ th hidden node is represented by the following equation:

$$f_L(\mathbf{x}) = \sum_{i \in L} g(\mathbf{a}_i \cdot \mathbf{x} + b_i) \cdot \beta_i, \quad \mathbf{a}_i \in \mathcal{R}^d, \beta_i \in \mathcal{R}$$

where  $g$  is the activation function (we use radial basis function),  $\mathbf{x}$  is the input feature vector,  $\mathbf{a}_i$  is the input weight vector for the  $i$ th hidden layer node,  $b_i$  is the bias weight, and  $\beta_i$  is the output weight for the  $i$ th hidden layer node. There are  $d$  input layer nodes. The basic idea is to assign input weight vector and bias weight randomly for all the hidden-layer nodes. Then calculate the output matrix,  $\mathbf{H}$ , of the hidden layer using  $N$  training samples. The output weights can be found by the following equation:

$$\beta = (C^{-1} + \mathbf{H}\mathbf{H}^T)^{-1}\mathbf{H}^T\mathbf{M}$$

where  $C^{-1}$  is a positive penalty parameter, and  $\mathbf{M}$  is the teacher matrix of the training samples.

In the proposed system, there are only two classes, which are normal and pathological. In a future system, we will also include the classes of voice pathologies. The steps of the proposed system are summarized below:

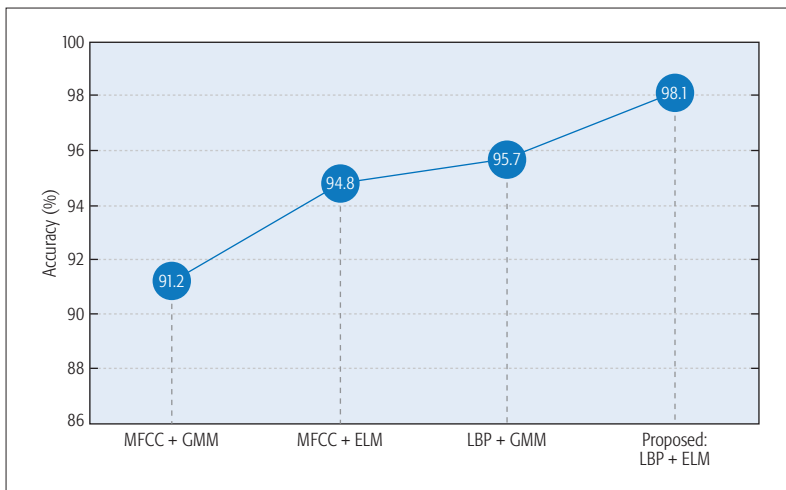


Figure 4. Voice pathology detection accuracy of different systems.

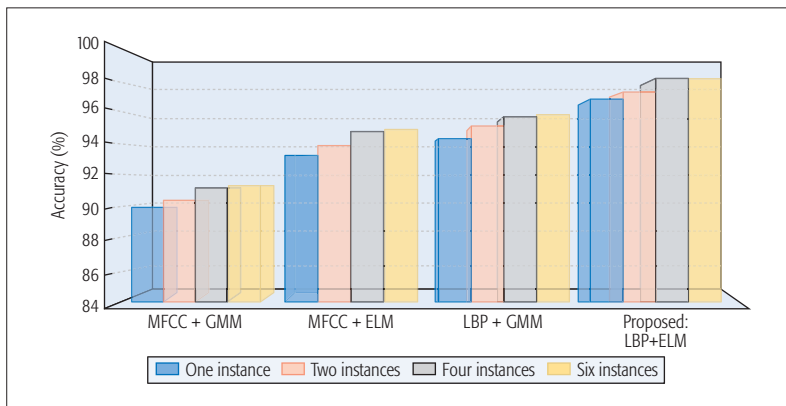


Figure 5. Voice pathology detection accuracy of different systems using different instances of cloud servers.

- Capture voice input by using IoT, and send.
- The patient identification number is inserted as a watermark into the voice signal.
- The signal is transferred to the cloud server, where the verification of the patient is first performed.
- In the cloud server, the time-domain signal is converted into the frequency-domain signal, and the LBP features are extracted.
- The ELM is used as the classifier, which determines whether the input voice signal is normal or pathological.
- The decision on the signal is sent to the registered doctors and nurses.

The other signals, which are body temperature, electrocardiogram, and ambient humidity, are not processed for feature extraction and classification, because they are the supportive information that may be required by the doctors. For example, if the humidity is high, the patient may feel difficulty in pronouncing or speaking, which may result in unnecessary performance degradation of the system; the electrocardiogram may indicate a tense or breathy condition of the patient; and so on.

We simulated the whole framework using some publicly available databases to find the accuracy of the proposed system. The databases used are the Massachusetts Eye and Ear Infirmary (MEEI) database and the Saarbrücken Voice Database (SVD). We used these two databases

because they are widely utilized in voice pathology detection research, and they are publicly available. The samples of the databases were played in a normal room at medium volume. A portable voice recorder with Bluetooth technology recorded the played samples at a sampling frequency of 11.25 kHz. The recorded samples were then transferred to a smartphone, where the watermark was embedded. There are 53 normal samples in the MEEI database and more than 1500 samples in the SVD, while those numbers for pathological samples are larger than 500 and 1500, respectively.

The samples of the two databases were mixed, because we did not want to develop a system for a specific database. The system should be universal irrespective of recording conditions and speakers. We compared the proposed system with some other systems. The compared systems are:

- Features: Mel-frequency cepstral coefficients (MFCCs); classifier: the Gaussian mixture model (GMM)
- Features: MFCC; classifier: ELM
- Features: LBP; classifier: GMM
- The proposed system.

The classification was performed using a 10-fold cross validation approach. In this approach, the whole dataset is randomly divided into 10 equal subsets; in each iteration, nine subsets are used for the training, while the other one is for testing. The final accuracy is obtained by averaging the 10 accuracies. Different instances of cloud servers were used to see the effect of the number of servers on the accuracy. Figure 4 shows the average accuracies of the systems using four instances of cloud servers, which achieved the optimal accuracy. From the figure, we see that the proposed system outperformed all the other systems that we compared. The best accuracy, which is 98.1 percent, was obtained by the proposed LBP + ELM system. In this case, the time required by the system per voice sample is 2.4 s for detection. Figure 5 shows accuracies using different instances of cloud servers. As we see from the figure, if we increase the number of instances, the accuracy increases; however, there is hardly any improvement after four instances. In addition to these experiments, we also performed experiments without embedding the watermark in the signal; however, we noticed that there was no significant degradation of performance when we embedded the watermark.

## CONCLUSIONS

A healthcare framework based on the IoT and the cloud is discussed and proposed. A voice pathology monitoring system inside the framework is developed using the LBP features and the ELM classifier. The proposed system experimentally proved to be accurate. There are several issues that need to be addressed before this type of system can be fully operative in a trustable manner. These issues include dynamic scalability, secured transmission, availability, ease of users, and interoperability. In the proposed system, interoperability and ease of users are solved. Secure transmission can be guaranteed by the service provider; however, we embed a watermark into the signal for authenticity.

A future research direction can be to achieve dynamic scalability of the system by integrating different input modalities of voice, for example, microphones, smart devices, high-speed cameras for recording vocal folds, and neck sensors. Another research direction is to create a framework of how to handle the big data of pathological and normal voice in the cloud; a recently proposed particle swarm optimization based solution in the mobile environment can be used in this aspect [12].

#### ACKNOWLEDGMENT

This work is supported by the Deanship of Scientific Research at King Saud University, Riyadh, Saudi Arabia, through the International Research Group Project No. 14-204.

#### REFERENCES

- [1] A. Botta *et al.*, "Integration of Cloud Computing and Internet of Things: A Survey," *Future Generation Computer Systems*, vol. 56, Mar. 2016, pp. 684–700.
- [2] M. S. Hossain and G. Muhammad, "Cloud-Assisted Industrial Internet of Things (IIoT)-Enabled Framework for Health Monitoring," *Computer Networks*, vol. 101, 2016, pp. 192–202, DOI: 10.1016/j.comnet.2016.01.009.
- [3] M. Henze *et al.*, "A Comprehensive Approach to Privacy in the Cloud-Based Internet of Things," *Future Generation Computer Systems*, vol. 56, Mar. 2016, pp. 701–18.
- [4] G. Muhammad *et al.*, "Voice Pathology Detection Using Interlaced Derivative Pattern on Glottal Source Excitation," *Biomedical Signal Processing and Control*, vol. 31, Jan. 2017, pp. 156–64, DOI: 10.1016/j.bspc.2016.08.002.
- [5] L. Hu *et al.*, "Software Defined Healthcare Networks," *IEEE Wireless Commun.*, vol. 22, no. 6, Dec. 2015.
- [6] M. Alhussein and G. Muhammad, "Watermarking of Parkinson Disease Speech in Cloud-Based Healthcare Framework," *Int'l. J. Distributed Sensor Networks*, vol. 2015, article ID 264575, 2015, 9 pages, DOI: 10.1155/2015/264575.
- [7] A. Al-Fuqaha *et al.*, "Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 4, 4th qtr. 2015, pp. 2347–76.
- [8] Y. Sun *et al.*, "Internet of Things and Big Data Analytics for Smart and Connected Communities," *IEEE Access*, vol. 4, 2016, pp. 766–73.

- [9] G. Muhammad, "Automatic Speech Recognition Using Interlaced Derivative Pattern for Cloud Based Healthcare System," *Cluster Computing*, vol. 18, no. 2, June 2015, pp. 795–802.
- [10] T. Ahonen, A. Hadid, and M. Pietikainen, "Face Description with Local Binary Patterns: Application to Face Recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 28, no. 12, 2006, pp. 2037–41.
- [11] G.-B. Huang, "What Are Extreme Learning Machines? Filling the Gap between Frank Rosenblatt's Dream and John von Neumann's Puzzle," *Cognitive Computation*, vol. 7, 2015, pp. 263–78.
- [12] M. S. Hossain *et al.*, "Big Data-Driven Services Composition Using Parallel Clustered Particle Swarm Optimization in Mobile Environment," *IEEE Trans. Service Computing*, vol. 9, no. 5, Sept./Oct. 2016.

#### BIOGRAPHIES

GHULAM MUHAMMAD [M'10] (ghulam@ksu.edu.sa) is an associate professor in the Computer Engineering Department, College of Computer and Information Sciences (CCIS), King Saud University, Riyadh, Saudi Arabia. He received his Ph.D. degree in 2006 from the Department of Electronic and Information Engineering at Toyohashi University of Technology, Japan. He has authored and co-authored more than 140 publications including refereed IEEE/ACM/Springer/Elsevier journals, conference papers, and book chapters. He owns a U.S. patent.

SK. MD. MIZANUR RAHMAN [M'10] is an assistant professor in the Information System Department, CCIS, King Saud University. He worked as a postdoctoral researcher at the University of Ottawa, University of Ontario Institute of Technology, and University of Guelph, Canada. His research interests include cryptography, software security, information security, privacy enhancing technology, and network security. He has published over 60 peer-reviewed journal articles, conference papers, and book chapters.

ABDULHAMEED AL ELAIWI [M'12] is an assistant professor in the Software Engineering Department, CCIS, King Saud University. He has authored and co-authored more than 40 publications. His research interests include software testing, cloud collaboration, multimedia cloud, sensor-cloud, mobile cloud, and E-learning systems.

ATIF ALAMRI [M'09] is an associate professor in the Information Systems Department, CCIS, King Saud University. He was a Guest Associate Editor of *IEEE Transactions on Instrumentation and Measurement* and a Co-Chair of the 10th IEEE International Symposium on Haptic Audio Visual Environments and Games, and serves as a Program Committee Member of many conferences in multimedia, virtual environments, and medical applications.

A future research direction can be to achieve dynamic scalability of the system by integrating different input modalities of voice, for example, microphone, smart devices, high-speed camera for recording vocal folds, and neck sensors.

## ENABLING MOBILE AND WIRELESS TECHNOLOGIES FOR SMART CITIES



Ejaz Ahmed



Muhammad Imran



Mohsen Guizani



Ammar Rayes



Jaime Lloret



Guangjie Han



Wael Guibene

Due to advancements in communication and computing technologies, smart cities have become a main innovation agenda of research organizations, technology vendors, and governments. To make a city smart, a strong communications infrastructure is required for connecting smart objects, people, and sensors. Smart cities rely on wireless and mobile technologies for providing services such as healthcare assistance, security and safety, real-time traffic monitoring, and managing the environment, to name a few. Such applications have been a main driving force in the development of smart cities. Without the appropriate communication networks, it is really difficult for a city to facilitate its citizens in a sustainable, efficient, and safer manner/environment. Considering the significance of mobile and wireless technologies for realizing the vision of smart cities, there is a need for conducting research to further investigate the standardization efforts and explore different issues/challenges in wireless technologies, mobile computing, and smart environments.

For this *IEEE Communications Magazine* Feature Topic (FT), we invited researchers from academia, industry, and government to discuss challenging ideas, novel research contributions, demonstration results, and standardization efforts on enabling mobile and wireless technologies for smart cities. After a rigorous review process, 17 papers have been selected to be published in this FT of *IEEE Communications Magazine*. Seven of these are published here in Part 1 of the FT.

LTE/LTE-A is one of the promising communication technologies for smart cities. M. S. Ali *et al.*, in "LTE/LTE-A Ran-

dom Access for Massive Machine-Type Communications in Smart Cities," present a review on recent advances in the random access (RA) mechanism of LTE/LTE-A. Based on the study, they highlight the key limitations of the RA mechanisms. Further, they propose a collision resolution-based RA (CRB-RA) model for massive machine-type communication over LTE/LTEA. The model focuses on managing the bursty and massive access attempts.

The Internet of Things is one of the enabling technologies for smart cities where the devices and applications running in them require energy management solutions. W. Ejaz *et al.*, in "Efficient Energy Management for Internet of Things in Smart Cities," briefly present an overview of energy management solutions and discuss the challenges in designing the energy management solutions for smart cities. They propose a framework of energy-efficient scheduling for IoTs in smart cities. Finally, two case studies on energy efficient scheduling and wireless power transfer in IoT are also discussed.

Geo-conquesting is an emerging computational advertising technology for the smart cities. B-W. Chen *et al.*, in "Geo-Conquesting Based on Crowdsourced Metatrails from Mobile Sensing," leverage the crowdsourced metatrails for geo-conquesting. A graph clustering approach is used to extract the sequential visiting patterns and affinity subnetworks of a city. The sequential patterns and affinity subnetworks are used to investigate the activities of crowd sequential. Lastly, an interesting discussion on the challenges in smart marketing is also provided for readers.

The key services in smart cities are based on media streaming, which imposes high bandwidth requirements

on mobile networks. J. M. Batalla *et al.*, in “Efficient Media Streaming with Collaborative Terminals for the Smart City Environment,” propose a collaborative framework for adaptive media streaming received by multi-path transmission. The solution combines network assisted device cooperation with adaptive streaming operations to guarantee optimized resource allocation in both ways: device-to-device and base-station-to-device.

Named Data Networking (NDN) is another promising technology for smart cities that can be integrated with Intelligent Transportation Systems (ITS) to meet the demands of data-intensive applications. S. H. Bouk *et al.*, in “Named Data Networking Based ITS for Smart Cities,” propose an architecture for NDN-based ITS in smart cities. The proposed architecture comprises naming, caching, and cache replacement policies, face management, content segmentation and reassembly, communication reliability, application services, forwarding strategy, management, and security components. The research challenges for enabling the NDN-based ITS in the smart cities are also highlighted in the article.

Interference, user mobility, and high energy consumption are some open issues hindering flawless connectivity in smart cities. I. Yaqoob *et al.*, in “Enabling Communication Technologies for Smart Cities,” conduct a study to explore the enabling communication technologies and highlight the issues that remain to be addressed. Recent research efforts are investigated by analyzing the strengths and weaknesses. Moreover, the literature is classified by devising a taxonomy. Capabilities of the modern communication technologies are also analyzed by performing comparisons based on important parameters. A few notable use cases are also discussed.

Security and privacy of data are among the most important concerns in the smart city applications. Kuan Zhang *et al.*, in “Security and Privacy in Smart City Applications: Challenges and Solutions,” investigate the security and privacy concerns in applications designed for smart cities. Further, the recent advances in addressing these challenges are also presented. Finally, the authors highlight several research directions with respect to security and privacy of applications in smart cities.

We would like to sincerely thank all the people, including the contributing authors, the anonymous reviewers, and the *IEEE Communications Magazine* publications staff, who have significantly contributed to this FT. We believe that the

research findings presented in this FT will stimulate further research and development ideas for mobile and wireless technologies of smart cities.

#### BIOGRAPHIES

EJAZ AHMED has worked as a researcher at C4MCCR, University of Malaya, Malaysia, CogNet Lab, NUST, and CoReNet, MAJU, Pakistan. He is an Associate Technical Editor of *IEEE Communications Magazine*, *IEEE Access*, *Springer MJCS*, and *Elsevier JNCA*. He has also served as a Lead Guest Editor for the *Elsevier FGCS Journal*, *IEEE Access*, *Elsevier Computers & Electrical Engineering*, *IEEE Communications Magazine*, *Elsevier Information Systems*, and *Transactions on Emerging Telecommunications Technologies*.

MUHAMMAD IMRAN is currently working at King Saud University and is a visiting scientist at Iowa State University. His research interests include MANETs, WSNs, WBANs, M2M/IoT, SDN, and security and privacy. He has published a number of research papers in refereed international conferences and journals. He serves as a Co-Editor-in-Chief for *EAI Transactions* and Associate/Guest Editor for *IEEE Access*, *IEEE Communications Magazine*, *Computer Networks*, *Sensors*, *IJDSN*, *JIT*, *WCWC*, *AHSWN*, *IET WSS*, *IJAACS*, and *IJITEE*.

MOHSEN GUIZANI [S'85, M'89, SM'99, F'09] received his B.S., M.S., and Ph.D. from Syracuse University. He is currently a professor and the ECE Department Chair at the University of Idaho. His research interests include wireless communications/mobile cloud computing, computer networks, security, and smart grid. He is the author of nine books and 400+ publications. He was the Chair of the IEEE Communications Society Wireless Technical Committee. He served as an IEEE Computer Society Distinguished Speaker.

AMMAR RAYES [S'85, M'91, SM'15] is a Distinguished Engineer focusing on the technology strategy for Cisco Services. His research interests include IoT, network management NMS/OSS, machine learning, analytics, and security. He has authored three books, over 100 publications in refereed journals and conferences on advances in software & networking related technologies, and over 25 patents. He received B.S. and M.S. degrees from the University of Illinois at Urbana and his D.Sc. degree from Washington University, all in electrical engineering.

JAIME LLORET [M'07, SM'10] received his M.Sc. in physics in 1997, his M.Sc. in electronic engineering in 2003, and his Ph.D. in telecommunication engineering in 2006. He is the head of the Communications and Networks research group of the Research Institute IGIC. He is Editor-in-Chief of *Ad Hoc and Sensor Wireless Networks* and *Network Protocols and Algorithms*. He has been general chair of 36 International workshops and conferences. He is an IARIA Fellow.

GUANGJIE HAN [S'01, M'05] is currently a professor with the Department of Information and Communication System, Hohai University, China. His current research interests include sensor networks, computer communications, mobile cloud computing, and multimedia communication and security. He has served on the Editorial Boards of up to 14 international journals, including *IEEE Access* and *Telecommunication Systems*. He has guest edited a number of Special Issues in IEEE journals and magazines. He is a member of ACM.

WAEEL GUIBENE has been a research scientist at Intel Labs since June 2015. He was awarded his Ph.D. from Telecom ParisTech in July 2013. He also holds an M.Eng. and a Master's degree in telecommunications obtained in 2009 and 2010, respectively. He worked at Eurecom as research engineer from 2010 to November 2013, and then joined Semtech to work on LoRa systems from 2013 to June 2015. His research activities include IoT, 5G, and wireless communications.

# LTE/LTE-A Random Access for Massive Machine-Type Communications in Smart Cities

Md Shipon Ali, Ekram Hossain, and Dong In Kim

The authors review the current state-of-the-art proposals to control massive random access of MTC devices in LTE/LTE-A networks. The proposals are compared in terms of five major metrics: access delay, access success rate, power efficiency, QoS guarantee, and the effect on HTC. To this end, the authors propose a novel collision resolution random access model for massive MTC over LTE/LTE-A.

## ABSTRACT

Massive MTC over cellular networks is expected to be an integral part of wireless smart city applications. The LTE/LTE-A technology is a major candidate for provisioning of MTC applications. However, due to the diverse characteristics of payload size, transmission periodicity, power efficiency, and QoS requirement, MTC poses huge challenges to LTE/LTE-A technologies. In particular, efficient management of massive random access is one of the most critical challenges. In the case of massive random access attempts, the probability of preamble collision drastically increases, and thus the performance of LTE/LTE-A random access degrades sharply. In this context, this article reviews the current state-of-the-art proposals to control massive random access of MTC devices in LTE/LTE-A networks. The proposals are compared in terms of five major metrics: access delay, access success rate, power efficiency, QoS guarantee, and the effect on HTC. To this end, we propose a novel collision resolution random access model for massive MTC over LTE/LTE-A. Our proposed model basically resolves the preamble collisions instead of avoidance and targets the management of massive and bursty access attempts. Simulations of our proposed model show huge improvements in random access success rate compared to the standard slotted-Aloha-based models. The new model can also coexist with existing LTE/LTE-A MAC protocol and ensure high reliability and time-efficient network access.

## INTRODUCTION

The term *smart city* represents an environment in which all a city's assets are virtually connected and electronically managed. Smart utility, e-health, online education, e-library, online surveillance, environment monitoring, and connected vehicles are some smart city applications. For such an application, a huge number of autonomously operated, low-cost devices (i.e., sensors, actuators) need to be connected to physical objects. The communications between these autonomously operated devices are called machine-type communications (MTC), and the MTC devices (MTCs) form an integral part of a smart city environment. On the other hand, due to the requirements of mobility, extended coverage area,

security, diverse quality of service (QoS), etc., a large percentage of MTCs will need to connect directly to cellular networks. The orthogonal frequency-division multiple access (OFDMA)-based LTE<sup>1</sup> technologies are major cellular technologies that will need to support the MTC applications in smart cities.

Random access (RA) is the first step to initiate a data transfer using an LTE network. According to Third Generation Partnership Project (3GPP) specifications, contention-based RA occurs in the following cases:

- Initial access to the network
- Recovering a radio resource connection (RRC)
- Data transfer and location identification during RRC-connected state when uplink is not synchronized

RA management is the most challenging task to support massive MTC in LTE systems. The medium access control (MAC) layer in LTE systems is based on the slotted Aloha protocol, and severe congestion during RA is generally expected due to the irregular and bursty nature of transmissions by MTCs.

To resolve the RA congestion in LTE systems, different solutions have been proposed. In this article, we provide a review of these proposals in terms of five key performance metrics: access delay, access success rate, QoS guarantee, energy efficiency, and the impact on HTC traffic. Nonetheless, most of the solutions are based on the collision avoidance technique, which simply restricts the arrival rate of access attempts. This results in large access delay, and therefore, the QoS requirements may not be satisfied for some MTCs. This motivates us to develop a novel collision-resolution-based RA approach, where an  $m$ -ary contention tree splitting technique [1] is applied to resolve collisions among preambles during random access. In this approach, the base station (BS), for example, the evolved node B (eNB) in an LTE network, resolves RA collisions by scheduling the collided MTCs into a set of reserved opportunities. In [2], a different tree splitting RA model was studied. Different from that in [2], our proposal is able to handle massive bursty traffic and can also coexist with the existing LTE MAC protocol without any major modifications.

<sup>1</sup> We use the term LTE to refer to both LTE and LTE-Advanced (LTE-A) technologies.

The rest of the article is organized as follows. We first review the contention-based RA process in LTE systems. Major limitations of the existing approaches are then presented, where a particular MTC application is studied to understand the limitation of slotted-Aloha-based RA protocol. Next, we provide a survey of the existing RA congestion control proposals, which is followed by our proposed collision resolution approach. Simulation results for the proposed approach are presented and compared to those for the standard LTE RA process.

## CONTENTION-BASED RANDOM ACCESS IN LTE

### RANDOM ACCESS PREAMBLE

Random access preambles are the orthogonal bit sequences, called digital signature, used by user equipments (UEs) to initiate an RA attempt. RA preambles are generated by cyclically shifting a root sequence such that every preamble is orthogonal to each other. There are 64 preambles in total, which are initially divided into two groups: contention-free RA preambles and contention-based RA preambles. The eNB reverses some preambles, say  $N_{cf}$ , for contention-free RA, and assigns distinct preambles to different UEs. The rest of the preambles ( $64 - N_{cf}$ ) are used for contention-based RA, where each UE randomly generates one preamble [3].

### RANDOM ACCESS SLOT

A random access slot (RA slot) refers to the LTE physical radio resources, called physical random access channel (PRACH), in which RA preambles are mapped and transmitted to the eNB. In frequency-division duplex (FDD) operation [Fig. 1], an RA slot consists of six physical resource blocks (RBs) in the frequency domain, while the time duration of each RA slot can be one, two, or three subframe(s) depending on the preamble format [3]. There are a total of 864 subcarriers in one RA slot, which are equally distant at 1.25 kHz. All 64 preambles are mapped into 839 centred RACH subcarriers, while the remaining 25 subcarriers are used as guard frequency [3].

In time-division duplex (TDD) operation, four different preamble formats are available based on preamble cyclic prefix duration ( $T_{CP}$ ), and preamble sequence duration ( $T_{SEQ}$ ) [3]. A UE can select an appropriate preamble under a specific format depending on the distance from the eNB, maximum delay spread, amount of transmission resource needed to transmit RRC request, and so on. On the other hand, the number of RA slots in each radio frame is defined by the preamble configuration index. For each preamble format 16 different indices are available, where the eNB allocates radio resource as PRACH. Depending on system bandwidth, some LTE systems may not be able to use some preamble configuration indices. However, systems using 20 MHz bandwidth are able to use all of the indices [3]. The eNB periodically broadcasts the preamble information as a part of system information block 2 (SIB2) message.

### CONTENTION-BASED RANDOM ACCESS PROCEDURE

When a UE is switched on or wakes up, it first synchronizes with the LTE downlink channels by decoding the primary and secondary syn-

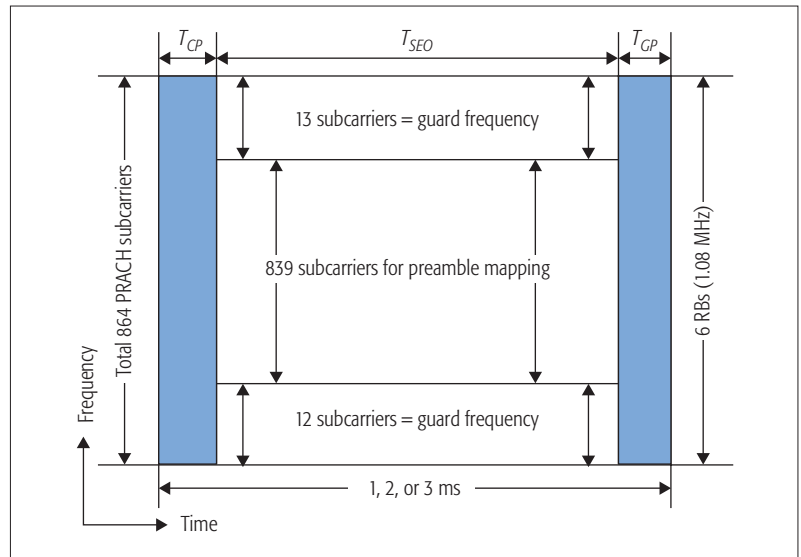


Figure 1. FDD-based RA slot in time-frequency resources.

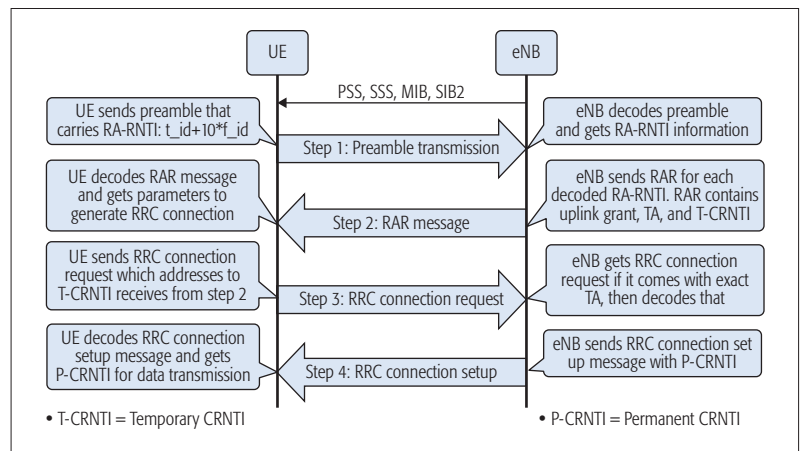


Figure 2. Contention-based RA procedure.

chronization signals (PSS and SSS). The UE then decodes the master information block (MIB), which contains information about the location of the downlink and uplink carrier configurations, and thus gets the information of SIBs. All the RA parameters, that is, RA slots, preamble formats, preamble configuration indices, and so on, are contained in SIB2. Therefore, after decoding SIB2, UEs can generate a contention-based RA attempt. The contention-based LTE RA procedure (Fig. 2) consists of four main steps as follows.

**Preamble Transmission from UE to eNB:** To initiate contention-based RA, the UE randomly generates one of the available contention-based preambles and sends that to the eNB at the next available RACH slot. Due to the orthogonal properties, different preambles can easily be decoded at the eNB unless multiple UEs transmit the same preamble at the same RA slot. After sending a preamble, a UE waits for an RA response (RAR) window.

**Random Access Response from the eNB to UE:** At the eNB, the received preambles are regarded as active/inactive based on their power delay profile (PDP) estimation. For each active preamble, the eNB decodes the specific

A UE can select an appropriate preamble under a specific format depending on the distance from eNB, maximum delay spread, amount of transmission resource needed to transmit RRC request and so on. On the other hand, the number of RA slots in each radio frame is defined by the preamble configuration index.

Proposal	Sub-proposal/mechanism	Reference	Access delay	Success rate	Energy efficiency	QoS assurance	Impact on HTC	Performance evaluation
Access class barring	Individual ACB	[4]	Varied	High	Medium	High	positive	No
	Extended ACB	[4, 5]	Varied	High	Medium	High	Positive	Yes
	Cooperative ACB	[6]	Varied	High	Medium	High	Positive	No
MTC-specific backoff	BI	[4, 7]	High	Low	Low	No	Positive	Yes
Resource separation	RACH split	[4, 7]	High	Low	Low	No	Positive	Yes
Dynamic RACH allocation	RACH add/drop	[4, 7]	Medium	Medium	Medium	No	Negative	Yes
Slotted access	Specific RA	[4]	High	Very high	Very high	Very low	Negative	No
Pull-based access	Individual paging	[4]	Medium	Medium	Medium	No	Negative	No
	Group paging	[4, 8]	Medium	Medium	Medium	No	Negative	No
	Group access	[8]	Low	High	Very high	No	Positive	Yes
Self-optimization	ACB, RA split add/drop	[9]	Low	High	High	High	Positive	No
Prioritized access	ACB, RACH split	[10]	Varied	Medium	Medium	High	Positive	No
Code-expanded	Code-wise access	[11]	Low	High	Very low	No	–	Yes
Spatial grouping	Preamble reuse	[12]	Low	High	Medium	No	–	Yes
Guaranteed access	Instant control	[13]	Low	High	Low	High	Positive	No
Non-Aloha-based RA	Analog fountain code	[14]	Low	High	Low	No	–	No

Table 1. Summary of various solutions of LTE random access congestion.

RA slot in which the preamble has been sent. After that, the eNB sends the RAR message to the decoded UEs. The RAR message contains all the necessary information, including a timing advance (TA) instruction for an RRC attempt. However, if multiple UEs transmit the same preamble at the same RA slot, they will receive the same RAR if the preamble is detected as active.

#### RRC Connection Request from UE to eNB:

After receiving the bandwidth assignment at Step 2, the UE sends an RRC connection request along with tracking area update and scheduling request. However, the colliding UEs (i.e., those that were not detected at Step 2) transmit RRC connection requests using the same uplink resources.

#### RRC Connection Setup from eNB to UE:

This step is called the contention resolution stage. After decoding the RRC request message, the eNB acknowledges this to UEs, and sends RRC contention setup messages. Successful UEs then proceed onto data transmission. However, the collided UEs, that is, those which had sent the RRC requests using the same uplink grant, will not receive feedback if their requests do not come with proper TA instruction. In this case, they will initiate a new RA procedure after a maximum number of attempts for retransmission.

## MAJOR LIMITATIONS OF LTE RANDOM ACCESS

In each RA slot, let us consider that 54 preambles are utilized for contention-based random access, and each radio frame contains two RA slots. Thus, the maximum number of RA opportunities per second is 10,800 ( $= 54 \times 2 \times 100$ ), while simultaneous RA opportunities (preambles per RA slot) are still bounded by 108. Also, if 30 percent of contention-based preambles are initially allocated for low data rate MTCs, the maximum number of RA opportunities for low data rate MTCs per second is 3240. In addition, since LTE MAC protocol is slotted-Aloha-based, the average RA success rate is around 37 percent. On the other hand, for massive MTC applications, a single event can drive several thousands of MTCs to access the network almost simultaneously, and consequently, huge preamble collisions are anticipated.

#### AN EXAMPLE SCENARIO

Consider an earthquake monitoring scenario in a densely populated urban area. Assume that MTCs are deployed in a cell of radius 2 km with a density of 60 MTCs/km<sup>2</sup>. Thus, the intensity of MTCs per cell is 754 ( $\approx \pi \times 22 \times 60$ ). Also, consider that the speed of seismic surface wave is 10 km/s, which will result in 754 access attempts by MTCs in



200 ms ( $= (2 \times 1000)/10$  ms). In this case, the probability of preamble collision is around 30 percent ( $\approx 1 - e^{-(754)/(10,800 \times .2)}$ ) with 10,800 RA opportunities per second. However, if 30 percent of the contention-based preambles are dedicated for low data rate MTCs, the probability of collision will be 69 percent ( $\approx 1 - e^{-(754)/(3240 \times .2)}$ ).

Since the collision rate of a slotted Aloha system increases exponentially with increasing rate of RA attempts, the random access in LTE networks is likely to be unstable for massive MTC applications.

## PROPOSALS TO IMPROVE LTE RANDOM ACCESS

In this section, we review major RA congestion solution proposals in LTE systems. The proposals are discussed under two classes: 3GPP specified solutions and non-3GPP specified solutions. Table 1 summarizes the proposals in terms of five key performance metrics: access delay, success rate, energy efficiency, QoS guarantee, and impact on HTC.

### 3GPP SPECIFIED SOLUTIONS

In [4], 3GPP specified the following six distinct solutions of LTE RA congestion due to massive MTC applications.

**Access Class Barring:** Access class barring (ACB) is a well-known tool to control RA congestion by reducing the access arrival rate. ACB operates on two factors: a set of barring access classes (ACs) in which devices are classified, and a barring time duration ( $T_b$ ). Depending on the RA congestion level, the eNB broadcasts an access probability  $p$ , and barring time duration  $T_b$  as a part of SIB2. The intended UEs generate their own access probability  $q$  accordingly to the AC to which they belong. If  $q \leq p$ , the UE gets permission to access the network; otherwise, it is barred for an ACB window  $T_b$ . To support massive MTC along with HTC, 3GPP specified separate AC(s) for MTCs [4]. 3GPP also defined two different ACB mechanisms for a massive MTC over LTE system as follows:

- *Individual ACB*, where each individual device or a group of devices having the same QoS requirements are classed together [4]
- *Extended ACB (EAB)*, where low-priority MTCs are dynamically barred and unbarred depending on the RA arrival rate [4, 5]

Apart from 3GPP specified improvements, the authors in [6] proposed cooperative ACB, where the cooperative eNBs jointly determine their ACB parameters, and thus the RA congestion is distributed among the cooperating eNBs.

**MTC-Specific Backoff:** A backoff mechanism is a common solution to control RA in cellular networks. The basic idea behind backoff scheme is that it discourages the UEs to seek the access opportunity for a time duration, called Backoff Interval (BI), if their first attempt failed due to collision or channel fading. If a device fails second time to get access, it will be subjected to a larger BI than the previous one. In MTC-specific backoff, MTCs are subjected to a larger BI compared to the HTCs [7].

**Dynamic Resource Allocation:** Dynamic allocation of RACH is a straightforward solution for

the RA congestion problem. Under this scheme, the eNB can increase the RACH resources in the frequency domain, time domain, or both, based on the RA congestion level [4]. However, if more uplink resources are utilized as PRACH, there might be a shortage of data channel. In [7], 3GPP evaluated the performance of a dynamic RACH allocation scheme and recommended it as the primary solution to the RA congestion problem for massive MTC.

**Slotted Random Access:** In a slotted RA scheme, each MTC is allocated a dedicated RA opportunity and only allowed to perform RA in its own dedicated access slot [4]. All the access slots comprise an RA cycle. However, for a large number of MTCs, the duration of the RA cycle is likely to be very large; thus, MTCs might experience long access delay. In addition, there is a strong possibility of all 64 access attempts being within a single RA slot, thereby giving rise to collision in a slotted-Aloha-based MAC system, while some other slots may remain underutilized.

**Separate RA Resources:** To save HTC devices (HTCs) from RA congestion, separate RACH for MTCs has been proposed. The separation of resources can be made by either allocating separate RA slots for HTCs and MTCs or splitting the available preambles into HTC and MTC subsets [4]. To ensure QoS guarantee for HTC, some studies proposed to utilize full resources by HTCs, whereas MTCs are restricted to their own subsets. Although the RACH separation scheme potentially reduces the negative impact on HTCs, MTCs might experience serious congestion because the available resources are reduced for MTCs, and the performance tends to be worse under high MTC traffic load.

**Pull-Based RA:** All of the above RA congestion solutions use a push-based approach, where the RA attempts are performed arbitrary by individual devices. The pull-based RA model [4] is an alternative approach where the devices are only allowed to perform RA attempts when they receive any paging message from the eNB. Therefore, it is a centralized approach in which the eNB can completely control the RA congestion by delaying the paging message. The pull-based RA model is suitable where the MTCs transmit information to their server on an on-demand basis. In addition, in order to reduce the paging load for massive MTC applications, 3GPP proposed a group paging method where a large number of MTCs are paged in one paging occasion [4]. However, all the MTCs under a group paging occasion simultaneously perform RA attempts. Therefore, the number of MTCs under a group paging occasion are bounded by the RACH resources.

### NON-3GPP RANDOM ACCESS SOLUTIONS

Besides the 3GPP specified solutions, different organization bodies also proposed LTE RA congestion solutions for massive MTC applications. Important proposals are reviewed below.

**Self-Optimization Overload Control RA:** The self-optimization overload control (SOOC) approach combines RA resource separation, dynamic RA resource allocation, and a dynamic access barring scheme [8]. Under this model, MTCs send RRC requests along with a counter value that indicates the number of RA attempts

Dynamic allocation of RACH is a straightforward solution for the RA congestion problem.

Under this scheme, the eNB can increase the RACH resources in frequency domain, time domain, or both, based on RA congestion level. However, if more uplink resources are utilized as PRACH, there might be shortage of data channel.

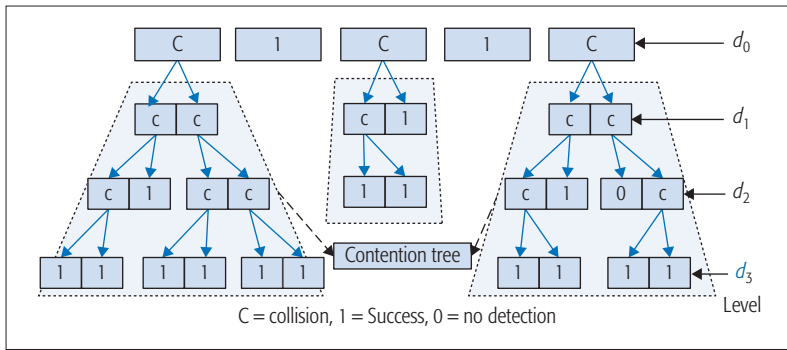


Figure 3. Illustration of the collision resolution RA model.

made before receiving a successful RAR message. By observing the counter value, the eNB estimates the RA congestion level. To control the RA congestion, the eNB either increases the RA resources, decreases the access probability of low-priority MTCs, or takes both actions together.

**Prioritized RA:** Prioritized RA is another optimization approach based on RA resource separation and the ACB mechanism. Here, applications are divided into five classes: HTC, high-priority MTC, low-priority MTC, scheduled MTC, and emergency service [9]. The available RACHs are virtually separated into three groups: HTC, random MTC, and scheduled MTC and emergency service [9]. A prioritized access algorithm is developed to ensure QoS guarantee for the application classes as well as virtual groups. Prioritization is achieved by introducing distinct backoff for different classes.

**Group-Based RA:** The group-based RA approach is an extension of the pull-based group paging RA model. In this scheme, MTCs under a group paging occasion form one or more access group(s). Formation of access groups can be based on different criteria, including belonging to the same server, having similar specifications and/or QoS requirements, being located in a specific region, and so on. However, the key aspect enabling the group access mechanism is that all group members are in close proximity such that TA estimation for the group delegate is valid for all group members [10]. In the group-based RA process, a single preamble is used for all MTCs of each access group, but only the group delegate is responsible for communicating with the eNB. The eNB selects the group delegate based on different metrics such as channel condition and transmission power.

**Code-Expanded RA:** In the code-expanded RA model, an RA attempt is initiated by sending a set of preamble(s) over a predefined number of RA slots instead of sending simply a single preamble at any arbitrary RA slot. In this method, a virtual RA frame is considered, which consists of a group of RA slots or a set of preambles in each RA slot. MTCs need to send multiple preambles over each virtual RA frame, thus making a codeword. At the receiver end, the eNB identifies the individual RA attempts based on the identical codeword perceived inside it [11]. The code-expanded RA scheme increases RA opportunities without significantly increasing any physical resources.

**Spatial-Group-Based Reusable Preamble Allocation:** The main idea behind this RA model is

to spatially partition the cell coverage area into a number of spatial group regions. The UEs in two different spatial group regions can use the same preambles at the same RA slot if their minimum distance is larger than the multi-path delay spread. It is possible due to the fact that the eNB is able to detect simultaneous transmission of identical preambles from different nodes if the distance between the detected picks is larger than the delay spread. In the RAR message, the eNB sends a distinct RAR for each of the detected UEs, where all the RARs are addressed to the same preamble but contain different TA values for different UEs. The UEs can detect the correct RARs by matching their estimated TA with the set of TAs in the RAR message [12].

**Reliability Guaranteed RA:** Generally, RA congestion is detected by the preamble collision rate, and the control schemes deal with high RA load by optimizing the control parameters. However, to activate these controlling scheme, the eNB takes up to 5 seconds (SIB2 broadcasting) [3]. To address this issue, the authors in [13] proposed a proactive approach, where the RA attempt is performed in two phases: the load estimation phase, which contains one RA slot per RA frame, and the serving phase, which contains the rest of the slots. MTCs are also sub-grouped according to their QoS requirements, and each sub-group is assigned different preambles in the estimation phase. All the MTCs need to perform RA attempts during the estimation phase. Based on the estimated collision rate, the eNB allocates the RA resources among the MTC groups. After that, the MTCs again send their RA requests in specific RA slots during the serving phase.

**Non-Aloha-Based RA:** Recently, the authors in [14] proposed an RA model based on the analog fountain code (AFC). AFC-based RA combines multiple access with resource allocation. In this model, multiple MTCs can send RA requests by using the same preamble, and then data transmission also occurs within the same RB. The RA process has two phases: the contention phase and the data transmission phase. In the contention phase, all the MTCs with the same QoS are grouped together and initiate an RA attempt by using predefined preamble(s). Depending on the received preamble power, the eNB estimates the number of contended MTCs per preamble, and broadcasts this information to all contending MTCs. The MTCs that sent the same preamble obtain the information about the total number of candidate MTCs for that preamble, then generate an orthogonal random seed and share it with the eNB. Therefore, both the eNB and MTC can construct the same bipartite graph to perform AFC encoding and decoding for subsequent communications.

## COLLISION-RESOLUTION-BASED RANDOM ACCESS MODEL

The basic idea behind the collision-resolution-based RA (CRB-RA) model is to ensure RA reattempts from a reserved set of preambles if the current attempt is detected as a collision. The number of preambles in each reserved set is optimized according to the rate of collision at each level. In this model, separate RA preambles are

1. Set collision threshold:  $x, y; x < y$
2. Set preambles per contention tree slot:  $m_0,^2 m_x, m_y;$   
 $m_0 < m_x < m_y$
3. Set additional RA slot:  $\Delta_x, \Delta_y; \Delta_x \leq \Delta_y$
4. Check preamble collision rate:  $k$
5. While  $k \neq 0$
6. If  $y > k \geq x$ , then set  $m = m_x$  and  $\Delta = \Delta_x$
7. Elseif  $y \leq k < x$ , then set  $m = m_y$  and  $\Delta = \Delta_y$
8. Else  $m = m_0$ , and  $\Delta$  unchanged
9. Reserve  $m$  preambles for each collision
10. Send RAR to collided MTCDs for reattempt RA
11. Broadcast the updated RA resources on SIB2.

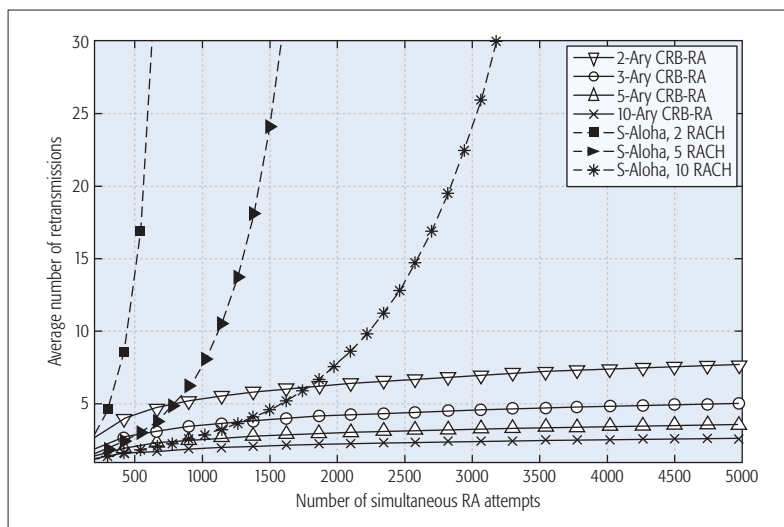
<sup>2</sup> Initial value of  $m$ ; for optimal resource utilization,  $m_0 = 3$ .

**Algorithm 1.** Collision-resolution-based random access.

used for HTC and MTC, where the collision resolution technique is only applicable for MTC. A number of RA slots form a virtual RA frame and the eNB broadcasts SIB2 at the end of each virtual RA frame. The eNB can allocate new RACH resources into the virtual RA frame if the collision rate is increased at certain thresholds; thus, the size of a virtual RA frame is optimized depending on the rate of preamble collision. Meanwhile, the duration of each virtual RA frame is adjusted for the QoS requirements of high-priority MTCDs. In addition, each contending UE (MTCD/HTCD) transmits its identity, UE-ID, along with a randomly generated preamble for an RA attempt [15]. Some RACH subcarriers are used to map UE-IDs such that the UE-IDs of different preambles are orthogonal to each other [15]. However, if multiple UEs transmit the same preamble at the same RA slot, the eNB is unable to decode their UE-IDs, and thus this is considered as collision. For each collided preamble, the eNB assigns a set of new preambles (say  $m$ ) to the collided UEs if the collided preamble belongs to the MTCDs. In the RAR message, the eNB instructs the collided MTCDs to retransmit on the reserved preamble set in the next available virtual RA frame. On the other hand, if the collided preamble arrives from HTCDs, the eNB does not send any RAR feedback; thus, the collided HTCDs initiate a new RA procedure at the next available RA slot.

In the next virtual RA frame, the collided MTCDs retransmit RA requests using preambles from the reserved set, while the others are not allowed to use that set. The eNB imposes this restriction by broadcasting the information as part of SIB2. If the collided MTCDs collide again within the preassigned  $m$  preambles, another set of preambles will be allocated accordingly. This process will continue until the eNB properly decodes each preamble with an individual UE-ID. Therefore, an optimistic  $m$ -ary splitting tree algorithm is developed for each collision. However, based on the collision rate, the eNB can also utilize dynamic ACB mechanisms to facilitate channel access for high-priority MTCDs.

Figure 3 illustrates our proposed CRB-RA model by using a binary ( $m = 2$ ) splitting tree algorithm. In this model, the basic splitting tree algorithm is slightly modified to resolve the RA problem in LTE. The root of the new model, where collisions initially occur, consists of the total



**Figure 4.** Average number of retransmissions for the proposed RA model and the slotted-Aloha-based RA model.

number of contention-based preambles (say  $q$ ) of a virtual RA frame. Let us denote the root as level 0. For each single collision at level 0, a new set of  $m$  preambles is reserved at level 1. Similarly,  $m$  preambles are also reserved at level 2 for each collision detected at level 1, and the process continues until the collision is resolved. Therefore, an  $m$ -ary tree is developed for every preamble collision detected at level 0, but the root of each individual tree is level 1.

In the CRB-RA model, the number of preambles in each reserved set ( $m$ ) is dynamically adjusted according to the collision rate. Also, each level of contention tree is resolved at an individual virtual RA frame. In a particular virtual RA frame, if the collision rate is sufficiently high, more reserved preamble sets are required where the value of  $m$  would also be high. For example, in the case of full collision, the maximum number of preambles required at any level is  $(m^d \times q)$ , where  $d$  indicates the level of the tree. However, if the value of  $m$  is set to high, resolution of each level of contention tree requires more time. On the other hand, if the value of  $m$  is set to low, the number of levels would be high. Therefore, the access delay of CRB-RA mainly depends on the proper selection of  $m$ . The general algorithm of our proposed CRB-RA model for two different collision thresholds is presented as Algorithm 1.

#### PERFORMANCE ANALYSIS

We evaluate the performance of our proposed CRB-RA model in terms of average number of preamble retransmissions and average outage probability. The results are compared to those for the standard slotted-Aloha-based RA model. The energy efficiency and access delay of the proposed CRB-RA model are also discussed based on the outage probability and average number of preamble retransmissions. It is assumed that massive access requests are attempted, that is, as in the earthquake monitoring scenario discussed before. Each preamble can be successfully detected (collision/active/ideal) at the eNB.

To simplify the simulation, misdetection, propagation delay, and device processing time are also not considered. In addition, we simulate our

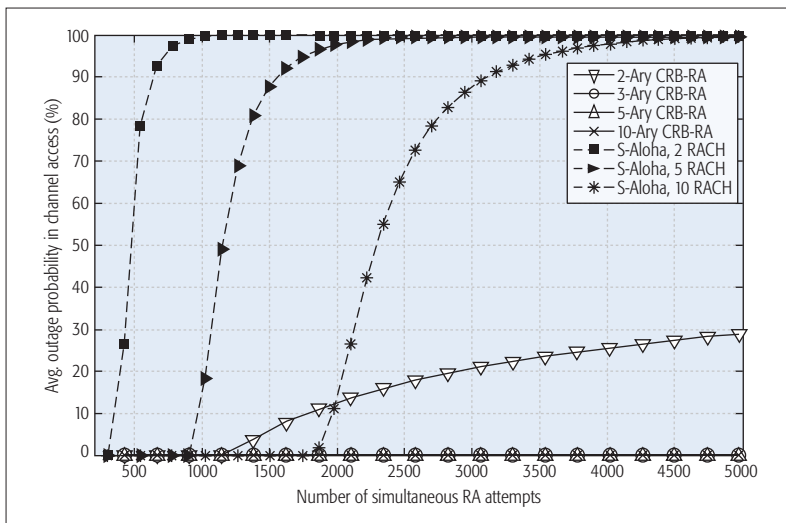


Figure 5. Average outage probability in channel access vs. number of simultaneous RA attempts.

proposed CRB-RA model by considering fixed contention slot size ( $m$  is fixed). All the simulations are done based on the 3GPP standard [3], where the initial PRACH configuration index is 6 (2 RA slots per radio frame), and the maximum RA retransmission limit is 10. In each initial RA slot, 30 contention-based preambles are used for MTC. Also, depending on the collision rate, the eNB allocates up to 10 RA slots per radio frame.

Figure 4 shows the average number of RA attempts required to successfully decode each MTC with respect to the number of simultaneous RA attempts.<sup>3</sup> It is clearly observed that for any arbitrary RA attempt, our proposed CRB-RA model ensures network access within a limited number of retransmissions, while a large number of retransmissions are required in the standard slotted-Aloha-based RA model. The slotted-Aloha-based RA scheme with peak preamble configuration index (10 RA slots per radio frame) needs on average more than 30 retransmission attempts for one successful access, when the number of simultaneous RA attempts is 3200 or higher. It is noted that in Fig. 4, the CRB-RA model utilizes only two RA slots in each radio frame.

Also, Fig. 5 shows the average RA outage probability of MTCs as a function of the number of simultaneous access attempts. It is evident that by setting appropriate number of preambles (value of  $m$ ) per contention slot, the CRB-RA model can reduce the outage in network access significantly. The standard slotted-Aloha-based RA system with 2 RA slots per radio frame shows an average outage rate of 70 percent if 500 simultaneous RA attempts arrive. In addition, with maximum RA slots per radio frame, the standard slotted-Aloha-based RA system shows an average outage probability of about 70 percent for 2500 RA attempts per radio frame. Therefore, massive multiple access by MTCs will make the system unstable. However, in contrast, with minimal preambles per contention slot ( $m = 2$ ), although the proposed CRB-RA model may result in a non-zero outage probability for a large number of simultaneous RA attempts, by optimizing the slot length ( $m$ ), the outage probability in channel access can be made very small.

<sup>3</sup> We use the term simultaneous RA attempts to refer to the number of RA attempts within one radio frame.

In addition, since the average number of RA retransmission requirement in the CRB-RA model is very low compared to slotted-Aloha-based RA, the proposed RA model is very efficient for power constrained MTC applications. For the same reason, the access delay of the CRB-RA model is also much lower in comparison to slotted-Aloha-based RA models.

## CONCLUSION

We have reviewed a wide range of LTE MAC layer congestion control proposals from the perspective of massive MTC for smart city applications. Many of the proposals are not capable of managing massive bursty access attempts. To solve this congestion problem in massive random access, we have proposed a novel collision-resolution-based RA model, which can effectively manage massive RA requests. Also, our proposed RA method can coexist with existing LTE MAC protocol without any modification. Simulation results have also shown that the collision resolution RA model provides reliable and time-efficient access performance.

Although we have simulated our model with a fixed size of reserved preamble set, the proposed model exhibits a multi-dimensional optimization problem, where the number of preambles per contention tree slot, and the size and duration of the virtual RA frame can be optimized based on the preamble collision rate, available radio resources, and delay constraints.

## ACKNOWLEDGMENT

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) and in part by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) (2014R1A5A1011478 and 2013R1A2A2A01067195).

## REFERENCES

- [1] A. J. Janssen and M. de Jong, "Analysis of Contention Tree Algorithms," *IEEE Trans. Info. Theory*, vol. 46, no. 6, 2000, pp. 2163–72.
- [2] G.C. Madueno et al., "Efficient LTE Access with Collision Resolution for Massive M2M Communications," *Proc. IEEE GLOBECOM Wksp.*, 2014, pp. 1433–38.
- [3] 3GPP TS 36.321 V12.7.0, "Evolved Universal Terrestrial Radio Access: Medium Access Control," France, Sept. 2015.
- [4] 3GPP TR 37.868 V11.0.0, "Study on RAN Improvements for Machine Type Communications," Sept. 2011.
- [5] 3GPP RAN WG2 #77, "Further Performance Evaluation of EAB Information Update Mechanisms," R2-120270, Intel, Germany, Feb. 2012.
- [6] S. Lien et al., "Cooperative Access Class Barring for Machine-to-Machine Communications," *IEEE Trans. Wireless Commun.*, vol. 11, Jan. 2012, pp. 27–32.
- [7] 3GPP TSG RAN WG2 #70: 70bis R2-103742, "RACH Overload Solutions," R2-103742, ZTE, Sweden, July 2010.
- [8] G. Farhadi and A. Ito, "Group-Based Signaling and Access Control for Cellular Machine-to-Machine Communication," *Proc. IEEE 78th VTC*, 2013.
- [9] A. Lo et al., "Enhanced LTE-Advanced Random-Access Mechanism for Massive Machine-to-Machine Communications," *Proc. 27th Wireless World Research Forum*, Oct. 2011.
- [10] J. P. Cheng, C. H. Lee, and T. M. Lin, "Prioritized Random Access with Dynamic Access Barring for RAN Overload in 3GPP LTE-A," *Proc. IEEE GLOBECOM Wksp.*, Dec. 2011, pp. 368–72.
- [11] N. K. Pratas et al., "Code Expanded Random Access for Machine-Type Communications," *Proc. IEEE GLOBECOM Wksp.*, 2012, pp. 1681–86.
- [12] H. S. Jang et al., "Spatial Group Based Random Access for M2M Communications," *IEEE Commun. Lett.*, vol. 8, no. 6, 2014, pp. 961–64.

- [13] G. C. Madueno *et al.*, "Massive M2M Access with Reliability Guarantees in LTE Systems," *Proc. IEEE ICC*, 2015, pp. 2997–3002.
- [14] M. Shirvanimoghaddam *et al.*, "Probabilistic Rateless Multiple Access for M2M Communication," *IEEE Trans. Wireless Commun.*, vol. 14, no. 12, Dec. 2015, pp. 6815–26.
- [15] N. Zhang *et al.*, "Resource Allocation in a New Random Access form M2M Communications," *IEEE Commun. Lett.*, vol. 19, no. 5, May 2015, pp. 843–46.

## BIOGRAPHIES

MD SHIPON ALI [S'16] received his B.Sc. degree in electronics and communication engineering with distinction from Khulna University, Bangladesh, in 2009. He was awarded the University Gold Medal for securing first position and obtaining the highest CGPA in Bachelor Examinations among all disciplines under the School of Science, Engineering and Technology. During 2010–2015, he worked on radio, backhaul, backbone, and packet core network operations for Grameenphone Ltd., and received the top performer awards in 2013 and 2014. Since September 2015, he has been working toward his M.Sc. degree at the University of Manitoba under the supervision of Prof. Ekram Hossain. He has been awarded the the University of Manitoba Graduate Fellowship and the Manitoba Graduate Scholarship. His research interests include multiple access for 5G and beyond 5G networks, coordinated multipoint transmissions, and heterogeneous wireless networks.

EKRAM HOSSAIN [F'15] is a professor (since March 2010) in the Department of Electrical and Computer Engineering at the University of Manitoba, Winnipeg, Canada. He is a member (Class of 2016) of the College of the Royal Society of Canada. He received his Ph.D. in electrical engineering from the University of Victoria, Canada, in 2001. His current research interests include design, analysis, and optimization of wireless/mobile communications networks, cognitive radio systems, and network economics. He has authored/edited several books in these areas (<http://home.cc.umanitoba.ca/~hossaina>). He serves as the Editor-in-Chief for *IEEE Communications Surveys & Tutorials* and an Editor for *IEEE Wireless Communications*. Also, he is a

member of the IEEE Press Editorial Board. Previously, he served as the Area Editor for *IEEE Transactions on Wireless Communications* in Resource Management and Multiple Access from 2009 to 2011, an Editor for *IEEE Transactions on Mobile Computing* from 2007 to 2012, and an Editor for the *IEEE Journal on Selected Areas in Communications Cognitive Radio Series* from 2011 to 2014. He has won several research awards including the IEEE VTC-2016 (Fall) "Best Student Paper Award" as a co-author, IEEE Communications Society Transmission, Access, and Optical Systems (TAOS) Technical Committee's Best Paper Award at IEEE GLOBECOM 2015, the University of Manitoba Merit Award in 2010, 2014, and 2015 (for Research and Scholarly Activities), the 2011 IEEE Communications Society Fred Ellersick Prize Paper Award, and the IEEE Wireless Communications and Networking Conference 2012 Best Paper Award. He was elevated to an IEEE Fellow "for spectrum management and resource allocation in cognitive and cellular radio networks." He was a Distinguished Lecturer of the IEEE Communications Society (2012–2015). Currently he is a Distinguished Lecturer of the IEEE Vehicular Technology Society. He is a registered Professional Engineer in the province of Manitoba, Canada.

DONG IN KIM [S'89, M'91, SM'02] received his Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, in 1990. He was a tenured professor with the School of Engineering Science, Simon Fraser University, Burnaby, British Columbia, Canada. Since 2007, he has been with Sungkyunkwan University, Suwon, Korea, where he is currently a professor with the College of Information and Communication Engineering. He served as an Editor and a Founding Area Editor of Cross-Layer Design and Optimization for *IEEE Transactions on Wireless Communications* from 2002 to 2011. From 2008 to 2011, he served as Co-Editor-in-Chief for the *Journal of Communications and Networks*. He served as the Founding Editor-in-Chief for *IEEE Wireless Communications Letters* from 2012 to 2015. From 2001 to 2014, he served as an Editor of Spread Spectrum Transmission and Access for *IEEE Transactions on Communications*, and then as an Editor-at-Large for *Wireless Communication*. He is a first recipient of the NRF of Korea Engineering Research Center (ERC) in Wireless Communications for Energy Harvesting Wireless Communications (2014–2021).

# Efficient Energy Management for the Internet of Things in Smart Cities

Waleed Ejaz, Muhammad Naeem, Adnan Shahid, Alagan Anpalagan, and Minhó Jo

The authors present a brief overview of energy management and challenges in smart cities. They then provide a unifying framework for energy-efficient optimization and scheduling of IoT-based smart cities. They also discuss the energy harvesting in smart cities, which is a promising solution for extending the lifetime of low power devices and its related challenges.

## ABSTRACT

The drastic increase in urbanization over the past few years requires sustainable, efficient, and smart solutions for transportation, governance, environment, quality of life, and so on. The Internet of Things offers many sophisticated and ubiquitous applications for smart cities. The energy demand of IoT applications is increased, while IoT devices continue to grow in both numbers and requirements. Therefore, smart city solutions must have the ability to efficiently utilize energy and handle the associated challenges. Energy management is considered as a key paradigm for the realization of complex energy systems in smart cities. In this article, we present a brief overview of energy management and challenges in smart cities. We then provide a unifying framework for energy-efficient optimization and scheduling of IoT-based smart cities. We also discuss the energy harvesting in smart cities, which is a promising solution for extending the lifetime of low-power devices and its related challenges. We detail two case studies. The first one targets energy-efficient scheduling in smart homes, and the second covers wireless power transfer for IoT devices in smart cities. Simulation results for the case studies demonstrate the tremendous impact of energy-efficient scheduling optimization and wireless power transfer on the performance of IoT in smart cities.

## INTRODUCTION

Smart city solutions use communication and networking technologies for dealing with the problems precipitated by urbanization and growing population. The Internet of Things (IoT) is a key enabler for smart cities, in which sensing devices and actuators are major components along with communication and network devices. The sensing devices are used for real-time detection and monitoring of city operations in various scenarios. It is projected that in the near future, common industrial, personal, office, and household devices, machines, and objects will hold the ability to sense, communicate, and process information ubiquitously [1]. However, it is challenging to design a fully optimized framework due to the interconnected nature of smart cities with different technologies. Further, smart city solutions have to be energy-efficient from both the users' and environment's points of view.

These challenges have forced network designers to consider a wide range of scenarios in different conditions for IoT-enabled smart cities. Thus, efficient deployment of sensors and an optimized operational framework that can adapt to the conditions is necessary for IoT-enabled smart cities. In other words, smart city solutions have to be energy-efficient, cost-efficient, reliable, secure, and so on. For example, IoT devices should operate in a self-sufficient way without compromising quality of service (QoS) in order to enhance the performance with uninterrupted network operations [2]. Therefore, the energy efficiency and life span of IoT devices are key to next generation smart city solutions.

We classify the energy management in smart cities into two main types: energy-efficient solutions and energy harvesting operations. This classification along with a few examples of research topics are shown in Fig. 1. Energy-efficient solutions for IoT-enabled smart cities include a wide range of topics such as lightweight protocols, scheduling optimization, predictive models for energy consumption, a cloud-based approach, low-power transceivers, and a cognitive management framework [3–5]. Energy harvesting allows IoT devices to harvest energy from ambient sources and/or dedicated RF sources. The aim of energy harvesting is to increase the lifetime of IoT devices. The research topics included within both types of energy harvesting are energy harvesting receiver design, energy arrival rate, placement of a minimum number of dedicated energy sources, scheduling of dedicated energy sources, and multi-path energy routing [2, 6].

Both academia and industry are focusing on energy management in smart cities. The IEEE in partnership with the International Telecommunication Union (ITU) has a smart cities community with the aim to provide assistance to municipalities for the transition to smart cities. Fujitsu suggested an approach to energy management for companies and has introduced an energy management system for smart buildings as cloud service [7]. In addition, companies such as IBM, Cisco, Honeywell, Intel, and Schneider Electric are involved in various energy-efficient solutions for smart cities. There have been various projects on energy-efficient smart cities sponsored by the Seventh Framework Programme (FP7) for research of the European Commission in the past

few years. For example, the main objectives of the “Reliable, Resilient, and Secure IoT for Smart City Applications” project are to develop, evaluate, and test a framework of IoT-enabled smart city applications in which smart objects can operate energy-efficiently [8]. The “ALMANAC: Reliable Smart Secure Internet of Things for Smart Cities” project focuses on IoT-enabled green and sustainable smart solutions [9]. Likewise, energy-saving solutions are developed for smart cities under the projects “Planning for Energy Efficient Cities (PLEEC)” and “NiCE — Networking Intelligent Cities for Energy Efficiency.”

In this article, we consider energy management for IoT in smart cities. An illustration of smart cities with the focus on smart homes is shown in Fig. 2. Our contributions can be summarized as follows:

- We provide an optimization framework for research in IoT-enabled smart cities. We present the objectives, problem type, and solution approaches for energy management.
- We cover energy-efficient solutions for IoT-enabled smart cities. A case study is presented to show the performance gains achieved by scheduling optimization in smart home networks.
- Next, we devote a section to energy harvesting for IoT-enabled smart city applications. A case study is provided to investigate the performance gains achieved by the scheduling of dedicated energy sources.
- Finally, the conclusions are drawn, and we provide future research directions for energy management in IoT-enabled smart cities.

## ENERGY MANAGEMENT AND CHALLENGES FOR SMART CITY APPLICATIONS

An urgent need for energy management has emerged all over the globe due to a continuous increase in consumption demands. Global warming and air pollution are serious threats to future generations. This is caused by the emission of fumes with volume increased with the increase in energy demand. On the other hand, according to the statistics provided by Cisco, there will be more than 50 billion IoT devices connected to the Internet by 2020 [10]. This explosion in devices will pose serious energy consumption concerns; thus, it is imperative to manage energy for IoT devices so that the concept of smart cities can be better realized in a sustained manner. Following are a few examples where we can reduce energy consumption by effective management.

**Home Appliances:** Home appliances are the major sources of energy consumption. Demand management is a key for customizing energy use by managing the lighting, cooling, and heating systems within residential units. On the other hand, the intelligent operation of activities can also facilitate the optimized management and operation of energy.

**Education and Healthcare:** Considering the importance of educational and healthcare services, it is difficult to dematerialize them. However, it is possible to demobilize services for the reduction of energy consumption; for example, exploiting remote healthcare by visualizing sensors and mobile phones, and distance education

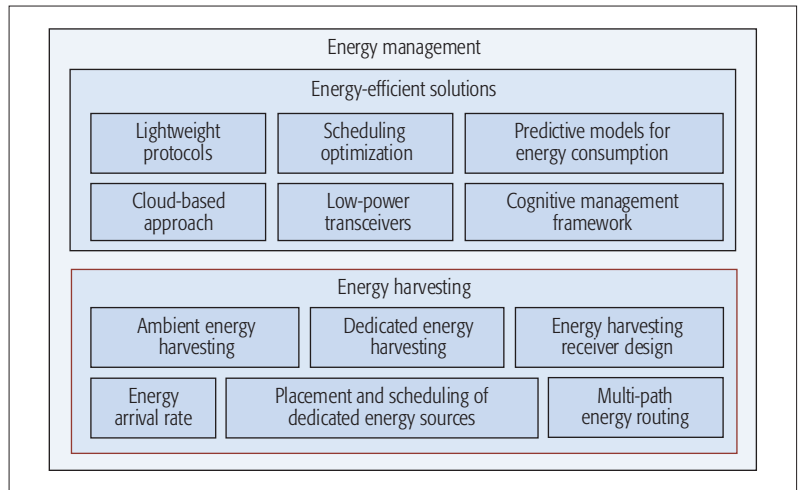


Figure 1. Classification of energy management for IoT in smart cities.

can create a significant reduction in energy consumption.

**Transportation:** The energy use for transportation includes public transport, daily commuting to work in personal vehicles, leisure travel, and so on. In addition to the energy consumed by public transport and personal vehicles, they are also a major cause of pollution in cities. IoT-enabled solutions can be employed for energy management, such as traffic management, congestion control, and smart parking. This can significantly reduce energy consumption as well as CO<sub>2</sub> emission.

**Food Industry:** Energy consumption in the food industry is not only related to the storage, purchase, and preparation of food; it also includes diners moving into restaurants in search of food. IoT-enabled solutions can be used here for making optimized choices in terms of food availability. On the other hand, the transportation of the food can also be optimized by incorporating intelligent means of transportation.

IoT devices are generally battery operated and have limited storage space. Concerning these fundamental limitations of sensors, it is difficult to realize the IoT solutions with prolonged network life. In order to efficiently utilize the limited sensor resources, an optimized energy-efficient framework is of paramount importance. It will not only reduce energy consumption, but also maintain the minimum QoS for the concerned applications.

A typical optimization framework for IoT-enabled smart cities is given in Fig. 3. This framework provides details of the objectives, problem types, and corresponding optimization techniques for energy management. For example, an optimization problem for minimizing the cost of electricity usage is presented in [11]. The authors developed an optimization-based residential energy management scheme for energy management of appliances. The authors in [12] presented an optimization framework for smart home scheduling of various appliances and assignment of energy resources. This results in a mixed integer combinatorial problem which is transformed into a standard convex programming problem. The goal of this study is to minimize cost and user dissatisfaction. In [13], the authors presented an energy-centered and QoS-aware services selec-

With the increase in IoT applications for smart cities, energy-efficient solutions are also evolving for low-power devices. There are some energy-efficient solutions that can either reduce energy consumption or optimize resource utilization.

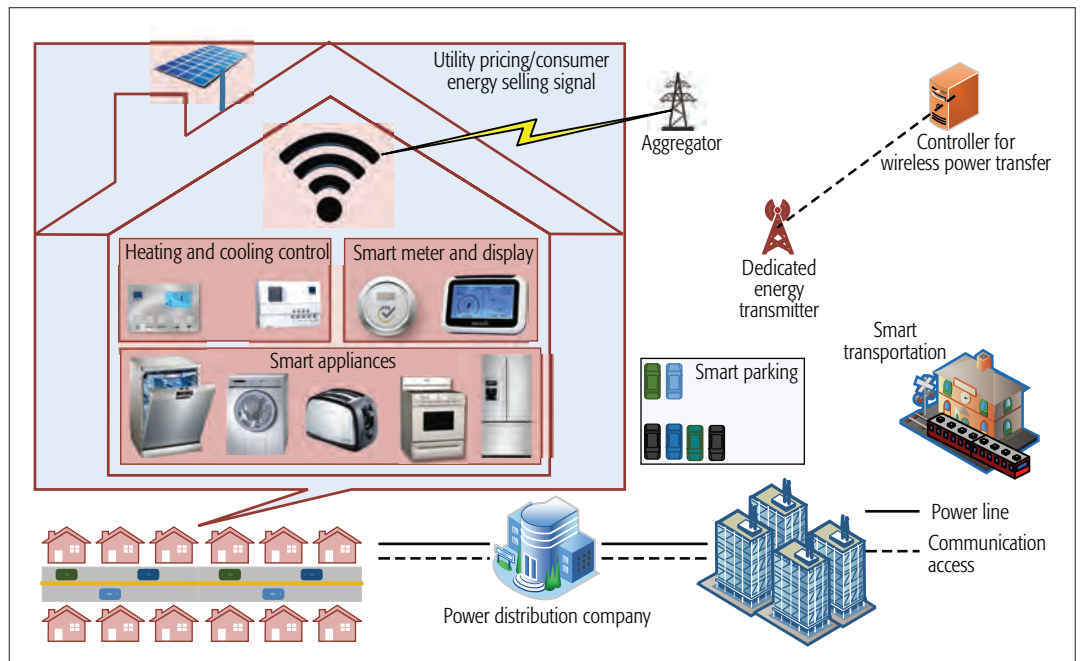


Figure 2. An illustration of smart cities focused on smart homes.

tion algorithm for IoT environments. The objective is to minimize energy consumption while satisfying QoS requirements. Similarly, the objectives shown in Fig. 3 can be considered, and the framework can be used as a guideline to solve the optimization problems.

## ENERGY-EFFICIENT SOLUTIONS FOR SMART CITIES

With the increase in IoT applications for smart cities, energy-efficient solutions are also evolving for low-power devices. There are some energy-efficient solutions that can either reduce energy consumption or optimize resource utilization. Following are some main research trends for energy-efficient solutions of IoT-enabled smart cities.

**Lightweight Protocols:** Lightweight means that a protocol causes less overhead. IoT-enabled smart cities have to use various protocols for communication. There are several existing protocols in the literature such as Message Queue Telemetry Transport (MQTT), Constrained Application Protocol (CoAP), Extensible Messaging and Presence Protocol (XMPP), Advanced Message Queue Protocol (AMQP), 6lowPAN, and Universal Plug and Play (UPnP) IoT. MQTT and CoAP are the most popular protocols. MQTT is a lightweight protocol that collects data from IoT devices and transmits to the servers. CoAP is designed for constrained devices and networks for web transfer (See [14] for IoT protocols). Each of these protocols is designed for specific scenarios and applications in which it performs well. In addition, protocol conversion is an important building block for IoT, which may require that the IoT devices be from different manufacturers or using different protocols.

**Scheduling Optimization:** Scheduling optimization for IoT-enabled smart cities refers to the optimization of resources with the aim of minimizing energy consumption and subsequently reducing electricity usage. In this regard, demand-side

management (DSM) is of prime importance; it refers to the manipulation of residential electricity usage by altering the system load shape and consequently reducing the cost. Broadly speaking, DSM comprises two main tasks: load shifting and energy conservation, where load shifting refers to the transfer of customers' load from high-peak to low-peak levels. By adopting this, electricity can be conserved and provide room for other customers.

**Predictive Models for Energy Consumption:** Predictive models for energy consumption in IoT-enabled smart cities are indeed of vital importance. They refer to the wide range of applications in smart cities, including predictive models for traffic and travel, predictive models for controlling temperature and humidity, and so on. Various prediction models such as neural networks and Markov decision processes can be incorporated here. Exploiting the predictive models will not only reduce the significant energy consumption but also lead to many societal benefits.

**Cloud-Based Approach:** Cloud computing has reshaped the computing and storage services, which can be used to provide energy-efficient solutions for IoT-enabled smart cities. More precisely, the cloud-based approach helps in managing the massive data center flexibility and in a more energy-efficient manner.

**Low-Power Transceivers:** Since the IoT devices in smart city applications operate on limited batteries, a low-power design architecture or operation framework is of superior importance for addressing the energy management in IoT-enabled smart cities. Mostly, the existing application protocols for IoT devices are not in accordance with the energy efficiency perspective. More specifically, the radio duty cycle for IoT devices is an important factor in energy efficiency, and researchers are exploring methods of reducing the radio duty cycle of IoT devices and subsequently to achieve the energy-efficient architecture.

**Cognitive Management Framework:** IoT



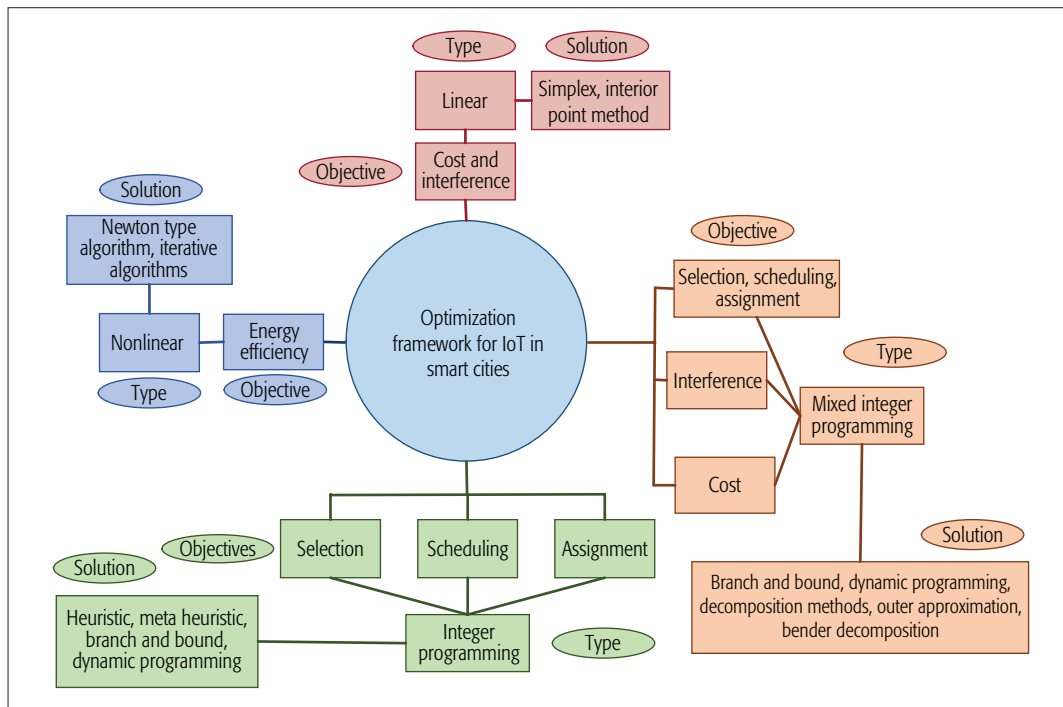


Figure 3. A typical optimization framework for IoT in smart cities.

To reduce electricity bills, smart home networks offer better life style, customized day to day schedule, and so on. The smart grid has provided the ability to keep the electricity demand in line with the supply during the peak time of usage. This is called demand-side management.

devices are heterogeneous in nature, and the associated services are unreliable. Therefore, it is important to investigate a cognitive management framework that adopts intelligence and cognitive approaches throughout the IoT-enabled smart cities. The framework should include reasoning and learning in order to improve decisions for IoT networks. A context-aware cognitive management framework was presented in [4], which made decisions regarding IoT devices (when, why, and how to connect) according to the contextual background.

### CASE STUDY ON SMART HOME NETWORKS

Smart home networks enable home owners to use energy efficiently by scheduling and managing appliances. In addition, to reduce electricity bills, smart home networks offer better lifestyles, customized day-to-day schedules, and so on. The smart grid has provided the ability to keep the electricity demand in line with the supply during the peak time of usage. This is called demand-side management. DSM reduces the electricity cost by altering/shifting the system load [5]. Generally, DSM is responsible for the demand response program and load shifting. In the demand response program, a customer's load can be reduced in peak hours by shifting it to off-peak hours. This helps to provide more electricity at less cost.

Home appliances are becoming smart with added features of connectivity that enable consumers to take advantage of the demand response program. The electric utility can contact consumers to reduce/shift their electricity consumption in return for certain monetary benefits. In smart home networks, appliance load can further be categorized into manageable and unmanageable loads. Here, we focus on the energy management of manageable appliance load in smart homes since it has high energy consumption and predictability in operations. The manageable load is

further divided into shiftable load (e.g., washing machine, dishwasher), interruptable load (e.g., water heater and refrigerator), and weather-based load (e.g., heating and cooling). An illustration of the smart home network model for appliance scheduling is given in Fig. 2.

We consider a smart home network in which  $N_A$  is the set of load types,  $A_n$  is the set of appliances in the  $n$ th load type, and  $A$  is the set that is a union of all appliances. We define  $T$ ,  $C^t$ , and  $P_{na}^t$  as number of time slots in a day, tariff/cost in dollars in time slot  $t$ , and  $P_{na}^t$  power of the  $n$ th load type's  $a$ th appliance in time slot  $t$ , respectively. We formulate a problem for scheduling of smart home appliances while considering the tariffs and peak load. The overall objective is to schedule the appliances in such a way that total cost is minimum, that is, minimize the  $\sum_{na} x_{na}^t C_t P_{na}^t$  for whole set of  $N_A$ ,  $A$  for all  $T$  time slots, where  $x_{na}^t$  is a binary variable with value 1 when the  $n$ th load type's  $a$ th appliance in time slot  $t$  is on; otherwise, 0. We consider practical constraints on time occupancy and time consecutiveness that need to be satisfied for realistic execution of appliance scheduling. The constraints ensure that each appliance should not occupy more time slots than required, and the time slots for shiftable loads are consecutive. The optimization problem here is integer programming; such problems are generally NP-hard and require very efficient algorithms. We solved the optimization problem using an efficient heuristic algorithm.

### PERFORMANCE ANALYSIS

For illustration purposes, we consider only four types of appliances: washing machine, dryer, dishwasher, and electric vehicle. Figure 4a shows the tariff, and slot time for appliances with (thick slots) and without DSM (thin slots). It is considered that a dryer cannot be activated before a washing machine. It is evident that with DSM the

appliances are activated when the tariff is low. However, without DSM, there is no scheduling for appliances, and they can be activated at any time. For instance, all the appliances are scheduled at the time when the tariff is low with DSM. In contrast, only the dryer is activated when the tariff is low in the absence of DSM. Similarly, Fig. 4b shows that the total load is less in the case of optimum energy management when the tariff is high. It is important to notice that at some times the total load for both optimum energy management and no energy management is the same. This is because there is no shiftable load at this time.

### ENERGY HARVESTING IN SMART CITIES

Energy harvesting is considered as a potential solution to increase the lifetime of IoT devices in smart cities. Energy harvesting can generally be classified into two categories:

- In ambient energy harvesting, IoT devices harvest energy from ambient sources such as wind, RF signals in the environment, vibration, and solar. However, harvesting from ambient sources depends on their availability, which is not always guaranteed.

- In dedicated energy harvesting, the energy sources are intentionally deployed in the surroundings of IoT devices.

The amount of energy harvested by each IoT device depends on the sensitivity of harvesting circuits, the distance between an IoT device and an energy source, the environment, and so on. Thus, the success of energy harvesting for IoT devices in smart cities has to face several challenges, which are discussed below.

**Energy Harvesting Receiver Design:** The harvesting circuit design is the primary issue in RF-based energy harvesting. The sensitivity required for the harvesting circuit is higher than for traditional receivers, which can result in fluctuations in energy transfer due to the environment and mobility (energy source and IoT devices). Therefore, efficient and reliable harvesting circuit design is required to maximize the harvested energy. In addition, RF-to-DC conversion is the fundamental ingredient of RF energy harvesting. Hence, circuit designers should enhance the efficiency of RF-to-DC conversion using advanced technologies.

**Energy Arrival Rate:** The level of uncertainty of the energy arrival rate is higher in energy harvesting from ambient sources than in dedicated energy harvesting. This is because the former uses renewable energy sources, whereas the latter uses dedicated energy sources the location of which is set by network designers based on the harvesting requirements of IoT devices. Accurate and detailed modeling of the energy arrival rate is indispensable in order to analyze the performance of energy harvesting systems in smart cities.

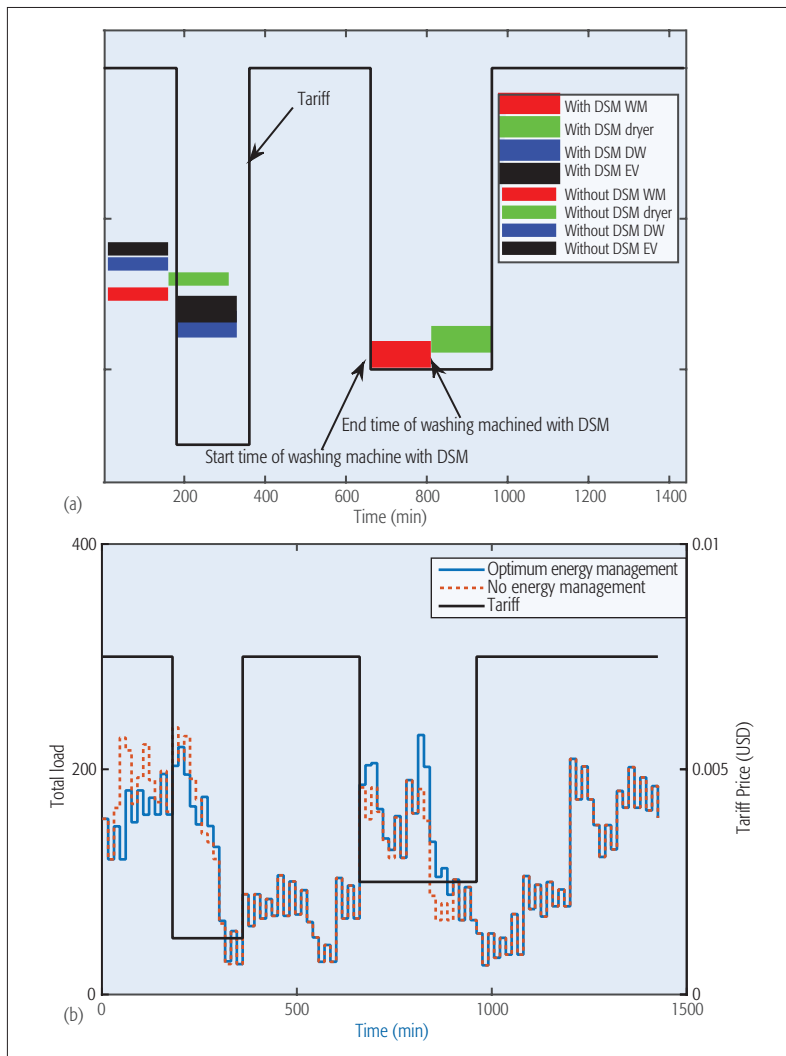
**Placement of a Minimum Number of Dedicated Energy Sources:** IoT devices that are spatially distant from energy sources can result in uneven energy harvesting. This can result in energy depletion of devices that are far from dedicated energy sources and thus reduce the lifetime of the network. We can do much in the case of ambient energy sources; however, optimal placement and number of dedicated energy sources are crucial issues in dedicated energy harvesting.

**Scheduling of Energy Transmitters:** Energy consumed by dedicated energy sources can be reduced by introducing task-based energy harvesting, where energy transmitters can be scheduled for RF power transfer based on the harvesting requirements of IoT devices. This requires a certain level of coverage and sufficient time to harvest. Therefore, scheduling of energy transmitters with guaranteed coverage and duration is vital for the energy efficiency of dedicated energy harvesting.

**Multi-Path Energy Routing:** Multi-path energy routing collects the scattered RF energy from different sources with the help of RF energy routers. Then these energy routers can transfer energy via an alternative path to IoT devices. Multi-path energy routing is based on the idea of multihop energy transfer in which relay nodes are deployed near IoT devices. This will help to reduce path loss between the relay node and the IoT devices, and also improve the RF-to-DC conversion efficiency.

### CASE STUDY: SCHEDULING OF ENERGY SOURCES IN DEDICATED ENERGY HARVESTING FOR IOT DEVICES

We consider a network in smart cities with dedicated RF energy transmitters that consists of  $N_t$  IoT devices (each device is equipped with a har-



**Figure 4.** Load pattern: a) appliances starting and ending times with and without DSM; b) load pattern of appliances while minimizing total electricity cost and tariff.

vesting circuit) and  $N_E$  energy transmitters, as shown in Fig. 5. It is assumed that energy transmitters have continuous power supply, and they can satisfy the requirements of all IoT devices in the area. The IoT devices can request power transfer from a harvesting controller, which is considered as task  $k$ . The harvesting controller is considered as a cloudlet controller, which is a centralized resource pool with information about the location of IoT devices and energy transmitters. The controller can assign multiple tasks from  $K$  ( $K$  is a set of tasks) to the energy transmitters. The transmit power of the  $e$ th energy transmitter is denoted by  $P_e$ . The energy transmitter  $e$  can transfer power to a task  $k \in K$  if the requesting IoT device is in the harvesting range of  $e$ . The harvesting range is denoted by  $\phi_{et}$ , which is 1 if task  $k$  is in the harvesting range of  $e$  and 0 otherwise. Let the energy consumption of the  $e$ th energy transmitter in active mode be  $\xi_{e,A}$  and in sleep mode  $\xi_{e,S}$ .

We propose a scheduling scheme for energy transmitters in dedicated energy harvesting for IoT devices, as shown in Fig. 5. IoT devices request power transfer from the controller by sending a request if their residual energy is less than a preset threshold  $\xi_{Th}$ . The threshold is set while considering that the node has sufficient energy for critical operations. The request packet contains the requesting node's ID, the controller's ID, and energy harvesting requirements. Here, we adopt the RF-medium access control (RF-MAC) protocol proposed in [15]. A sensor node with residual energy less than a preset threshold can send RFP for instant charging through an access priority mechanism (for details about this mechanism, see [15]), which ensures that the node with residual energy  $\leq \xi_{Th}$  gets channel access before data transmission by other sensor nodes. The nodes that have data to transmit are forced to freeze their backoff timers as data transmission is not possible at this time. The controller receives this packet and processes it to activate the energy transmitter(s). The harvesting controller receives this request for task  $k$  and calculates  $\phi_{et}$  for all energy transmitters. An energy transmitter can be activated for harvesting the target IoT device(s) if and only if task  $k$  is within the harvesting range of  $e$ , that is,  $\phi_{et} = 1$ ; and task  $k$  is scheduled/activated on  $e$ . We define a binary variable  $\psi_e$ , which is 1 if the energy transmitter  $e$  is scheduled/activated and 0 otherwise.

The objective here is to activate the minimum number of energy transmitters to minimize the energy consumed by dedicated energy transmitters, that is,  $\psi_e \xi_{e,A} + (1 - \psi_e) \xi_{e,S}$ . This is subject to constraints on coverage  $\phi_{et}$ , duration of energy harvesting  $\delta_{et}$ , and target harvesting energy  $\bar{E}_C$ . One way to get an optimal solution is to enumerate over all possible combinations of  $\psi_e$ , which is computationally expensive and unrealistic for a large number of energy transmitters and tasks. Therefore, we consider a branch and bound algorithm for the scheduling of dedicated RF energy sources. Once the activation of energy sources is optimized at the controller, a grant for a power transfer packet is sent to the energy transmitters that are selected for RF power transfer. Finally, the energy source(s) send the acknowledgment packet to the IoT device(s) that requested the power

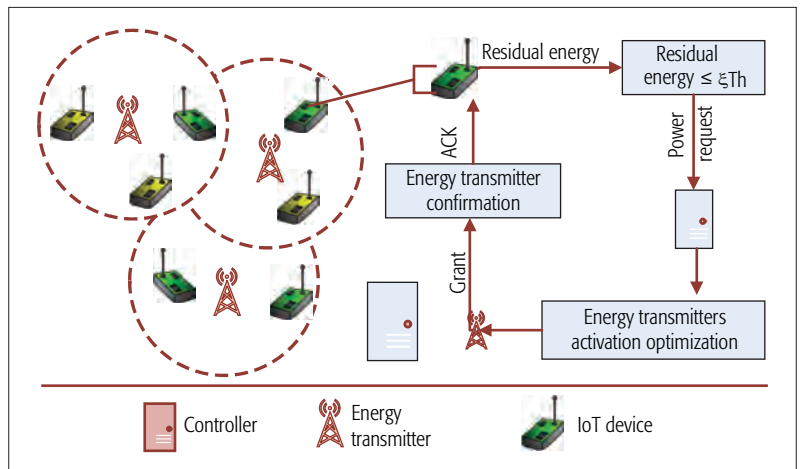
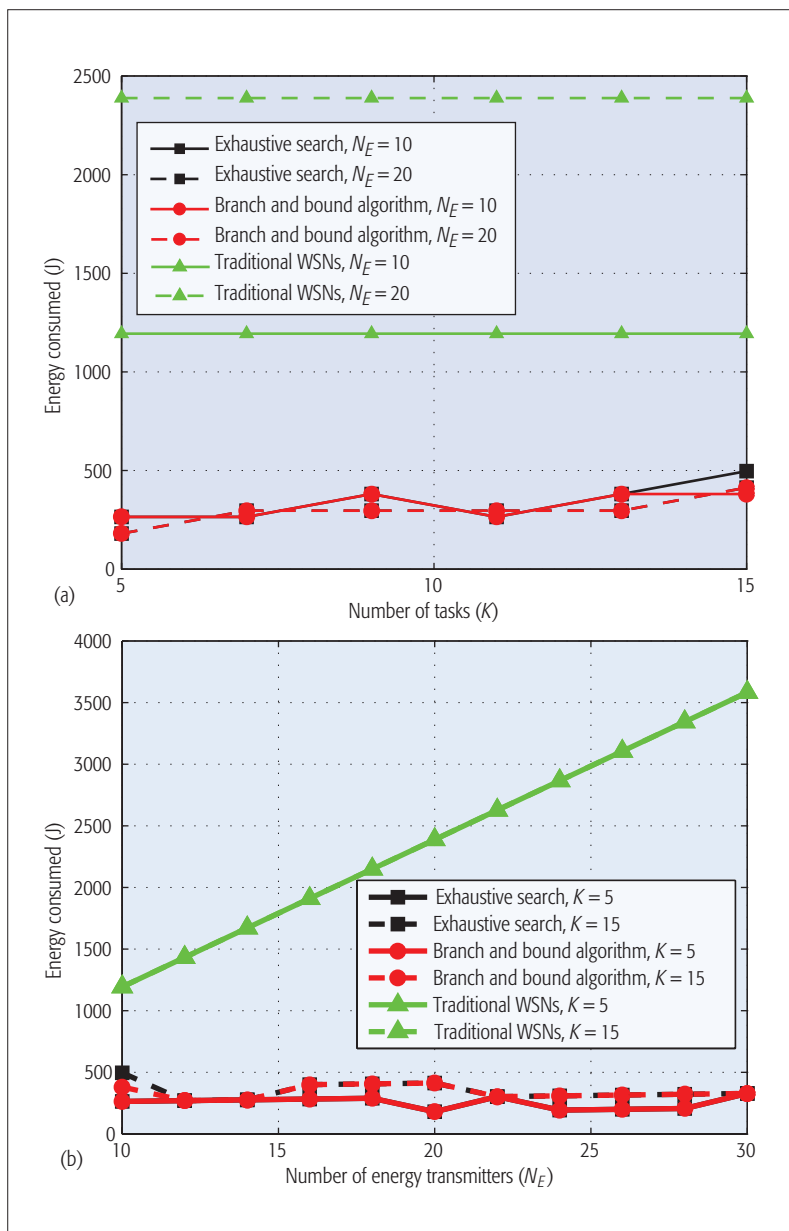


Figure 5. A mechanism for scheduling of energy transmitters.

transfer. This packet has the information of the central frequency of the energy transmitter and the duration of energy charging.

**Performance Analysis:** We evaluate the performance of energy-efficient scheduling of energy transmitters. We consider omnidirectional energy transmitters that radiate waves with power 46 dBm. The proposed schemes can be modified to use with directional energy transmitters to overcome path losses, which can certainly help to improve the charging efficiency. The transmit and receive energy for IoT devices are considered from MICA2 specifications. We consider  $N_I = 200$  IoT devices, which are randomly distributed in a rectangular field of  $100 \text{ m} \times 100 \text{ m}$ .

Figures 6a and 6b illustrate the impact of a number of tasks and energy transmitters on energy consumption, respectively, for an energy-efficient scheduling scheme (branch and bound, exhaustive search, and a traditional wireless sensor network [WSN]). Figure 6a shows that the energy consumption is increased slowly with the increase in the number of tasks in an energy-efficient scheduling scheme (for a given number of energy transmitters, i.e.,  $N_E = 10$  and  $20$ ). This is because energy transmitters are activated based on the number of tasks and their location instead of a total number of energy transmitters. We may need a different number of active energy transmitters if requesting devices are far from or close to each other. The energy consumption in traditional WSNs is constant regardless of the number of tasks, that is, all the energy transmitters are activated all the time. Thus, the energy consumption is doubled when  $N_E = 20$  compared to the case when  $N_E = 10$ . The energy consumption in the proposed scheme is reduced at the cost of overhead and delay due to the exchange of packets among IoT devices, controller, and energy transmitters. From Fig. 6b, it can be noted that the energy consumption for efficient scheduling schemes is not much affected by the increase on energy transmitters  $N_E$  for a given number of tasks ( $K = 5$  and  $K = 15$ ). We consider a small network size for which the probability that tasks are spatially nearby is high. Thus, for different numbers of tasks, we may need to activate the same number of energy transmitters based on their location. Therefore, curves are superimposed. In contrast, traditional



**Figure 6.** Impact of: a) number of tasks  $K$  on energy consumption; b) energy transmitters ( $N_E$ ) on energy consumption for different numbers of tasks  $K$  for different numbers of energy transmitters ( $N_E$ ).

WSNs in IoT-enabled smart cities activate all the energy transmitters regardless of the number of tasks, which results in a linear increase in energy consumption. Moreover, results of the branch and bound algorithm are very similar to exhaustive search with less complexity.

## CONCLUSIONS AND FUTURE WORK

Energy management in smart cities is an indispensable challenge to address due to rapid urbanization. In this article, we first present an overview of energy management in smart cities, and then present a unifying framework for IoT in smart cities. Energy management has been classified into two levels: energy-efficient solutions and energy harvesting operations. We cover various directions to investigate energy-efficient solutions and energy harvesting for IoT devices in smart cities. Furthermore, two case studies have been presented to

illustrate the significance of energy management. The first case study presents appliance scheduling optimization in smart home networks where the objective is to reduce the electricity cost. The second case study covers efficient scheduling of dedicated energy sources for IoT devices in smart cities. Simulation results are presented to show the advantage of energy management in IoT for smart cities. Possible future directions for energy management in smart cities are:

- Energy-efficient mechanisms for software-defined IoT solutions, which can provide scalable and context-aware data and services.
- Directional energy transmission from dedicated energy sources for wireless power transfer.
- Energy efficiency and complexity of security protocols are crucial aspects for their practical implementation in IoT; thus, it is important to investigate robust security protocols for energy constraint IoT devices.
- Fog computing can lead to energy saving for most of the IoT applications; therefore, it is important to study energy consumption of fog devices for IoT applications.

## ACKNOWLEDGMENT

This research was supported by the Korea-China Joint Research Center Program (2016K1A3A1A20006024) through the National Research Foundation (NRF), South Korea.

## REFERENCES

- [1] A. El-Mougy et al., "Reconfigurable Wireless Networks," *Proc. IEEE*, vol. 103, no. 7, July 2015, pp. 1125–58.
- [2] P. Kamalinejad et al., "Wireless Energy Harvesting for the Internet of Things," *IEEE Commun. Mag.*, vol. 53, no. 6, June 2015, pp. 102–08.
- [3] W. Ejaz et al., "Energy and Throughput Efficient Cooperative Spectrum Sensing in Cognitive Radio Sensor Networks," *Trans. Emerging Telecommun. Technologies*, vol. 26, no. 7, July 2015, pp. 1019–30.
- [4] P. Vlachas et al., "Enabling Smart Cities through a Cognitive Management Framework for the Internet of Things," *IEEE Commun. Mag.*, vol. 51, no. 6, June 2013, pp. 102–11.
- [5] F. Qayyum et al., "Appliance Scheduling Optimization in Smart Home Networks," *IEEE Access*, vol. 3, Nov. 2015, pp. 2176–90.
- [6] D. Mishra et al., "Smart RF Energy Harvesting Communications: Challenges and Opportunities," *IEEE Commun. Mag.*, vol. 53, no. 4, Apr. 2015, pp. 70–78.
- [7] T. Ozawa et al., "Smart Cities and Energy Management," *Fujitsu Scientific and Technical J.*, vol. 50, no. 2, Apr. 2015, pp. 49–57.
- [8] H. C. Pohls et al., "RERUM: Building a Reliable IoT upon Privacy-and Security-Enabled Smart Objects," *IEEE Wireless Commun. Networking Conf. Wksp.*, Istanbul, Turkey, Apr. 2014, pp. 122–27.
- [9] D. Bonino et al., "ALMANAC: Internet of Things for Smart Cities," *Proc. Int'l. Conf. Future Internet of Things and Cloud*, Rome, Italy, Aug. 2015, pp. 309–16.
- [10] D. Evans, "The Internet of Things: How the Next Evolution of Internet Is Changing Everything," Cisco tech. rep., Apr. 2011.
- [11] M. Erol-Kantarci and H. T. Mouftah, "Wireless Sensor Networks for Cost-Efficient Residential Energy Management in the Smart Grid," *IEEE Trans. Smart Grid*, vol. 2, no. 2, June 2011, pp. 314–25.
- [12] K. M. Tsui and S.-C. Chan, "Demand Response Optimization for Smart Home Scheduling under Real-Time Pricing," *IEEE Trans. Smart Grid*, vol. 3, no. 4, Dec. 2012, pp. 1812–21.
- [13] M. E. Khanouche et al., "Energy-Centered and QoS-Aware Services Selection for Internet of Things," *IEEE Trans. Automation Science Eng.*, vol. 13, no. 3, July 2016, pp. 1256–69.
- [14] A. Al-Fuqaha et al., "Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 4, Nov. 2015, pp. 2347–76.

- [15] M. Y. Naderi, P. Nintanavongsa, and K. R. Chowdhury, "RF-MAC: A Medium Access Control Protocol for Re-Chargeable Sensor Networks Powered by Wireless Energy Harvesting," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, July 2014, pp. 3926–37.

## BIOGRAPHIES

WALEED EJAZ [S'12, M'14, SM'16] (waleed.ejaz@ieee.org) is a senior research associate in the Department of Electrical and Computer Engineering, Ryerson University, Toronto, Canada. Prior to this, he was a postdoctoral fellow at Queen's University, Kingston, Canada. He received his Ph.D. degree in information and communication engineering from Sejong University, Republic of Korea in 2014. He earned his M.Sc. and B.Sc. degrees in computer engineering from the National University of Sciences & Technology, Islamabad, Pakistan and the University of Engineering & Technology, Taxila, Pakistan, respectively. His current research interests include IoT, energy harvesting, 5G cellular networks, and mobile cloud computing.

MUHAMMAD NAEEM [SM'16] (muhammadnaeem@gmail.com) received his B.S. (2000) and M.S. (2005) degrees in electrical engineering from the University of Engineering and Technology, Taxila, Pakistan. He received his Ph.D. degree (2011) from Simon Fraser University, British Columbia, Canada. From 2012 to 2013, he was a postdoctoral research associate with WINCORE Lab at Ryerson University. Since August 2013, he has been an assistant professor with the Department of Electrical Engineering, COMSATS Institute of IT, Wah Campus, Pakistan, and a research associate with WINCORE Lab. His research interests include optimization of wireless communication systems, non-convex optimization, resource allocation in cognitive radio networks, and approximation algorithms for mixed integer programming in communication systems.

ADNAN SHAHID (adnan.shahid@intec.ugent.be) received his B.Eng. and M.Eng. degrees in computer engineering with communication specialization from the University of Engineering and Technology in 2006 and 2010, respectively, and his Ph.D. degree in information and communication engineering from Sejong University, South Korea, in 2015. He is currently working as a postdoctoral researcher at iMinds/IBCN, Department of Information Technology, University of Ghent, Belgium. From September 2015 to June 2016, he was with the Department of Computer Engineering, Taif University, Saudi Arabia. From March 2015 to August 2015, he worked as a postdoctoral researcher at Yonsei University, South Korea. He was also the recipient of the prestigious BK 21 plus Postdoc program at Yonsei University. His research interests include device-to-device communication, self-organizing networks, small cells networks, cross-layer optimization, and integration of WiFi and small cells.

ALAGAN ANPALAGAN [SM] (alagan@ee.ryerson.ca) received his B.A.Sc. M.A.Sc., and Ph.D. degrees in electrical engineer-

ing from the University of Toronto, Canada. He joined the Department of Electrical and Computer Engineering at Ryerson University in 2001 and was promoted to full professor in 2010. He served the department as graduate program director (2004–2009) and interim electrical engineering program director (2009–2010). During his sabbatical (2010–2011), he was a visiting professor at the Asian Institute of Technology and a visiting researcher at Kyoto University. His industrial experience includes working at Bell Mobility, Nortel Networks, and IBM Canada. He directs a research group working on radio resource management and radio access and networking within the WINCORE Lab. His current research interests include cognitive radio resource allocation and management, wireless cross-layer design and optimization, cooperative communication, M2M communication, small cell networks, energy harvesting, and green communications technologies. He serves as an Associate Editor for *IEEE Communications Surveys & Tutorials* (2012–) and *Springer Wireless Personal Communications* (2009–), and a past Editor for *IEEE Communications Letters* (2010–2013) and the *EURASIP Journal of Wireless Communications and Networking* (2004–2009). He also served as Guest Editor for two *EURASIP Special Issues on Radio Resource Management in 3G+ Systems* (2006) and *Fairness in Radio Resource Management for Wireless Networks* (2008), and a *MONET Special Issue on Green Cognitive and Cooperative Communication and Networking* (2012). He co-authored or edited three books: *Design and Deployment of Small Cell Networks* (Cambridge University Press, 2014), *Routing in Opportunistic Networks* (Springer, 2013), and *Handbook on Green Information and Communication Systems* (Academic Press, 2012). He is a registered Professional Engineer in the Province of Ontario, Canada.

MINHO JO [SM'16] (minhojo@korea.ac.kr) is a professor in the Department of Computer and Information Science, Korea University, Sejong Metropolitan City. He received his Ph.D. in industrial and systems engineering from Lehigh University in 1994 and his B.A. in industrial engineering from Chosun University, South Korea, in 1984. He was one of the founding members and Founding Captain of the IT Team, LCD Division, Samsung Electronics. He has extensive experience in wireless communications and software development in industry for more than 15 years. He is Founding Editor-in-Chief of *KSII Transactions on Internet and Information Systems*. He served as an Editor of *IEEE Network*. He is now an Editor of *IEEE Wireless Communications* and an Associate Editor of the *IEEE Internet of Things Journal*. He is an Associate Editor of *Security and Communication Networks* and *Wireless Communications and Mobile Computing*. He was Vice President of the Institute of Electronics and Information Engineers and is now Vice President of the Korean Society for Internet Information. His current research interests include IoT, LET-Unlicensed, cognitive radio, mobile cloud computing, network security, heterogeneous networks in 5G, wireless energy harvesting, AI in IoT, optimization in wireless communications, and massive MIMO.

# Geo-Conquesting Based on Graph Analysis for Crowdsourced Metatrails from Mobile Sensing

Bo-Wei Chen, Wen Ji, and Seungmin Rho

The authors investigate graph analysis for intelligent marketing in smart cities, where metatrails are crowdsourced by mobile sensing for marketing strategies. Unlike most works that focus on client sides, this study is intended for market planning, from the perspective of enterprises.

## ABSTRACT

This article investigates graph analysis for intelligent marketing in smart cities, where metatrails are crowdsourced by mobile sensing for marketing strategies. Unlike most works that focus on client sides, this study is intended for market planning, from the perspective of enterprises. Several novel crowdsourced features based on metatrails, including hotspot networks, crowd transitions, affinity subnetworks, and sequential visiting patterns, are discussed in the article. These smart footprints can reflect crowd preferences and the topology of a site of interest. Marketers can utilize such information for commercial resource planning and deployment. Simulations were conducted to demonstrate the performance. At the end, this study also discusses different scenarios for practical geo-conquesting applications.

## INTRODUCTION

As the advancement in communication and computing technologies stimulates the progress of smart cities, trillions of sensors are deployed in every corner of a city, subsequently forming a huge sensor network. Data collected by various sensors range from buildings, streets, and transportation systems to natural spaces. These heterogeneous data delineate the characteristics of a city. Among these sensors, mobile sensing is becoming popular due to its mobility and pervasiveness. Furthermore, it reflects human social behavior and shows the interaction between persons and a city.

According to a survey conducted by [www.statisticbrain.com](http://www.statisticbrain.com), the total number of worldwide cellular phone subscriptions was as high as 6.9 billion in 2014. The total app downloads for iOS and Android smartphones reached 29 billion and 31 billion, respectively.

With such a tremendous number of subscribers using a variety of apps, crowdsourced data from mobile sensing become a valuable resource for market planning. For marketers, user behavior reasoning is an important subject for targeting potential customers. Analysis on mobile sensing data provides different clues about user dynamics. One of the feasible analyses is geo-conquesting.

Geo-conquesting is a newly emerging technology in computational advertising. Such a phrase originates from a combination of two words: geography and conquering. According to the definition of conquering mentioned in

[1], it means to deploy an advertisement next to competitors or the products of rivals. Recent location-aware technologies, such as geo-locationing and geo-fencing, have upgraded the scale of conquering to cyber-enabled geo-conquesting. With outdoor and indoor positioning, for example, iBeacon and IndoorAtlas, user locations in geographical areas of interest can be pinpointed more accurately than before. Marketers can use geo-conquesting to hyper-target consumers proactively. Ambient services or products are fed back in real time to users, depending on their locations. This success is attributed to advances in portable and sensing technologies, which have brought the traditional advertising industry into a new era. Conventional demographics (e.g., residential density and population density) no longer keep up with city dynamics. Demographics are static, but geo-conquesting counts on live dynamic data.

Geo-conquesting is not merely smart advertising. It represents proactive market strategic planning. No advertisement spamming is involved. Take retail store locating, for example. Selecting a store site is the decisive factor in profit making. Storeowners know that large traffic flows along the side of a road dominate in-store visiting. However, which side should they choose? What type of stores should they operate? The observation in Fig. 1 reveals an interesting example. We use convenience stores, 7-11, as a case study. The area in Fig. 1 is 1.62 km × 1.20 km, and there are 11 stores. The finding shows that eight stores are on the right side. The explanation is that when people leave their homes and head downtown, they have higher chances of visiting convenience stores if they temporarily need something. As opposed to convenience stores, business operators of supermarkets can select a site based on such a principle. When people head for suburban areas back to their houses, hypermarkets that offer daily necessities attract more customers.

In the cyber age, metadata collected via mobile sensors facilitate geo-conquesting, for mobile sensing reveals customer activities and environmental conditions. Marketers can benefit from these crowdsensed data by carefully investigating regional dynamics.

Crowd behavior can be further manifested when mobile crowdsensing is applied. Based on communication types [2], mobile sensing can be classified into two categories: direct and indirect

sensing. The former involves direct communication between devices (e.g., device-to-device and machine-to-machine approaches) or direct communication between terminals and base stations (e.g., radio networks). Signals during transmission can be directly used for power management and resource planning, and subsequently utilized for estimating the number of cellular phone subscribers or user densities. The latter, indirect sensing, relies on intermedia, for example, social websites and clouds, where the captured data via mobile sensors can be downloaded by analysts and further processed.

For market analytics, indirect sensing is more convenient than direct sensing because the indirect mechanism avoids negotiations with wireless service providers and telecommunication companies. Data can be accessed via dedicated apps or harvested from social websites.

To analyze crowd behavior, urban monitoring is a feasible approach for sampling the dynamics of a city. Relevant research on mobile crowdsensing, like [3, 4], used vehicular sensors to collect the characteristics of a city. Lee *et al.* [3] proposed a vehicular sensor network, MobEyes, for gathering information in an urban environment. Each vehicle represented a sensor node and carried standard equipment for sensing, communication, and storage. Their system was capable of generating street images, recognizing license plates, and sharing information with other drivers and the police. Zheng *et al.* [4] analyzed the traces generated by the Global Positioning Systems (GPSs) in taxicabs to detect flawed urban planning in a city, such as traffic problems in the regions of interest. They used taxi trajectories and the partitioned regions of a city for modeling passenger transitions. Frequent transition patterns were extracted by mining a large transition graph. Transitional time was examined for finding flawed regions. The advantage of vehicular sensing is that the hardware specifications of sensors are usually better than those of mobile phones. This provides better quality during data collection. However, the routes of vehicles are fixed, and the traces are limited to the places accessible to vehicles.

Reades *et al.* [5] developed a measure of bandwidth usage for radio networks, called Erlang. A single unit of Erlang was defined as one person-hour of phone use. They monitored the Erlang data in a region of 47 km<sup>2</sup> over four months as these data could provide bandwidth consumption data and insights into the spatial and temporal dynamics of a city. By observing cumulative Erlang data, salient regions were revealed on the map. Like [5], Calabrese *et al.* [6] also employed cellular phones as the media for urban monitoring. The researchers developed the Localizing and Handling Network Event System (LoCHNES) to monitor traffic conditions and pedestrian/vehicle movements. The LoCHNES employed the radio network database, the antenna database, and the signal-propagation model, all provided by a local telecommunication company, to locate subscribers. When events occurred (e.g., call-in process, message sending, and handover), the system was capable of tracking subscribers. As mentioned earlier, urban monitoring based on radio networks requires participation of telecommunication companies. For third-party marketers



**Figure 1.** Example of geo-conquesting by monitoring in-town and out-of-town flows. The direction of downtown is northwest. Eight out of 11 convenience stores, marked as red circles, are located on the right side. Notably, cars drive by the right side of the road in this city.

(e.g., an advertising agency), such data need to be purchased, and this increases costs.

Based on a recent survey carried out by [www.statisticbrain.com](http://www.statisticbrain.com), more and more smartphone users prefer using apps for communication, so urban sensing via mobile apps has become another way to collect data. Furthermore, the captured data can be forwarded and uploaded to cloud sides stealthily without interfering with user operations. Closely examining the metadata crowdsourced from mobile users shows different demands of customers. This is beneficial for both sides, marketers and customers.

Related innovations are frequently published in the literature. For example, Lu *et al.* [7] devised a pattern recognition approach for acoustic signal processing. The signal, including human voices and ambient sound, was collected using phone receivers. Typical classifiers were used to recognize sound events. Likewise, Xu *et al.* [8] also used machine learning to analyze sounds, but their application focused on speaker counting. The systems developed by Kanjo [9] and by Rana *et al.* [10] were designed for monitoring the noise level of a city by using mobile phones. Actually, these aforementioned creative systems can be furthered to monitor the background sound (e.g., babble) by analyzing phone conversations. With GPS trails, marketers are capable of learning the location information of crowds. Thus, direct marketing becomes effective.

Rather than using acoustic information, [11, 12] concentrated on images collected from mobile social networks. The former discovered the people, activity, and context (e.g., indoor/outdoor and location names) in a picture, whereas the latter involved semantic interpretation and understanding on flyers. Both adopted machine learning and social network analyses. Compared to acoustic signals, images tell more tales than audio does. For instance, crowdsourced photographs taken in a place at a certain moment

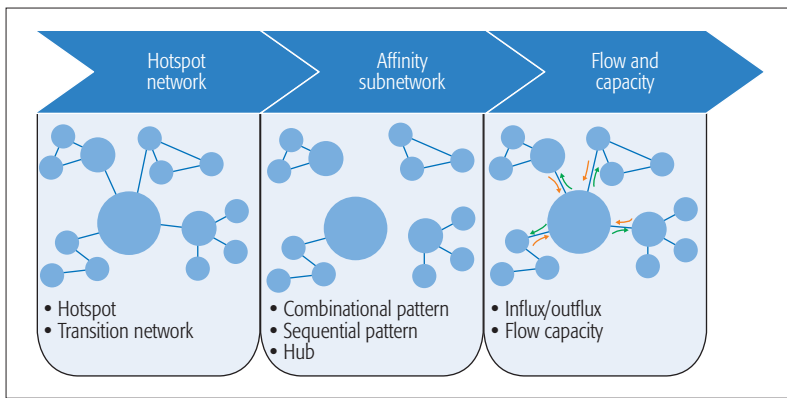


Figure 2. Illustration of the system.

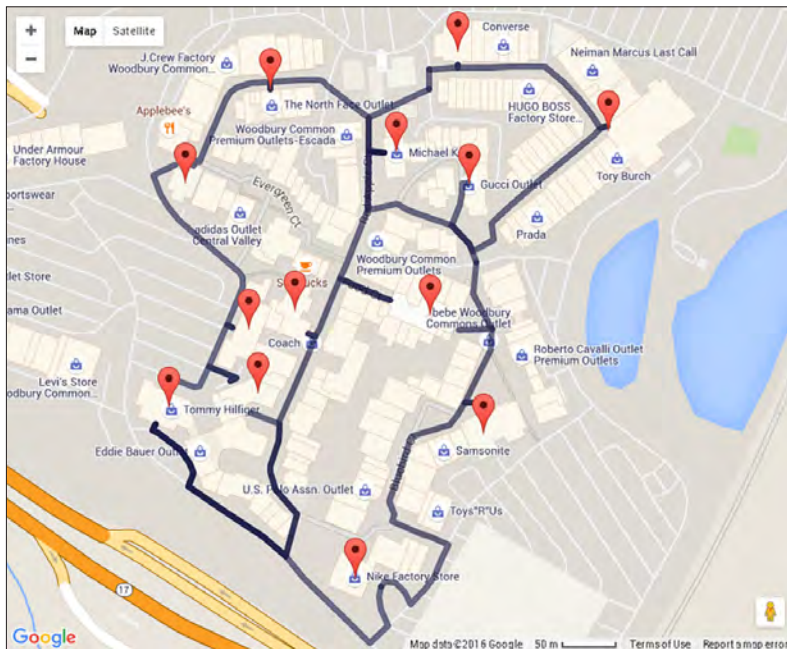


Figure 3. Example of hotspot networks (generated by gmaps.js): mall.

probably indicate a tourist attraction or a popular social activity. With geospatial and temporal analyses, marketers can publicize their products easily. Such methods as sales force allocation technology for electronic advertisement insertion or retail location analysis for physical store deployment can enhance the product image.

At present, much effort has been devoted to research on consumer purchasing behavior and patterns. Nevertheless, few IT approaches have been proposed for maximizing the interest of enterprises. To this end, this study concentrates on the side of marketers. In the remainder of this article, several novel ideas are examined to deal with the challenge of geo-conquesting in smart cities.

### CROWDSOURCED FEATURES

Customer trails, as mentioned earlier, provide useful clues for market analysis. Previous works [5, 6, 13] usually focused on generation of heatmaps for displaying regions of interest in a city. However, valuable information is hidden in the crowdsourced trails, for instance, sequential visiting patterns, transition flows, and site combinations.

To extract such hidden information that reflects customer behavior, this study investigates the dynamics behind crowdsourced trails, including hotspot networks, transitions, and affinity subnetworks by using graph analysis. Figure 2 shows the overview. Heterogeneous data transmission while different apps are running is used in the model. Such information is another indicator of user dynamics, especially when the data are co-displayed with trajectories. Typically, two types of data are formed while apps are running. One is the trails based on app types (e.g., instant messaging apps), and the other is crowdsensed multimedia. This study highlights the former type rather than the latter because crowdsensed multimedia involve pattern recognition, and this study focuses on graph analysis.

As these data are generated along with trails, app types and their timestamps are labeled in the trajectories. We use the term “metatrails” to represent them. Metatrails contain more geospatial and temporal characteristics that feature the preference of customers than GPS trails do. Different metatrails render various activities of customers. More importantly, metatrails are embedded with mobile geosocial networks.

### HOTSPOT NETWORK

This feature is inspired by GPS navigation systems, widely used in our daily lives. For a tourist attraction, it is usually a frequently visited spot or a finish point. Observations show that a waypoint or start point on the route is sometimes another frequently visited spot. With sufficient data, there is a frequently visited route between two hotspots. Large-scale metatrails collected from crowds can substantiate such an observation. This feature is applicable in a city or a site. The following steps show how to build a hotspot network.

Let an edge denote a road, and let a vertex represent a point of interest (i.e., a waypoint, a start point, or a finish point). According to graph theory, a walk is defined as a sequence of alternating vertices and edges. Let us also define the distance between two directly connected vertices as one.

Given a frequently visited spot  $r$  on a map, a tree is created by tracing all the routes (i.e., walks) of which the distance is one. Next, removing all the vertices of which the visiting frequencies are lower than a predefined threshold generates a new pruning tree. Iteratively selecting a vertex in this tree yields a network  $G$ .

Based on all the vertices and edges in this network, a matrix of transition probabilities is formed by calculating the in-degrees and the out-degrees of the vertices. Notably, a user trajectory is a sequence of coordinates with timestamps. Therefore, the system can compute the in-degree and the out-degree of a vertex. For example, assuming there are three vertices  $v_2-v_4$  adjacent to vertex  $v_1$ , the out-degrees from  $v_1$  to  $v_2-v_4$  are one, two, and three, respectively. Thus, the transition probabilities from  $v_1$  to  $v_2-v_4$  are  $1/6$ ,  $2/6$ , and  $3/6$ , respectively.

Compared to heatmaps that display frequently visited areas without showing connections between them, hotspot networks use transition probabilities to present user preferences and flows between places.



An example of hotspot networks is shown in Fig. 3. This hotspot network displays frequently visited stores (i.e., red marks) in a mall, where the black lines are frequently traversed routes.

### AFFINITY SUBNETWORKS: COMBO-SITE MINING

The transition matrix of a hotspot network represents the preference of mobile crowds when they move between places. A visiting pattern can be discovered by performing subgraph analysis. One feasible way is graph clustering. When graph clustering is applied to a hotspot network, hotspots are grouped together, subsequently forming several subnetworks. As the objective of graph clustering is to group vertices that present high connectivity, we denote the resulting subnetworks as *affinity subnetworks*. Herein, affinity is used to describe crowd preferences. Affinity subnetworks further the analysis for retail site selection from single sites to combo-site selection. Combo sites take multiple effects and yield more profits than single sites.

At present, many graph clustering approaches have been proposed. Among these approaches, spectral clustering, Markov clustering, minimum cut, and  $K$ -means are most related to our case. Notably, crowdsourced data can reach trillions. To process billions of vertices in a network, complexity is of prior concerns. Algorithms that involve matrix decomposition like Principal Component Analysis and Singular Value Decomposition take up too much computational time. Thus, they are inappropriate in our case. In the following content, we use Markov clustering as a case study because it is directly related to transition probabilities.

Markov clustering was derived from flow simulation by Stijn van Dongen [14]. It used the ideas of Markov chains and random walks within a graph by iteratively computing transition matrices based on edge weights. The intuition behind Markov clustering is that if a graph possesses a clustered structure, random walks between vertices lying in the same cluster are more likely than those between vertices located in different clusters [14]. This finding is based on the equilibrium distribution of Markov chains. Let  $\pi$  represent the matrix of initial probabilities for all the vertices in a hotspot network  $G$ . By multiplying  $\pi$  by a transition-probability matrix  $T$  within finite times, the resulting product becomes stable. Furthermore, regardless of start points, the equilibrium distribution is the same. Markov clustering employs two major operations — expansion and inflation — for graph clustering. The former tests connectivity between vertices when taking the Hadamard product. The latter increases tightness of clusters. Eventually, iterations result in separation of the network.

After clustering, each cluster forms an affinity subnetwork, that is, a combination of frequently visited hotspots, as shown in Fig. 4. An affinity subnetwork indicates that people prefer visiting these places in combination. This is because Markov clustering simulates people randomly walking in these hotspots based on their interest (i.e., transition probabilities) when sufficient crowdsourced trails are collected. Therefore, strong connectivity is created among hotspots.

A hub in an affinity subnetwork, that is, a vertex with the highest degree, is a pivotal place that

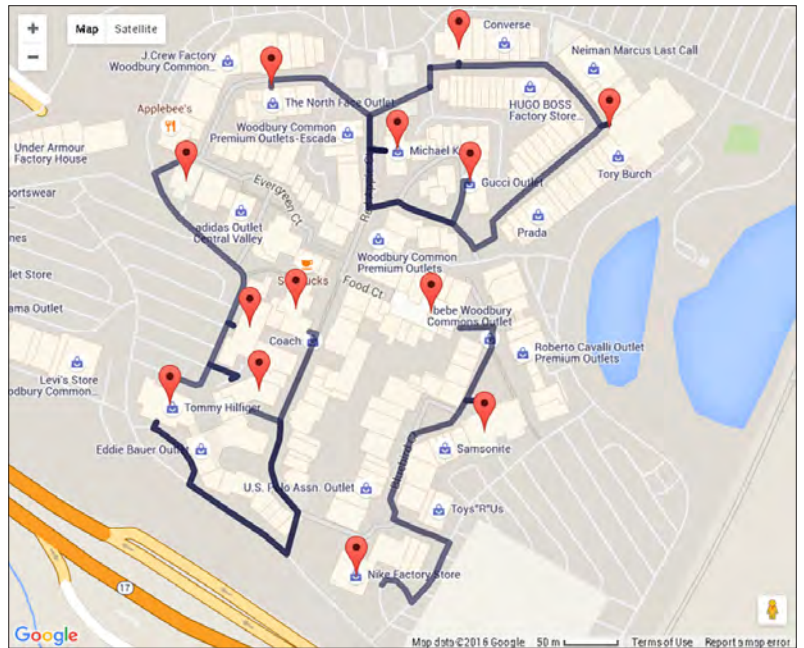


Figure 4. Example of affinity subnetworks. Three subnetworks were discovered by graph mining when crowd dynamics were considered. Each represents a combination of frequently visited hotspots. People prefer staying in the same subnetwork and visiting these places in combination.

people at adjacent hotspots have more chances to visit. Mining affinity subnetworks and hubs helps marketers discover combo stores and combo hotspots.

Each affinity subnetwork can be contracted to a vertex. Thus, connections between clusters are easily visualized.

### SEQUENTIAL VISITING PATTERN: BETWEEN HOTSPOTS AND BETWEEN SUBNETWORKS

Mining the sequential patterns of mobile footprints is conducive to predicting crowd preferences and arranging store locations. Nevertheless, it is difficult to analyze large-scale trails because of complexity. Besides, sequential patterns of long duration are not practical for market planning. Fortunately, affinity subnetworks reduce computational time since graph clustering breaks a large graph into small components.

Unlike affinity subnetworks that concentrate on combo stores, this feature highlights the order of patterns. There are two types of sequential visiting patterns. One is the order of the hotspots in an affinity subnetwork, and the other is the order of subnetworks.

To analyze the order of hotspots, first all the trails in an affinity subnetwork are extracted. Then the system rearranges the vertices based on their latest timestamps in the trails. When more than one sequential pattern is generated, it means at least two types of orders exist in an affinity subnetwork.

When the traversal order of different subnetworks is focused, an entire subnetwork is viewed as a vertex. That is, all the vertices are contracted to generate one vertex. The whole hotspot network  $G$  becomes a new graph, of which each vertex represents an affinity subnetwork. The approach for analyzing the order of subnetworks is similar to that for the order of hotspots.

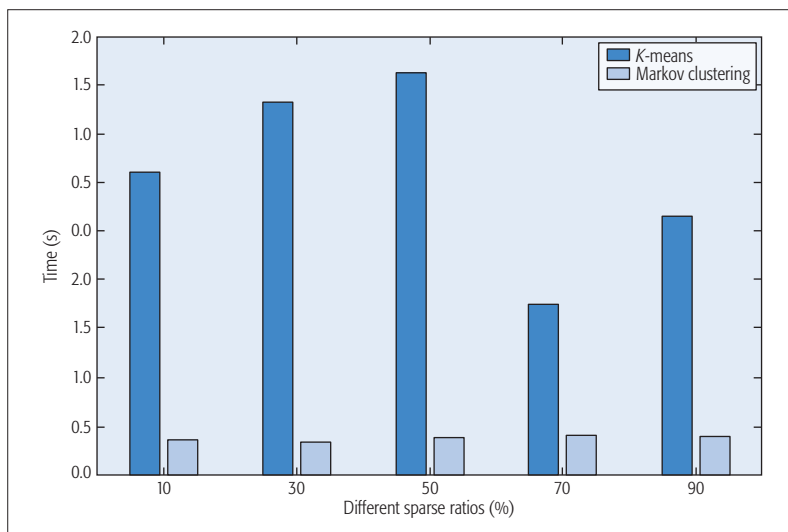


Figure 5. Graph clustering for combo-site analysis.

### FLows AND CAPACITIES OF THE HOTSPOT NETWORK

In-store visits involve two factors. One is the directions of traffic flows, and the other is the flow capacity. The former dominates store types that are highly related to customer preferences when they move, as mentioned earlier in Fig. 1. The latter increases store visibility if the location of a store is right next to a major flow. Therefore, monitoring flow directions and flow capacities in a hotspot network is important. Flow directions can easily be manifested by computing in-degrees and out-degrees. However, flow capacities require a more complicated mechanism because both overflows and underflows in a route have influences on in-store visits and visibility. This relatively affects site rentals, product prices, and profits.

The information on the maximum and minimum flow capacity is useful when marketers select sites. Their requirements for maximum and minimum capacities can be converted into upper and lower bounds (i.e., constraints) during profit optimization. Assume that low flows bring few in-store visits and subsequently low rentals. There is a balancing relation between flows and site rentals. However, this is still a challenge of multi-side profit optimization.

### GEO-CONQUESTING

Unlike many works that focus on the user side by mining their transaction records, browsing histories, and social connections, geo-conquesting concentrates on the marketer side. The following section discusses two different scales of geo-conquesting strategies by using the crowdsourced features in the previous section. One is from the perspective of mall management, and the other is for metropolitan planning.

For mall operators, as hotspot networks display frequently visited stores, the difference in rentals between stores with high and low in-store visits can be adjusted dynamically. When rentals decrease, the cost is reflected in the price of retail products, subsequently attracting more customers. There is mutual interest among mall operators, store owners, and customers.

Traffic flow management is another ben-

efit brought by hotspot networks. Mall operators can examine transition probabilities, crowd directions, and flow capacities to redeploy mall facilities and balance flows. Compared to crowd flows or heatmaps in the unstructured open space [13], flows detection based on hotspot networks generates more concrete information on store connectivity. As various app users generate different routes, electronic billboards or digital billboards can be set up in the hallway for displaying advertisements. Additionally, advertising content can be proactively changed based on flow types for targeting customers. Mall operators can dynamically charge advertising agencies according to flow amounts. When combined with sequential visiting patterns, deployment of stores and digital billboards can be reshaped to fit the flow directions, subsequently bringing in more customers.

For mall operation, how to select a group of stores and deploy them together is an important topic since an appropriate combination of stores creates a weighting factor in in-store visits. With the use of graph clustering and transition flows, the entire hotspot network is separated into several affinity subnetworks. Each subnetwork represents combo stores. With such combo information, mall managers can preallocate space for lease. Store owners can join an affinity subnetwork and open a store that fits this subnetwork.

For conquering in a city, the main focus is on profitable regions of interest or points of interest. As large traffic flows increase visibility of stores, a hotspot network provides a good indicator for retail store locationing. Moreover, affinity subnetworks can further store location analysis from individual sites to combo-site selection. Advertising agencies can utilize affinity subnetworks to project related-product images to customers because affinity subnetworks represent customer preferences. When geo-conquesting meets city planning, it becomes city marketing. City planners can improve public transportation and infrastructures by exploring affinity subnetworks. Metropolitan branding will become more effective via dynamic crowdsourcing.

### NUMERICAL RESULT

As the combo-site analysis is important to geo-conquesting, we conducted a simulation to test the performance of generating affinity subnetworks. A directed weighted graph was randomly synthesized with 500 vertices. The sparse ratios were from 10 to 90 percent with a separation of 20 percent. Two typical algorithms for rapid graph clustering — Markov clustering and *K*-means — were benchmarked. Both involved no matrix decomposition. The former used the source codes (developed by Daniel A. Spielman, Yale University) with our modifications. The latter followed the algorithm in [15]. Both the iterations were fixed at 500. The Hadamard power was two, and the number of clusters in *K*-means was 10.

Figure 5 shows the computational performance. The vertical axis signifies the computational time, whereas the horizontal axis displays the two different approaches applied to various data. As shown in Fig. 5, when Markov clustering was

used, the performance was higher than that of K-means. The computational time of Markov clustering was 0.1518 s on average, faster than that of K-means (1.2365 s). This is conducive to the data analysis as the computational time is favorable when crowdsourced data are collected.

## CONCLUSION

This study examines crowdsourced features generated from mobile metatrails for geo-conquesting. To reflect the dynamics of a city, graph analysis is used for discovering hotspot networks, transition probabilities, and flow capacities. Subsequently, affinity subnetworks and sequential visiting patterns are extracted by using rapid graph clustering. Affinity subnetworks (i.e., combo stores) and sequential patterns allow marketers to analyze crowd sequential activities, between stores and even between subnetworks. Such a discovery creates a weighting factor for in-store visits.

Different scales of geo-conquesting strategies, ranging from mall operations to metropolitan business targeting, are also presented in the discussion. With these features, intelligent marketing becomes feasible because store locationing is based on dynamic city characteristics instead of static demographics.

Interesting challenges arise in intelligent marketing, for example, combo-site locationing, sales force allocation theory, and multiside profit optimization. In the future, there will be systematic mathematical equations for modeling these challenges.

## REFERENCES

- [1] E. Steel, "Marketers Try 'Conquesting' to Get on Rivals' Nerves," *Wall Street J.*, June 7, 2007.
- [2] P.-Y. Chen *et al.*, "When Crowdsourcing Meets Mobile Sensing: A Social Network Perspective," *IEEE Commun. Mag.*, vol. 53, no. 10, Oct. 2015, pp. 157–63.
- [3] U. Lee *et al.*, "MobEyes: Smart Mobs for Urban Monitoring with a Vehicular Sensor Network," *IEEE Wireless Commun.*, vol. 13, no. 5, Oct. 2006, pp. 52–57.
- [4] Y. Zheng *et al.*, "Urban Computing with Taxicabs," *Proc. 13th Int'l. Conf. Ubiquitous Computing*, Beijing, China,

2011, Sept. 17–21, pp. 89–98.

- [5] J. Reades *et al.*, "Cellular Census: Explorations in Urban Data Collection," *IEEE Pervasive Computing*, vol. 6, no. 3, July–Sept. 2007, pp. 30–38.
- [6] F. Calabrese *et al.*, "Real-Time Urban Monitoring Using Cell Phones: A Case Study in Rome," *IEEE Trans. Intell. Transp. Sys.*, vol. 12, no. 1, Mar. 2011, pp. 141–51.
- [7] H. Lu *et al.*, "SoundSense: Scalable Sound Sensing for People-Centric Applications on Mobile Phones," *Proc. 7th Int'l. Conf. Mobile Systems, Applications, Services*, Kraków, Poland, June 22–25, 2009, pp. 165–78.
- [8] C. Xu *et al.*, "Crowdsensing the Speaker Count in the Wild: Implications and Applications," *IEEE Commun. Mag.*, vol. 52, no. 10, Oct. 2014, pp. 92–99.
- [9] E. Kanjo, "NoiseSPY: A Real-Time Mobile Phone Platform for Urban Noise Monitoring and Mapping," *Mobile Networks and Applications*, vol. 15, no. 4, Aug. 2010, pp. 562–74.
- [10] R. K. Rana *et al.*, "Ear-Phone: An End-to-End Participatory Urban Noise Mapping System," *Proc. 9th ACM/IEEE Int'l. Conf. Info. Processing in Sensor Networks*, Stockholm, Sweden, Apr. 12–15, 2010, pp. 105–16.
- [11] C. Qin *et al.*, "TagSense: Leveraging Smartphones for Automatic Image Tagging," *IEEE Trans. Mobile Computing*, vol. 13, no. 1, Jan. 2014, pp. 61–74.
- [12] B. Guo *et al.*, "FlierMeet: A Mobile Crowdsensing System for Cross-Space Public Information Reposting, Tagging, and Sharing," *IEEE Trans. Mobile Computing*, vol. 14, no. 10, Oct. 2015, pp. 2020–33.
- [13] F. Girardin *et al.*, "Digital Footprinting: Uncovering Tourists with User-Generated Content," *IEEE Pervasive Computing*, vol. 7, no. 4, Oct.–Nov. 2008, pp. 78–85.
- [14] S. V. Dongen, *A Cluster Algorithm for Graphs*, Nat'l. Research Inst. Mathematics and Comp. Sci., Amsterdam, The Netherlands, May 2000.
- [15] S.-Y. Kung, *Kernel Methods and Machine Learning*, Cambridge Univ. Press, June 2014.

## BIOGRAPHIES

BO-WEI CHEN [M'14] is with the School of Information Technology, Monash University. His research interests include data analytics, machine learning, and audiovisual sensor networks. He serves as the Chair of the Signal Processing Chapter, IEEE Harbin Section.

WEN JI [M'09] is with the Institute of Computing Technology, Chinese Academy of Sciences. Her research areas include communications, video coding, information theory, and optimization.

SEUNGMIN RHO [M'08] is with the Department of Media Software, Sungkyul University, South Korea. His research interests include databases, big data analysis, music retrieval, and knowledge management.

Interesting challenges arise in intelligent marketing, for example, combo-site locationing, sales-force allocation theory, and multiside profit optimization. In the future, there will be systematic mathematical equations for modeling these challenges.

# Efficient Media Streaming with Collaborative Terminals for the Smart City Environment

Jordi Mongay Batalla, Piotr Krawiec, Constandinos X. Mavromoustakis, George Mastorakis, Naveen Chilamkurti, Daniel Négru, Joachim Bruneau-Queyreix, and Eugen Borcoci

The authors introduce the architectural modules required for collaboration streaming inside the radio access network and end user's device, and they propose enhancements in HTTP-compliant adaptive streaming protocols in order to become suitable in a multipath collaborative scenario.

## ABSTRACT

Among multiple services delivered over future mobile networks, the most demanding (from the required bandwidth point of view) are related to media streaming, which is a key component in smart applications (entertainment, tourism, surveillance, etc.). Such applications have to exploit a considerable amount of data, which is difficult to achieve especially in dense urban environments. In this context, the article presents a new solution for HTTP-compliant adaptive media streaming applicable to future 5G mobile networks, aimed at increasing bandwidth availability through the use of multiple radio access technologies and direct connections between devices if they are in proximity of each other. The proposed solution considers a scenario in which a high-quality media stream is received by multipath transmission through the radio access network. Collaboration of neighboring devices is exploited by using direct device-to-device links. Thus, proxy nodes can be inserted between a given media receiver and an access network. Toward ensuring optimized resource allocation at both levels, base station-to-device and device-to-device, this article introduces the architectural modules required for collaboration streaming inside the radio access network and end user's device, and proposes enhancements in HTTP-compliant adaptive streaming protocols in order to become suitable for a multipath collaborative scenario.

## INTRODUCTION

Multimedia emerging applications including advanced video definition (4K, 8K) together with other features such as augmented reality, 3D, or multi-angle vision will flood mobile networks in the near future. Multimedia delivery in a large metropolitan area will require wireless network infrastructure to handle much higher traffic volume than today. These requirements are targeted by the novel fifth generation (5G) mobile network. The 5G design aims to enhance the system capacity in urban areas, mainly to be achieved by:

- Higher spectral efficiency thanks to massive multiple-input multiple-output (MIMO) techniques
- Infrastructure densification by widespread

deployment of small (micro/pico/femto) cells, jointly with support of heterogeneous radio access technologies (RATs)

- Usage of additional spectrum, particularly from the millimeter-wave region, which may provide high bandwidth, but with relatively short-range coverage due to line of sight propagation with minimal refraction

In addition to the capacity increase, the radio access network (RAN) in urban areas — characterized by high density, in terms of population, buildings, and traffic demand — should introduce new mechanisms to efficiently manage the resources. The bandwidth available for users can vary substantially even in neighboring locations. On the other hand, the urban environment increases the probability of finding networking devices, which could act as intermediate nodes between media consumer and media source.

Collaboration is commonly considered as a way to improve the efficiency of resource usage in scenarios with physically dense networks, such as smart cities. Some research studies have proposed a collaborative framework for reducing congestion at the network backhaul by caching popular media content at base stations or mobile devices [1, 2]. The advantage of such solutions is their simplicity; however, they do not solve the problem of reduced bandwidth access for a single end user. Another scope refers to improvement of streaming capacity thanks to aggregation of downlink rates available for collaborating terminals. Many publications propose overlay solutions installed on top of the network and, possibly, profiting from facilities (e.g., server capacity) of the network. These solutions suffer high response delays and inefficiency in resource usage. Moreover, they are constrained by a supplementary service agreement in addition to the one with the mobile operator. Examples of such solutions are [3–6]. Proposals [3, 4] are solutions for HTTP direct download, which is not compliant with current network conditions where adaptive streaming protocol is required. Only [5, 6] present overlay solutions with adaptive streaming. To the best of our knowledge, only Syrivelis *et al.* proposed a network-native solution for collaborative streaming by the use of software defined networking [7]; however, it does not consider adaptive streaming,

*Jordi Mongay Batalla and Piotr Krawiec are with the National Institute of Telecommunications and Warsaw University of Technology; Constandinos X. Mavromoustakis is with the University of Nicosia; George Mastorakis is with the Technological Educational Institute of Crete; Naveen Chilamkurti is with La Trobe University – Melbourne; Daniel Négru and Joachim Bruneau-Queyreix are with the University of Bordeaux; Eugen Borcoci is with the University Politehnica of Bucarest.*

Digital Object Identifier:  
10.1109/MCOM.2017.1600225CM

which excludes the solution from implementation in real networks.

The solution presented in this article proposes a collaborative framework managed directly by the RAN, which is responsible for governing uplink/downlink access, but also controls the state of the device-to-device (D2D) links. Our solution merges network-assisted device cooperation with modern adaptive streaming functionality to ensure optimal radio resource allocation jointly with the highest video quality perceived by users.

The rest of the article is organized as follows. The next section presents the architecture for collaborative streaming management in dense heterogeneous networks (e.g., cities) specifying the proposed modules and entities. Then we show the enhancements in the end devices required for making collaborative media streaming feasible and discuss the resource allocation in the system. After that we present the solution for adaptive streaming in the multipath collaborative scenario. Results of the implemented solution in contrast to other uni- and multipath proposals are included. At last, the final section concludes the article.

## ACCESS NETWORK ARCHITECTURE FOR COLLABORATIVE MEDIA STREAMING

Highly dense heterogeneous networks (HetNets) may increase the efficiency of radio spectrum usage by dividing the space into micro- and picocells. Multiple antennas and multiple RATs may increase the capacity of a RAN only if the system is capable of locating each mobile device and executing appropriate algorithms to automate the cooperation procedures. Such functionalities are generally located in the remote radio head (RRH) within micro/picocells. On the other hand, the use of cloud methods for control and management of the radio resources may greatly help with the introduction of such complex algorithms [8]. These cloud facilities are located in the baseband units (BBUs). BBUs are responsible for the signal processing and layer 2/layer 3 functions, detaching these functions from the conventional base stations and moving them to a centralized location. The distributed conventional base stations now become simple RRH modules that perform conversion between digital baseband signals and analog signals transmitted/received by antennas.

In centralized RAN (C-RAN) [8], the BBUs run as a virtualized pool of processing resources in a dedicated data center or using cloud services, and can be shared between different RRHs. In this way, the network operators can reduce energy consumption in the mobile infrastructure, decrease both deployment and operational costs of BBU, and also facilitate network upgrades and future migrations to new solutions. Furthermore, emerging mobile edge computing [9] servers deployed on RAN premises can offer spare cloud computing capabilities available at the RAN to authorized third parties, using open interfaces. It allows additional functionalities to be launched at the RAN, such as caching and transcoding in the case of media streaming.

We propose to increase cooperation of technologies in HetNets for ensuring high-quality multimedia streaming anywhere in the city. In a basic scenario an end user intends to download a high

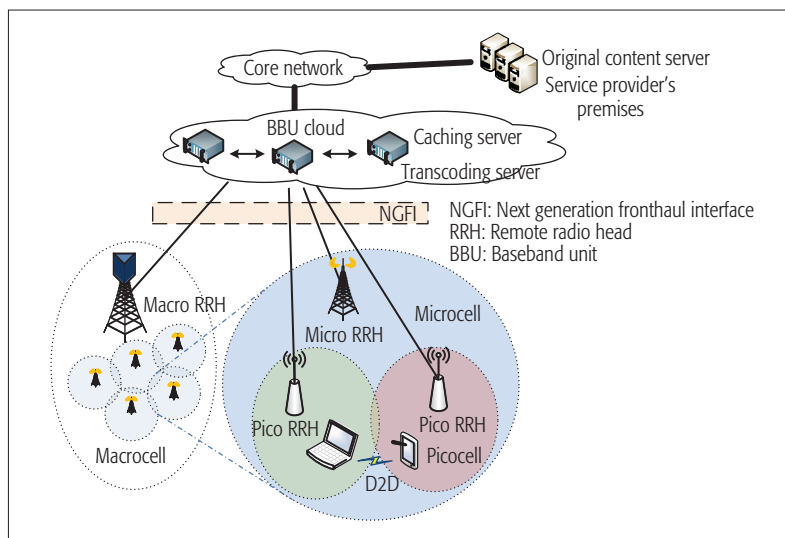


Figure 1. C-RAN architecture.

bandwidth stream (in general, video) by using an HTTP adaptive connection (Fig. 1). To this aim, a collaborating relay device (if possible) may download portions of the media, exploiting the downlink bandwidth available to it; next, it retransmits the downloaded media to the receiver (end user) via a direct D2D link. An efficient collaborative streaming should involve different RATs with coordinated spectrum usage (e.g., LTE for downloading media from the network, and WiFi, Bluetooth, or LTE-Direct to transfer the media directly between user devices) as well as make use of current adaptive streaming protocols for adapting the media bit rate to the overall download rate.

A centralized BBU architecture is more consistent with collaborative media streaming than edge computing, given the availability at the BBU level of device contextual information that is important for making appropriate collaboration decisions. Although various schemes of splitting the baseband functionality between BBU and RRH have been investigated [10], depending on fronthaul capacity and delay requirements, the most promising is the fully centralized option, with BBU performing baseband processing at three layers from layer 1 (i.e., physical) to layer 3. It makes the most efficient use of the radio resources based on cross-layer optimization while considering multi-RAT availability. BBU centralization makes cooperation and synchronization of RRHs feasible for sending data to users positioned in close pico-, micro-, and macrocells with coordinated multi-point (CoMP), which enables better resource utilization due to reduction of inter-cell interference.

To enable *collaborative media streaming*, this work introduces new functional modules into the BBU to exploit any possible transmission method available at a given time, even by using the D2D connection between neighboring terminals. The BBU functional architecture extended with the proposed collaborative media streaming modules is presented in Fig. 2. It covers modules that perform protocol processing for the supported RAT family controlled by a *radio resource allocation* (RRA) entity. RRA implements multilayer (L1 to L3) and multi-RAT coordination for attaining high spectral efficiency and throughput as well as load

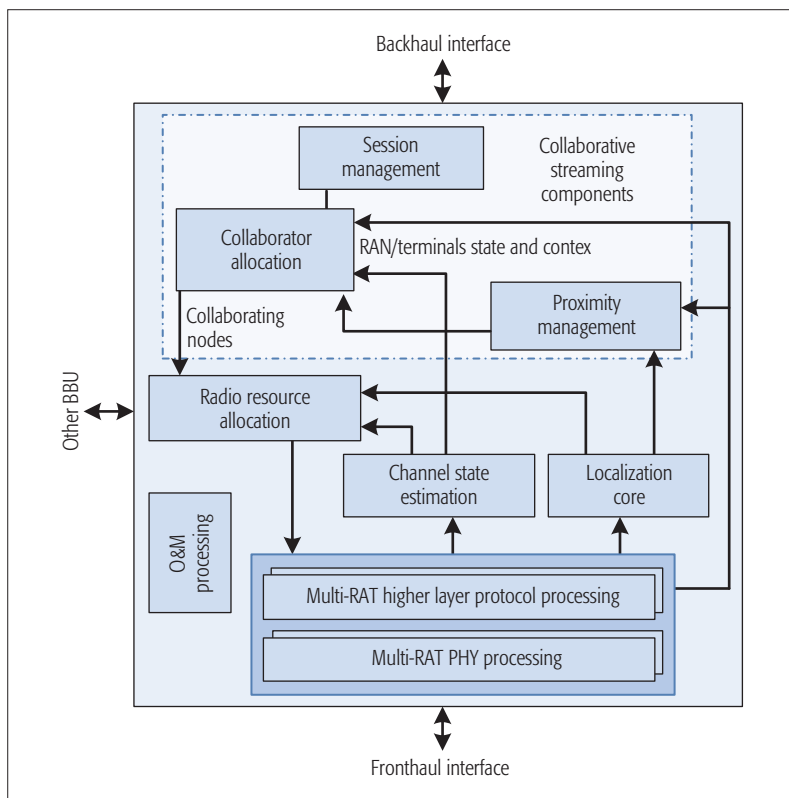


Figure 2. Enhancement of BBU architecture for collaborative media streaming.

balancing mechanisms between cells, and exploits information about terminals' physical locations, obtained from a *localization core*, which may provide highly accurate device geolocation [11].

The new components control the whole collaborative media streaming process, from finding the appropriate collaborator to assigning data to the proper path (directly to the destination terminal of the neighboring device). The collaborative streaming control layer in BBU involves a *proximity management* module that dynamically decides which devices/terminals can operate as neighbors, considering their physical location, but also the devices' capabilities, including available wireless interfaces (usable for direct connections), battery status, and so on. Moreover, collaborative streaming components on different BBUs of the pool interact to discover which terminals are associated with different base stations and which ones could establish a direct connection with each other.

When *proximity management* determines the set of potential neighbors, it passes information about them to *collaborator allocation entities*. Furthermore, the BBU sends (to *collaborating terminals*) a command initiating the monitoring of the state of potential D2D connections. Measurement reports from the terminals are collected by a *channel state estimation* module, which provides (to *collaborator allocation*) information about current channel state for each D2D link (registered in the *collaborator allocation* with a separate ID). Consequently, the *collaborator allocation* decides which devices, from a set of neighbors, can be currently used as collaborators in media streaming. The *collaborator allocation* module indicates to the RRA which devices are paired as neighbors.

The *session management* module controls media sessions at the BBU. Media sessions are supposed to be transported by HTTP-based adaptive streaming protocols, which are the most common solutions in the current Internet (e.g., Adobe Dynamic Streaming, Apple's HTTP Live Streaming, Microsoft's IIS Smooth Streaming and Dynamic Adaptive Streaming over HTTP – DASH). Their most interesting feature is the capacity of adapting the streaming bit rate to the state of the transportation path for the flow. The file ordering all the HTTP segments (the fragments of video are independently callable with their own URLs) is the manifest file.

*Session management* identifies a new media session start by recognizing the manifest file downloaded by a user; it analyzes the file, and becomes aware of the different fragments of video (called chunks or segments) and the bandwidth necessary for their transmission. Next, when the module detects a segment request from the user, it asks *collaborator allocation* if additional bandwidth for a given user could be available through collaboration. After checking the potential collaborators and selecting the appropriate one/s, *Session management* informs the user about how to request the content fragments from collaborating devices and, concretely, sends information containing the URLs to request the content fragments through different paths. Let us remark that the content fragments are short in HTTP-based adaptive streaming (2 to 10 seconds), which avoids problems of device mobility due to very short session duration. When a collaborator receives the request through a D2D link, it redirects the request to a BBU, and the request is further handled by the *session management* module in the BBU.

To speed up the media streaming process, *session management* may work as a caching entity and, based on information provided by the manifest file, downloads (in advance) a small number of portions of video with best quality (higher bit rate) from the content server. In this way, the module responds to a user or collaborators just after receiving media requests, using a priori cached data, to reduce media delivery latency. This solution is also open to other caching solutions [9, 12].

## TERMINAL ARCHITECTURE FOR COLLABORATIVE APPLICATIONS

This section presents the functionalities required in the end users' terminals in order to provide collaboration in media streaming upon acceptance of the end users (agreement for collaboration in media streaming).

In the proposed solution, the process of finding adjacent devices and establishing D2D links is fully BBU controlled; the BBU has a view of the entire access domain, so it can optimize the selection of neighbors and RAT used for D2D connection. In this way, D2D connection establishment is fast and resource-efficient, whereas neighbor discovery performed solely by terminals (as occurs in overlay solutions) is energy- and time-consuming. Last but not least, the network operator performs authentication of collaborating devices, which is important from the point of view of security and privacy. Nonetheless, the set of neighboring devices is restricted to only a given operator's cli-

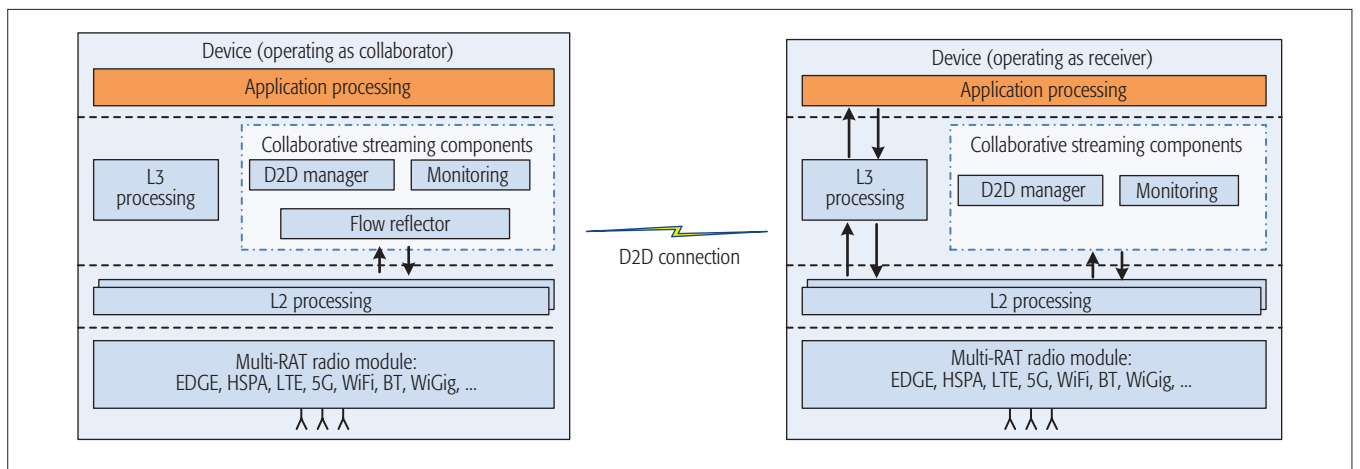


Figure 3. Architecture of client devices for collaborative streaming.

ents, which limits the number of potential collaborators. This drawback can be overcome by proper cooperation and collaboration of different network operators (cellular and wireless), intended to form one ecosystem to cover a smart city area.

Figure 3 shows the high-level architecture of the 5G user device (each might be capable to simultaneously use multiple RATs) for collaborative streaming, which can play two roles: a collaborator or a media receiver. The *D2D manager* establishes and terminates D2D links through the RAT indicated by a BBU and manages D2D connections in all the lower layers. For example, at the medium access control (MAC) layer, the *D2D manager* sends and receives measurement packets, in the appropriate RAT, for monitoring direct link conditions. The *monitoring* module receives this information and estimates the quality of service (QoS) metrics for a given direct link, such as estimated bandwidth and packet losses. Measured QoS metrics, as well as contextual information (mainly available wireless interfaces and state of the battery), are transmitted from the *monitoring* module to the associated BBU.

A collaborator contains a *flow reflector* to play the role of a proxy in communication between the receiver and associated BBU. The module parses the request generated by the receiver, replaces its own address with the BBU's address in the request, and then transfers it to the BBU. Next, it redirects the BBU's response, which carries the requested content, to the receiver using D2D connection.

### RESOURCE CONTROL

Resource control functionality is crucial for the performance of collaborative media streaming due to the high requirements of multimedia content. This section presents the framework between BBU and terminals in order to support efficient collaboration between terminals. Collaborative streaming is a different scenario comparing to conventional D2D approach, where two devices (the source and the destination) control the end-to-end communication. On the contrary, collaborative streaming is managed by an entity different to the devices (i.e., the BBU), and the content source is usually in the Internet. Consequently, all involved devices (i.e., the collaborators and the receiver) have to set up a connection with

their base stations. In the in-band approach the resource control algorithm takes care to allocate separate resources (frequency, time slots) to D2D links and conventional cellular links. If the out-of-band D2D solution is adopted, the resource control algorithm should prefer in the collaborator selection process those terminals that are able to establish direct links by RATs other than the RAT used in the primary downlink (from the base station) to avoid co-channel interference.

The sequence of operation is as follows. First, the BBU selects a set of possible collaborators (after receiving information on the localization of terminals) and sends information about them to the target end user's terminal (receiver). Moreover, the BBU establishes D2D connections with indicated collaborators and sets parameters for the D2D link such as maximum allowed signal strength and parameters of D2D radio in case of licensed spectrum (frame number, system bandwidth, synchronization information, etc.). The 5G radio resource control (RRC) signaling protocol is used for this communication, and each active collaborator is in an RRC-connected state with bidirectional communication with the BBU. The collaborating terminal measures direct link parameters (estimated bandwidth and wireless interface usage) and sends them to the BBU. The collaborating terminal sends the BBU the D2D connection measurements together with the following device information: RATs supported by the device, CPU usage, available energy, estimated bandwidth available in the downlink (from base station to terminal), last locations of the terminal for predicting the terminal trajectory, and terminal activity (i.e., if the terminal is involved in other collaborative streaming sessions). At last, the BBU processes all the information and performs selection of terminals in collaboration with the scope of the given media session. In advanced RAN architectures [9], RRA might be supported by information provided by external actors through open interfaces provided by mobile edge computing technology in order to select the best collaborators [13].

The optimization objectives of the terminal selection algorithm can vary according to the service provider's policies and/or user requirements and expectations. Example policies are: to maximize overall media throughput, to maximize throughput while minimizing power con-

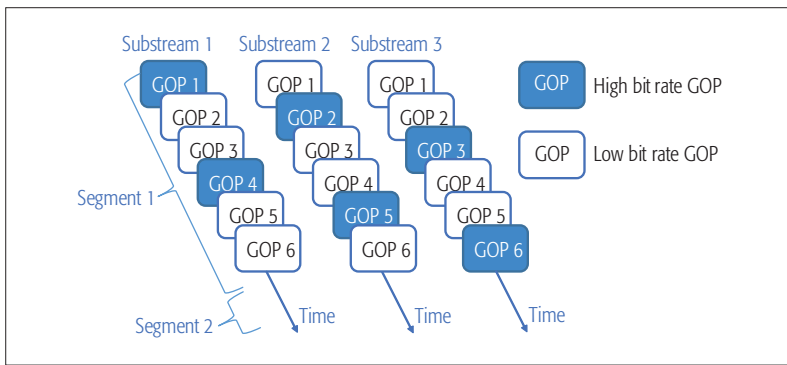


Figure 4. MD-DASH scheme: substream creation strategy.

sumption (in this case, the algorithm should avoid power-consuming D2D connections, even if they provide higher bandwidth), and to maximize overall user quality of experience, QoE (e.g., one should avoid too frequent variation of media streaming bit rate in adaptive streaming).

The resource control algorithm implemented by the *collaborator allocation module* is a complex optimization task performed by the BBU pool. We use a two-phase evolutionary multi-objective optimization algorithm that we previously proposed in [14]. This algorithm enables the finding of several solutions belonging to the Pareto frontier (several collaborators) instead of one unique solution (other multi-criteria decision algorithms aim to find one unique solution).

## CONTENT ENCODING FOR COLLABORATIVE MEDIA STREAMING

This section presents the solution adopted for multipath streaming, which is necessary in a collaborative scenario. The multipath streaming protocol is installed in the server and the receiver at the application level over HTTP and runs independently of the other modules of collaborative streaming (only the *collaborator allocation* informs the proxy server about the state of the paths). The protocol is transparent to the collaborators.

The proposed solution is compliant with adaptive streaming. In fact, proposed Multiple Description – Dynamic Adaptive Streaming over HTTP (MD-DASH) is an extension of DASH and fully compliant with the standard. The main advantage of MD-DASH in comparison to the other multi-path streaming protocols is that it retains an adaptive feature, which makes implementation in current networks feasible. Moreover, MD-DASH is easily implementable in current applications (since only the adaptation protocol library should be extended) and does not require excessive overhead.

DASH defines the video content as being divided into a subset of chunks (called segments) and encoded at multiple bit rates (all the content segments encoded with the same nominal bit rate belong to the same representation). The client application requests each segment independently (new HTTP request), and for each segment, it selects the best representation that may be streamed in current network conditions by analyzing the download of the last segments.

The idea of MD-DASH is to interleave groups of pictures (GOPs) that belong to higher and lower representations, creating a substream, with

mean bit rate between lower and higher representative bit rates. Each substream is fully decodable and viewable (so synchronization of substreams is not mandatory); however, the playing out of one unique substream offers suboptimal QoE to the viewer since some parts of the content will be played at higher quality than others. The playout of all different substreams together offers optimal QoE due to the capacity of the MD-DASH client to select from each substream the best-quality GOPs, discarding the lower-quality ones.

The GOPs are encoding operation units used in current (H.264) and upcoming (H.265) codecs. One GOP contains a video of the same scene, such that the encoding operations inside one GOP are highly correlated, which improves the possibility-to-compression ratio. Figure 4 shows MD-DASH triple substream creation based on two different representations, each substream transmitted through a different path within the network.

The high bit rate GOPs contain high-definition information of the video, whereas low bit rate GOPs contain the minimum video information necessary for playing the video at lower quality. Therefore, the overhead introduced by MD-DASH is low (and equals the relation between low and high bit rate GOPs, as indicated in [15]). The adaptation algorithm in the client application decides about the representation to be downloaded for each segment, and this request is sent to the BBU.

Substream creation is performed on the servers of the BBU cloud when the *collaborator allocation* reports the number of available collaborators (say  $N$ ) and the available bandwidth of each D2D link, in such a way that the bit rate of the best quality GOPs selected from the substreams corresponds to the representation bit rate indicated by the client application (at the receiver) included in the HTTP request message. In the case that all  $N$  paths have similar downlink capacity, the  $N$  substreams will contain a similar number of high bit rate GOPs. However, in the case in which the  $N$  paths have different bottleneck bit rates, the substream for each path will be adapted to the path bottleneck. Substream creation is a lightweight operation that may be performed in real time by a medium class server.

We evaluated our MD-DASH implementation by considering the perceived quality at the consumer's side (i.e., QoE). To this end, we monitored and evaluated two criteria considered as essential for the QoE of video streaming services: the number of rebuffering events and the quality distribution throughout the streaming session. Collaborative multipath streaming with MD-DASH is compared to two other scenarios: unipath adaptive streaming (no collaboration; the content is downloaded directly from the BBU server; this is the case of current downloading on the Internet) and multipath streaming without adaptation (the content is downloaded through three different paths, but the client application is not able to adapt the content bit rate to the network state). In this last case, the 10-minute Big Buck Bunny movie was encoded in three substreams (containing 33 percent of GOPs at 6 Mb/s and 66 percent of GOPs at 200 kb/s). In the other cases, the movie was encoded at 7 different bit rates in H.264 (200 kb/s, 1 Mb/s, 1.5 Mb/s, 2 Mb/s, 3 Mb/s, 4 Mb/s, and 6 Mb/s), and the client



#1 Mb/s (ms;%)	#2 Mb/s (ms;%)	#3 Mb/s (ms;%)	#4 Mb/s (ms;%)	#5 Mb/s (ms;%)	#6 Mb/s (ms;%)
5.0 (38;0.09)	5.0 (13;0.81)	5.0 (11;1.00)	9.0 (25;0.06)	9.0 (10;0.40)	9.0 (6;1.00)
4.0 (50;0.08)	4.0 (18;0.63)	4.0 (13;1.25)	4.0 (50;0.07)	4.0 (50;0.08)	4.0 (13;1.25)
3.0 (75;0.06)	3.0 (28;0.44)	3.0 (15;1.50)	2.0 (75;0.10)	2.0 (150;0.03)	2.0 (20;1.50)
2.0 (88;0.09)	2.0 (58;0.21)	2.0 (20;1.75)	1.0 (100;0.16)	1.0 (200;0.07)	1.0 (25;2.00)
1.5 (100;0.12)	1.5 (200;0.03)	1.5 (25;2.00)	2.0 (75;0.10)	2.0 (150;0.03)	2.0 (20;1.50)
2.0 (88;0.09)	2.0 (58;0.21)	2.0 (20;1.75)	4.0 (50;0.07)	4.0 (50;0.08)	4.0 (13;1.25)
3.0 (75;0.06)	3.0 (28;0.44)	3.0 (15;1.50)			
4.0 (50;0.08)	4.0 (18;0.63)	4.0 (13;1.25)			

Table 1. Network profiles (from DASH standard).

adaptation algorithm selected the best acceptable representation for each segment download. In multipath adaptive (i.e., MD-DASH), the lower bit rate GOPs were 200 kb/s-encoded in order to minimize overhead. The MD-DASH substreams were created based on the information about the state of the three paths between the server and the end user's terminal (i.e., the mean encoding bit rate of the substream to be downloaded was equal to the bandwidth in the path). Such information was provided by the *collaborator allocation* module. The client implementation of unipath adaptive, multipath non-adaptive, and MD-DASH streaming clients were based on dash.js player.

The DASH Industry Forum provides benchmarks for various aspects of the DASH standard. The benchmarks include 12 different network profiles (NPs) featuring different bandwidths, delays, and packet loss. We used only six of them (marked as #Number in Table 1). Each profile spends 30 s for each step described in Table 1, then starts back at the beginning. For each experiment, a specific NP is associated with all paths between the client and the BBU server. A random time offset is set to each assigned NP to represent bandwidth diversity and variability in the network. Each video was repeated 40 times per application and per NP, and a total playback time of 120 h was performed.

Unipath adaptive streaming delivered content with lower quality than multipath streaming due to single-path bandwidth limitations; nonetheless, it reached no rebuffering situations during the 10 min experiments thanks to the bit rate adaptation feature. Instead, multipath non-adaptive streaming suffered constant rebuffering, making the viewing of the content very unpleasant. In fact, multipath the non-adaptive client observed between 0.12 and 1.22 rebuffering situations per streaming session on NPs #1-#2-#3 and between 2.76 and 4.08 on NPs #4-#5-#6 (Fig. 5b).

MD-DASH performed a trade-off between content quality and network availability in multiple paths. Besides, the adaptation mechanisms in MD-DASH allowed the client to take benefit of download bandwidth and eventually provided the 6 Mb/s top quality for 84 percent of the time on average on all NPs (Fig. 5a). In terms of quality distribution over the streaming session, MD-DASH performed significantly better than unipath thanks to the simultaneous usage of collaborators and even better than multipath non-adaptive streaming since MD-DASH adapted the bit rate to each path condition. Moreover, MD-DASH avoided buffer depletion on all NPs.

These results demonstrate the asset of mul-

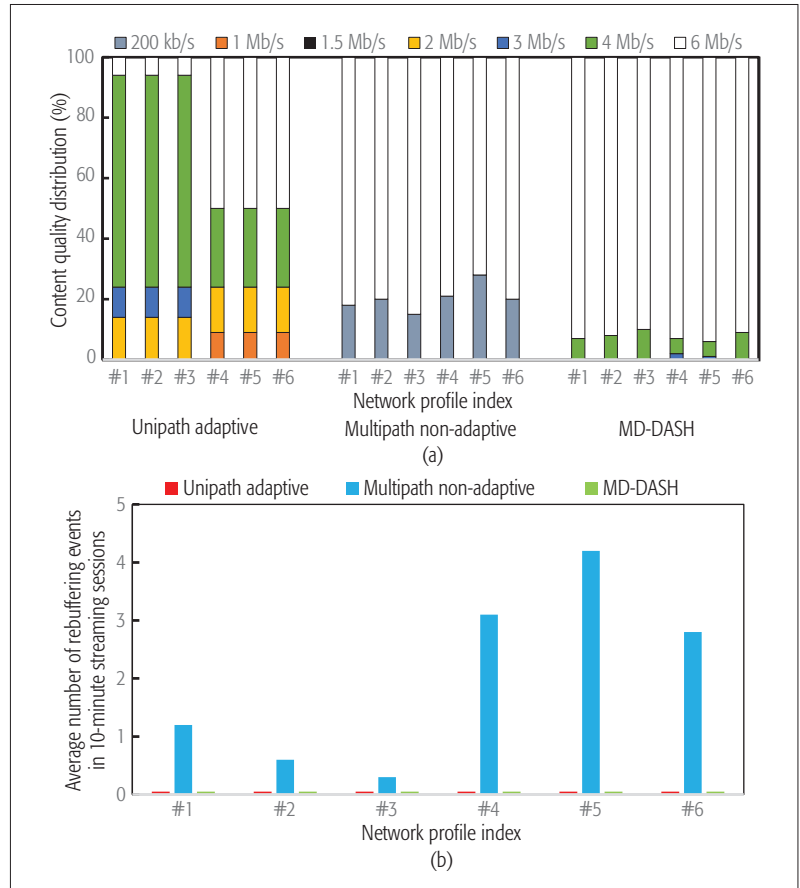


Figure 5. Quality of experience results: a) quality representation throughout streaming session; b) averaged number of rebuffering events.

tipath collaborative media and adaptive streaming in terms of QoE in contrast to current media downloading on the Internet.

## CONCLUSIONS

This article presents a solution for growing media delivery throughput in dense heterogeneous wireless networks. It is based on the collaboration of terminals in proximity for providing multipath delivery of media segments. Multipath transmission is well suited to increase efficiency in future 5G systems, which respond to smart city requirements for communication infrastructure. We describe new modules designated for user devices as well as for centralized RAN to allow collaboration. At the client side, such modules are responsible for establishing and managing D2D links. They also implement mechanisms of relay-

With such a network-aware approach, content characteristics can be fit to current radio conditions. The results obtained in the implementation of the system show that multipath adaptive streaming clearly overcomes both unipath adaptive and multipath non-adaptive streaming.

ing data between the media receiver and the network. On the other hand, modules introduced to the BBU are responsible for selecting appropriate collaborators and optimal resource allocation.

Multipath delivery in C-RAN enables the use of novel MD-DASH streaming technology, which assumes generating a number of substreams transferred to the receiver through different paths. The more substreams reach the user's terminal, the higher QoE will be perceived by the user during media consumption. In the presented system, the process of substream creation is performed using computing capabilities provided by the C-RAN, in accordance with the instructions from the BBU collaborator allocation. With such a network-aware approach, content characteristics can be fit to current radio conditions. The results obtained in the implementation of the system show that multipath adaptive streaming clearly overcomes both unipath adaptive and multipath non-adaptive streaming.

#### ACKNOWLEDGMENTS

The work presented in this article is co-funded by the European Union, Celtic-Plus Programme, under the project C2015/4-5, MONALIS: monitoring and control of QoE in large-scale media distribution architectures.

#### REFERENCES

- [1] R. Yu *et al.*, "Enhancing Software-Defined RAN with Collaborative Caching and Scalable Video Coding," *Proc. IEEE ICC*, Kuala Lumpur, Malaysia, 2016, pp. 1–6.
- [2] N. Golrezaei *et al.*, "FemtoCaching and Device-to-Device Collaboration: A New Architecture for Wireless Video Distribution," *IEEE Commun. Mag.*, vol. 51, no. 4, 2013, pp. 142–49.
- [3] M. Zhong *et al.*, "ColStream: Collaborative Streaming of On-Demand Videos for Mobile Devices," *Proc. IEEE 15th Int'l. Symp. on a World of Wireless, Mobile and Multimedia Networks*, Sydney, Australia, 2014, pp. 1–7.
- [4] E. Toledano *et al.*, "CoCam: A Collaborative Content Sharing Framework Based on Opportunistic P2P Networking," *Proc. IEEE 10th Consumer Commun. Networking Conf.*, Las Vegas, NV, 2013, pp. 158–63.
- [5] Y. Zhang, L. Chao, and S. Lifeng, "DECOMOD: Collaborative DASH with Download Enhancing Based on Multiple Mobile Devices Cooperation," *Proc. 5th ACM Multimedia Sys. Conf.*, Singapore, 2014, pp. 160–63.
- [6] L. Zhang *et al.*, "Green and Cooperative DASH in Wireless D2D Networks," *Wireless Personal Commun. J.*, vol. 84, no. 3, 2015, pp. 1797–1816.
- [7] D. Syrivelis *et al.*, "Bits and Coins: Supporting Collaborative Consumption of Mobile Internet," *Proc. IEEE INFOCOM*, Kowloon, China, 2015, pp. 2146–54.
- [8] A. Checko *et al.*, "Cloud RAN for Mobile Networks — A Technology Overview," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 1, 2015, pp. 405–26.
- [9] M. Patel *et al.*, "Mobile-Edge Computing Introductory Technical White Paper," ETSI, 2014; [https://portal.etsi.org/Portals/0/TBpages/MEC/Docs/Mobile-edge\\_Computing\\_-\\_Introductory\\_Technical\\_White\\_Paper\\_V1%2018-09-14.pdf](https://portal.etsi.org/Portals/0/TBpages/MEC/Docs/Mobile-edge_Computing_-_Introductory_Technical_White_Paper_V1%2018-09-14.pdf) (accessed September 2016)
- [10] M. Sheng *et al.*, "Video Delivery in Heterogenous CRANs: Architectures and Strategies," *IEEE Wireless Commun.*, vol. 22, no. 3, 2015, pp. 14–21.
- [11] A. Hakkarainen *et al.*, "High-Efficiency Device Localization in 5G Ultra-Dense Networks: Prospects and Enabling Technologies," *Proc. IEEE 82nd VTC-Fall*, Boston, MA, 2015, pp. 1–5.
- [12] X. Wang *et al.*, "Cache in the Air: Exploiting Content Caching and Delivery Techniques for 5G Systems," *IEEE Commun. Mag.*, vol. 52, no. 2, 2014, pp. 131–39.
- [13] J. O. Fajardo, I. Taboada, and F. Liberal, "Improving Content Delivery Efficiency through Multi-Layer Mobile Edge Adaptation," *IEEE Network*, vol. 29, no. 6, 2015, pp. 40–46.
- [14] J. M. Batalla *et al.*, "Evolutionary Multiobjective Optimization Algorithm for Multimedia Delivery in Critical Applications through Content Aware Networks," *J. Supercomputing*, 2016, pp. 1–24.

- [15] J. Bruneau-Queyrex *et al.*, "Multiple Description-DASH: Pragmatic Video Streaming Maximizing End-Users' Quality of Experience," *IEEE ICC*, Kuala Lumpur, Malaysia, 2016, pp. 1–7.

#### BIOGRAPHIES

JORDI MONGAY BATALLA (J.Mongay@itl.waw.pl) is the head of the Internet Technologies and Applications Department at the National Institute of Telecommunications, Poland. He is also with Warsaw University of Technology (WUT), where he is an assistant professor and provides research on Internet protocols and applications, especially multimedia delivery, Internet of Things, and cloud/edge computing. He has written more than 100 research papers in renowned international journals and conferences, and is on the Editorial Boards of several journals.

PIOTR KRAWIEC (P.Krawiec@itl.waw.pl) received a Ph.D. degree in telecommunications from WUT in 2011. He is an assistant professor in the Internet Technologies and Applications Department at the National Institute of Telecommunications and concurrently holds a research position at WUT. His research interests lie in the areas of multimedia delivery, streaming protocols, wireless communications, and the Internet of Things.

CONSTANTINOS X. MAVROMOUSTAKIS [SM] (mavromoustakis.c@unic.ac.cy) is currently a professor in the Department of Computer Science at the University of Nicosia, Cyprus, where he leads the Mobile Systems Lab. He has been a Vice-Chair of the IEEE/RB regional Cyprus section since January 2016, and since May 2009 has served as the Chair of the C16 Computer Society Chapter of the Cyprus IEEE section. He has a dense research work outcome (more than 220 papers) in distributed mobile systems and spatio-temporal scheduling, consisting of numerous refereed publications including several books, and he has served as Track Chair and Co-Chair of various IEEE international conferences.

GEORGE MASTORAKIS (mastorakis@gmail.com) received his B.Eng. degree from the University of Manchester, his M.Sc. from University College London, and his Ph.D. from the University of the Aegean. He is an associate professor at the Technological Educational Institute of Crete and a research associate at the Centre for Technological Research of Crete. His research interests include cognitive radio networks, network traffic analysis, and radio resource management.

NAVEEN CHILAMKURTI (N.Chilamkurti@latrobe.edu.au) is currently acting head of the Department of Computer Science and IT, La Trobe University, Melbourne, Australia. He is also the Inaugural Editor-in-Chief for the *International Journal of Wireless Networks and Broadband Technologies*. He is currently serving as a Technical Editor for *IEEE Wireless Communications* and *IEEE Communications Magazine*. He has published about 190 journal and conference papers. His current research areas include cybersecurity, IDS, authentication, wireless sensor networks, and more.

DANIEL NEGRU (daniel.negru@labri.fr) received his Ph.D. from the University of Versailles Saint Quentin, France, in 2006 in the field of communication networks. In 2007, he became an associate professor at the University of Bordeaux, France, specializing in multimedia and networking. He has participated in more than 10 collaborative research projects at the European level and published more than 60 papers, including high-level ones in prestigious journals and conferences, such as *IEEE Communications Magazine*, *IEEE Multimedia*, *GLOBECOM*, *ICC*, and *FIA*.

JOACHIM BRUNEAU-QUEYREIX (jbruneauqueyrex@viotech.net) received his M.Sc. in telecommunications at ENSEIRB-MATMECA graduate school of engineering, Bordeaux, in 2014. Since then, he has pursued his Ph.D. at the University of Bordeaux in the field of multi-criteria optimization for content delivery within the future media Internet. His research interests include video codecs, streaming protocols, systems and architectures, multimedia streaming prototyping, and network functions virtualization. He has joined three European research projects in the field of ICT (FP7, CHIST-ERA, and EUREKA).

EUGEN BORCOCI (Eugen.Borcoci@elcom.pub.ro) is a professor in the Department of Electronics, Telecommunications, and Information Technology of University Politehnica of Bucharest, Romania. His expertise comprises theoretical and/or experimental knowledge in telecommunication/computer systems and networks: architectures and protocols, multicast and multimedia services, and network and services management. His recent areas of interest are future Internet architectures, networked media and QoS/QoE, content-aware networking, software defined networking and network functions virtualization, and 5G technologies. He has participated as team leader and researcher in many FP5, FP6, FP7, and Chist-Era European projects.

# Named-Data-Networking-Based ITS for Smart Cities

Safdar Hussain Bouk, Syed Hassan Ahmed, Dongkyun Kim, and Houbing Song

## ABSTRACT

A smart city enhances the quality of its citizens' lives by providing ease of access to ubiquitous services through integration using communication systems at the foundation. Additionally, ITS plays a major role in making a metropolitan area into a smart city. The current IP-based solutions for ITS have slanted the performance due to high demand for data on the move, especially when the consumers become the producers. Meanwhile, NDN has evolved as a promising future Internet architecture and is being investigated extensively. In this article, we discuss the core functionality of NDN followed by our new architecture proposed for ITS in smart cities. Also, we highlight the current and future research challenges for NDN-enabled ITS in the context of smart cities.

## INTRODUCTION

Information and communications technologies (ICT) started with the simple concept of communications and became a necessity and part of our everyday lives. ICT has a vital role in enabling ubiquitous connectivity to users with services as well as the things around them. These services include health, transportation, emergency response, shopping, utilities, economy, weather, and so on, and are referred as smart services in this article. Information related to smart services is ubiquitously made available to citizens through varying underlying technologies that make these citizens smart to proactively deal with any forthcoming situations. In addition, inter-service information exchange is requisite to make these services smarter to shape a smart city (Fig. 1).

In the context of this article, we define a *smart city* as a *collection of entities (living and non-living) in an urban area that is always connected, fully aware, auto-managed, self-secure, adaptive, and well-informed*. Likewise, smart services are the enablers of a smart city, and ICT provides the platform to achieve those services. In short, the ICT infrastructure is a backbone for any city to be transformed into a smart city. Consider the example of the intelligent transportation system (ITS) or smart mobility, which is one of the vital pillars of citizens' quality of life. ITS in a smart city consists of private, government, and public transport related applications. A smart city must keep track of all transportation on the move to efficiently manage mobility and avoid traffic con-

gestion as well. To achieve this, all vehicles, either on the road, in aerospace, or maritime, are connected through various communication technologies, including third/fourth generation (3G/4G), LTE-Advanced (LTE-A), LTE-Unlicensed (LTE-U), DSRC/WAVE, and wireless LAN (WLAN) technologies (802.11TGax, 802.11AX, 802.11p, etc.) [1].

At present, vehicles are capable of sharing useful critical information including their current location, direction, passengers or goods they are carrying, speed, and so on, with their neighboring vehicles on the road through vehicle to vehicle (V2V) communication, and also pedestrians in their proximity via vehicle-to-pedestrians (V2P) communication, to avoid any possible hazards and collisions. In addition, this information can also be retrieved by nearby traffic or government center(s) through vehicle-to-infrastructure (V2I) communication to get a city-wide or regional mobility view. Furthermore, this information is also made available to individuals to plan trips, companies to track their goods, and the government to regulate traffic in the city. In short, smart transportation is a key strength of all the services expected in smart cities. Those services are, but are not limited to, a centralized fleet management system, real-time traveler information, a smart mass transportation system (SMT), a citywide transportation system, variable speed limit, smart parking, and smart electric vehicle charging. Nevertheless, to enable the aforementioned services, the ICT empowers the electronic devices with connectivity capabilities [2].

On the whole, ITS requires a plethora of information shared by every connected device(s) to provide all the preceding services to the consumers. However, the consumers are interested in small chunks of information, regardless of the location and identity of the provider(s). Notably, most of the communications take place on the move, where consumers are maneuvering while retrieving the required data. Unfortunately, the current ITS relies on the current IP-based Internet architecture for communication support among all devices, which fails to efficiently disseminate contents in the mobile environment. Additionally, it poses other issues such as inefficient IP assignment to mobile devices, intermittent connectivity, IP-dependent data, inappropriate interface selection, scalability of services, incompetent routing in disruptive networks, and so on.

These challenges argue the need for new communication architecture that inherently overcomes

The authors discuss the core functionality of the NDN followed by their new architecture proposed for ITS in smart cities. They also highlight the current and future research challenges for NDN-enabled ITS in the context of smart cities.

These smart services include city-wide traffic management and monitoring, smart parking assistance, public transportation, information services (e.g., bus, train, taxi, plane), logistics, real-time traffic, and road speed limit monitoring and management, among others.

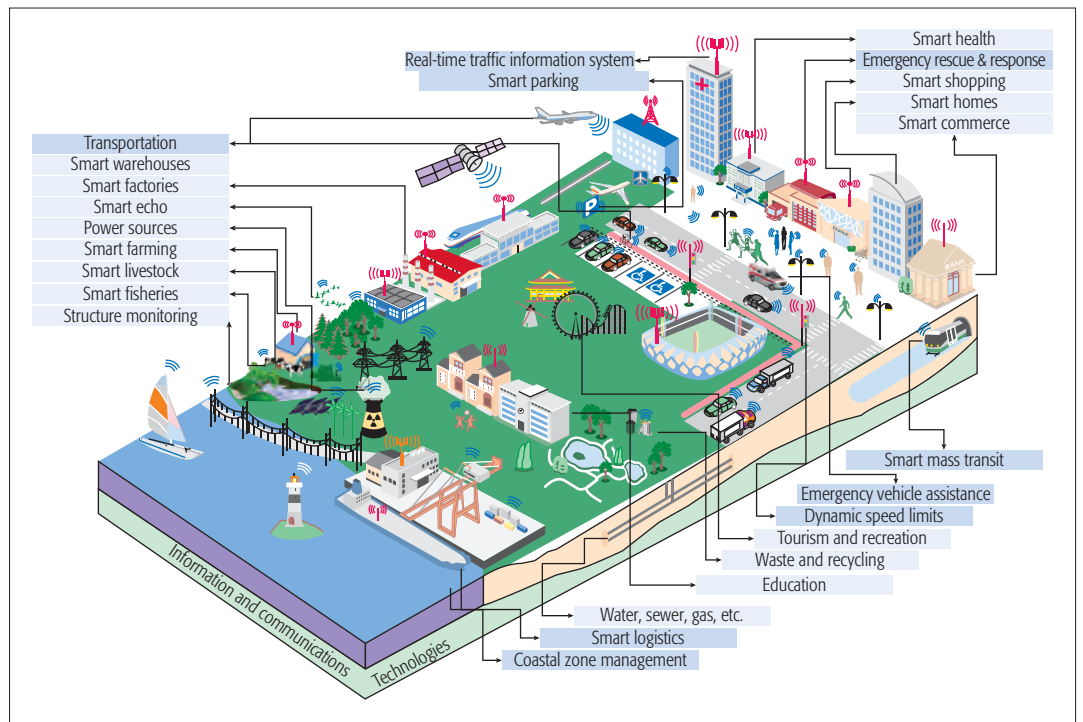


Figure 1. Services offered by smart cities.

all the shortcomings and provides a paradigm shift from IP/host-centric to information-centric communications. In this regard, several architectures have been proposed under the umbrella of information-centric networking (ICN). A few examples of those architectures are data oriented networking architecture (DONA), network of information (NetInf), content-centric networking (CCN), and named data networking (NDN) [3]. NDN as an extension of CCN, relinquishes the location dependency of the data, provides inherent data security, and supports network mobility, scalability, and intermittent connectivity. In NDN, every piece of data is identified by a unique ID (i.e., name), contrary to IP-based networks, where the location of provider and medium of communication matters.

In short, the application of NDN in ITS can be a potential solution to make robust, secure, scalable, and reliable communications between connected mobile devices. Along with this objective, the key contributions of this article are as follows. First, we provide an overview of the ITS areas of potential in smart cities. Second, we propose a detailed future Internet architecture for ITS and discuss the intended functionality of each component of our proposed architecture. Furthermore, we identify the research challenges that require adequate attention from the research community to realize the NDN-based ITS for smart cities.

### THE ROLE OF ITS IN SMART CITIES

In urban areas, transportation is the nervous system, since it is used by the public to reach their workplaces and transport goods. In the past two decades, various applications are supported by the ITS to enhance mobility, reduce carbon emissions, achieve fuel efficiency, improve safety, and save traveling time. Additionally, there are different ITS applications that require connectivity

through multiple interfaces. These smart services include city-wide traffic management and monitoring, smart parking assistance, public transportation information services (e.g., bus, train, taxi, plane), logistics, real-time traffic, and road speed limit monitoring and management, among others. A few of the known ITS use cases are enlightened as follows:

#### Real-Time Traffic Information System (RTIS):

It provides information to travelers about the current transportation status in the city as well as the public transportation schedule. The same information is also provided to the drivers onboard so that they should be aware of any possible congestion and traffic conditions ahead on the desired route. This information is continuously collected and disseminated by the different sensors deployed on the routes and the vehicles. Further, it is also used by government transportation departments to avoid city-wide congestion by dynamically adjusting the traffic signals [4].

#### Emergency Vehicle Assistance (EVA):

This system includes the information dissemination by the emergency vehicle within directional proximity to indicate its upcoming route to ensure prioritized access on the road. Additionally, the traffic signals are also regulated to guarantee the arrival of emergency response vehicles in time at the desired location(s).

#### Smart and Priority Parking (SPP):

The main objective of this application is to rapidly find a parking space to achieve fuel economy. It involves many sensing devices in parking areas and streets to collect parking space status and share that information instantaneously with approaching vehicles.

#### Dynamic Speed Limits (DSL):

Most of the highways and roads in a city have electronic sign boards that warn drivers about the traffic status and speed limits. The onboard units (OBUs) in

vehicles also tell drivers about the current road speed limits. Based on the current traffic conditions, the speed limits can be varied by traffic engineers and displayed on digital traffic signs and OBUs to avoid the congestion in peak hours or caused by incidents.

**Smart Mass Transit (SMT):** It is a transportation system that is customized based on the passenger as well as the goods' requirements and needs. SMT is designed with the objectives of minimum cost, fast and safe transit, minimized traffic overhead, and ease of access to all walks of life. The system is mostly envisioned as a rail or bus system; however, it can vary from onshore, offshore, to airspace options within the city and between states in the future.

**Smart Logistics (SL):** It involves transport vehicles that interact with the market or customers and the goods that they are moving, to transfer goods cost effectively, on time, on demand, and with high customer satisfaction. Vehicles involved in smart logistics may range from small drones to mass transit vehicles in order to reduce traffic volume and logistics cost.

**Smart Electric Vehicle Charging (SEVC):** Future transportation is projected to have zero carbon emission and be environment-friendly, and the first step toward this goal is electric vehicles (EVs). SEVC lets EVs charge batteries smartly and wirelessly. The smart option is that charging can be carried out in low demand electricity hours to reduce the charging cost. The charging options can be available on the signals, car parking, resting areas on highways, and so on.

Regardless of the applications' motivation (i.e., safety or non-safety), the main purpose of connecting vehicles is to share data to fulfill the applications' requirements. However, due to mobility, which is an intrinsic feature of vehicular networks, it is difficult to communicate data reliably and efficiently using the existing communication standards directly in ITS. The main reason is that the current standards were originally proposed for static and quasi-static environments. Henceforth, an ITS specific communication standard has been proposed, called Wireless Access in the Vehicular Environment (WAVE), which works over dedicated short-range communication (DSRC). WAVE/DSRC collectively enables communication between vehicles, roadside units (RSUs), and pedestrians. Although these standards tend to support mobility and fast data delivery in ITS networks, the applications still require a destination address to communicate data. Hence, the communication is contingent on the vehicle's identity (IP and/or MAC address). Data delivery to the farthest vehicles in a network also requires identities of intermediate nodes to establish the path. Path establishment, maintenance, and identity assignment in dynamic topology-based vehicular networks is challenging and requires much overhead [5].

From the application's standpoint, it requires data irrespective of the identity and location of the actual provider or producer. The vehicles require data regardless of the underlying communication technologies. Additionally, guaranteed and secure connectivity in the intermittent vehicular network is quite difficult. Therefore, the NDN is the most suitable option to cater to the information needs of the ITS in the smart cities.

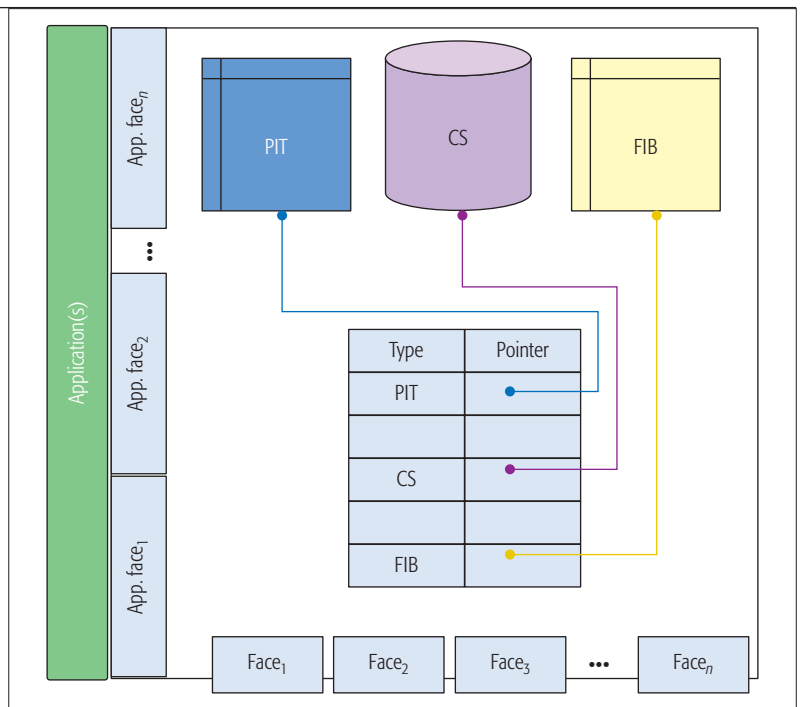


Figure 2. Named data networking core components.

## NAMED DATA NETWORKING IN SMART CITIES

### NAMED DATA NETWORK

The NDN architecture separates the application execution from the type and nature of the lower-layer technologies, which makes it a suitable candidate for a clean slate or overlay architecture [6]. Contents are uniquely identified and communicated by NDN using the hierarchical name structure where each name component is of arbitrary size and separated by the character “/” — thus resembling the Uniform Resource Identifier (URI). Due to the hierarchical name structure, it is easy to aggregate an identical group of names under a uniform name prefix [7]. Simplified pull-based communication is perpetrated by the NDN using two simple messages, Interest and Data. The consumer initiates an Interest message carrying the content name as well as some additional information, including content publisher key, digital signature, content attributes, and so on, to find and ensure the exact matching content. The provider node(s) having the required content simply communicates the content through a Data message back to the consumer.

Moreover, NDN uses some additional data structures to accomplish the multihop Interest-Data forwarding functionality, which includes the pending interest table (PIT), forwarding information base (FIB), and content store (CS), as shown in Fig. 2. A PIT keeps a record of interests (an Interest's receiving interface or application face [App. Face], name prefix, Interest number/nonce value, etc.) received from the consumer (downstream). The consumer first compares the name prefix in the PIT to check whether or not the received Interest has already been processed. In the case of a new Interest, the node searches the desired content in the CS. The Interest is forwarded further in the network (upstream), and a PIT record is created if the desired content is not available in the CS. The CS is a cache mem-

The proposed NDN enabled ITS architecture shows that ITS aided devices in the smart city can either use the legacy WAVE/DSRC plane, IP-based communication, or NDN plane to find the desired data within the plethora of devices connected through ICT in the smart city scenario.

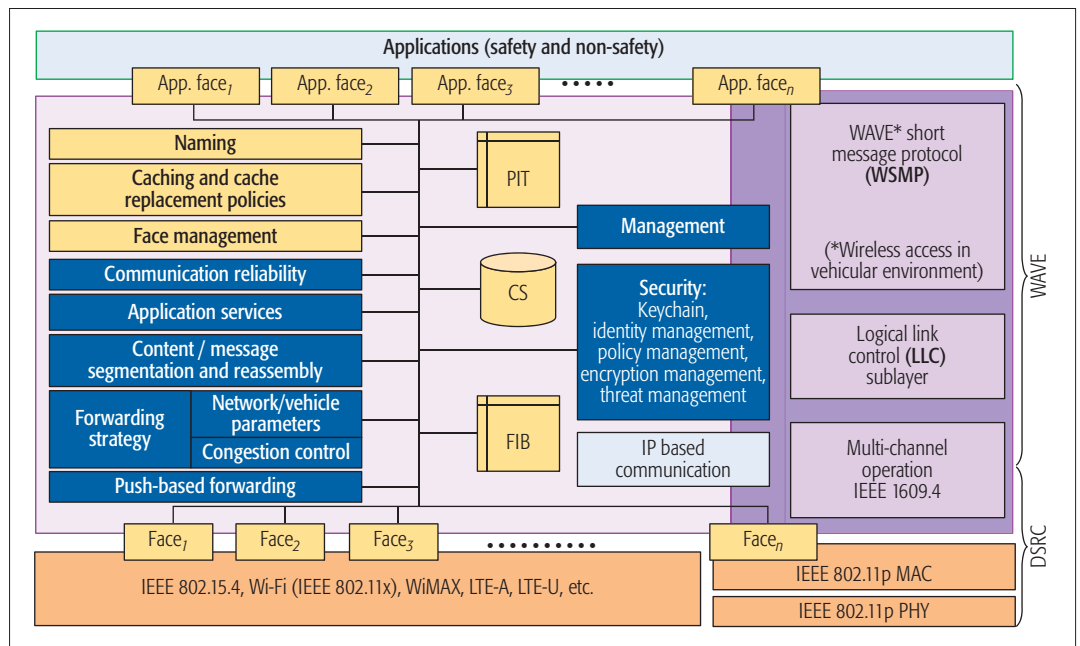


Figure 3. NDN enabled ITS architecture for smart cities.

ory that stores the contents which were either generated or requested by other vehicle(s). If the requested content is not in the CS, the Interest's record is created in the PIT. This Interest's record is held in PIT for a specified period, and once the Data is received or the entry's time expires, the entry is removed from the table [8]. To forward an Interest upstream, the outgoing face information from the FIB is used; it contains the name prefixes and their associated outgoing face(s). When the requested content is found in the CS of a node (producer), it replies to the consumer with a Data message. Any intermediate node that receives the Data message and has a valid PIT record forwards the Data message and may store the Data/content in its CS. Content storage, holding duration, and replacement with new content are contingent on the content caching or replacement policy in use, such as no caching, least frequently used (LFU), least recently used (LRU), first-in first-out (FIFO), and random caching [9].

### NDN-BASED ITS FOR SMART CITIES

Recently, many researchers have been reconnoitering the viability of NDN in several network scenarios ranging from wired networks to wireless ad hoc, sensors, vehicular networks, and the Internet of Things (IoT). Here, we focus on the applicability of NDN in ITS for smart cities to enrich applications, gratify mobility, and ensure security. ITS applications, whether targeted for safety or non-safety, require data without concern regarding the identity-location of its provider. That data can be spatial, temporal, or spatiotemporal in nature. For instance, an application may require spatiotemporal information such as road condition, weather, parking availability, traffic volume information, litter level/status of dumpsters for garbage trucks, a patient's current/previous vital signs information for an ambulance, the status of crosswalk/pedestrians, and so on, for a specific period and/or location. Here we discuss our proposed architecture for

the NDN-enabled ITS supporting smart services in smart cities (Fig. 3).

The proposed NDN-enabled ITS architecture shows that ITS aided devices in a smart city can use either the legacy WAVE/DSRC plane, IP-based communication, or the NDN plane to find the desired data within the plethora of devices connected through ICT in the smart city scenario. Following is a brief discussion about each component of the proposed architecture.

**Naming:** NDN uses hierarchical naming to route and identify the data or chunk of data in the network. There are a few naming schemes typically proposed for vehicular networks [10] and smart city scenarios [11]. Here we propose the modified naming structure for ITS. The smart city (SC:) is the identifier for the naming scheme for the smart city environment. The service identifier name component signifies the type of services offered by ITS, including RTIS, CTMM, EVA, SPP, and so on. Then the spatiotemporal information is used in a name that indicates the spatial (three-dimensional location), temporal (time-date encoded in UNIX format), or spatiotemporal scope (~) information. In the last, there are different attribute-value pairs (Temperature:75F, entity:dumpster, destination:2km, AirDir:NW, etc.) that are used to precisely recognize the multi-dimensional data.

SC://service-id/spatio\_and\_or\_temporal\_scope.../attribute:value/

This naming scheme can be used to request road traveler, traffic condition, and related information for a commuter.

**Caching and Cache Replacement Policies:** The caching policies define whether the content destined for either the current device or the remote device should be stored or not and, if stored, then how long it should be kept in the CS. The caching feature enables each NDN node to serve as a provider to efficiently disseminate device information in the intermittent and disjoint network architectures.

**Face Management:** It monitors all the communication over faces, and based on the communications statistics (number of satisfied Interests, face down counts, face active duration, etc.), it grades the faces. The NDN forwarding strategies prioritize the higher graded faces to communicate the priority data. The priority of any face depends on the network as well as the face parameters, including number of satisfied Interests, dropped messages, PIT size, vehicle speed, rate of change in the neighborhood, and so on. Forwarding messages using the face statistics may reduce message loss, and provide guaranteed and timely information to the consumers.

**Content/Message Segmentation and Reassembly:** Currently, there are different technologies used to provide wireless communication faces, and each face has different frame or maximum transmission unit (MTU) size ranging from hundreds of bytes to tens of thousands of bytes. NDN-based ITS devices use these diversified faces to communicate contents using Interest and Data messages. This NDN element handles the segmentation and reassembly of the large content, and Data and Interest messages because the size of the content or messages may vary and does not fit the MTU size. Therefore, segmentation is inevitable, and each segment is properly numbered by the provider or the intermediate node to facilitate reassembly at the consumer. Additionally, the segmentation may facilitate the distributed content caching in the network.

**Communication Reliability:** NDN does not establish a connection before initiating the communication. For every Interest, the consumer waits for the data chunk. This element can be used in our proposed architecture to ensure that either the requested chunk or sequence of chunks is received successfully by sending acknowledgment or negative acknowledgment messages.

**Application Services:** This NDN element provides spatiotemporal information and NDN communication services for ITS applications, and may include a vehicle's location, speed, direction, distance, communication synchronization, localization information, vehicle type and priority, incident or emergency level, and so on.

**Forwarding Strategy:** The plain vanilla NDN architecture simply broadcasts the Interest on the wireless face(s) and node(s) with matching content, replying with Data message(s). With a large number of wireless devices, it is not feasible to simply broadcast the messages. Therefore, this element is responsible for reducing the message broadcast and forwarding the Interest/Data messages according to the node/network's parameters (i.e., directional or selective forwarding).

**Push-Based Forwarding:** In plain vanilla NDN, any node that receives a Data message first checks its PIT. If there is no PIT entry for that Data message, it is considered as unsolicited and dropped. To be precise, every producer requires an Interest message from a consumer before it can send the required Data message. Even though, this mechanism hinders the retrieval of unsolicited Data messages and secures an overall network from being overwhelmed, the behavior of vanilla NDN is more "consumer-centric" than "content-centric."

Our proposed architecture will support push-based forwarding of critical data, which can be any emergency situation either sensed by sensors or recorded by vehicles. The main objective of this element is to differentiate between the solicited and unsolicited critical nature of the Data and forward it accordingly.

**Management:** This element is common for the legacy WAVE/DSRC communication standard as well as the NDN stack. The main objective of this element in NDN is to manage the NDN data structures (FIB, PIT, and CS) to occupy less memory space and perform faster search, add, and delete operations. The PIT, FIB, and CS are the most frequently used data structures in the NDN, and their size may vary instantly in the ITS scenario. Therefore, efficient management of the data structure is required to achieve robust communication of Interest and Data messages.

**Security:** NDN is a very simple architecture that has been proposed to support content security instead of connection security. This is only possible when all the communicated contents are signed. Hence, most of the sub-elements of the security element of ITS architecture should provide that functionality. The security element contains a keychain that stores public and private keys and certificates. The information in the keychain sub-element is used to sign packets, identify the content provider or application that provides the signed content, and encrypt or decrypt the messages exchanged by NDN. Therefore, the identity management sub-element should be there to facilitate identities, keys, and certificate management along with packet signing. To verify the identity of the content provider, there should be a concerted signing and verification policy. In case of diverse signing policies being in use, the policy management sub-element is responsible for handling the task. To provide data security, the encryption management sub-element offers the data encryption, decryption, and symmetric key management functionalities.

## FUTURE ASPECTS OF NDN IN SMART CITIES

In this section, we provide a broader picture of challenges to be faced when applying NDN into smart cities in general as well as in ITS.

### CONTENT NAMING

Naming is pivotal in NDN, since the content discovery and forwarding solely depend on the namespace. NDN supports the hierarchical naming; however, there are flat and attribute-based naming schemes as well. The main advantage of hierarchical naming is its aggregation and extendability. However, it fails to ensure the advantages of the content search flexibility and linkage between the public key and real-world entity offered by attribute-based and flat naming schemes [12]. Another important aspect that is ignored in the hierarchical naming scheme is the dimensionality/modality of the requested data (e.g., the data units). For example, the temperature, speed, pressure, location, and other information requested in an Interest from the ITS devices should be clearly specified in the name. The requested information is from that available in the cache; for example, location may be in terms of GPS coordinates/block, street/road

To verify the identity of the content provider, there should be a concerted signing and verification policy. In case of diverse signing policies being in use, the policy management sub-element is responsible for handling the task. To provide data security, the encryption management sub-element offers data encryption, decryption, and symmetric key management functionality.

The amount of content in a smart city scenario can be enormous, which can make it challenging to achieve CS space and management efficiency. Therefore, content caching and replacement policies should be optimized for ITS assisted nodes to avoid CS overhead.

name; speed can be represented in meters per second, kilometers per hour, miles per hour; temperature in or Celsius/Kelvin/Fahrenheit; and pressure in bar/pascal/atmosphere. Therefore, the naming scheme should explicitly indicate and follow the common dimensionality of the information.

#### INTEREST/DATA FORWARDING

Every content or its chunk is sent in a Data message as a response to an Interest message by NDN architecture. ITS applications may send/receive Interest/Data messages to and from any device in the smart city. Message forwarding is achieved through PIT and FIB structures and face(s) information [13]. The scalability of Interest/Data traffic may vary, which requires the PIT to be optimized to achieve fast search and management. Selection of the suitable face(s) is also important in the Interest/Data forwarding.

#### CONTENT CACHING

The efficiency of NDN's content delivery mostly depends on in-network content caching. NDN nodes spare some storage for CS. The Data packets received by an NDN node are first checked in CS; if not found, before forwarding the data is stored in CS depending on the caching policy. The amount of content in a smart city scenario can be enormous, which can be challenging to achieve CS space and management efficiency. Therefore, content caching and replacement policies should be optimized for ITS assisted nodes to avoid CS overhead. For example, vehicles are not required to store any information that is out of spatiotemporal scope, for example, information about past incidents, and road conditions where they traveled in the past. Therefore, the caching and replacement policies for ITS applications must consider the spatiotemporal and significance of content in processing.

#### SCALABILITY AND QUALITY OF SERVICE

There may be a surfeit number of heterogeneous devices in the NDN based ITS in a smart city, which may be interested in different or similar contents. Due to the assorted nature of these devices, having diverse processing, memory, bandwidth, and communication capabilities, they may not be able to handle and cache that amount of NDN traffic. Hence, the forwarding, caching, processing, and other NDN related operations may be assigned to the nodes relative to their potential. It must be made sure that the scalability and non-homogenous device nature should not influence the quality of service of ITS applications.

#### PUSH SUPPORT IN NDN FOR SMART CITIES

The Vanilla NDN architecture does not support PUSH-based data retrieval, which is important if any node and/or vehicle has critical data or notifications to disseminate to its neighbors. There are solutions that enable provider(s) to push some sort of short information within the Interest message by replacing the content name. In this case, the data producer generates the Interest, and it has to add some component in the name to differentiate between a name and data in the Interest message. The other way is that a producer sends an Interest to consumers.

#### SECURITY FRAMEWORK

It is worth noting that the NDN ensures content provenance and integrity, and it requires private-public key pairs. The challenging part of NDN security, which it inherits from the legacy trust mechanism, is to find linkage between public key and the content provider entity that NDN assumed to be handled by the public key infrastructure (PKI). However, it is claimed that NDN does not rely on any PKI, while the distributed trust management enables consumers to specify their own trust policies and trust anchors. Additionally, the authentication of long-lived data is also one of the challenges, which include the verification, construction, and reconstruction of the digital certificates. These important issues raise more questions that need to be addressed. For example, who will verify, construct, and reconstruct the certificates? How long should the certificate be valid? Should it be valid throughout the data's lifetime, or should it be reconstructed periodically for long-lived data? Most importantly, are the trust anchors really trustworthy? Do all ITS related heterogeneous devices apply the same trust schema? In short, the more flexible security infrastructure comes with more challenges, and still the security in NDN requires more attention.

#### NDN-ITS DEPLOYMENT ISSUES

NDN can be deployed in parallel with the IP networks and as a clean slate. An NDN deployment as a layer over the IP networks requires some customizations to fully work in smart city use cases or applications because the working principles of the current Internet and the NFD are completely different. For instance, we need to install the NFD on the ITS related devices including RSUs, OBU, servers, parking sensors, and so on. In addition, the absence of sophisticated emulators for NDN-ITS due to its early stage makes it difficult for engineers to test real-time NDN-ITS performance in smart cities.

#### CONCLUSION

We have drawn a broader picture of NDN enabled ITS in future smart cities in this article. The preliminary discussion highlights that the previous ICT solutions are not feasible for the scenario under consideration. Furthermore, the role of ITS has been discussed, and a contemporary detailed architecture of the NDN-enabled ITS has also been proposed. Although NDN research is rapidly growing and approaching maturity, it still requires more attention from research community to pave its foundation and guarantee its effectiveness in smart cities.

#### ACKNOWLEDGMENT

This study was supported by the BK21 Plus project (SW Human Resource Development Program for Supporting Smart Life) funded by the Ministry of Education, School of Computer Science and Engineering, Kyungpook National University, Korea (21A20131600005).

#### REFERENCES

- [1] F. Y. Wang, "Scanning the Issue and Beyond: Transportation and Mobility Transformation for Smart Cities," *IEEE Trans. Intelligent Transportation Systems*, vol. 16, no. 2, Apr. 2015, pp. 525–33.



- [2] S. Djahel *et al.*, "A Communications-Oriented Perspective on Traffic Management Systems for Smart Cities: Challenges and Innovative Approaches," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 1, 1st qtr. 2015, pp. 125–51.
- [3] G. Tyson *et al.*, "A Survey of Mobility in Information-Centric Networks," *Commun. ACM*, vol. 56, no. 12, 2013, pp. 90–98.
- [4] V. Kostakos, T. Ojala, and T. Juntunen, "Traffic in the Smart City: Exploring City-Wide Sensing for Traffic Control Center Augmentation," *IEEE Internet Computing*, vol. 17, no. 6, Dec. 2013, pp. 22–29.
- [5] L. Wang *et al.*, "Data Naming in Vehicle-to-Vehicle Communications," *IEEE INFOCOM Wksp.*, Orlando, FL, 2012, pp. 328–33.
- [6] L. Zhang *et al.*, "Named Data Networking," *ACM SIGCOMM Comp. Commun. Rev.*, vol. 44, no. 3, 2014, pp. 66–73.
- [7] M. F. Bari *et al.*, "A Survey of Naming and Routing in Information-Centric Networks," *IEEE Commun. Mag.*, vol. 50, no. 12, Dec. 2012, pp. 44–53.
- [8] S. H. Bouk *et al.*, "DPEL: Dynamic PIT Entry Lifetime in Vehicular Named Data Networks," *IEEE Commun. Lett.*, vol. 20, no. 2, Feb. 2016, pp. 336–39.
- [9] M/ Zhang, H. Luo, and H. Zhang, "A Survey of Caching Mechanisms in Information-Centric Networking," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 3, 3rd qtr. 2015, pp. 1473–99.
- [10] S. H. Bouk, S. H. Ahmed, and D. Kim, "Hierarchical and Hash Based Naming with Compact Trie Name Management Scheme for Vehicular Content Centric Networks," *Computer Commun.*, vol. 71, 1 Nov. 2015, pp. 73–83.
- [11] G. Piro *et al.*, "Information Centric Services in Smart Cities," *J. Systems Software*, vol. 88, Feb. 2014, pp. 169–88.
- [12] H. Zhang *et al.*, "Uniform Information with a Hybrid Naming (HN) Scheme," IETF tech. rep. 01, 2014.
- [13] S. H. Ahmed, S. H. Bouk, and D. Kim, "RUFs: RobUst Forwarder Selection in Vehicular Content-Centric Networks," *IEEE Commun. Lett.*, vol. 19, no. 9, Sept. 2015, pp. 1616–19.

#### ADDITIONAL READING

- [1] J. Wan *et al.*, "Mobile Crowd Sensing for Traffic Prediction in Internet of Vehicles," *Sensors*, vol. 16, no. 1, 88, 2016.

#### BIOGRAPHIES

SAFDAR HUSSAIN BOUK [SM'16] received his B.E. degree in computer systems from Mehran University of Engineering and Technology, Jamshoro, Pakistan, in 2001, and his M.S. and Ph.D. in engineering from the Department of Information and Computer Science, Keio University, Yokohama, Japan in 2007 and 2010, respectively. He worked as an assistant professor in the Department of Electrical Engineering, COMSATS Institute of Information Technology, Islamabad, Pakistan, from 2010 to 2014. Currently, he works as a research professor at Kyungpook National University, Daegu, Korea. His research

interests include wireless ad hoc, sensor networks, underwater sensor networks, vehicular networks, and information-centric networks.

SYED HASSAN AHMED [S'13] received his B.S. in computer science from Kohat University of Science and Technology, Pakistan. Later on, he joined the School of Computer Science and Engineering, Kyungpook National University, Korea, where he completed a Master's and is pursuing his Ph.D. in computer engineering at the Monet Lab. In 2015, he was a visiting researcher at the Georgia Institute of Technology, Atlanta. Since 2012, he has published over 60 international journal and conference papers in the field of wireless communications. Along with several book chapters, he also authored two short Springer books. For three consecutive years, 2014–2016, he won the Best Research Contributor award at the Workshop on Future Researches of Computer Science and Engineering, KNU, South Korea. In 2016, he also won the Qualcomm Innovation Award at KNU, South Korea. Moreover, he is a recipient of travel grants for prestigious conferences like IEEE CCNC and ACM ICN 2016. He is an active IEEE/ACM member, serving several well reputed conferences/journals as a TPC member and reviewer. His research interests include sensor and ad hoc networks, cyber-physical systems, vehicular communications, and future Internet.

DONGKYUN KIM [M'16] received his B.S. degree from the Department of Computer Engineering, Kyungpook National University, Daegu, Korea. He also received his M.S. and Ph.D. degrees from the School of Computer Science and Engineering, Seoul National University, Korea. He was a visiting researcher at the Georgia Institute of Technology, Atlanta, in 1999. He also performed a postdoctorate program in the Computer Engineering Department, University of California at Santa Cruz, in 2002. He has been on the Organizing Committees or Technical Program Committees of many IEEE and ACM conferences. He received the Best Paper Award from the Korean Federation of Science and Technology Societies in 2002. He has been engaged in many editorial activities in several well reputed international journals. Currently, he is a professor in the School of Computer Science and Engineering, Kyungpook National University. His current research interests include connected cars, vehicular ad hoc networks, the Internet of Things (M2M/D2D), WiFi networks (including WiFi Direct), wireless mesh networks, wireless sensor networks, and future Internet.

HOUBING SONG [M'12, SM'14] received his Ph.D. degree in electrical engineering from the University of Virginia, Charlottesville, in August 2012. In August 2012, he joined the Department of Electrical and Computer Engineering, West Virginia University, Montgomery, where he is currently an assistant professor and the founding director of the Security and Optimization for Networked Globe Laboratory (SONG Lab, www.SONGLab.us). He was the very first recipient of the Golden Bear Scholar Award, the highest faculty research award at West Virginia University Institute of Technology (WVU Tech), in 2016.

# Enabling Communication Technologies for Smart Cities

Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Yasir Mehmood, Abdullah Gani, Salimah Mokhtar, and Sghaier Guizani

The authors discuss the enabling communication and networking technologies used in smart cities. The similarities and differences among different communication technologies based on the important parameters are also analyzed. Moreover, a taxonomy is devised by classifying the literature based on future and emerging technologies, modern communication technologies, IEEE wireless technology standards, objectives, network classes, and mode of operations.

## ABSTRACT

Tremendous advancements in heterogeneous communication technologies have enabled smart cities objects to interact with each other while ensuring network connectivity. However, these communication technologies cannot provide flawless connectivity in smart cities due to the coexistence of thousands of devices, which brings about several problems. In this article, we discuss the enabling communication and networking technologies used in smart cities. The similarities and differences among different communication technologies based on the important parameters are also analyzed. Moreover, a taxonomy is devised by classifying the literature based on future and emerging technologies, modern communication technologies, IEEE wireless technology standards, objectives, network classes, and mode of operations. Furthermore, some reported case studies of different cities (Barcelona, Stratford, Singapore, and Porto) are also presented. Lastly, several research challenges, such as interference management, scalable wireless solutions, interoperability support among heterogeneous wireless networks, mobility management, and high energy consumption that remain to be addressed for enabling unimpaired connectivity in smart cities are discussed as future research directions.

## INTRODUCTION

Over the past few years, communication technology innovations have emerged as a widely recognized trend and are expected to play a crucial role in terms of providing connectivity in smart cities. The term smart city is used to refer to technology-intensive cities that can offer collection, analysis, and distribution of information so as to transform services offered to citizens, increase operational efficiency, and entail better decisions at the municipal level [1]. The vision of the future smart city paradigm is based on the notion of connectivity. Certainly, connectivity plays a major role in smart cities to enable interoperable access and interconnection among heterogeneous smart city objects [2]. Furthermore, in smart cities, telecommunication infrastructures can also provide efficient delivery of services and high-quality information through a large number of digital devices with the involvement of various technologies, such as wireless sensor networks, machine-to-ma-

chine (M2M) communication, vehicle-to-vehicle (V2V) communication, network virtualization, and gateways, to name a few [3, 4].

Recently, the analysts have forecasted that the total number of devices will reach 50 billion by the end of 2020 (<http://www.cisco.com/web/solutions/trends/iot/portfolio.html>). No doubt by that time the desire for automation will be satisfied in most cities. Apparently, at that time providing error-free connectivity will become the main challenge because most of the existing communications technologies can potentially be exposed to interference. Figure 1 illustrates several communication technologies and smart cities applications. Although a number of solutions, based on Bluetooth, ZigBee, WiFi, NFC, Z-Wave, LoRaWAN, and 6LoWPAN, exist, their capabilities in terms of throughput and transmission range are very limited. Moreover, the advanced technologies of the Third Generation Partnership Project (3GPP), such as WiMAX, LTE, and LTE-Advanced (LTE-A), will also not be suitable because of high energy consumption, as most of the connected devices in smart cities have battery constraints [5]. This article aims to qualitatively look into whether the state-of-the-art wireless technologies are able to provide unimpaired connectivity for a huge number of devices in smart cities.

The contributions of the article are numerous. First, we investigate the credible state-of-the-art research efforts directed at the smart city paradigm from communication and networking perspectives. With the aim of classifying literature based on communication and networking aspects of smart cities, the article introduces a taxonomy. We qualitatively evaluate the capabilities of modern communication technologies using important parameters. A discussion of reported case studies is provided. Finally, we discuss future research challenges. These contributions are given in separate sections.

## MOTIVATION

With increasing miniaturization of mobile phones, computers, and sensors, attention in smart cities has increased substantially toward communication technologies with the aim of enabling error-free connectivity. Smart cities depend completely on the network connection, which not only demands high speed, high reliability, and availability, but also features that are required in today's networks.

*Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Abdullah Gani, and Salimah Mokhtar are with Centre for Mobile Cloud Computing Research, Faculty of Computer Science and Information Technology, University of Malaya; Yasir Mehmood is with the University of Bremen; Sghaier Guizani is AlFaisal University. The corresponding authors are Ibrar Yaqoob and Abdullah Gani.*

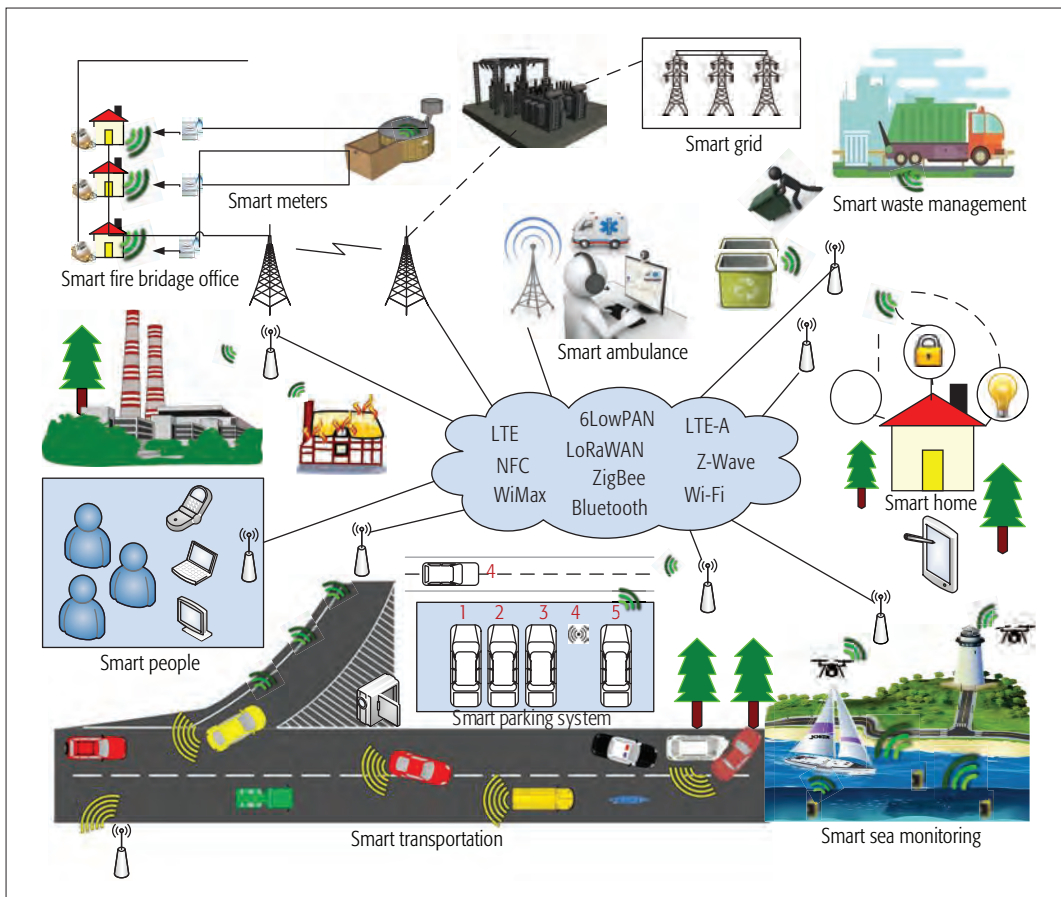


Figure 1. Enabling communication technologies for smart cities.

It has been demonstrated that deployment of small cell technology can help to meet the communication and networking requirements of smart cities applications in terms of interoperability, robustness, limited power consumption and multi-modal access to improved quality of experience.

Moreover, other devices with new requirements need to be connected to each other effectively [6]. The motivation for enabling communication technologies for smart cities can be viewed in [7], which presents how much the demands of networking and communication technologies are increasing to provide a wide variety of connectivity services in smart cities.

Reference [8] stated that 66 percent of the world's population will be urban until 2050. Pike Research has reported that from 2010 to 2020 the investment in smart cities technology infrastructure will reach a total of \$108 billion (<http://www.smartgridnews.com/press-releases/global-investment-smart-city-technology-infrastructure-total-108-billion-20>). Another report indicated that 152 million cars will be connected to the Internet by 2020 (<http://blog.atmel.com/2013/11/19/analysts-see-152-million-connected-cars-by-2020/>). Apparently, in the future, providing flawless connectivity will become a real challenge due to the coexistence of a huge number of devices that have multi-radio capabilities in smart cities. The importance of error-free connectivity can be seen in the smart transportation scenario, where network delay due to impaired connectivity can cause serious accidents by preventing drivers applying brakes instantly. On the other hand, in the precision-based applications of smart cities, error-free connectivity will also be required. The next-generation networks must be connected to LTE, LTE-A, WiMAX, 3G, Bluetooth, ZigBee, Z-Wave, and LoRaWAN to support the various applications of smart cities. Such technologies

provide high transmission rates. However, their capabilities are limited by the number of devices connected at the same time in busy cities [5]. Figure 2 shows that the titles of most published research works in the smart city paradigm are mainly focused on wireless technologies, sensors, and urban networks, to name a few.

### STATE OF THE ART IN SMART CITIES

In this section, we look into recent research efforts made in the smart city paradigm.

A discussion on smart cities applications from the perspective of communication technologies was provided in [9]. In this context, the role of LTE-A, which aims to increase the bandwidth coverage using small cell technology, was explained by the authors. The small cell is a low-power and low-cost radio base station that can provide superior cellular services to enterprises. Moreover, it has also been demonstrated that deployment of small cell technology can help to meet the communication and networking requirements of smart cities applications in terms of interoperability, robustness, limited power consumption, and multi-modal access to improve quality of experience. Although the discussed technology can satisfy many connectivity requirements of smart cities, higher rate of energy consumption is one of the limitations.

A survey of the novel WiFi technology based on IEEE 802.11ah that is currently under development for the smart city has been conducted in [5]. The objective of the survey was to discuss the communication technologies that are used in



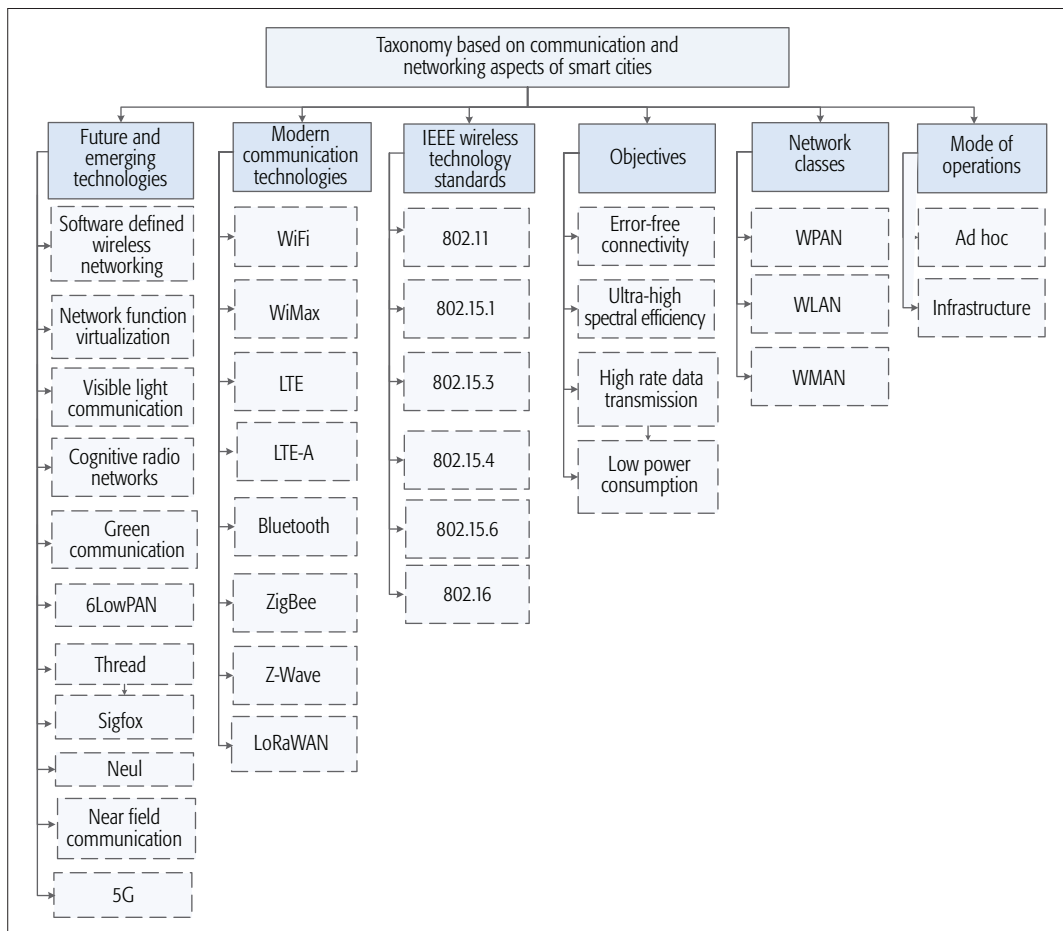


Figure 3. Taxonomy of enabling communication and networking technologies for smart cities.

The aim of future and emerging technologies is to enable high rate data communication, provide better networking infrastructure, low signal attenuation, ensure efficient spectrum utilization, high scalability, high coverage, low cost, robustness, high authentication, and agile encryption mechanisms.

terms of controlling and monitoring. Moreover, the authors also explained that IoT is the future of smart city applications. Despite many advantages of the architecture, such as low cost, more granular security, and centralized network provisioning, code complexity and lack of friendly features are some of the limitations.

In [15] research efforts were carried out to show that near field communication (NFC) technology can play an important role in the development of smart cities. To turn the smart city into reality, it is required to develop a system that supports the applications of NFC. In this context, cloud architecture based on NFC technology was proposed by the authors. The architecture employs a unified interface to receive a message and process it on a server that is based on cloud computing technology. Moreover, a resource scheduling model that aims to satisfy the feature of NFC application was presented. Despite many advantages of NFC technology, such as convenience and versatility, weak security is a major remaining concern that can hinder the deployment of NFC in smart cities.

## TAXONOMY

Figure 3 presents the taxonomy devised by classifying the literature based on the parameters, that is, future and emerging technologies, modern communication technologies, IEEE wireless technology standards, objectives, network classes, and modes of operation.

### FUTURE AND EMERGING TECHNOLOGIES

The emerging networking and communication technologies — software defined wireless networking (SDWN), network functions virtualization (NFV), visible light communication (VLC), cognitive radio networks (CRNs), green communication (GC), 6LowPAN, Thread (IP-based IPv6 networking protocol), Sigfox, Neul, and NFC — can play an important role in enabling connectivity in smart cities. Among these technologies, some of them, such as SDWN, NFV, CRN, and GC, are enabling technologies (that do not have data rate, communication range, and so on) and are separate from communication protocols. The aim of future and emerging technologies is to enable high rate data communication, better networking infrastructure, low signal attenuation, efficient spectrum utilization, high scalability, high coverage, low cost, robustness, high authentication, and agile encryption mechanisms. In addition, fifth generation (5G) technologies can also provide numerous benefits, such as 10 times more capacity than others, expected speed up to 1 Gb/s, global accessibility, and lower cost. These emerging technologies can be used in different smart cities applications, such as smart home, smart industry, and smart grid, to name a few.

### MODERN COMMUNICATION TECHNOLOGIES

Communication technologies enable connectivity among heterogeneous smart devices. These key technologies include WLAN (WiFi), WiMAX, LTE, LTE-A, Bluetooth, Zigbee, Z-Wave, and

Communication technology	Standard/ governing bodies	Frequency	Range (approximately)	Data rates	Topology
Bluetooth	IEEE 802.15.1	2.4 GHz	1–100 m	1 Mb/s	Point-to-point
Z-Wave	–	900 MHz	100 m	9.6–100 kb/s	Star, cluster, mesh
ZigBee	IEEE 802.15.4	2.4 GHz	10–20 m	250 kb/s	Mesh
LoRaWAN	LoRa Alliance	867–869 MHz (Europe)	2–5 km	290 b/s–50 kb/s	Star
WiFi	IEEE 802.11 (a/b/g/n)	2.4 GHz, 3.6 GHz, 4.9 GHz, 5 GHz, 5.9 GHz	100 m	1–5 4Mb/s	Star
WiMAX	3GPP	3.5 GHz	50 km	75 Mb/s	Point-to-multipoint, mesh
LTE	3GPP	2.5 GHz, 5 GHz, 10 GHz	30 km	300 Mb/s(DL), 75 Mb/s(UL)	Star
LTE-A	3GPP	2.5 GHz, 5 GHz, 10 GHz, 15 GHz, 20 GHz	30km	1Gb(DL), 500Mb/s(UL)	Point-to-point

**Table 1.** Comparison of modern communication technologies used in smart cities.

LoRaWAN. WLAN enables the mobile user to connect a local network using a wireless connection. The IEEE 802.11 group specifies the technologies and also offers different flavors of it (IEEE 802.11 a/b/g/n/p/aka/ah). The 802.11 standards use carrier sense multiple access with collision avoidance (CSMA/CA) for path sharing and WEP as an encryption algorithm. Moreover, WiMAX (IEEE 802.16), LTE (3GPP), and LTE-A (3GPP) also provide portable mobile broadband connectivity across cities. Bluetooth (802.15.1), and Zigbee (802.15.4) are considered as low-range communication technologies and more suitable for personal-area-network-based applications. On the other hand, Z-Wave is a low-power RF communications technology designed for home automation; LoRaWAN is designed to provide low-power WANs with features specifically needed to support low-cost mobile secure bidirectional communication. These communication technologies can be used in most smart cities applications, such as smart grids and metering, smart street lighting, smart homes, smart health monitoring, and smart transportation. Tables 1 and 2 present comparative summaries of the modern communication technologies based on the important parameters.

#### IEEE WIRELESS TECHNOLOGY STANDARDS

Wireless technologies are used to establish the connection among different devices. Commonly used standards for wireless technologies in smart cities are named IEEE 802.11, IEEE 802.15.1, IEEE 802.15.3, IEEE 802.15.4, IEEE 802.15.6, and IEEE 802.16. IEEE 802.11 can be used in different applications of smart cities, such as smart transportation, smart homes, and smart waste management. On the other hand, IEEE 802.15.1, IEEE 802.15.3, IEEE 802.15.4, and IEEE 802.15.6 have relatively shorter coverage than IEEE 802.11. In addition, these standards are more suitable for smart health monitoring and smart lighting applications. The IEEE 802.16 standard defines several technologies that support long-range communication. The technologies of IEEE 802.16 can be used in the smart grid, one of the applications of smart cities.

#### OBJECTIVES

The key motivation behind the transformation of a city into a smart city is providing facilities to the inhabitants' lives in different situations, and it is only possible when flawless connectivity is ensured to users. Some of the precision-based applications of smart cities particularly demand high-quality communication technologies. In the case of smart transportation, one of the applications of smart cities, network latency can raise serious issues. Therefore, some objectives — error-free connectivity, ultra-high spectral efficiency, high-rate data transmission, and low power consumption — are defined for smart cities communication technologies.

#### NETWORK CLASSES

Smart city networks can be categorized into three classes: WPAN, WLAN, and wireless metropolitan area network (WMAN). WPAN is based on IEEE 802.15 and is used for interconnecting the individual's workplace devices. ZigBee and Bluetooth are some examples of WPAN-based technology, whereas WLAN connects two or more devices to each other using a wireless distribution method, orthogonal frequency-division multiplexing (OFDM), within a short range (e.g., smart home, and smart parking). Moreover, IEEE 802.11 standards define most WLAN technologies and are known as WiFi. However, WMAN is intended to cover large areas in terms of connectivity (range approximately the size of a city). WMANs are point-to-point or point-to-multipoint networks with individual links. A WMAN is usually owned by one entity, such as an Internet service provider (ISP), a government, or an enterprise. Access to a WMAN is only through subscribing to the service. IEEE 802.16 standards define several technologies for WMANs, and WiMAX is one of its examples.

#### MODE OF OPERATIONS

The connectivity of smart city objects relies on different types of networks and communication technologies to perform the collaborative tasks for making the lives of inhabitants more comfort-

Modern communication technologies	Advantages	Disadvantages
WiFi	(a) Lack of wires (b) User can move, no need to be stuck at one place	(a) High signal attenuation (b) Limited service radius (c) Less stable compared to wired connections (WiMAX)
WiMAX	(a) High-speed wireless Internet (b) Broad coverage area	(a) Expensive to install
LTE	(a) Backward compatibility and future-proofing (b) High spectrum efficiency (c) Reduce the problem of lagging in Internet connection	(a) Higher cost due to the usage of additional antennas at network base stations for data transmission
LTE-A	(a) High data rates (b) Particularly elevated voice excellence	(a) High cost (b) Accessible in convinced cities only
Bluetooth	(a) Cheap (b) Easy to install	(a) Short-range communication (b) Security flaws
ZigBee	(a) Power saving (b) Collision avoidance (c) Low cost	(a) A bit slower
Z-Wave	(a) A lot simpler than ZigBee	(a) Mobility management is very difficult (b) Security flaws
LoRaWAN	(a) Low power consumption (b) Secure bidirectional communication (c) Low cost	(a) Short-range communication

**Table 2.** Advantages and disadvantages of the modern communication technologies.

able. In this context, the modes of operation used in wireless network communication are of two types: ad hoc and infrastructure-based. Wireless communication technologies based on infrastructure are cellular networks, WiMAX, and digital video/audio broadcast (DVB/DAB). On the other hand, WLAN, infrared, millimeter-wave, ZigBee, Bluetooth, and NFC are infrastructure-less technologies.

### CASE STUDIES OF SMART CITIES

This section describes a number of reported case studies provided by different cities. The purpose of this section is to discuss what types of communication and networking technologies are being used by smart cities. A summary of these case studies is provided in Table 3.

#### BARCELONA

Barcelona (<http://smartcity.bcn.cat/en/bcn-smart-city.html>) has made significant reforms to turn itself into a smart city. To meet the initiated objectives of a smart city, Barcelona is using information and communications technologies (ICT) for transforming companies, institutions, specific spaces, universities, technological centers, incubators, residences, dissemination, and entrepreneurs toward automation. With the aim of supporting smart city initiatives in terms of connectivity, Barcelona is using 3G and 4G technologies, a WiFi mesh network, a sensor network, a public WiFi network, a new mobility plan, new heating and cooling systems, new energy networks, and underground galleries. Moreover, for enabling better city services in terms of quick response, optical fiber is deployed that covers 325 km of

the city. Recent research indicates that the transformation of Barcelona into a smart city is succeeding magnificently.

#### STRATFORD

Recently, Stratford (<http://www.smartcitiesassociation.org/showcase/case-studies/smart-city-stratford.html>) has risen as a “smart city” by initiating a program of smart metering with the aim of facing new energy conservation regulations and stimulating economic growth. To meet the initiation objectives, Motorola’s 802.11n mesh wide area network (MWAN) technology was deployed in the city. The technology helped in enabling a smart metering program and also ensured high-speed mobile Internet access to residents. The Motorola AP 7181 802.11n was used as an outdoor access point, and a GPON AXS1800 system was used to transmit the encrypted smart meter data. At the start, the meters were read manually once a month. To perform the testing, smart meters were placed in 200 homes based on a mesh network of 40 access points. During the trial period, Fextival Hydro was used to access the meters remotely on a daily basis with the objective of determining how to reduce electricity usage. The results of the smart metering program were very attractive. In this project, Rhyzome Networks and Fextival Hydro were involved as stakeholders.

#### SINGAPORE

Singapore (<http://www.smartnation-forbes.com/>) has engaged in the journey toward more sustainable urban development and smarter city. One of the key motivations behind the transformation

The connectivity of smart city objects relies on different types of networks and communication technologies to perform the collaborative tasks for making the lives of the inhabitants more comfortable. In this context, the modes of operations used in wireless network communication are of two types: ad hoc and infrastructure based.

Case study	Devices	Business needs	Solution	Company Involved	Country
Barcelona	IP cameras, wearable devices, traffic monitoring systems	Automation of the companies, institutions, specific spaces, universities, technological centers, incubators, residences, dissemination, and entrepreneurs	Fiber optical 3G, 4G, and WiFi	Endesa, Cisco, Orange, Telvent, UPC, Telefónica, Urbiotica, and T-Systems	Spain
Stratford	Smart meters, smart readers, access points	To meet the energy conservation regulations and stimulate economic growth	Motorola's 802.11n mesh wide area network	Rhizome Networks and Festival Hydro	Canada
Singapore	Cameras, tiny sensors, GPS, access points, and RFID cards	To enable the intelligent transportation system	Ultra-high-speed 1 Gb/s nationwide broadband access, and wireless broadband infrastructure	LTA and A*STAR	Singapore
Porto	Access points, sensors, and cellular operator	To address the mobility and transportation challenges	NetRider, multi-networks, on-board units, fiber, and WiFi	Grupo Galmes, Cisco	Portugal

**Table 3.** Comparison of the case studies.

of the city is the need for the intelligent transportation system to overcome some of the city constraints, such as land and the absence of natural resources. The recent research indicates that in Singapore the roads are already occupied by 12 percent of the land area, and the number of cars has increased to 62 percent of the 9,70,000 vehicles compared to buses, taxis, and two-wheelers, which have increased by 2, 3, and 15 percent, growing with foreign firms and laborers, respectively. To implement the intelligent transportation system, sensors are deployed by leveraging Singapore's ultra-high-speed 1 Gb/s nationwide broadband access and wireless broadband infrastructure. Additionally, to accomplish the initiation objectives, cameras, GPS devices, and a network of sensors are deployed on taxicabs. These smart technologies not only help in monitoring traffic, but also enable capabilities of predicting future congestion that can result in optimal route management. Moreover, to facilitate disabled people, RFID cards are used to extend crossing times when tapped against traffic light poles. This intelligent transportation initiation is helping Singapore to rise as a leading city in the world.

#### PORTO

With the aim of addressing mobility and transportation challenges, Porto (<http://www.yumpu.com/cs/document/view/55647191/creating-the-worlds-largest-network-of-connected-vehicles-for-smart-cities/2?>) has planned to turn into a smart city. The problems, such as unconnected municipal services and underutilization of resources, were the motivations behind the transformation of the city toward automation. In addition, at that time, the research indicated that 413 service and public vehicles traveled a distance of 28 km/year. Twenty-five percent of this travel was estimated to be unnecessary, resulting in waste of fuel and money, and city pollution, to name a few. In order to address these issues, Porto used Veniam's solution for deploying a city-scale vehicular network that was based on existing fiber and WiFi infrastructure. To connect the different types of vehicles that provide transportation for passengers, Veniam developed a multi-network onboard unit (OBU) equipped with WiFi/DSRC/cellular interfaces, called NetRider. NetRider steers vehi-

cles into WiFi hotspots that help deliver Internet access to people in and around vehicles. Moreover, the NetRider access point has also been developed by Veniam to connect the vehicles to the wired infrastructure of heterogeneous network providers and cloud. Veniam's development in Porto has turned it into the biggest WiFi-in-motion network in the world.

### OPEN RESEARCH CHALLENGES

This section discusses the open research challenges. The purpose of the discussion is to provide research directions to new researchers in the domain.

#### INTERFERENCE MANAGEMENT

Due to the unprecedented proliferation of wireless devices, coexistence of devices is growing at a tremendous rate, leading to an interference problem that causes frequent data communication errors. Interference can impede the successful deployment of sensors, WLANs, and other equipment in smart cities. Interference management is one of the key challenges to ensure that these devices work without facing interference. Off-the-shelf interference management models and mechanisms for wireless networks can monitor the extent of interference and provide certain solutions to cope with it. However, these models and mechanisms will not be able to solve the interference problem completely and optimally in smart cities due to the volume of connected devices, which makes it quite complex. Therefore, robust interference management services are required over smart cities' networks.

#### SCALABLE WIRELESS SOLUTIONS

As futurologists have predicted that a huge number of devices will be connected to the Internet with the aim of turning the city into a smart city, the need for scalable wireless solutions has arisen. Although the state-of-the-art wireless technologies, such as RFID, ZigBee, Bluetooth, LoRaWAN, Z-Wave, and other WPAN already support low-power device communication, their capabilities are limited in terms of a number of devices, throughput, and transmission range [5]. Moreover, modern wireless technologies like 802.11 (WiFi) will also not be suitable for smart



cities' communication because these were originally designed to offer high throughput to a limited number of stations located indoors at a short distance from each other. Therefore, it is required to pay much attention to enabling smart communication in the smart city environment.

### INTEROPERABILITY SUPPORT AMONG HETEROGENEOUS WIRELESS NETWORKS

Smart city networks are usually deployed using different wireless network technologies, such as WiFi, WiMAX, mobile ad hoc networks, and wireless mesh networks. However, interoperability among these heterogeneous wireless networks has become a serious concern. To enable communication among different wireless networks requires addressing the challenges associated with interoperability. The interworking of the diverse wireless technologies for efficient delivery of value-added applications and services leads to several challenging issues, mainly related to architecture, resource allocation, mobility management, quality of service (QoS) provisioning, and security. Thus, it is required to pay considerable attention to addressing these challenges in the future.

### MOBILITY MANAGEMENT

Smart cities offer numerous services for mobile users, such as e-health, intelligent transportation systems, and logistics. These applications significantly rely on heterogeneous mobile technologies and thus demand various services ranging from non-real-time (low data rate) applications to real-time (high-speed) multimedia applications offered by various access networks. Therefore, one of the major research challenges for the upcoming mobile systems is designing intelligent mobility management techniques that take advantage of various wireless access technologies to achieve global roaming. Moreover, integration and interoperation of contemporary mobility management techniques in the heterogeneous access networks is required for the integration of forthcoming wireless technologies in smart cities.

### HIGH ENERGY CONSUMPTION

Due to the deployment of resource constrained devices in smart cities, communication and networking from an energy point of view have gained serious attention. Although advanced communication technologies, WiMAX and LTE-A, have facilitated users in terms of tremendous downloading and uploading speed, the energy consumption rate is significantly high and can impede the realization of these technologies. In the future, devices may have high specifications in terms of battery life, but the energy consumption rate of these modern communication technologies will still be considered as higher. The reasons for high energy consumption in modern communication technologies are as follows: enhancement of the radio network to attain good quality signals, support of multiple parallel transmission, and powerful data transmission, to name a few. Thus, future smart cities must ensure control and optimization of renewable energy sources and demand side management programs by delivering real-time information.

## CONCLUSIONS

The error-free connectivity requirement in smart cities has attracted the attention of the IT community and industry toward communication technologies. The coexistence of a high number of intelligent devices in smart cities has brought several problems in terms of connectivity that can impede the realization of existing communication technologies. In this article, we provide a tutorial on research efforts made so far from the perspective of communication and networking technologies in smart cities. A discussion on state-of-the-art enabling communication technologies used in smart cities is presented to help the reader in comprehending the recent efforts in this direction. We also classify the literature by devising a taxonomy based on emerging technologies, communication technologies, IEEE wireless technology standards, objectives, network classes, and modes of operation. Moreover, some reported case studies of different cities are presented. Furthermore, several open research challenges are discussed as future research directions. Finally, we conclude that the utilization of the existing communication technologies cannot provide error-free connectivity in smart cities because these technologies are designed only for a limited number of devices and supported for a specific range of communication. Therefore, in the future, much attention must be paid to enabling unimpaired connectivity in smart cities.

### ACKNOWLEDGMENTS

This work is funded by the Bright Sparks Program and the Research Grant from the University of Malaya under references BSP/APP/1689/2013, RP012C-13AFR, and UM.C/625/1/HIR/MOE/FCSIT/03.

### REFERENCES

- [1] I. A. T. Hashem *et al.*, "The Role of Big Data in Smart City," *Int'l. J. Info. Mgmt.*, vol. 36, no. 5, 2016, pp. 748–58.
- [2] F. Theoleyre *et al.*, "Networking and Communications for Smart Cities Special Issue Editorial," *Computer Commun.*, vol. 58, 2015, pp. 1–3.
- [3] Y. Mehmood *et al.*, "Mobile M2M Communication Architectures, Upcoming Challenges, Applications, and Future Directions," *EURASIP J. Wireless Commun. Networking*, vol. 2015, no. 1, 2015, p. 1.
- [4] S. Djahel *et al.*, "Toward V2I Communication Technology-Based Solution for Reducing Road Traffic Congestion in Smart Cities," *Proc. 2015 IEEE Int'l. Symp. Networks, Computers Commun.*, 2015, pp. 1–6.
- [5] E. Khorov *et al.*, "A Survey on IEEE 802.11 ah: An Enabling Networking Technology for Smart Cities," *Computer Commun.*, vol. 58, 2015, pp. 53–69.
- [6] R. Wenge *et al.*, "Smart City Architecture: A Technology Guide for Implementation and Design Challenges," *Commun.*, China, vol. 11, no. 3, 2014, pp. 56–69.
- [7] K. Hamaguchi *et al.*, "Telecommunications Systems in Smart Cities," *Hitachi Rev.*, vol. 61, 2012, pp. 152–58.
- [8] U. DESA, "United Nations, Department of Economic and Social Affairs, Population Division: World Urbanization Prospects, the 2009 Revision: Highlights," 2010.
- [9] A. Cimmino *et al.*, "The Role of Small Cell Technology in Future Smart City Applications," *Trans. Emerging Telecommun. Technologies*, vol. 25, no. 1, 2014, pp. 11–20.
- [10] J. M. Duarte, E. Cerqueira, and L. A. Villas, "Indoor Patient Monitoring through wi-fi and Mobile Computing," *Proc. 2015 IEEE 7th Int'l Conf. New Technologies, Mobility Security*, 2015, pp. 1–5.
- [11] S. Ayub *et al.*, "A Practical Approach of VLC Architecture for Smart City," *Proc. 2013 IEEE Antennas Propagat. Conf.*, 2013, Loughborough, U.K., 2013, pp. 106–11.
- [12] J. Wan *et al.*, "M2M Communications for Smart City: An Event-Based Architecture," *Proc. 2012 12th IEEE Int'l. Conf. Computer Info. Technology*, 2012, pp. 895–900.

The authors conclude that the utilization of the existing communication technologies can not provide error-free connectivity in smart cities because these technologies are designed only for a limited number of devices and supported for a specific range of communication. Therefore, in the future, high attention must be paid for enabling unimpaired connectivity in smart cities.

- [13] R. Fernandes *et al.*, "Flexible Wireless Sensor Network for Smart Lighting Applications," *Proc. 2014 IEEE Int'l. Conf. Instrumentation and Measurement Tech.*, 2014, pp. 434–39.
- [14] M. M. Mazhar *et al.*, "Conceptualization of Software Defined Network Layers over Internet of Things for Future Smart Cities Applications," *Proc. 2015 IEEE Int'l. Conf. Wireless Space Extreme Environments*, 2015, IEEE, pp. 1–4.
- [15] Y. Wang and Y. Zhou, "Cloud Architecture Based on Near Field Communication in the Smart City," *Proc. 2012 IEEE 7th Int'l. Conf. Computer Science & Education*, 2012, pp. 231–34.

### BIOGRAPHIES

IBRAR YAQOOB received his B.S. (Hons.) degree in information technology from the University of the Punjab, Gujranwala campus, Pakistan, in 2012. Currently, he is pursuing his Ph.D. degree in computer science at the University of Malaya, Malaysia, since November 2013. He won a scholarship for his Ph.D. and is also working as a Bright Spark Program research assistant. He has published a number of research articles in refereed international journals and magazines. His numerous research articles are very famous and often downloaded in top journals. His research interests include big data, mobile cloud, the Internet of Things, cloud computing, and wireless networks.

IBRAHIM ABAKER TARGIO HASHEM is currently a Ph.D. candidate at the Department of Computer Systems, University of Malaya. He has been working on big data since 2013. He has published a number of research articles in refereed international journals. His numerous research articles are very famous and often downloaded in top journals. His main research interests include big data, cloud computing, distributed computing, and computer networks.

YASIR MEHMOOD completed his Master's in electrical (telecom) engineering from the National University of Science and Technology Islamabad, Pakistan. He is currently a doctoral researcher at the Communication Networks research group, University of Bremen, Germany, in the framework of the International Graduate School (IGS) for Dynamics in Logistics (a doctoral training group at the University of Bremen). His major research area

includes cellular communications, mobile M2M communications, and the cellular Internet of Things.

ABDULLAH GANI [M'01, SM'12] is a full professor at the Department of Computer System and Technology, University of Malaya. He received his Bachelor and Master 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, a 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 level of study:- Bachelor, Master, and Ph.D. His interest in research includes self-organized systems, reinforcement learning, and wireless-related networks. He is now working on mobile cloud computing with a High Impact Research Grant of US\$1.5 million for the period of 2011–2016.

SALIMAH MOKHTAR is an associate professor in the Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya. Her research interests are in the area of information systems for education, blended learning, scholarship of teaching and learning (SoTL), spiritual intelligence, big data, and data science. Her most recent interest centers on learning analytics.

SGHAIER GUIZANI received his Ph. D degree from the University of Quebec, Canada, in 2007. He is currently an assistant professor in the Electrical Engineering Department at Alfaisal University, Riyadh, Saudi Arabia. His research interests include communication networks and security (particularly wireless ad hoc, sensor networks, quality of service, wireless sensor networks security, and RFID/NFC application and security) and IoT. He has published a number of research papers in refereed international conferences and journals. He has served/serves as an Associate Editor for *Security and Communication Networks* (Wiley), the *International Journal of Sensor Networks* (InderScience), and the *Journal of Computer Systems, Networking, and Communications*. He has been involved in a number of conferences and workshops in various capacities.

## CALL FOR PAPERS

*IEEE COMMUNICATIONS STANDARDS MAGAZINE*

# REAL-TIME COMMUNICATIONS IN THE WEB: CURRENT ACHIEVEMENTS AND FUTURE PERSPECTIVES

## BACKGROUND

Web Real-Time Communication is a joint standardization effort between the Internet Engineering Task Force (IETF) and the World Wide Web Consortium (W3C). Since 2011 the "Real-Time Communication in WEB-browsers" (RtcWeb) Working Group has been working on key aspects like the overall communication infrastructure, the protocols and API (Application Programming Interface) requirements, the security model, the media formats (and related media codecs), as well as advanced functionality like congestion/flow control and interworking with legacy VoIP equipment. While the W3C WebRTC wg has conducted a parallel activity on the definition of a whole set of APIs exposing functions like exploration and access to device capabilities, capture of media from local devices, encoding/processing of media streams, establishment of peer-to-peer connections between web-enabled devices, decoding/processing of incoming media streams and delivery of such streams to the end-user in an HTML5-compliant fashion.

To date, the two mentioned working groups have done a tremendous amount of work, which has brought us close to what can undoubtedly be considered an unprecedented milestone in the field of real-time multimedia communications: the so-called "WebRTC 1.0" standards suite. The idea behind WebRTC 1.0 is to allow all of the involved stakeholders (browser vendors, telecommunication providers, application providers, web developers, etc.) to converge onto a well-defined set of protocols and APIs to be leveraged in order to allow wide-spread deployment on the market of interoperable products offering end-users a media-rich, web-enabled real-time experience.

WebRTC has also had to confront itself with alternative views. In fact, since the beginning of 2014, a brand new initiative has seen the light in the W3C, the ORTC (Object Real-time Communications) Community Group, which has initially been identified as a clear opponent to WebRTC. Fortunately, after the unavoidable initial friction, the international standardization community has decided not to disperse precious energies and has eventually come to a sort of compromise. The idea has been to rapidly converge to a unique agreed-upon solution by allowing the ORTC community to contribute to the finalization of the first version of the standard. At the same time, a common decision has been taken as to welcome most of the key concepts proposed with ORTC's low-level object API to be adopted for the so-called "Next Version" of the standard, which nonetheless has backward compatibility with the 1.0 release among its foundational requirements. This is exactly where we stand now. We are a step away from completing WebRTC 1.0, with our minds already looking at the rising WebRTC-NV initiative.

In light of these trends, this FT (Feature Topic) will both focus on the current state of the art with respect to WebRTC 1.0 completion and introduce the envisioned work program for the second generation of the standard. In doing so, we invite contributions dealing with the hot topics and illustrating how the community has successfully coped with most of them. We also await papers making some projections as to what the standardization community imagines will represent the upcoming milestones and open issues. Tutorial-oriented contributions shedding some light on the current status of standardization, with special focus on the upcoming final release of the so-called WebRTC 1.0 standard ecosystem are most welcome. This FT will take the stock of the situation with respect to topics like, e.g., codecs, session description, data-channel and stream multiplexing. It will also illustrate how standard bodies are dealing with seamless integration of WebRTC with the initially competing ORTC effort. Finally, it will take a look at the future by welcoming contributions about the forthcoming initiative informally known as WebRTC Next Version (WebRTC-NV).

Stated in one sentence, this Feature Topic aims at presenting the consolidated results achieved so far in the area of standardization of real-time communications in the Web, by presenting a comprehensive view of the numerous challenges researchers have had to face before arriving at the first release of an agreed-upon WebRTC standard suite, while also providing useful hints on the upcoming standardization efforts.

Original contributions previously unpublished and not currently under review, are solicited in relevant areas including (but not limited to) the following:

- WebRTC 1.0 standard protocols and APIs
- WebRTC 1.0 JavaScript programming patterns and APIs
- Practical experiences with WebRTC 1.0: testbeds and business cases
- WebRTC Security Architecture: signaling, consent, privacy, communications security, peer authentication, trust relationships
- ORTC low-level object oriented approach to real-time web communications
- WebRTC support in the browsers
- WebRTC business readiness and industry adoption
- WebRTC gateways for interoperability with legacy VoIP architectures
- Performance monitoring/evaluation of WebRTC architectures
- WebRTC Next Version: milestones and work plan

## SUBMISSIONS

Articles should be tutorial in nature and written in a style comprehensible and accessible to readers outside the specialty of the article. Authors must follow the *IEEE Communications Standards Magazine's* guidelines for preparation of the manuscript. Complete guidelines for prospective authors can be found at <http://www.comsoc.org/comstandardsmag/author-guidelines>.

It is important to note that the *IEEE Communications Standards Magazine* strongly limits mathematical content, and the number of figures and tables. Paper length should not exceed 4500 words. All articles to be considered for publication must be submitted through the IEEE Manuscript Central site (<http://mc.manuscriptcentral.com/commag-ieee>) by the deadline. Select "Standards Supplement" from the drop-down menu of submission options.

## IMPORTANT DATES

- Manuscript Submission Date: January 15, 2017
- Decision Notification: March 15, 2017
- Final Manuscript Due Date: April 15, 2017
- Publication Date: June 2017

## GUEST EDITORS

Dr. Salvatore Loreto  
Ericsson AB, Stockholm, Sweden

Prof. Simon Pietro Romano  
University of Napoli Federico II, Italy

Prof. Carol Davids  
Illinois Institute of Technology, USA

# Security and Privacy in Smart City Applications: Challenges and Solutions

Kuan Zhang, Jianbing Ni, Kan Yang, Xiaohui Liang, Ju Ren, and Xuemin (Sherman) Shen

The authors investigate security and privacy in smart city applications. Specifically, they first introduce promising smart city applications and architecture. Then they discuss several security and privacy challenges in these applications. Some research efforts are subsequently presented to address these security and privacy challenges for intelligent healthcare, transportation, and smart energy.

## ABSTRACT

With the flourishing and advancement of the IoT, the smart city has become an emerging paradigm, consisting of ubiquitous sensing, heterogeneous network infrastructure, and intelligent information processing and control systems. A smart city can monitor the physical world in real time, and provide intelligent services to both local residents and travelers in terms of transportation, healthcare, environment, entertainment, and energy. However, security and privacy concerns arise, since smart city applications not only collect a wide range of privacy-sensitive information from people and their social circles, but also control city facilities and influence people's lives. In this article, we investigate security and privacy in smart city applications. Specifically, we first introduce promising smart city applications and architecture. Then we discuss several security and privacy challenges in these applications. Some research efforts are subsequently presented to address these security and privacy challenges for intelligent healthcare, transportation, and smart energy. Finally, we point out some open issues for future research.

## INTRODUCTION

With the rising economy and social transformation, people have been moving from the country to cities, resulting in the largest wave of urbanization throughout the world. By 2030, the urban population is estimated to reach 5 billion (about 60 percent of the world population), which produces massive opportunities for the economic and social development of cities [1]. Due to the ever growing demands of local residents, the development of fundamental infrastructure and policies are not correspondingly ensured. Moreover, this unplanned and overly fast urban growth brings excessive burdens to climate, energy, the environment, and even living. These problems slow down the sustainable development of urban cities as a consequence. To mitigate the problems of rapid urbanization, it is urgent to improve governance and service delivery, offer swift seamless mobility, and achieve easy access to urban public facilities, affordable housing, quality healthcare, education, and living in highly populated areas [2]. A special spotlight is needed, covering urbanization trends in innovative management of urban operations and a variety of "smart" services to

local residents, visitors, and the government to satisfy the ever increasing and diverse demands [3]. The advancement and flourishing of the smart city shed light on materializing these value-added services and tackling the problems of urbanization.

As an emerging paradigm, the smart city leverages a variety of promising techniques, such as the Internet of Things (IoT), cyber-physical systems, big data analysis, and real-time control, to enable intelligent services and provide comfortable life for local residents [4]. It integrates ubiquitous sensing components, heterogeneous network infrastructure, and powerful computing systems to sense the physical changes from cities and feed back to the physical world. Specifically, RFID devices, sensors, and versatile wearable devices are promoted to offer real-time monitoring and ubiquitous sensing, from energy to environments, from road traffic to healthcare, from home area to public venues, and so on. Then this sensing information is transmitted to a control center via heterogeneous networks. This control center takes comparative advantage of powerful computing systems, such as cloud servers, to process and analyze the collected data. Fueled by human intelligence, the control center makes optimal decisions and manipulates the urban operations via feedback components, such as actuators [3]. Having the advanced information, communication, and control technologies as backbones, a smart city can offer various applications, including intelligent transportation, smart energy, intelligent healthcare, and smart homes. Not only can this up-and-coming connected city quickly identify the demands of people and a city, but it can also manipulate urban operations to improve urban living quality in an intelligent and sustainable way. It is expected that the global smart city market will exceed US\$1200 billion by 2020, which is almost triple that in 2014 [1].

When cities become smarter, people may suffer from a series of security and privacy threats due to the vulnerabilities of smart city applications [5]. For example, malicious attackers may generate false data to manipulate sensing results such that services, decisions, and control in a smart city are influenced and not "intelligent" enough. Moreover, these malicious attackers could also launch denial-of-service attacks, disrupting the sensing, transmission, and control to degrade the quality of intelligent services in a smart city.



Figure 1. Smart city applications.

In addition, the pervasive video surveillance in a smart city captures a tremendous number of images and video clips, which may be utilized to infer local residents' trajectories and inherently endanger their privacy. The home area information collected and managed by smart home applications may pave the way to disclosing residences' highly privacy-sensitive lifestyle and even cause economic loss. Although some off-the-shelf techniques (encryption, authentication, anonymity, etc.) and policies might be directly applied to avert these problems [5], the emerging "smart" attackers could still infer and violate privacy in many other ways, such as side channel attack and cold boot attack [6]. Without sufficient security and privacy protections, users may refrain from accepting the smart city, which would remain as a far-off futuristic idea.

These emerging trends motivate our research investigating the not-for-profit global initiative of security and privacy for the smart city. In this article, we first introduce smart city applications and a heterogeneous architecture. Then we discuss several challenging security and privacy issues, including privacy leakage, secure information processing, and dependability in control. Some innovative research efforts are presented to address these challenges in various smart city applications. Finally, we point out several open research directions and the outlook of the smart city from the security and privacy perspective.

## SMART CITY APPLICATIONS AND ARCHITECTURE

As a smart city connects the physical world and the information world, many intelligent applications are emerging, from local to global, from sensing to control, as shown in Fig. 1. In this section, we introduce smart city applications and the heterogeneous architecture.

### SMART CITY APPLICATIONS

Smart city applications benefit people and the city in a variety of aspects: energy, environment, industry, living, and services.

We introduce several key applications as follows.

**Smart Energy:** Exploiting the widely deployed sensors to monitor energy generation, transmission, distribution, and consumption, smart energy [7] leverages utility usage, electric vehicle charging, smart grid, and so on. Not only can it reduce the energy consumption in many aspects, but it can also prevent blackout of power grid and failure of individual energy usage.

**Smart Environment:** The smart environment is promoted to support a comfortable climate and sustainable environment for the smart city. Ubiquitous sensing and intelligent climate management are jointly applied in smart environment applications [4]. They can monitor waste gas, greenhouse gas, city noise, air and water pollution, forest conditions, and so on, to afford intelligent and sustainable development.

**Smart Industry:** With the main driver of industrial sustainable development in the smart city, smart industry is being rolled out to optimize industrial production and manufacturing, while achieving efficiency and robustness. On one hand, it curtails the material and resource consumption (e.g., labor, time, and production lines) during the industrial process; on the other hand, it prevents industrial heat and gas waste from excessive emission. Both sensing and control are equally arresting components in smart industry, which requires real-time feedback and precise operations. Finally, servo actuators, motors, and robots are adopted to enable precise control and operations of consequence in smart industry.

**Smart Living:** In home areas, smart living offers intelligent management of various appliances and utilities to create comfortable homes and improve energy efficiency simultaneously [6]. It can enable remote control of home appliances, climate adjustment, energy saving, surveillance, entertainment, and education. In the community (or building), smart living applications also intelligently manage waste recycling, social networking, and parking to provide a smart community (or building) with comfortable lives, intimate service, wonderful experiences, and sustainable environment and energy.

To achieve ubiquitous sensing and finesse city management, smart city manipulates the information sensed from the physical world, the information transmitted in the communication world and the information processed in the information world for intelligent services.

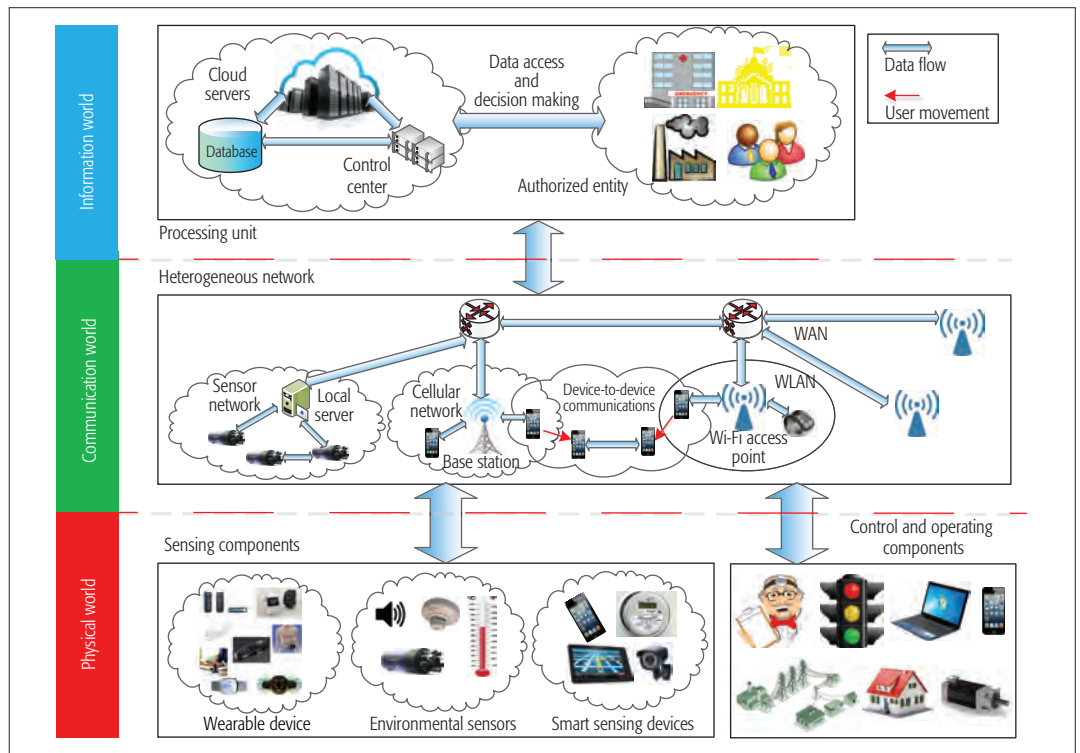


Figure 2. Architecture for smart city: physical world, communication world, and information world.

**Smart Service:** Smart service enables the public facilities and services to benefit people in a wide range of aspects [1]. For example, intelligent transportation [8] can help local residents and travelers to avoid road traffic congestion, enable road navigation, discover points of interests, manage the travel planning, and so on. The road traffic information can be collected by deployed sensors, cameras at the intersections, GPS, smartphones from people on the road, and so on. The control center adjusts travelers' road plans and feedback to their smartphones or GPS. In addition, road traffic can be adjusted by managing the traffic light and public transportation tools, such as buses, trains, and shared bicycles.

To provide quality healthcare, intelligent healthcare enables continuous health monitoring and timely diagnosis (including health warnings) to the people in a smart city [9]. It relies on wearable devices and medical sensors to measure users' health conditions, and sends health data to the processing unit for doctors' further diagnosis. It also provides easy access to a user's historical comprehensive health information, considerably increasing the chance to diagnose chronic or infectious diseases in the early stage. In addition, intelligent healthcare contains various health-related applications, such as home care, emergency alarm, and intelligent fitness and training.

#### SMART CITY ARCHITECTURE

To achieve ubiquitous sensing and finesse city management, the smart city manipulates the information sensed from the physical world, the information transmitted in the communication world, and the information processed in the information world for intelligent services. It incorporates sensing components, heterogeneous network infrastructure, processing units, and control and operating components as shown in Fig. 2.

**Sensing Components:** Sensing components exploit wearable devices, industrial sensors, and smart devices (e.g., smartphones, smart meters, and video surveillance cameras [4]) to measure information from the physical world and transmit this information to the processing unit for decision making. In other words, sensing components are the bridge connecting the physical and information worlds. The sensing devices are either deployed by the government, departments, and companies, or carried by users as discussed above. In addition, due to the limitations of device size, battery, and processing capabilities, these resource-constrained sensing devices usually pre-process or compress the real-time and granular data before sending it to the network.

**Heterogeneous Networks:** With the coexistence of massive sensing devices and various applications [9], the sensing information is collected in different ways such that the heterogeneous network infrastructure plays an instrumental role in supporting the smart city. Heterogeneous networks incorporate cellular networks, wireless local area networks (WLANs), wide area networks (WAN), device-to-device (D2D) communications, millimeter-wave communications, sensor networks, and so on, and enable seamless switching among different types of networks. Heterogeneous networks represent the communication world in a smart city to connect the physical and information worlds.

**Processing Unit:** The processing unit exploits the powerful cloud computing servers, abundant databases, and dedicated control systems to analyze and process the collected sensing information from the physical world for decision making. The processing unit manages the information world in a smart city. Authorized entities, such as the government, hospitals, factories, users, and so

on, have certain privileges and authorizations to access the collected information. They can also determine the requirements or policies for decision making and control in a smart city.

**Control and Operating Components:** Leveraging the optimization and decisions of the processing unit, a smart city feeds back to manipulate the physical world via the control and operating components, such as servo actuators or smartphones. These control and operating components optimize and make adjustments to the physical world such that a good quality of life can be offered in a smart city. They also implement the two-way flow of the smart city (i.e., sensing and control). Not only can his two-way flow acquire the knowledge about the physical world; it can also monitor and manage every device or component in a smart city to make it operate properly and “smart.”

## SECURITY AND PRIVACY ISSUES IN A SMART CITY

Although cities are seeking to become “smarter,” smart city applications raise a series of concerns and challenges in terms of security and privacy. As an information and networking paradigm, the smart city should be able to defend the involved information from unauthorized access, disclosure, disruption, modification, inspection, and annihilation. Underlying security and privacy requirements, including confidentiality, integrity, non-repudiation, availability, access control, and privacy [5], should be satisfied in the information, communication, and physical worlds. Besides these general requirements, securing a smart city still faces a set of unique challenges. On one hand, a smart city collects granular-scale and privacy-sensitive information from people’s lives and environments; on the other hand, it processes this information, and manipulates and impacts people’s lives. Due to these unique characteristics, security and privacy issues become challenging and prevent the smart city from being tempting enough to encourage more use.

### PRIVACY LEAKAGE IN DATA SENSING

A smart city is vulnerable to privacy leakage and information inferring by outside attackers, since private information is collected, transmitted, and processed. The disclosed privacy in a smart city may contain a user’s identity and location in transportation, health condition in healthcare, lifestyle inferred from intelligent surveillance, smart energy, home and community, and so on. It would be a major oversight to disclose this privacy-sensitive information to untrusted or unauthorized entities in both the physical and communication worlds. To preserve user privacy during data sensing, some off-the-shelf security and privacy techniques, such as encryption, anonymity, and access control, can be applied [10, 11]. Martinez *et al.* [5] propose a set of privacy concepts and general privacy requirements toward smart city applications. The privacy of identity, query, location, footprint, and owner is identified and provided with some basic ideas to solve the general problems.

However, a portion of private information may still be unconsciously disclosed to untrusted entities. For example, intelligent surveillance may capture local residents’ daily life hints, style, or

even privacy, although it was originally designed for monitoring criminal behaviors in the real and cyber worlds. Similarly, a smart home also utilizes a surveillance camera to detect theft or abnormal events. The intruding attackers in a smart home may acquire private information about the home area, which is prejudicial to the residence’s privacy. Most existing security and privacy protection [10] are developed against outside eavesdroppers and attackers. But potential inside attackers, such as agents, employees, and security guards, who can access surveillance records may either steal users’ data or leave a gap for outside attackers. In addition, the data in a smart city are on a highly granular scale and of diverse types such that the privacy requirements vary with different types. It is challenging to develop adjustable privacy protection mechanisms in a smart city to balance the trade-off between privacy and efficiency.

### PRIVACY AND AVAILABILITY IN DATA STORAGE AND PROCESSING

As a smart city takes comparative advantage of powerful cloud servers for data storage and processing in the information world, it faces security threats due to the untrusted cloud servers. If the smart city data are in clear text during storage and processing, they are directly revealed to the cloud server [12]. An alternative is to encrypt the smart city data and send ciphertexts to the cloud server for storage and processing [13]. Although this method prevents the untrusted cloud server from directly accessing the collected data, the cloud server cannot process the encrypted data and perform effective analytical operations for smart city applications. The latest breakthrough on fully homomorphic encryption sheds light on the processing, such as summation and comparison, over encrypted data. The computational overhead poses another impending challenge in terms of efficiency, especially when massive data are involved in a smart city.

Another challenging issue of securing a smart city is data sharing and access control. For example, road traffic data can be collected by deployed cameras or travelers’ smartphones and GPS in a crowdsourcing way. During global road planning, it is challenging to define the access policy and enable privacy-preserving data sharing among the collaborators. Therefore, smart city data storage and sharing require extensive research efforts.

### TRUSTWORTHY AND DEPENDABLE CONTROL

A smart city, having a two-way control flow, relies on the control system and actuators to materialize the operations determined by the control center. The control and feedback systems in the physical world, especially public and industrial infrastructure, become highly attractive targets for attackers, criminals, and even terrorists [14]. Denial-of-service attacks, spoofing attacks, malicious data injection, and so on would disrupt the smart city such that the management, control, and operation are either biased and incorrect or disabled. Most of these malicious attacks and misbehaviors are detected based on third party inspection and auditing. In [15], data integrity functionality and digital signatures are adopted in software defined networks to achieve data integrity, access control, and so on. Meanwhile, trusted computing is

Although cities are seeking to become “smarter,” smart city applications raise a series of concerns and challenges in terms of security and privacy. As an information and networking paradigm, a smart city should be able to defend the involved information from unauthorized access, disclosure, disruption, modification, inspection, and annihilation.

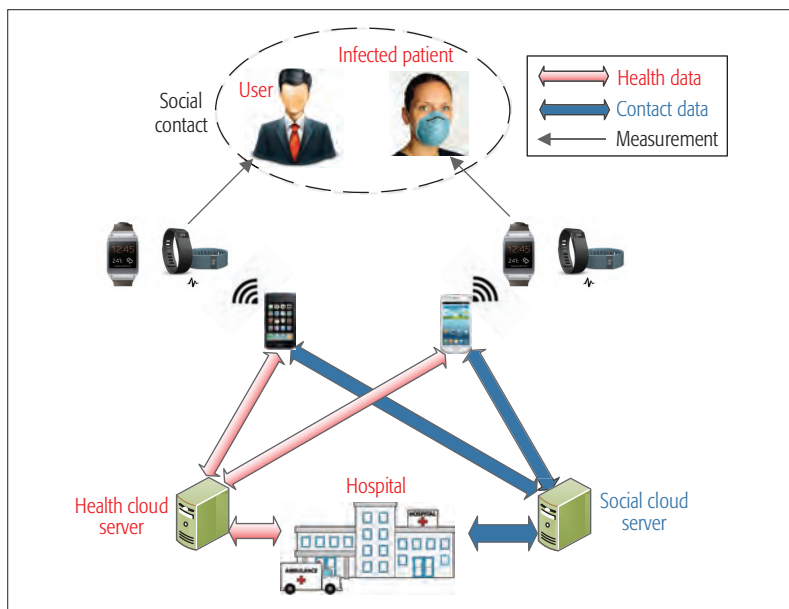


Figure 3. Intelligent healthcare integrating social networking and health data for infection analysis.

a state-of-the-art solution to resist operating system and software framework alterations. However, these schemes consume large latency and a high false rate to detect “smart” attacks in a smart city. As dependability of control is considered as the topmost priority in a smart city, efficient and fast detection of malicious attacks and misbehaviors becomes challenging, requiring collaborative efforts among various parties and stakeholders.

## SECURITY SOLUTIONS FOR SMART CITY PARADIGMS

To materialize the notion of security and privacy in a smart city, balanced and pragmatic solutions are desired. In this section, we introduce state-of-the-art security and privacy protection schemes for several emerging smart city paradigms, including intelligent healthcare, intelligent transportation, and smart grid.

### PRIVACY-PRESERVING INFECTION SPREAD ANALYSIS FOR INTELLIGENT HEALTHCARE

Intelligent healthcare, fueled by connected biomedical sensors, and health data storage and processing units, provides preventive, curative, and palliative health services. It can collect a wide range of real-time health data from users, and analyze and defend severe healthcare issues city-wide, such as infectious disease spread. Infectious diseases (e.g., Ebola, flu, and acute respiratory infection) could be rapidly spread in the population via human-to-human contact, especially when the infected patients cough and sneeze in a crowd. People having frequent contact or strong social relationships with a patient (e.g., students studying in the same classroom or families living in the same house) are usually considered susceptible from the perspectives of biomedicine and sociology. An old-fashioned prevention approach is to isolate the susceptible people for a certain period. However, this approach does not consider their health condition and susceptibility in terms of negative impacts, including massive

healthcare expense, economic loss of the isolated people, and panic or anxiety among the society.

To tackle the infection spread problem, intelligent healthcare would provide efficient diagnosis and health condition (or emergency) warning, by analyzing in real time the infectiousness during outbreak season. Suppose a junior school student, Bob, is continuously monitored from both the health and social perspectives during the outbreak of an infectious disease. Once Bob’s immunity goes very low and he frequently contacts an infected student, he may be inferred as a susceptible patient in the early stage. In general, the spread of infectious disease depends on users’ social contacts and health conditions. Specifically, this spread process can be affected by several key factors of infection, that is, susceptibility of the infected patient, immunity strength of the contacted user, contact duration, and social ties.

The fusion of social network data together with real-time health data facilitates a novel paradigm of infection analysis, as shown in Fig. 3. On one hand, a social network employs a variety of applications to mine users’ social contacts during their social interactions. For example, the Wechat friend discovery program can find users in physical proximity and record social interactions; speech recognition can detect if some people cough or sneeze; a face-tagging function can identify a user’s face from images. On the other hand, wearable devices and medical sensors can measure a user’s real-time health condition [3, 9].

However, health and social network data are collected by multiple independent service providers, such as hospitals and social network vendors (e.g., Facebook and Wechat). The collaboration of these service providers is the key challenge of enabling this enhanced infection analysis, and poses a series of security issues. Both social and health cloud servers are considered to be honest but curious [5] in intelligent healthcare applications. To preserve the user’s data privacy and achieve data availability, homomorphic encryption [9] can be adopted to make both social network and health data invisible to the untrusted cloud servers. The collaboration of different untrusted cloud servers is achieved via the authorized entity (i.e., a hospital authorized by users as shown in Fig. 3). However, when the hospital queries the infected patient’s data on the social cloud server, the social cloud server may infer that the queried user is infected even though the query content is still invisible. In addition, any entity without the authorization of the data owner should not be able to query the owner’s data. State-of-the-art security and privacy protections are essential for intelligent healthcare. Without effective protections, users may not be willing to share their social and health data with others such that the infection analysis would be disabled.

To this end, conditional oblivious transfer protocol is developed for the privacy-preserving data query [9]. On one hand, it allows an authorized entity, such as a doctor, to access a patient’s social network data from the social cloud server; on the other hand, it prevents the social cloud server from accessing the data and inferring any information about the query, such as the patient’s identity. Users or data owners are able to grant authorization to the trusted entity before the query. Any



entity without user authorization cannot query any data. In addition, secure multi-party computation based on homomorphic encryption [9] is utilized to prevent the untrusted health cloud from learning any private social and health data.

### SECURE NAVIGATION FOR INTELLIGENT TRANSPORTATION

A smart city offers intelligent transportation services to local residents and visitors in various aspects, including road traffic adjustment, navigation, point of interest recommendation, parking, and so on. As an integral part of intelligent transportation, navigation attracts intensive attention [5]. Existing GPS devices can provide static navigation by showing the route on pre-downloaded maps. However, it lacks real-time road traffic adjustment such that the calculated fastest route may be delayed by dynamic congestion. Dynamic navigation exploits human intelligence and dynamic road traffic sensing from travelers on the road and roadside units (RSUs) in a crowdsourcing way [3].

As shown in Fig. 4, a querier, that is, the querying vehicle in the navigation service, sends a navigation query to the closest RSU. The query contains the current location, destination, and expired time. Then the RSU forwards this query to the RSU that covers the destination through the network among RSUs. Upon receiving the navigation query, RSUs send the crowdsourcing task to the vehicles in its coverage area to find the fastest driving route for the querier. The querier retrieves a response from the RSUs when entering the coverage area of each RSU, and finally reaches the destination.

During this type of distributed navigation, the private location information of both the querying vehicle and responding vehicles may be disclosed. To this end, the Elgamal and Advanced Encryption Standard (AES) schemes [8] are utilized to encrypt the querier's location and destination in each hop from the querier to the last RSU, preserving a vehicle's query privacy. To prevent RSUs from linking the navigation query and retrieving query to a specific vehicle, each vehicle randomizes the credential issued by the trusted authority to generate a group signature. In addition, to prevent the sensitive information in the navigation response being disclosed, the driving route is encrypted by Elgamal and AES schemes associated with a zero-knowledge range proof, which proves that the time cost is less than the given threshold, without exposing the exact value [8]. Finally, the traceability of group signature allows the trusted authority to trace any malicious vehicle that does not honestly follow the rules.

In summary, this privacy-preserving navigation scheme relies on the distributed RSUs to complete the road planning task in a crowdsourcing way. During the querying, crowdsourcing, and navigation phases, both querier and responding vehicles can preserve location privacy.

### ADAPTIVE KEY MANAGEMENT FOR SMART GRID

Smart grid relies on millions of smart meters to measure the real-time power consumption in residential areas or buildings, as shown in Fig. 5. These metering data are aggregated to the control center to optimize the power distribution in return. However, a series of attacks attempt to tamper smart meter records and upload modified

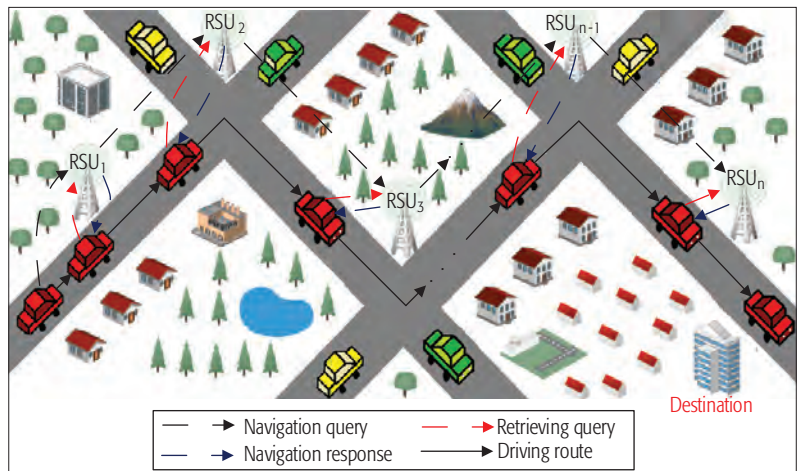


Figure 4. Intelligent navigation with privacy preservation.

data to the control center [15]. Moreover, the ever increasing volume of metering data poses a new challenging issue of managing secret keys for each device [6]. Predominantly, the data integrity and authentication should be achieved during the aggregation of smart metering. In addition, the metering data of a home area may reflect the residence's lifestyle, condition (e.g., very low power consumption over a long duration indicates that residents are out), and preferences [6]. If the untrusted aggregators learn and reveal this private information, the residents' privacy would be jeopardized, and economic loss may even be caused.

In [7], Zhang *et al.* propose a privacy-preserving aggregation scheme (PARK) to improve the computational efficiency and protect smart metering data from disclosure to untrusted aggregators. An adaptive key management scheme is developed based on bidirectional hash chains, generating the encryption keys for every smart meter during each period. The trusted authority calculates the decryption key for the aggregator as the summation of encryption keys from a group of  $N$  smart meters. It is only when having all  $N$  ciphertexts that the aggregator can decrypt the summation of  $N$  smart meters. If no smart meter joins or leaves the smart grid, every smart meter's encryption key is automatically updated. The aggregator's decryption key is updated in a synchronous way. The trusted authority determines the length of hash chains, which reflects the reputation of smart meters. A meter with a high reputation receives a key with long expiry time. When some meters join or leave the smart grid, the trusted authority only needs to update the aggregator's decryption key. The revocation overhead is mainly from the re-distribution of the decryption key. As shown in Fig. 6, the proposed PARK scheme costs one-time key distribution in every key update, while other schemes (distributed key management and a naive scheme [7]) cost higher key update overhead. In addition, forward and backward secrecy is achieved based on the security of a one-way hash function.

### FUTURE RESEARCH DIRECTIONS

Since some off-the-shelf security and privacy solutions [4] may not conquer all the challenges in a smart city, we discuss several open research directions including, but not limited to, the following.

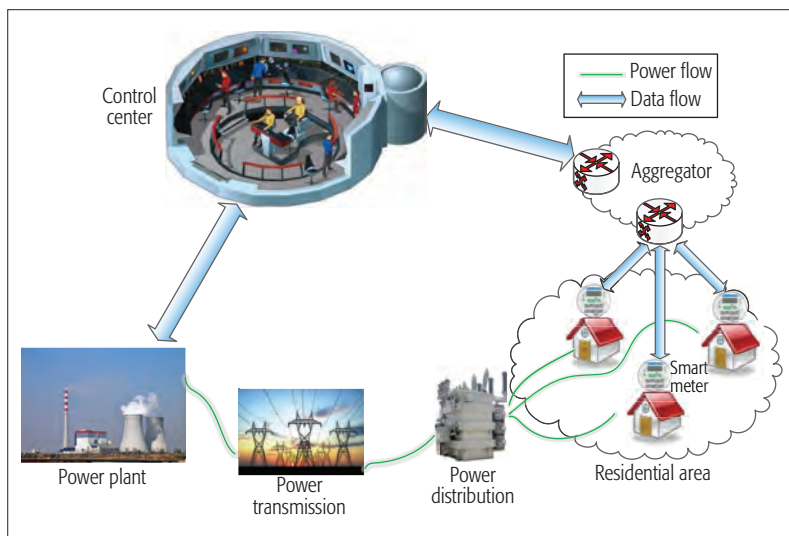


Figure 5. Smart grid architecture, including power flow and data flow.

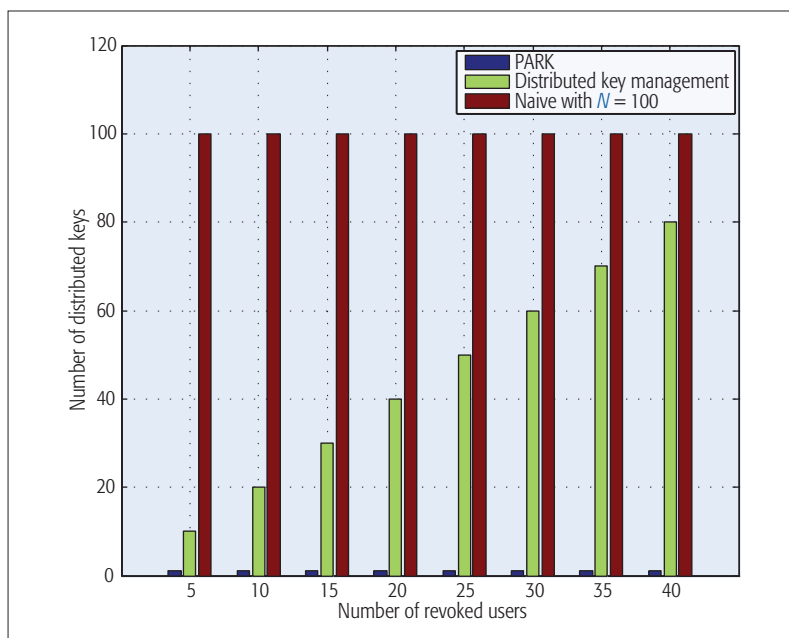


Figure 6. Comparison of key management overhead in smart grid.

First, crowdsensing, which exploits smart sensing devices of local residents, can provide improved sensing capability for the smart city rather than purely relying on pre-deployed fixed sensors. However, the crowdsensing accuracy may vary with a contributor's knowledge, preference, selfishness, and so on. An initial idea of stimulating citizens to contribute for crowdsensing is to develop incentives for them. Moreover, trustworthiness should also be considered when designing incentive schemes. In addition, crowdsensing contributors' privacy hidden in the sensing results may be jeopardized by "smarter" attackers. In particular, when multiple contributors pool their sensing results together, an individual contributor's private information is likely to be collaboratively inferred by others. Therefore, how to achieve incentive and privacy remains as a challenge for crowdsensing in smart city.

Second, a smart city is vulnerable to false data injection in both sensing and control phases. Dig-

ital signature techniques [9] cannot prevent the data from being tampered from the origination. An insight into detecting false data injection is to leverage machine learning and data mining to come up with a boundary of reasonable sensing data. Abnormal detection techniques may be an alternative to identify the false data. However, it is still an open issue requiring multidisciplinary knowledge and efforts to address.

Last but not least, the ever growing volume of data and devices in a smart city poses open problems for intelligent services and privacy. Inside attackers exploit human intelligence and have access to big data such that the privacy of data owners may be inferred and violated; even the traditional cryptographic schemes have been applied to big data. An alternative to detect these inside attackers is to enhance the traceability and allow a trusted third party to monitor and audit. Meanwhile, collaborative efforts among municipalities, regulation departments, industry, academia, and business companies are necessary to set up privacy policies and regulations. In addition, data privacy, availability, and management should be achieved simultaneously.

## CONCLUSIONS

In this article, we have investigated the smart city, and discussed the security and privacy challenges in emerging smart city applications. We have first introduced smart city applications in different aspects and discussed the architecture. Then we have presented the general security and privacy requirements and identified several security challenges for the smart city. In addition, we have dwelled in greater depth on state-of-the-art security and privacy solutions for smart city applications. Several open research directions are also discussed.

We hope this article sheds more light on the security and privacy for smart cities, where more ground-breaking research efforts along this emerging line will be seen in the future.

## ACKNOWLEDGMENT

This research has been supported by a research grant from the Natural Science and Engineering Research Council (NSERC) and Care In Motion (CIM) Inc., Canada.

## REFERENCES

- [1] P. Neirotti et al., "Current Trends in Smart City Initiatives: Some Stylised Facts," *Cities*, vol. 38, 2014, pp. 25–36.
- [2] R. G. Hollands, "Will the Real Smart City Please Stand Up? Intelligent, Progressive or Entrepreneurial?" *City*, vol. 12, no. 3, 2008, pp. 303–20.
- [3] J. Liu et al., "Software-Defined Internet of Things for Smart Urban Sensing," *IEEE Commun. Mag.*, vol. 53, no. 9, Sept. 2015, pp. 55–63.
- [4] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of Things for Smart Cities," *IEEE Internet of Things J.*, vol. 1, no. 1, 2014, pp. 22–32.
- [5] A. Martinez-Balleste, P. Perez-Martinez, and A. Solanas, "The Pursuit of Citizens' Privacy: A Privacy-Aware Smart City Is Possible," *IEEE Commun. Mag.*, vol. 51, no. 6, June 2013, pp. 136–41.
- [6] X. Li et al., "Smart Community: An Internet of Things Application," *IEEE Commun. Mag.*, vol. 49, no. 11, Nov. 2011 pp. 68–75.
- [7] K. Zhang et al., "PARK: A Privacy-Preserving Aggregation Scheme with Adaptive Key Management for Smart Grid," *Proc. IEEE ICC*, 2013, pp. 236–41.
- [8] J. Ni et al., "Privacy-Preserving Real-Time Navigation System Using Vehicular Crowdsourcing," *Proc. VTC-Fall*, 2016, pp. 1–6.

- [9] K. Zhang *et al.*, "Security and Privacy for Mobile Healthcare Networks — From Quality-of-Protection Perspective," *IEEE Wireless Commun.*, vol. 22, no. 4, Aug. 2015, pp. 104–12.
- [10] A. S. Elmaghraby and M. Losavio, "Cyber Security Challenges in Smart Cities: Safety, Security and Privacy," *J. Advanced Research*, vol. 5, no. 4, 2014, pp. 491–97.
- [11] R. H. Weber, "Internet of Things — New Security and Privacy Challenges," *Computer Law & Security Review*, vol. 26, no. 1, 2010, pp. 23–30.
- [12] M. Naphade *et al.*, "Smarter Cities and Their Innovation Challenges," *IEEE Computer*, vol. 44, no. 6, 2011, pp. 32–39.
- [13] J. Gubbi *et al.*, "Internet of Things (IoT): A Vision, Architectural Elements, And Future Directions," *Future Gen. Comp. Sys.*, vol. 29, no. 7, 2013, pp. 1645–60.
- [14] R. Roman, J. Zhou, and J. Lopez, "On the Features and Challenges of Security and Privacy in Distributed Internet of Things," *Comp. Net.*, vol. 57, no. 10, 2013, pp. 2266–79.
- [15] A. Akhunzada *et al.*, "Securing Software Defined Networks: Taxonomy, Requirements, and Open Issues," *IEEE Commun. Mag.*, vol. 53, no. 4, Apr. 2015, pp. 36–44.

## BIOGRAPHIES

KUAN ZHANG [S'13] received his B.Sc. degree in communications engineering and M.Sc. degree in computer science from Northeastern University, China, in 2009 and 2011, respectively. He received his Ph.D. degree in electrical and computer engineering from the University of Waterloo, Canada, in 2016. Currently, he is a postdoctoral fellow with the Department of Electrical and Computer Engineering, University of Waterloo. His research interests include security and privacy for mobile social networks, e-healthcare systems, and cloud computing.

JIANBING NI received his Bachelor and Master degrees from the University of Electronic Science and Technology of China in 2011 and 2014, respectively. Currently, he is pursuing his Ph.D. degree at the Broadband Communications Research (BBCR) Group, Department of Electrical and Computer Engineering, University of Waterloo. His research interests include security and privacy for crowdsourcing, vehicular ad hoc networks, cloud computing, and fog computing.

KAN YANG received his B.Eng. degree in information security from the University of Science and Technology of China in 2008 and his Ph.D. degree in computer science from the City University of Hong Kong in August 2013. From September 2013 to July 2014, he was a postdoctoral fellow in the Department of Computer Science, City University of Hong Kong. From July 2014 to June 2016, he was a postdoctoral fellow and the coordinator of the security group of the BBCR Group in the Department of Electrical and Computer Engineering, University of Waterloo. He will join the Department of Computer Science at the University of Memphis soon. His research interests

include cloud security, big data security, mobile security, applied cryptography, and distributed systems.

XIAOHUI LIANG [M'15] received his Ph.D. degree from the Department of Electrical and Computer Engineering of the University of Waterloo, and his Master and Bachelor degrees from the Computer Science Department of Shanghai Jiao Tong University. He was also a postdoctoral researcher at the Department of Computer Science, Dartmouth College, New Hampshire. Since 2015, he has been an assistant professor with the Computer Science Department at the University of Massachusetts Boston. His research interests include security, privacy, and trustworthiness in medical cyber physical systems, cyber security for mobile social networks, and applied cryptography.

JU REN [S'13] received his B.Sc., M.Sc., and Ph.D. degrees, all in computer science, from Central South University, China, in 2009, 2012, and 2016, respectively. From August 2013 to September 2015, he was a visiting Ph.D. student with the Department of Electrical and Computer Engineering, University of Waterloo. Currently, he is a Distinguished Professor with the School of Information Science and Engineering, Central South University. His research interests include wireless sensor networks, mobile sensing/computing, transparent computing, and cloud computing. He is the corresponding author of this article.

XUEMIN (SHERMAN) SHEN [M'97, SM'02, F'09] received his B.Sc. degree from Dalian Maritime University, China, in 1982, and his M.Sc. and Ph.D. degrees from Rutgers University, Newark, New Jersey, in 1987 and 1990, respectively, all in electrical engineering. He is a professor and University Research Chair, Department of Electrical and Computer Engineering, University of Waterloo. He was the Associate Chair for Graduate Studies from 2004 to 2008. His research focuses on resource management in interconnected wireless/wired networks, wireless network security, social networks, smart grid, and vehicular ad hoc and sensor networks. He was a recipient of the Excellent Graduate Supervision Award in 2006 and the Outstanding Performance Award in 2004, 2007, 2010, and 2014 from the University of Waterloo; the Premier's Research Excellence Award in 2003 from the province of Ontario; and the Distinguished Performance Award in 2002 and 2007 from the Faculty of Engineering, University of Waterloo. He served as the Technical Program Committee Chair/Co-Chair for ACM MobiHoc '15, IEEE INFOCOM '14, IEEE VTC-Fall '10, the Symposia Chair for IEEE ICC'10, the Tutorial Chair for IEEE VTC-Spring '11 and IEEE ICC '08, and the Technical Program Committee Chair for IEEE GLOBECOM '07. He also serves/has served as Editor-in-Chief for *IEEE Network*, *Peer-to-Peer Networking and Application*, and *IET Communications*. He is a registered Professional Engineer of Ontario, Canada, an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering Fellow, a Royal Society of Canada Fellow, and a Distinguished Lecturer of the IEEE Vehicular Technology and Communications Societies.

The collaborative efforts among municipalities, regulation departments, industry, academia and business companies are necessary to set up privacy policies and regulations. In addition, data privacy, availability and management should be achieved simultaneously.

## NEXT GENERATION 911: WHERE ARE WE? WHAT HAVE WE LEARNED? WHAT LIES AHEAD?



Carol Davids



Vijay K. Gurbani



Salvatore Loreto



Ravi Subramanyan

IP-based telecommunications services are a fact of life today, embraced by private enterprise, telecommunications carriers, and the public: IP-PBXs have replaced the traditional enterprise-based telecommunications infrastructure, mobile carriers are introducing voice over LTE, and individuals are subscribing to an increasing number of voice over IP (VoIP) telephone offerings from cable operators and over-the-top providers such as Skype. The move to IP-based telecommunications expands the functions and features that telecommunications applications and services can provide. The public is accustomed to ordering pizza, finding routes, texting, and placing video calls using their smartphones. People are generally not aware, however, that the 911 services on which they rely in emergencies are not capable of using these communications modes.

The 911 system that provides a universal emergency number for citizens of the United States has been in place since the 1960s. It is designed to route emergency calls to a public safety answering point (PSAP) where they are answered by telecommunicators who dispatch first responders (firefighters, police, EMS) to a caller's location. The caller's location is used by the switched infrastructure to route a call to the closest PSAP. The caller's location is also displayed on a screen at the telecommunicators' workstations in the PSAP to enable dispatch even in the case where the caller does not know his/her location or is unable to speak. This location, presented on the screen, is sent over a different path than the path taken by the call itself, however, since the trunks that route the call are not capable of carrying the volume of data needed to describe the caller's location.

The service, both in the United States and around the globe, was originally designed to operate over the circuit-switched network of a single telecommunications carrier. Much has changed since the service was first engineered; multiple carriers now provide access to the circuit-switched infrastructure, cellular networks now carry a majority of all emergency calls in the United States, and telecommunication is increasingly conducted over the packet-switched Internet.

To address this change and to deliver next generation emergency services, the National Emergency Numbers Association (NENA) in the United States and its global affiliates have specified a set of standards. The NENA requirements, known as the NENA i3 Standard [1] define a managed IP backbone network known as the Emergency Services IP Backbone Network (ESINet), and a set of next generation (NG) core services, also known as functional elements. Figure 1 illustrates the flow of calls and information into and through the ESINet. Session Initiation Protocol (SIP) signaling elements are shown in blue, data-

bases in yellow, and public switched telephone network (PSTN) elements in beige. Calls that arrive from IP networks use SIP for signaling and Real-Time Transport Protocol (RTP) for voice and video. Calls arriving from the PSTN traverse a legacy network gateway (LNG) that translates the signaling and media associated with the call to SIP and RTP, respectively.

In steps 1 and 2, a SIP proxy uses the Domain Name Service (DNS) to locate its emergency call routing function (ECRF), a location service that identifies the emergency services routing proxy (ESRP) through which the call signaling will flow. In steps 3 and 4, the proxy routes the call to the ESRP on the ESINet through a border control function (BCF) that protects the ESINet from attacks of various kinds. At the ESRP, an ECRF is queried (step 5) to identify which PSAP should receive the call. Note that in case the designated PSAP is not i3-compliant, the call must route back to a circuit-switched network through a legacy PSAP gateway (LPG).

Progress toward implementing NG communications services based on these specifications is slow. Many interesting engineering and operational challenges will need to be addressed, and many policy and financial ones as well. While the policy and financial issues remain unresolved, private companies have sprung up to provide emergency services to their own customers, not to the general public.

The evolution of 911 services as described above widens the scope of interest to all communications engineers, who should understand the ramifications of new technologies on what has historically been considered basic and foolproof service. This Feature Topic aims to introduce readers to the state of the art in next generation 911 (NG911), the challenges and the solutions being implemented as we bring NG emergency communications to our communities, as well as new challenges in validation and new services such as in-vehicle emergency services. Practical aspects of deployment and testing, including lessons learned, are emphasized in this set of articles in keeping with the charter of the *IEEE Communications Magazine Design and Implementation* series, in order to give practitioners examples of real-world experiences in NG emergency service deployments.

The articles in this Feature Topic discuss a variety of interesting topics related to NG911. The article by Kemp describes in detail the planning and deployment of NG911 in one instance in the United States. It shows how collaboration between industry, academia, and standards and regulatory bodies was effectively used to successfully modernize emergency services in a rural area. This example provides a template for similar efforts that are expected to take place in the future.

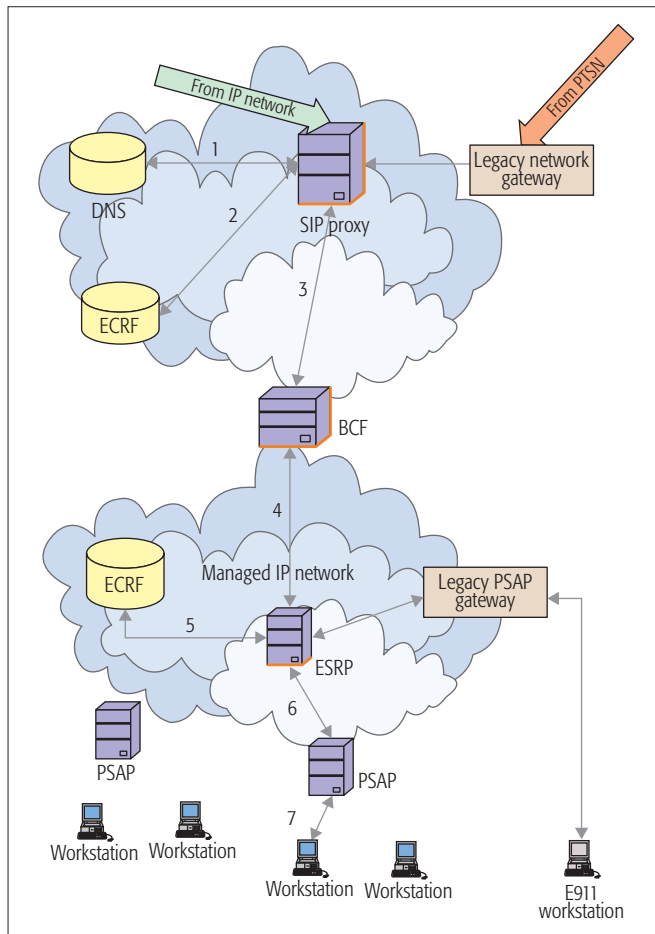


Figure 1. Access, ESInet and PSAP.

The articles by Liberal *et al.* and Markakis *et al.* discuss the challenges presented by modern communications technologies in the context of next generation emergency services. Reviewing two aspects of the European NG112 service, the article provides a view of emerging architectures for NG emergency services. Liberal *et al.* provide an overview of the architecture of the European Emergency Numbers Association (EENA) NG112 service and subtle differences in international approaches.

Markakis *et al.* present a more user-centric view of how new technologies enable non-traditional (i.e., not voice-based) emergency communication based on text, image, and video messaging within the context of a robust emergency network (as opposed to social media, where these methods are already in widespread use as long as connectivity is available).

Magnusson *et al.* discuss the extremely important aspect of validation of the new service paradigms. Together with the article by Kemp, the article describes initial experiences using university-based test labs.

Finally, the article by Öörni *et al.* discusses in-vehicle emergency calling. The article reviews the challenges that arise due to differences in national standards and ways of handling specific forms of messaging within systems that are broadly interoperable but not exactly the same. In so doing, the article illustrates the challenges of meeting customer expectations as a highly mobile population moves from country to country, and the importance of educating people about such differences in situations in which mere minutes could make the difference between life or death.

## REFERENCES

- [1] NENA, "Detailed Functional and Interface Standards for the NENA I3 Solution," 1st ed., 2016, accessed 22 Sept. 2016; [https://www.nena.org/default.asp?page=i3\\_Stage3](https://www.nena.org/default.asp?page=i3_Stage3).

## BIOGRAPHIES

CAROL DAVIDS is a professor at the Illinois Institute of Technology, Chicago, and the founder and director of the IIT Real-Time Communications (RTC) Lab. Her research includes the design and development of real-time communications applications and systems, and studies of the performance characteristics and interoperability of such systems. She chairs the annual IIT Real-Time Communications Conference and the IPTComm Conference, annual international events that foster collaboration between industry and academia. Before joining the faculty at IIT, she worked in the telecommunications industry for over 30 years, where she held positions at AT&T, Motorola, and Tellabs, leading integration test teams and developing open standards. She earned her B.S. in engineering mathematics from Columbia University, School of Engineering and Applied Science, New York, and her M.S. in information technology from Illinois Institute of Technology.

VIJAY K. GURBANI is a Distinguished Member of Technical Staff at Bell Laboratories' End-to-End Mobile Network Research Department in Nokia Networks. He holds a B.Sc. in computer science with a minor in mathematics and an M.Sc. in computer science, both from Bradley University; and a Ph.D. in computer science from Illinois Institute of Technology. His current work is focused on scalable analytic architectures and algorithms for autonomic 5G networks. His research has resulted in products that are used in national and international service provider networks. He has over 60 publications in peer-reviewed conferences and journals, 5 books, 7 granted U.S. patents, and 19 IETF RFCs.

SALVATORE LORETO [M'01, SM'09] ([salvatore.loreto@ieee.org](mailto:salvatore.loreto@ieee.org)) has more than 15 years of experience in a variety of information and communication companies, and has been working in networking and telecommunications since 1999. He works as strategic product manager within the business unity Media at Ericsson in Stockholm, Sweden. He has made contributions in Internet transport protocols (e.g., TCP, SCTP), signal protocols (e.g., SIP, XMPP), VoIP, IP-telephony convergence, conferencing over IP, 3GPP IP Multimedia Subsystem, HTTP, WebRTC, and web technologies. He is also a quite active contributor to the IETF, where he has co-authored several RFCs and has served as co-chair for several working groups. For the IEEE Communications Society, he serves as the Design and Implementation Series Co-Editor and Associate Technical Editor for *IEEE Communications Magazine*. He received an M.S. degree in engineering and computer science and a Ph.D. degree in computer networking from Napoli University in 1999 and 2006, respectively. In 2014 he received an executive M.B.A. from SDA Bocconi in Italy.

RAVI SUBRAHMANYAN ([ravi.subrahmanyan@ieee.org](mailto:ravi.subrahmanyan@ieee.org)) received M.S. and Ph.D. degrees in electrical engineering from Duke University, a B.Tech. from the Indian Institute of Technology Bombay, and an M.B.A. from MIT. He has over 50 refereed journal articles and conference publications, and holds over 20 issued patents. He has worked on various aspects of telecommunications, including hardware design and system architectures for data and video transport. He is a synchronization expert, was an Editor of an *IEEE Communications Magazine* Feature Topic on Synchronization in NG Networks, and was a presenter on the ComSoc Webinar on Next Gen Synchronization Networks. He has served on various GLOBECOM and ICC conference committees, and on ComSoc's TAOS TC since 2008, and is a Technical Editor for *IEEE Communications Magazine*.

# European NG112 Crossroads: Toward a New Emergency Communications Framework

Fidel Liberal, Jose Oscar Fajardo, Cristina Lumbreras, and Wolfgang Kampichler

Different initiatives worldwide are addressing the need for specifying a stable IP-based next generation emergency communications framework. The authors provide a general overview of the different international approaches with special focus on the European perspective, where the NG112 architecture has been specified, and it has now entered the testing and evaluation phase.

## ABSTRACT

Nowadays, citizens are getting used to new ways of agile communications supporting media-enriched and context-aware information. However, the adoption of these evolved technologies in emergency communications between citizens and public authorities faces a series of barriers, including the lack of harmonized and interoperable solutions. Different initiatives worldwide are addressing the need for specifying a stable IP-based next generation emergency communications framework. This article provides a general overview of the different international approaches with special focus on the European perspective, where the NG112 architecture has been specified, and it has now entered the testing and evaluation phase.

## INTRODUCTION

It is estimated that approximately 320 million emergency calls are made every year in the European Union (EU). Among all the different numbers still available, more than 135 million [1] used the unified 112 one (note that some member states do not provide such statistics). At the same time, multimedia applications and voice over IP (VoIP)-based devices have become commonplace, and citizens use them to conveniently communicate, sending and receiving multi-modal information. Among several standardized and proprietary over-the-top (OTT) VoIP solutions available today, voice over LTE (VoLTE) seems to be playing a significant role in the near future, as predominant access to emergency services from fourth generation (4G) broadband mobile networks. In fact, more than 60 percent of emergency calls are already from mobile devices in the EU [2].

Unfortunately, for the time being, most European emergency service organizations (ESOs) can only be reached by voice and through the public switched telephony or mobile networks. Meanwhile, different kinds of text messaging and video and picture sharing apps have become more common communication means, and social networks have indeed become a new medium by themselves. Modern mobile phones from which an emergency call might be placed have the potential to transmit life saving location information simultaneously with the call.

As a result, the way citizens and emergency

services interact is undergoing significant changes in terms of communication means and content. Similarly, the number of stakeholders involved in emergency communications and the technologies used for interacting among them has significantly increased, leading to a complex heterogeneous ecosystem. Nevertheless, the existing legacy emergency services infrastructure (circuit switched telephony for 112 telephone calls, not data) is not designed in a way that enables interaction with enhanced services, or current and future communications and operational requirements to be met. In such a landscape, more rapidly changing than ever, ESOs are struggling to identify a consistent and longstanding set of solutions that would leverage their current legacy systems while keeping systems interoperable, scalable, and technically stable.

The concept of Next Generation 911/112 (NG911/NG112) has been identified as a potential answer to such demands, since it combines the definition of a set of international standards with the scalability and flexibility of IP connections. Several initiatives all over the world have been initiated for both defining and testing such a concept.

NG911/NG112 systems are designed to close the gap between the quickly evolving technologies (fixed and mobile IP-based communications) and the more conservative approaches required by the emergency communications industry (including public administrations). Such systems will then enable citizens to contact emergency services in different ways, using the same types of technologies as they use to communicate every day. They will also make it possible for public safety answering points (PSAPs) to receive more effective and richer information about emergencies of all magnitudes through a stable communications framework, which improves interoperability between emergency services. Consequently, response times and operational costs will be reduced, while effective response will increase significantly.

In this article we analyze the current status of next generation emergency communications with special focus on Europe, where the first European-wide industry-driven interoperability initiative was recently launched. This interoperability event was hosted by the European Telecommunications Standards Institute (ETSI) and takes NG112 archi-

ture as a basis for next generation emergency communications. NG112 was originally proposed by the European Emergency Number Association (EENA) a European NGO dedicated to promoting high-quality 112 emergency services, similar to the National Emergency Number Association (NENA) in the United States. Among all the different boards, it is especially the Next-Generation 112 Committee (NG112) that is working on designing the future of IP-based emergency services.

The article foresees the challenges and the role of the NG112 structure, identifying the border elements as key enablers of smooth evolution in a scenario where evolving access technologies and heterogeneous use cases are called to become key elements in the emergency ecosystem.

The structure of the article is as follows. The following section identifies the novel user requirements concerning emergency communications, highlighting the main work performed in different standards development organizations (SDOs). Then we give an overview of the most relevant approaches by different SDOs worldwide for defining the next generation emergency communications framework. Following that we summarize the NG112 architecture proposed by EENA and briefly describe the main functional nodes. Based on the outcomes of the first industry interoperability event, we then identify the current status concerning different types of access networks. Finally, we provide the conclusions to the article.

## EVOLVING REQUIREMENTS FOR NEXT GENERATION COMMUNICATIONS BETWEEN CITIZENS AND PSAPs

User requirements in the design of a technology have always been of paramount importance in the deployment of emergency communications. Therefore, the identification of user requirements is the initial step prior to the technical definition of a system. In recent years, a wide number of international SDOs and research projects have performed surveys for obtaining user requirements for emergency communication systems.

The European Telecommunications Standards Institute (ETSI) Special Committee (SC) on Emergency Communications (EMTEL) has been specifically focused on emergency communications, including emergency call services, caller location enhanced emergency services, and public safety communication systems. As a result, the requirements between citizens and PSAPs and between PSAPs during emergency communications, at both the functional and operational levels, were collected in [3]. The EENA NG112 Technical Committee also conducted a series of surveys of different members of European emergency services regarding emergency services requirements. The outcomes of these surveys are publicly available in [4, 5]. Additionally, user requirements have also been taken into account in the definition of the Third Generation Partnership Project (3GPP) Non Voice Emergency Services (NOVES) characteristics [6]. More specifically, among other requirements, location information should be provided by users at call setup and instantly updated. NOVES services shall be free of charge, as any

other emergency call, while emergency communications have to be prioritized over other communications. At the same time, in cases lacking 4G coverage, voice and location should always be available.

In recent years, a large number of “SOS” and “help” applications have been created. Almost all European emergency services have been contacted by developers who wanted to send data and establish a voice connection directly to 112 [7]. Different solutions have been built on heterogeneous technologies that are not generally interoperable. This may explain why some public authorities have already developed their own official applications that can only be used by citizens living in a certain geographic area and may not work properly if they are used outside the boundaries of a certain PSAP.

The use of standardized or industry-adopted technologies may help to overcome this heterogeneity in emergency apps and OTT VoIP services. Different technologies can be identified as prevailing candidate solutions, including the use of webRTC or the mobile industry supported Rich Communications Suite (RCS). However, none of the solutions have really gained the required wide support in the emergency communications world.

To add more complexity to the unified emergency communications playground, the use of crowd sourcing and social networks is becoming more and more frequent in emergency situations [8]. Public authorities are trying to get prepared to seamlessly include this kind of communication means, mainly based on private mobile applications, in their daily operations. However, the lack of a standardized architecture forces public authorities to integrate ad hoc solutions for the different technologies. Additionally, the proliferation of emergency-related online user groups and volunteers indicates the need for a harmonized media-enhanced communication framework between citizens and PSAPs. The Internet of Things (IoT) will soon add new players to the emergency communications ecosystem

eCall can well be considered the first major industry initiative in this sense. 3GPP introduced the requirement of identification of eCall in the support of emergency calls in 2007 and approved the in-band modem solution in 2009. On 28 April 2015, the European Parliament voted in favor of eCall regulation, which requires that all new cars be equipped with eCall technology from April 2018. eCall specifications have now been stable for several years, covering circuit-switched 112 eCall over 2G and 3G mobile networks. However, a more evolved version of eCall is already being defined, taking advantage of the IP-based 3G and 4G communication technologies.

In summary, public authorities and PSAP professionals are aware of the benefits that modern mobile technologies can provide to the emergency communications framework. However, there is a lack of a unified stable technological framework that eases the adoption of the heterogeneous sources of information, which would be undoubtedly valuable for emergency management operations. Most of the available approaches are based on ad hoc solutions, generally not interoperable among them and with current PSAPs.

User requirements in the design of a technology have always been of paramount importance in the deployment of emergency communications. Therefore, the identification of user requirements is the initial step prior to the technical definition of a system.

Public authorities will rely on a stable technological framework for incoming emergency communications. The different emergency communication details specific to the access technologies will be mapped to the ESInet protocol suite by the defined border controllers.

## STANDARDS FOR IP-BASED EMERGENCY CITIZEN-TO-PSAP COMMUNICATIONS

Concerning next generation citizen-to-authority emergency communications, several relevant standardization initiatives have been launched worldwide. In general, all the initiatives tend toward IP-based scenarios, with Session Initiation Protocol (SIP) as the key technology for communication signaling and control.

In the Internet Engineering Task Force (IETF), the Emergency Context Resolution with Internet Technologies (ECRIT) working group has been mostly working on SIP-based citizen-to-authority emergency communications. Since its creation in 2005, ECRIT has released a significant number of Internet Drafts and RFCs including the definition of the framework [9], the specification of the signaling protocol details [10], the specification of call routing by means of the emergency service routing proxy (ESRP) node, the inclusion of location capabilities (in cooperation with the GEOPRIV WG) within the call [11], routing by means of the location-to-service translation (LoST) protocol [12], and so on.

Concerning the 3GPP, public safety (individual to/from authority) communications include eCall, public warning systems (PWs), multimedia priority service (MPS), and so on. Specifically, 3GPP TS 23.167 defines the architecture and procedures for establishment of citizen-to-PSAP emergency services in IMS since Release 7 (June 2006), by means of introducing a SIP session control node named the emergency call session control function (E-CSCF). Enhanced functionalities have been added through new releases; for example, LTE-specific support for IMS emergency services was introduced in Release 9, and enhanced emergency calling through WLAN is defined in Release 13. Additionally, 3GPP defines the architecture for location services (LCS) where the location resource function (LRF) is the network element responsible for providing user location information to other entities. More specifically, 3GPP TS 24.229 describes different methods to include location information in IMS signaling, while enhanced user location reporting (indoor and outdoor) is being defined in 3GPP Release 14. It must be stated that many of the IMS emergency protocol specifications are mainly based on IETF ECRIT's RFCs adapted to the IMS procedures.

In the scope of ETSI, the more relevant groups are EMTel and lately the Technical Committee on Network Technologies (NTECH). While the former is more related to user requirements and general specification of the emergency calling context, the latter works on the specifications of the interfaces surrounding the network architecture and protocol details to support location in emergency calling. In general, ETSI originally adopted the 3GPP's specifications involving IMS emergency architecture, but is currently further working on specific requirements due to mandate M/493.

NENA, back in 2000, already detected the need to develop, expand, and improve emergency communications in North America. NENA has been working since 2006 on its own research and development initiative to promote NG911, defining the system architecture and a transition plan that comprises costs, responsibilities, sched-

ule, and benefits derived from the deployment of a nationwide evolved emergency network. This NG911 standardized system permits the transmission of both voice and non-voice multimedia data from various devices: wired, wireless, VoIP, sensors, and so on. NG911 utilizes an IP-based network technology to connect different emergency agencies and citizens to a system capable of offering a wide range of emergency services and access to advanced data. The so-called Emergency Services IP Network (ESInet) comprises a broadband packet-switched core network. The Functional Interface Standards for NG911 (i3) comprise a set of standards that define the core IP functionality of the NG911 system based on standards from IETF and other organizations (e.g., SIP for session control). Some examples of the functional entities included in the NENA i3 architecture are: the location information server (LIS), providing the location of the endpoints; the emergency call routing function (ECRF), which is based on the location of the call, provides the information to contact the corresponding PSAP and ESRP; and a SIP proxy server that routes the calls using location and policy rules.

It is also worth mentioning the efforts being performed within the ATIS Next Generation Emergency Services Subcommittee (NGES) of the Emergency Services Interconnection Forum (ESIF), which is working on closing the gap between the NENA i3 architecture and the 3GPP IMS standards for emergency calling through commercial mobile broadband networks.

In Europe, the production of the Long-Term Definition Document (LTD) by EENA [13] was the first comprehensive attempt to describe the technicalities and potentialities of a structured approach. Due to the relevance of this initiative for the NG112 ecosystem in the EU, it is further analyzed in the following section.

Almost at the same time (May 2011), the European Commission (EC) sent the M/493 standardization mandate to the European standards organizations, referring in particular to Article 26 of the Universal Service Directive 2002/22/EC on emergency services and the single European emergency call number as amended by Directive 2009/136/EC. In early 2012 ETSI created a work item to address the M/493 requirements taken over by the ETSI project End-to-End Network Architectures (E2NA). ETSI published in 2015 the "Functional architecture to support European requirements on emergency caller location determination and transport" ES 203 178 [14], where the requirements and functional architecture are described. The stage 3 document, "Protocol specifications for emergency service caller location determination and transport," is currently being drafted by the Technical NTECH working group.

In summary, it can be observed that the major standardization efforts concerning next generation citizen-to-PSAP communications are based on IP-based networks and SIP communications, with different flavors and architectural specifications.

### EENA NG112 LONG TERM DEFINITION

The EENA NG112 committee released the first public version of the Next Generation 112 Long Term Definition document in 2012. To ensure global interoperability, EENA reused existing expe-



periences from other regions. In particular, the work from NENA was adapted to European PSAPs. The current NG112 LTD document, released in 2013, defines a long-term architecture for European emergency services and remains voluntarily close to the NENA i3 standard.

The NG112 LTD document describes the end state that has been reached after migration from legacy circuit-switched telephony, and the legacy E112 system built to support it, to an all-IP-based telephony system with a corresponding IP-based emergency services IP network.

The high-level NG112 LTD architecture and main functional elements (FEs) are illustrated in Fig. 1. Comprehensive message flows explaining how the emergency calls arrive at the appropriate PSAP are collected in [13].

Various originating networks and heterogeneous devices are able to trigger emergency communication toward the PSAPs, which are inter-connected through the NG112 ESInet. The different access networks considered include OTT VoIP providers, IMS/VoLTE operators, enterprise networks using unified communications (UC), as well as legacy public switched telephone networks (PSTNs). The standardization of emergency calling through these access networks is outside the scope of the NG112 LTD document. The NG112 LTD document focuses on clearly specifying a limited subset of protocol headers, messages, and procedures to be considered within the ESInet. In that way, public authorities will rely on a stable technological framework for incoming emergency communications. The different emergency communication details specific to the access technologies will be mapped to the ESInet protocol suite by the defined border controllers.

The main FEs included in the NG112 architecture are:

- ESRP, which is the SIP entity that makes decisions about the call routing by using location information.
- ECRF, which is the FE that provides the PSAP address to route an emergency call.
- The border control function (BCF), which is actually in charge of adapting the incoming emergency calls from the different access networks to the ESInet requirements. Additionally, the BCF acts as a border controller in both the signaling and media planes.
- The legacy network gateway (LNG), which acts as a border controller for legacy PSTN networks, converting the emergency calls to SIP.
- The location information server (LIS), which provides the user location functions in the scope of the ESInet.

The specific details of all the FEs and protocols involved in the different interfaces are clearly described in the NG112 LTD document [13].

Providing a converged network for different access networks, the NG112 ESInet supports several variations of end-to-end emergency communications with a series of objectives that need to be supported to ensure interoperability.

**Connectivity:** The NG112 system shall cover basic connectivity between FEs at the network and application levels. The application level refers to signaling and media transport protocols in use. This feature may require protocol translation to

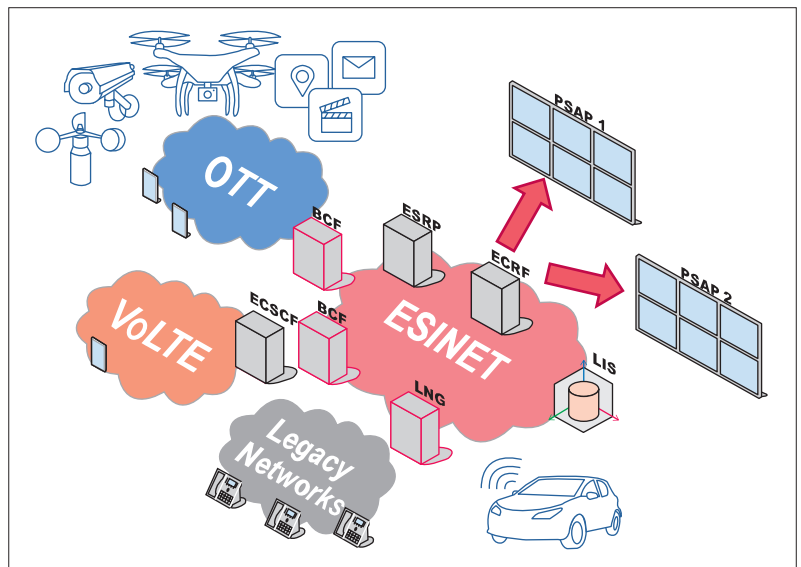


Figure 1. NG112 LTD high level architecture.

the ESInet SIP subset, including different security and privacy schemes.

**Routing:** The NG112 system shall cover variants of location-based emergency call routing. These include different methods to assess user location and how this information is delivered to emergency services. Location by value (LBV) and location by reference (LBR) are the two alternative ways supported for location conveyance. When required, the BCF shall adapt the incoming location information or include basic location information when not provided in the incoming emergency call.

**Media:** The NG112 system shall cover different media types in order to contact emergency services, including audio, video, text, messaging, and additional data. When needed, media transcoding shall be supported by the BCF.

**Policy:** The system shall cover variants of policy-based emergency call routing. A major strength of next generation emergency communication is advanced call routing features that allow re-targeting emergency calls based on time of day, call volume, and queue or element state.

**Quality:** The NG112 system shall cover quality aspects with respect to emergency calling. These are, among others, successful call setup, call setup time, and media quality including the use of SIP preconditions.

**Logging and Recording:** The NG112 system shall cover logging and recording aspects with respect to emergency calling. These are, among others, successful media recording and event logging.

It must be noted that the scopes of the EENA NG112 and ETSI NTECH working groups are different. ETSI NTECH is focused on the standardization of a general solution for emergency caller location acquisition and transport, which is valid for heterogeneous deployment alternatives including current and next generation communications systems. While NENA and EENA architectures are based on the concept of a unified ESInet with a common SIP-based signaling suite, ETSI NTECH develops its solution taking into account the complex deployment context where each European

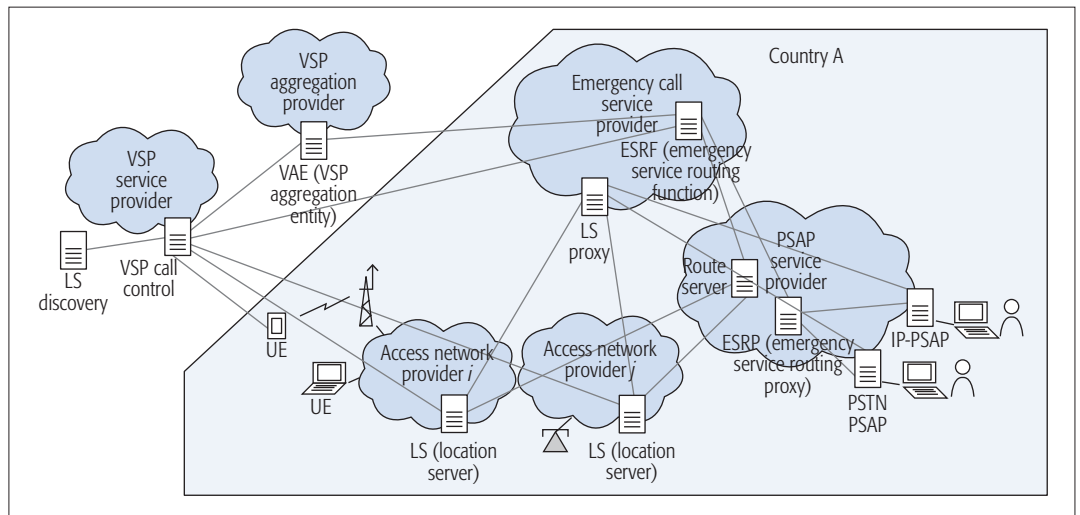


Figure 2. ETSI 203 178 architecture in response to M/493.

### IP-BASED ACCESS TO THE ESINET

This configuration is used for basic emergency call routing where calls originate from an IP network that connects to a PSAP, as shown in Fig. 3.

Any UE registers with the SIP Proxy, and the SIP Proxy forwards emergency calls to a configured BCF in the ESInet. This node comprises both signaling and media interfaces, and, as it remains in both the signaling and media path, events can be logged and media can be recorded.

These emergency calls are routed to the corresponding PSAP based on location information, which is retrieved from an LIS by either the UE or any capable ESInet FE. LTD references two different protocols to retrieve the location information from the LIS server: HELD or SIP using the presence event package. The location information is provided either by value or by reference. If LVB is provided, geodetic or civic location information is included in the call establishment signaling messages. When LVR is used, the reference added in the call establishment should be dereferenced in order to obtain the actual location information.

If the IP network does not include an LIS, the UE may be able to interface with the LIS deployed at the ESInet to provide location-related information.

However, the major issue confronted when accessing the ESInet from heterogeneous IP networks is the variety of IP/VoIP protocols used by different 112 applications which should be handled correctly by BCF. Currently, LTD defines the output of BCF toward the ESInet, but not all the different inputs.

Facing the lack of a harmonized European-wide approach for 112 applications, EENA promotes the clear definition of a Pan-European Mobile Emergency Application (PEMEA), which provides a functional architecture, and defines roles and responsibilities as well as data exchange formats and a general security model so that PSAPs can be sure of the veracity of the information being provided, and application users can be sure that information is not being misused. The additional caller information and the use of NG112 data formats foster the use of PEMEA toward the implementation of NG112 emergency services.

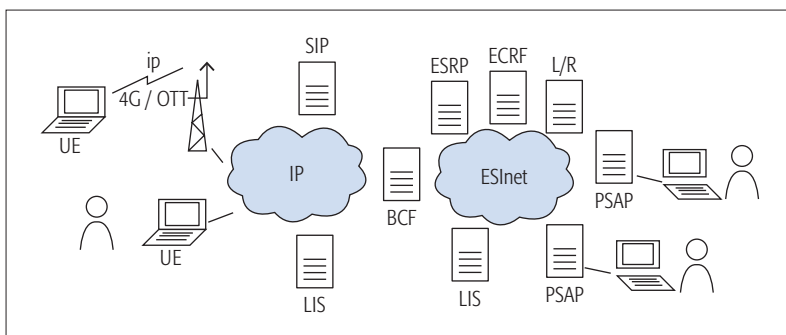


Figure 3. IP-based access to the NG112 ESINET.

PSAP may be served by one or more national network operators (Fig. 2). Due to national regulatory requirements, network-provided location information is used, and the voice service provider (VSP) gets the routing information from the location server (LS) based on a new extension to HTTP-enabled location delivery (HELD) [15].

### EUROPEAN-WIDE INTEROPERABILITY TESTING INITIATIVE

In March 2016, ETSI and EENA co-organized the first emergency communications interoperability test in Europe, the first of a series of planned NG112 Emergency Services Plugtest events supported by the EC. Thirteen vendors participated in the tests, providing different FEs including different types of user equipments (UEs) and PSAPs, mobile apps, IP/IMS/UC/PSTN access networks and different ESInet elements such as BCF, LIS, ESRP, and ECRF. In addition to NG112 industry partners, seven external international observers attended the event in order to get first-hand experience from the current situation of the NG112 developments.

In order to validate the NG112 architecture and to evaluate the maturity of the different commercial solutions already available, a wide range of tests over different access networks were performed.

The main use cases and the preliminary outcomes from the first test event are provided in the following paragraphs.

### VOLTE-BASED ACCESS TO THE ESINET

This configuration is used for basic emergency call routing where calls originate from an IMS that connects to an ESInet, as shown in Fig. 4.

Any UE registers with the IMS (emergency bearer), and the IMS E-CSCF forwards emergency calls to a configured BCF, which acts as a border controller at signaling and media planes. The specific SIP messages and data formats included between the IMS E-CSCF and the ESInet BCF are not fully standardized, and they heavily depend on the security and trust associations between the two networks.

Location information is either provided by the IMS or retrieved from the LIS by any capable FE within the ESInet. The actual 3GPP standardized node for location conformance and PSAP selection is the LRF. Emergency calls are detected in the IMS network and forwarded to the E-CSCF node, which contacts the LRF to validate the location provided in the call establishment messages and to select the appropriate PSAP. However, it is not yet standardized in 3GPP how to forward emergency calls to non-IMS PSAPs. Another challenge to be faced is when location is provided by reference, since 3GPP has not defined how to access to LRF from external PSAPs (Le) if possible.

### UC-BASED ACCESS TO THE ESINET

This configuration is used for basic emergency call routing where calls originate from a UC that connects to an ESInet, as shown in Fig. 5.

Any UE registers with the enterprise UC (soft switch), and the UC forwards emergency calls to a configured BCF. Again, the interface between the UC nodes and the ESInet BCF are not specified, and the BCF is required to understand the specific SIP flavor deployed in the enterprise to implement the signaling and media plane.

In this case, location information may or may not be provided by the UC nodes. Location information is either provided by the UC or retrieved from the LIS by any capable node within the ESInet.

### PSTN-BASED ACCESS TO THE ESINET

This configuration is used for basic emergency call routing where emergency calls originate from a PSTN that connects via an LNG to an ESInet, as shown in Fig. 6.

Any UE-triggered circuit-switched emergency call terminates at the LNG, and the LNG is responsible for forwarding the call to a configured BCF based on SIP/IP communications. At the signaling and media planes, the LNG is split into a protocol interworking function (PIF) and a NG112 interwork function (NIF). Depending on the compliance of the NIF to the NG112 specifications, the role of the BCF may differ among different implementations.

Additionally, the LNG includes a location interwork function (LIF) that provides user location information to be included in the signaling plane messages or accessed by the ESInet nodes.

## CONCLUSIONS

Mobile broadband technologies are quickly evolving, adding the possibility for end users to adopt enhanced multi-modal communications in their daily communications. However, the use of these novel multimedia capabilities is hardly incorporat-

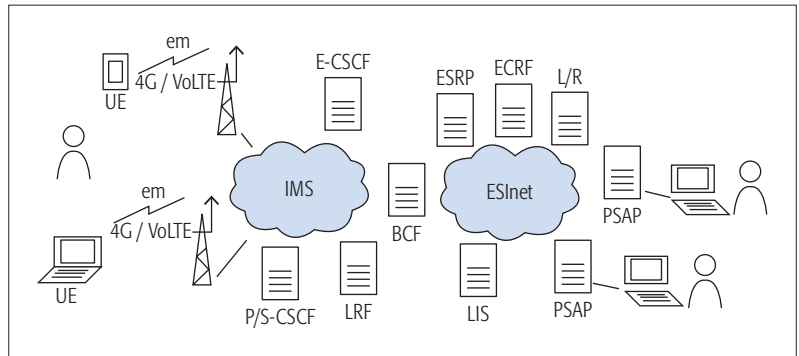


Figure 4. VolTE-based access to the NG112 ESInet.

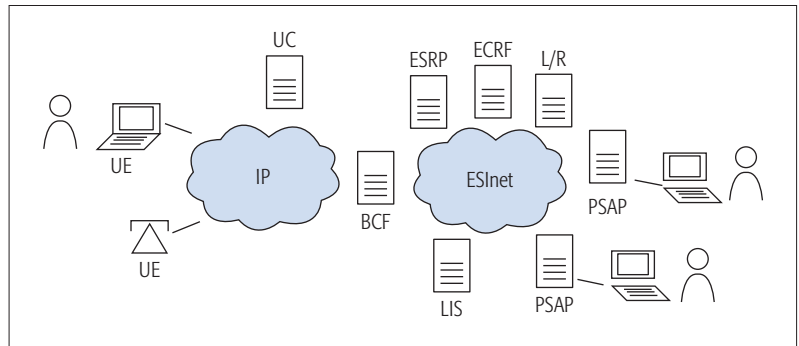


Figure 5. UC-based access to the NG112 ESInet.

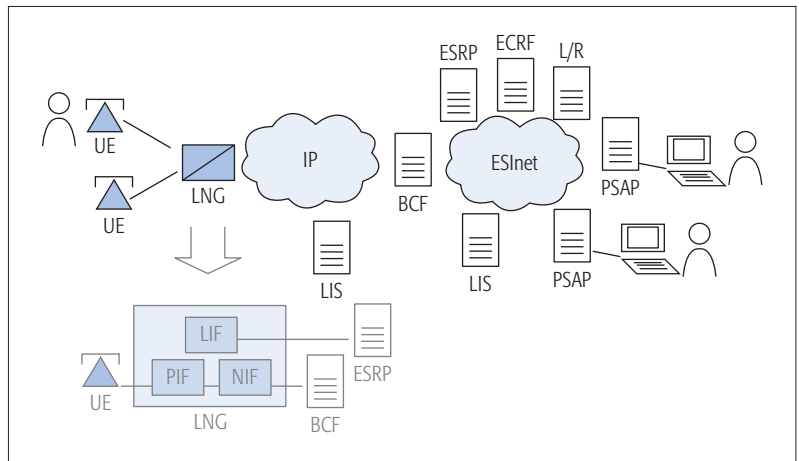


Figure 6. PSTN-based access to the NG112 ESInet.

ed into the overall management of emergency services due to the lack of standardized solutions. Public administrations generally require more solid communications frameworks with medium-/long-term stability.

In the last few years, different international standardization initiatives have been aimed at specifying a common playground for next generation emergency communications based on IP and SIP communications.

In Europe, the European Emergency Number Association released the NG112 Long Term Definition document, which is targeted to close the gap between the evolved needs of the end users and the public authorities in emergency management. The proposed NG112 architecture looks for the highest possible compatibility with international standards in order to foster interoperability between the involved players.

The EENA NG112 LTD solution has now been passed to a first testing phase, where different interoperability events will be organized under the auspices of ETSI and with the support of the EC. As a result of these interoperability events, it is expected to validate the maturity of the architectural solution and the associated commercial products.

The EENA NG112 LTD solution has now passed to a first testing phase, where different interoperability events will be organized under the auspices of ETSI and with the support of the EC. As a result of these interoperability events, it is expected to validate the maturity of the architectural solution and the associated commercial products. Additionally, these events will provide valuable feedback to re-design and fine-tune the NG112 architecture.

As a step toward the adoption of the NG112 LTD document as a European-wide solution, and after the experimental validation of the proposal, it is expected that the document will be submitted for consideration as an ETSI standard. In this sense, the NG112 LTD solution needs to fulfill the specifications related to caller location procedures provided by the ETSI NTECH working group, with special focus on the ESInet-based deployment.

#### ACKNOWLEDGMENTS

The authors would really like to thank the collaborative effort of all the contributors to EENA's technical documents, and especially Sebastian Mueller from ETSI for his support and valuable comments during NG112 plugtest and wrap-up sessions.

#### REFERENCES

- [1] Directorate-General for Communications Networks, Content and Technology, EC, "COCOM 13-04 REV1 – Implementation of the European Emergency Number 112 – Results of the Ninth Data-Gathering Round," Mar. 2013; [http://ec.europa.eu/newsroom/dae/document.cfm?doc\\_id=1674](http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=1674), accessed 8 Sept. 2016.
- [2] EENA, "Annual Report 2015," Jan. 2016; [http://www.eena.org/download.asp?item\\_id=163](http://www.eena.org/download.asp?item_id=163), accessed 8 Sept. 2016.
- [3] European Telecommunications Standards Institute (ETSI), "TR 102 180 V1.3.1 – Emergency Communications (EMTEL); Basis of Requirements for Communication of Individuals with Authorities/Organizations in Case of Distress (Emergency Call Handling)," Sept. 2011.
- [4] ETSI, "TS 102 181 V1.2.1 – Requirements for Communication between Authorities/Organizations during Emergencies," Feb. 2008.
- [5] ETSI, "TS 102 182 V1.4.1 – Requirements for Communications from Authorities/Organizations to the Citizens during Emergencies," July 2010.
- [6] 3GPP, "TR 22.871 V11.3.0 – Study on Non-Voice Emergency Services," Oct. 2011.
- [7] EENA, "Operations Document, 112 Smartphones Apps," Feb. 2014; [http://www.eena.org/uploads/gallery/files/operations\\_documents/2014\\_02\\_25\\_112smartphoneapps.pdf](http://www.eena.org/uploads/gallery/files/operations_documents/2014_02_25_112smartphoneapps.pdf), accessed 8 Sept. 2016.

- [8] A. L. Hughes, L. Palen "The Evolving Role of the Public Information Officer: An Examination of Social Media in Emergency Management," *J. Home. Sec. Emerg.*, vol. 9, no. 1; June 2012.
- [9] B. Rosen *et al.*, "Framework for Emergency Calling Using Internet Multimedia," IETF RFC 6443, Dec. 2011.
- [10] H. Schulzrinne, "A Uniform Resource Name (URN) for Emergency and Other Well-Known Services," IETF RFC 5031, Jan. 2008.
- [11] J. Polk *et al.*, "Location Conveyance for the Session Initiation Protocol," IETF RFC 6442 (Proposed Standard), Dec. 2011.
- [12] T. Hardie *et al.*, "LoST: A Location-to-Service Translation Protocol," IETF RFC 5222, Aug. 2008.
- [13] EENA, "Long Term Definition Document," Apr. 2012; [http://www.eena.org/uploads/gallery/files/pdf/eena\\_ng112\\_longtermdefinition.pdf](http://www.eena.org/uploads/gallery/files/pdf/eena_ng112_longtermdefinition.pdf), accessed 8 Sept. 2016.
- [14] ETSI, "ES 203 178 – Functional Architecture to Support European Requirements on Emergency Caller Location Determination and Transport," Feb. 2015.
- [15] J. Winterbottom *et al.*, "A Routing Request Extension for the HTTP-Enabled Location Delivery (HELD) Protocol," IETF RFC 7840, May 2016.

#### BIOGRAPHIES

FIDEL LIBERAL received his Ph.D. from the UPV/EHU in 2005. He was the project coordinator in FP7-SEC-GERYON (2012-2014) [www.sec-geryon.eu](http://www.sec-geryon.eu) dealing with public safety interoperability problems. He currently works as a lecturer and researcher in the Faculty of Engineering of Bilbao and cooperates in different international projects. His research interests include next generation emergency networks and 5G, and he has co-authored more than 35 conference and journal papers.

JOSE OSCAR FAJARDO received his Ph.D. degree from the UPV/EHU in 2016, where he works as a research fellow in the Department of Communications Engineering at the Faculty of Engineering in Bilbao. He works in adaptive management of mobile multimedia services under the framework of IMS, with special interest in mission-critical communications and 5G. He has co-authored more than 35 journal and conference papers since 2005.

CRISTINA LUMBREAS holds an M.Sc. in computer science and an M.A. in marketing, both from the Autonomous University of Madrid. She joined EENA in October 2010, where currently she coordinates all of the technical and emergency services' operations activities. She also represents EENA in the international standardization development organizations. Prior to joining EENA, she served as IT manager of Madrid's 112 Centre where she designed the IT components.

WOLFGANG KAMPICHLER received his Ph.D. from the Vienna University of Technology in 2002. He currently contributes to standardization working groups in Public Safety and Air Traffic Management, and is a recognized expert on NG9-1-1/NG112 and a speaker at public and consumer events addressing topics related to VoIP and emergency services with Internet technology. He has also chaired and co-chaired the Planning Committee for EENA Industry Collaboration Events, and since 2014, he has also chaired the EENA Technical Committee.

# EMYNOS: Next Generation Emergency Communication

Evangelos K. Markakis, Asimakis Lykourgiotis, Ilias Politis, Anastasios Dagiuklas, Yacine Rebahi, and Evangelos Pallis

## ABSTRACT

Current emergency systems and 112 services are based on legacy telecommunication technologies, which cannot cope with IP-based services that European citizens use every day. Some of the related limitations are partial media support, the lack of integration of social media, and the use of an analog modem for providing eCall services with limited data amounts. As most operators have started migrating toward broadband IP-based infrastructures, current emergency systems also need to be upgraded and adapted in order to fulfill regulatory requirements in terms of next generation emergency services. This article presents the EMYNOS project, the goal of which is the design and implementation of a next generation platform capable of accommodating rich-media emergency calls that combine voice, text, and video, thus constituting a powerful tool for coordinating communication among citizens, call centers, and first responders. Additionally, issues such as call routing/redirection to the closest available call center, retrieval of caller location, support for people with disabilities, and integration of social media are detailed.

## INTRODUCTION AND CONTEXT

Telecommunication networks are currently the primary infrastructure for providing emergency services. These emergency systems are based on old-fashioned telecommunication technologies that cannot cope with the IP-based services that the average European citizen uses every day. Furthermore, most telecommunication operators and providers have decided to migrate from circuit-switched networks to packet-switched networks after realizing the tangible benefits, which include convergence, rich services, cheaper maintenance, and improved user satisfaction. As next generation networks (NGNs) are replacing the current telecommunication networks, it follows that the current emergency systems need to be upgraded as well in order to fulfill the NGN regulatory requirements in terms of emergency services.

The NGN technologies make use of the best of both worlds: the flexibility, efficiency, and innovativeness of IP networks, and the quality of service (QoS), security, reliability, and customer-friendly features of legacy networks. The transition from circuit-switched telephony to IP telephony requires the provision of the same functionalities

already offered in circuit-switched networks. This applies, in particular, to emergency services. As public switched telephone networks (PSTNs) will be removed in the future (this is expected to be achieved by 2020), operators are obliged to provide emergency services in IP networks as well. In many countries, this is already regulated by the government or on the way to being regulated.

In this respect, this article presents EMYNOS (<http://www.emynos.eu/>), a next generation emergency management platform capable of accommodating rich-media emergency calls that combine voice, text, and video, thus constituting a powerful tool for coordinating communication among citizens, call centers, and first responders. Additionally, issues such as call routing/redirection to the closest available call center, retrieval of the caller location, hoax call prevention, support for people with disabilities, and integration of social media are addressed.

## BACKGROUND AND RELATED WORK

### CURRENT EMERGENCY

#### COMMUNICATION SCENE AND BEYOND

The International Telecommunications Union Telecommunication Standardization Sector (ITU-T), in Recommendation Y.2001 [1], states that an NGN is a packet-based network able to provide telecommunication services and able to make use of multiple broadband, QoS-enabled transport technologies, and in which service-related functions are independent of underlying transport-related technologies.

Most network operators and providers are migrating to replacing the current telecommunication networks, thus removing today's limitations, which are summarized as follows:

- There is no standard underlying technology for separate emergency systems.
- There is a lack of international access to national emergency centers.
- There is no interconnection among public safety answering points (PSAPs), which unfortunately limits the transfer of calls in case of congestion and network outage.
- Media limitation means that currently only voice calls and sometimes SMS are accepted.
- No unified platform leads to emergency warning systems currently being completely separate from the 112 emergency centers.

The authors present the EMYNOS project, the goal of which is the design and implementation of a next generation platform capable of accommodating rich-media emergency calls that combine voice, text, and video, thus constituting a powerful tool for coordinating communication among citizens, call centers, and first responders.

As NGNs are replacing the current telecommunication networks, it follows that the current emergency systems need to be upgraded as well in order to fulfill the NGN regulatory requirements in terms of emergency services.

- There are no advanced features such as caller location and support of end users with special needs (e.g., disabled people).
- Emergency calls are unidirectional; they are established from the end users toward the PSAP.
- No non-telecommunication platform is available as a backup in case the telecommunication infrastructure is not operational.
- There is no integration of social media: handling an emergency situation should not only be the task of rescue teams. Involving citizens, especially through social media (Twitter, Facebook, etc.) in monitoring events and sharing information will lead to better management
- The eCall (the emergency solution for vehicles in case of crash) technology is based on the Global System for Mobile Communications (GSM), which limits the amount of emergency data that can be sent.

As NGNs are replacing the current telecommunication networks, it follows that the current emergency systems need to be upgraded as well in order to fulfill the NGN regulatory requirements in terms of emergency services. As a consequence, next generation emergency services have the following needs:

- Improved natural disaster management, including the prevention of and response to potential terrorist actions
- Full support of new communications and information technology for emergency services, especially since today, millions of cell phone subscribers and commercial vehicles with Global Positioning System (GPS) and communications systems can provide precise locations and verbal descriptions of emergencies
- Enhancement of emergency systems with the appropriate security mechanisms in order to face attackers and prevent them from generating automated emergency calls and carrying out attacks against the network
- Improved accessibility and increased compatibility to ensure that all citizens have access to the emergency response system, including those with disabilities

#### CURRENT STANDARDIZATION IN NEXT GENERATION EMERGENCY COMMUNICATIONS

The National Emergency Number Association (NENA; <http://www.nena.org/>) and European Emergency Number Association (EENA; <http://www.eena.org/>) are organizations promoting a universal emergency service number in the United States and Europe, respectively. To provide guidance to standards development organizations, NENA issued two main documents about the Next Generation 911 (NG 911) architecture known as i2 and i3. The i2 specification describes the short-term architecture for 911 systems. It deals with the migration of emergency services where the access network is an IP network, and the emergency service provider network (PSAP's network) is still circuit-switched. The i3 specification [2] describes a complete redesign of the entire 911 systems toward NGNs (i.e., NG911). It deals with the long-term architecture, where both the access network and the emergency service provider network are based on IP.

On the other hand, in Europe, a significant step toward achieving the vision of NG112 is the EENA long-term definition (LTD) [3] of a European emergency services architecture. LTD is based on the NENA i3 architecture, which was achieved in cooperation with the Internet Engineering Task Force (IETF) Emergency Context Resolution with Internet Technologies (ECRIT) Working Group [4] and describes a framework based on existing protocols for emergency calling using Internet multimedia. Additionally, 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.

In this respect the core routing functional entities needed for NG112 call delivery include a border control function (BCF), an emergency services routing proxy (ESRP), and an emergency call routing function (ECRF). The BCF is the component that will be deployed between external networks and the ESInet, and between the ESInet and the agencies' networks. The BCF will be used as a border firewall and a session border controller to perform network edge control and Session Initiation Protocol (SIP) message handling. The ESRP is the base routing function for emergency calls. The function of the ESRP is to route a call to the next hop. It might be possible that one or more intermediate ESRPs will exist at various hierarchical levels in the ESInet. Finally, in NG112, emergency calls will be routed by the ECRF to the appropriate PSAP based on the location of the caller. In short, the ECRF takes the location information and service uniform resource name (URN) received in a routing query and maps it to the destination URI for the call. To do so, it uses the Location to Service Translation (LoST) protocol [5] used by both NENA's i3 and EENA's NG112 LTD, making it a widely accepted solution for emergency service resolution.

Finally, a crucial factor in the context of emergency services is location information, which can be either inserted by the user himself or made available to the device through location configuration protocols (LCPs). When it comes to manual configuration, there is always a risk that the user will not insert the location information when he configures his/her phone or does not update it if he/she uses the device somewhere else. As far as automatic location configuration is concerned, IETF extended the Dynamic Host Configuration Protocol (DHCP)[6] and developed HTTP Enabled Location Delivery (HELD) [7]. Another solution that the ECRIT emergency architecture supports uses the medium access control (MAC) layer protocol and is named the Link Layer Discovery Protocol for Media Endpoint Device (LLDP-MED)[8]. According to this architecture, network operators must support at least one of the IETF location configuration protocols (HELD or DHCP). In the context of EMYNOS and to accommodate a wide range of scenarios, the location methods described above will be implemented and integrated with the VoIP infrastructure. Lately, the concept of advanced mobile location (AML) [9] was introduced by British Telecom and HTC. When an emergency call is made using an AML-enabled smartphone, the phone automatically activates its location service, and uses GPS, WiFi, and the location position compar-

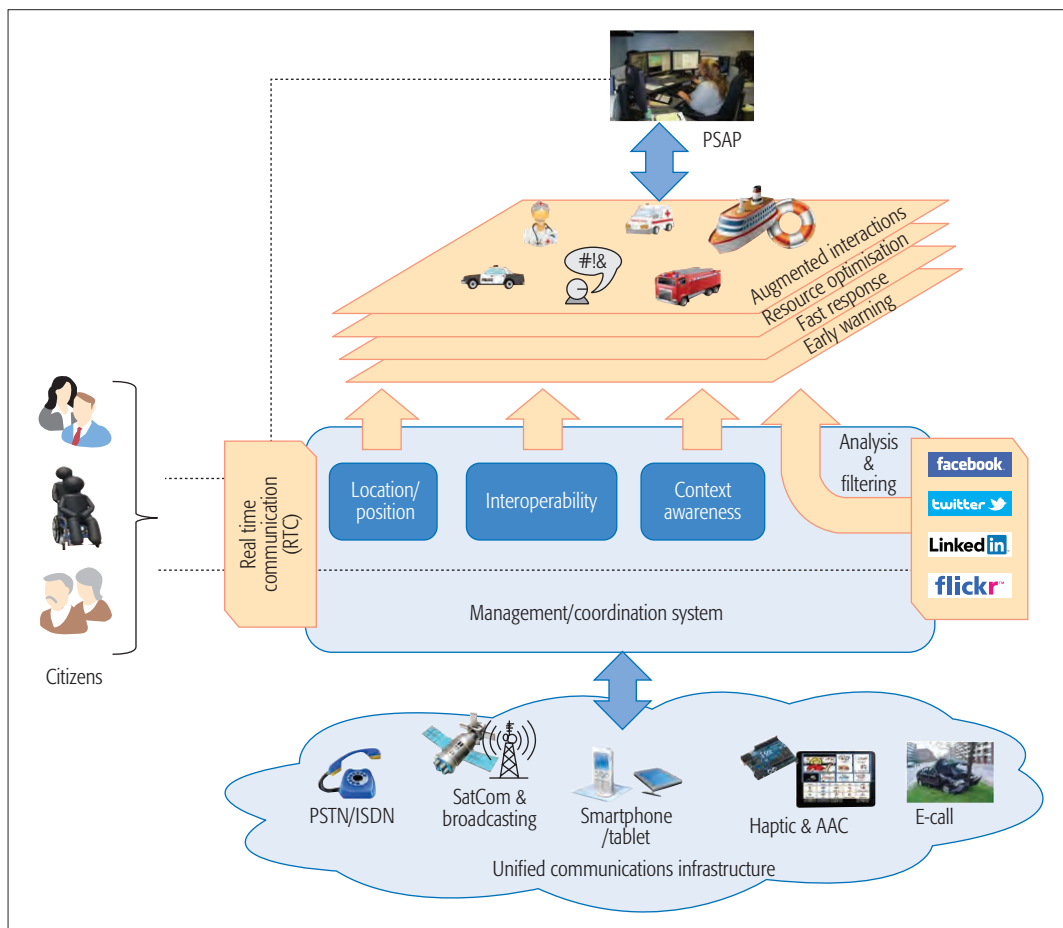


Figure 1. EMYNOS high-level architecture.

ing to the network cell ID information, achieving the best accuracy and sending a text message to the PSAP. The AML concept requires changes on the operator network and the PSAP systems. AML is seen as an enhancement of the legacy emergency systems of Europe.

## CHALLENGES FOR ADOPTION

### GENERIC DESCRIPTION OF EMYNOS ARCHITECTURE

One of the main characteristics of next generation emergency systems is the ability to use context awareness so that emergency alerts and messages can be initiated by either civilians or PSAPs. Moreover, the emergency call will be personalized, taking into account device capabilities (e.g., TV, tablet, PDA) and end user's profile (e.g., disabled people's hearing, vision, and cognitive impairments). As a consequence, this framework will allow the addition of advanced features such as automatic routing for end-user language preferences, automatic routing of emergency calls, emergency services mapping, location information retrieval, and support for people with disabilities. Moreover, it will therefore enable acceptance and handling of advanced information from citizens, including voice, video, photos, and text messages. Equally important is the provision of highly reliable voice/video service originating from mobile devices toward an IP infrastructure serving heterogeneous broadband access technologies.

In addition to that, the EMYNOS high-level architecture (Fig. 1) will focus on the integration of social media with emergency systems. This

will create a continuous channel among the citizens themselves, and between the citizens and the emergency management teams. In fact, social media such as Facebook and Twitter, which are becoming increasingly important in daily life, will play a special role in this context. It has been shown that citizens use this type of communication for hazard prevention in the context of disasters, major incidents, and planned events for their own safety, for family members, or in conjunction with volunteers using innovative crowdsourcing approaches to help people in need. In particular, we demonstrate how the eCall concept can be enhanced by and benefit from the IP technologies. This can be accomplished by allowing audio-video calls toward the PSAPs and sending location information, photos, and videos.

Future rescue coordination centers that follow this concept will be capable of managing complex emergency data in the form of voice, video, photos, real-time text messages, diverse local information visualization, social media information, eCall information, and additional medical data to be processed. The EMYNOS architecture intends to use innovative approaches to help increase the safety of citizens in highly populated areas as well as less populated ones, and to strengthen the resilience of the inhabitants in case of emergency. Optimizing the coordination between citizens and the public actors will provide the fastest way to deal with a disaster and save lives. Furthermore, next generation emergency systems must incorporate some basic functionalities, explained below:

One of the main characteristics of next generation emergency systems is the ability to use context-awareness so that emergency alerts and messages can be either initiated by the civilians or the PSAPs. Moreover, the emergency call will be personalized, taking into account device capabilities and end-user's profile.

The EMYNOS architecture intends to use innovative approaches to help increase the safety of citizens in high populated areas as well as non-high populated ones, and to strengthen the resilience of the habitants in case of emergency. Optimizing the coordination between citizens and the public actors will provide the fastest way to deal with a disaster and save lives.

- Location Support in IP:** Caller location information is crucial for emergency services. It is mainly needed for the following two purposes:
- Determine the appropriate PSAP that can serve the user fast and efficiently. This means the PSAP has to serve the area where the user's device is currently located.
  - Enable the PSAP to get more accurate or updated location information of the device, which leads to a faster and more efficient dispatching operation.

**Multimodality to Provide Access to People with Disabilities:** A next generation emergency system is important to offer people with special needs full access to emergency services. It is significant to integrate communication methods and technologies used so far by persons with disabilities in their everyday life, which will enable them to communicate effectively in an emergency situation. Typical systems should include the following.

**Deaf and hard of hearing users:** The support of people with hearing difficulties necessitates a friendly user interface supporting audio, video, and real-time text messaging through a unified communication platform. Real-time text is an improvement over traditional instant messaging, which is not adequate for intensive conversational situations such as reporting an emergency. This includes, for example, the communication between a person with disabilities related to hearing or speech and the PSAP.

**Ambient assisted living (AAL):** For blind and elderly users [10], as they are able to use voice communication, the focus will be put on a solution for fast and reliable triggering of a call from a mobile device. Providing a safe, effective mechanism to establish and maintain a call with a PSAP (automatic re-dialing, callback mechanism) is extremely important for all disabled users. Easy and reliable access to emergency calls will be granted by defining dedicated buttons or gestures.

**Augmentative and alternative communication (AAC) users:** AAC assists disabled persons to form sentences by supporting the selection of individual words. Electronic communication aids such as special keyboards or dynamic communication grids allow the user to choose picture symbols to create messages that can later be transferred to text or synthesized speech, as illustrated in Fig. 2.

**Haptics:** Although audio-visual systems provide a user with a satisfactory impression of being present in a remote environment, physical interaction and manipulation are not supported. True immersion in a networking environment requires the ability to physically interact with remote objects and to literally get in touch with other people. This can be accomplished by adding haptic modality to audio-visual systems. Haptic communications is a relatively young area of research that has the potential to substantially improve human-human and human-machine interaction. Haptic devices differ in their kinematics, which include provided degrees of freedom, output capability (e.g., displayed force/torque, velocity, and acceleration), sensorial capability, and accuracy.

**Social Media in NGN112:** A next generation emergency system is important to benefit from

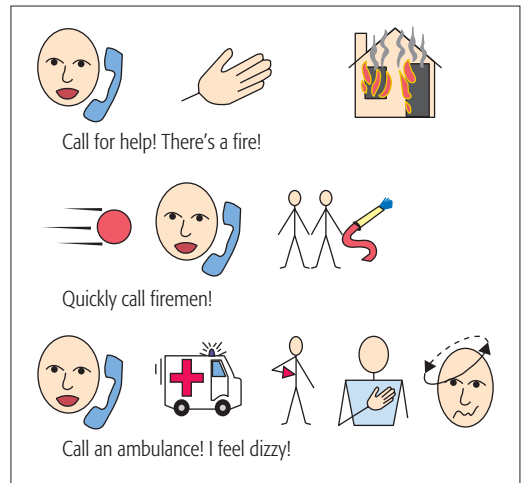


Figure 2. Examples of messages composed with the use of picture symbols in AAL.

continuous live channels between emergency centers and citizens, and among citizens themselves. The emergency centers need to keep the population informed, while the citizens expect to stay connected with friends, family, and services. Social media can be seen as the perfect solution not only for informing citizens but also as a backup solution in case the telecommunication infrastructure is not operational. Based on a survey provided by the American Red Cross [11], 18 percent of the people that participated in the survey mentioned that they would turn to social media if calls to 911 were unsuccessful. According to this survey, 69 percent said that the emergency call takers should regularly monitor their websites and social media networks in order to respond in time to requests. In this context, social media will play a crucial role, especially if integrated with a new technology such as WebRTC. For this reason, EMYNOS intends to particularly design and develop a crowdsourcing mechanism that collects social media information for detecting emergency situations and techniques for summarizing and aggregating emergency information retrieved from the posted messages and a mechanism for classifying messages according to their importance (infrastructure damage, scream for help, etc.).

These requirements are translated to the EMYNOS architecture functional block, which deciphers all of the above services in three main parties, discussed below.

## EMYNOS ARCHITECTURE FUNCTIONAL BLOCKS

EMYNOS will develop a next generation platform for enabling European citizens to make IP-based emergency calls (to police, ambulance and fire brigade). In fact, the EMYNOS intersects the NG112 architecture and implements some of the related functionalities according to the above requirement. The EMYNOS functional blocks are depicted in Fig. 3. This architecture, which already takes into account the requirements and the potential scenarios discussed above, includes various functional blocks that can be grouped into three main parts (or steps).



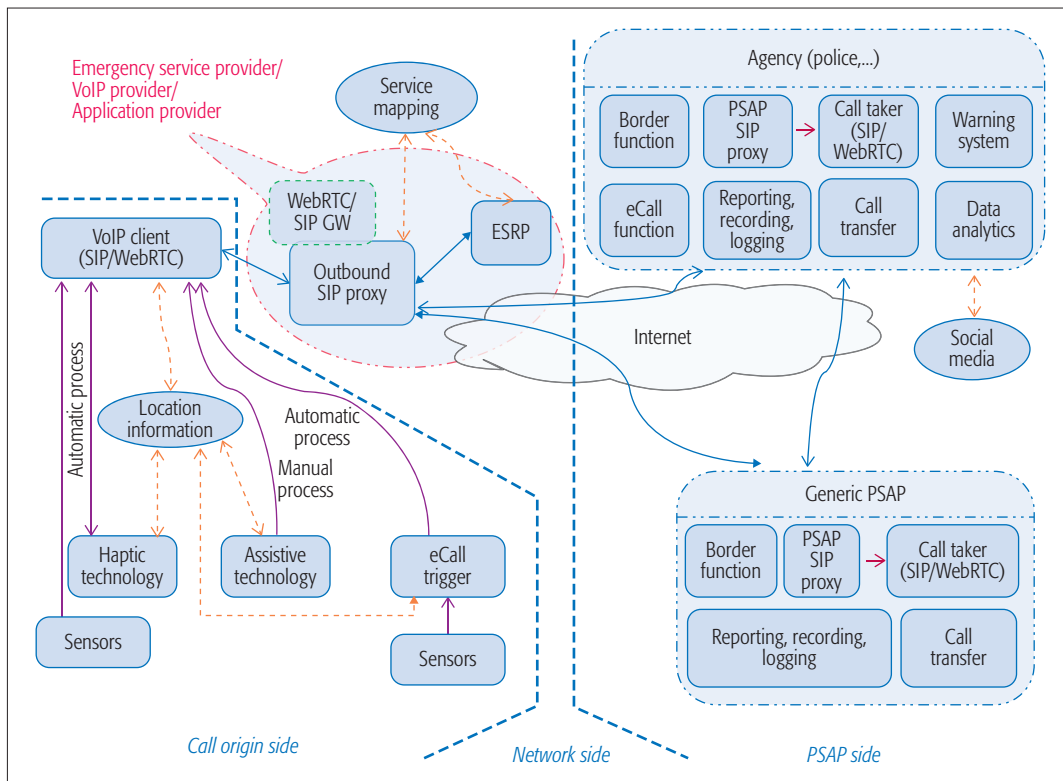


Figure 3. EMYNOS functional blocks.

### THE CALLER PART

This side reflects the initiation of the emergency call. Three main types of calls are considered: emergency calls (calls to civil protection agency, police, ambulance, or fire brigade) initiated by people without special needs, calls initiated by people with special needs, and eCalls. Although all these types are of an emergency nature and might overlap, the classification was introduced because call triggering could vary from one type to another. In the case of persons with disabilities, the caller might initiate the call either by himself (manually) or in an automatic way through an assistive technology. As far as eCall (automotive emergency call) is concerned, the call will be triggered manually or in an automatic way based on some information provided by sensors. The location information is a crucial aspect in emergency services, and the idea is to deliver it when establishing the call.

### THE NETWORK PART

This part is first of all about emergency call routing to the appropriate PSAP. This requires functionalities such as call identification, call classification based on the emergency type (e.g., eCall, call from a person with a disability) and call-to-service mapping to route the call to the appropriate PSAP. This part also includes routing policies in case a PSAP is not reachable (e.g., overloaded). If a call is issued from a WebRTC-enabled browser, a WebRTC-SIP gateway is required so that signaling translation is carried out, allowing the call to be routed appropriately. As far as deployment is concerned, the network part might span various stakeholders like a telecom operator, a VoIP provider, an emergency service provider (which operates an ESINet), and an application provider

if the client that initiated the call is a mobile application developed for emergency service purposes. The network part also includes components for location information provision.

### THE PSAP PART

This entity will be in charge of answering the emergency calls, reporting about them, and transferring them to the appropriate agency if needed. In fact, on the PSAP side, we are considering a generic PSAP with tasks that are more call taking, reporting, and call transfer; and agencies that are endowed with other functionalities such as dispatching and warning. The agencies and the generic PSAP could be seen as different parts of the same entity. In this part, we are also considering a data analytics functionality that will support the monitoring of an extreme situation by providing near-real-time social media data collection and analysis.

Based on this, EMYNOS will support a number of use cases that deal with the above identified technologies and solutions.

## REAL USE CASE SCENARIOS

The EMYNOS use cases conform to various next generation services that use either a web browser or a SIP client, or even automated calls initiated by people with disabilities. The following use cases are considered.

### EMERGENCY CALL TO 112 FROM A VOIP CLIENT/BROWSER

Location information is crucial in emergency services. This information is required to reach the caller as soon as possible, and to route the call to the appropriate PSAP. SIP-enabled terminals might support GPS or even offer a friendly interface that allows the caller to insert his/her current

The user opens a WebRTC-enabled browser, which takes advantage of the native support of video and audio through an HTML5 browser to initiate a call toward the PSAP. The communication platform will utilize the user's current location as a reference in order to select the nearest PSAP.

As the number of the IoT devices has grown rapidly, communication among the IoT devices and the huge amount of data that are transmitted to the cloud via the Internet have increased, stretching the network and cloud infrastructure so they cannot satisfy all the requirements of QoS and resource allocation.

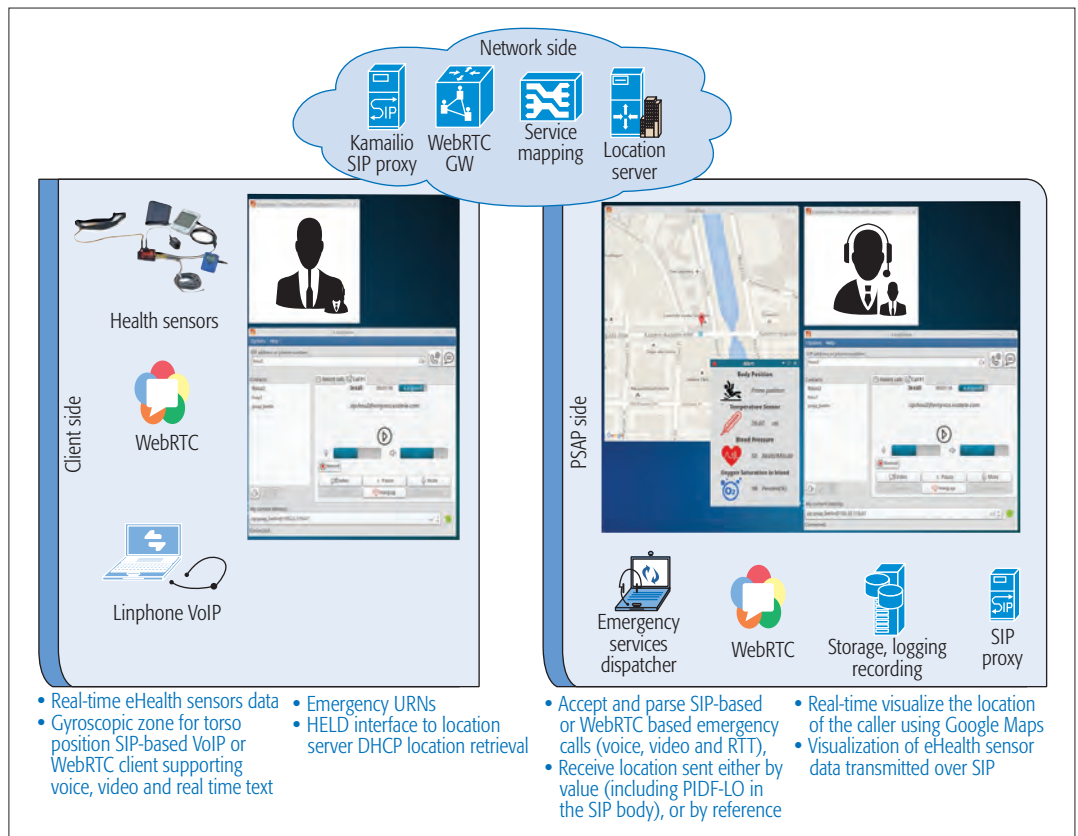


Figure 4. EMYNOS automated generated call with sensor data.

location data and send it within the emergency call to the PSAP.

#### EMERGENCY CALL TO 112 FROM A BROWSER

There are cases when a user (either at home or outside), using a fixed or portable device, experiences an emergency situation and needs to call the local emergency services directly using the common emergency telephone number 112 (or 911). The user opens a WebRTC-enabled browser, which takes advantage of the native support of video and audio through an HTML5 browser to initiate a call toward the PSAP. The communication platform will utilize the user's current location as a reference in order to select the nearest PSAP.

#### EMERGENCY CALL FROM A PERSON WITH A DISABILITY

The caller is an AAC user employing a device that allows them to control, using eye movements, a computer with AAC software with speech synthesis and a symbol dashboard. This is his/her main way to communicate directly and indirectly, sending and receiving emails, SMS, and calls. There are hundreds of predefined grid sets, which include a special grid with the ability to send text messages to emergency services with some details, such as his/her address and a few alternative possible life-threatening situations. Using the eye movement control device, the computer, and the alarm grid, he initiates an emergency call. The PSAP operator receives the call and information on the disability characteristics of the user. The user is able to answer important questions using a communication grid with predefined symbols and speech synthesis. The PSAP operator also receives text messages transferred from symbols sent by

the user. If the caller has difficulty understanding speech, the PSAP operator may also use text to ask simple questions, which are seen by the caller as symbols generated by his/her software. Alternatively, haptic devices integrated over the user's terminal alert the person with disabilities about an emergency situation in the vicinity, and through the haptic interface the user sets up a call to the nearest PSAP or receives a call (audio, text, or visual stimuli) from the PSAP.

#### AUTOMATIC GENERATED CALLS

The home environment of the caller is monitored by a network of environmental sensors (i.e., temperature, humidity, air pressure, etc.), which aggregate data in real time to either a home gateway or the cloud. Specifically, the scenario involves the integration of different types of Internet of Things (IoT) sensor devices (e.g., temperature/humidity, acceleration, light). As the number of IoT devices has grown rapidly, communication among the IoT devices and the huge amount of data that are transmitted to the cloud via the Internet have increased, stretching the network and cloud infrastructure so that they cannot satisfy all the requirements of QoS and resource allocation. For this reason, fog computing provides tasks such as data computing, data storage, local management of sensors, and mobility. Fog computing, a term coined by Cisco Systems, is also referred to as mobile edge computing (MEC) by the European Telecommunications Standards Institute (ETSI). It refers to the adaptation of cloud computing to the mobile environment in an anywhere and anytime manner, where data is stored and processed outside mobile devices. Some of the most critical

issues related to fog computing include network latency and limited bandwidth in the mobile network in order to handle the heterogeneity issues from the different devices. The IoT emergency gateway must adopt a service oriented approach in order to orchestrate the devices according to their behavior and determine the order of the exchanged messages. 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. An emergency call is initiated automatically by the alarm system, which triggers the PSAP. Figure 4 illustrates the graphical user interface at the client side and at the PSAP side, along with eHealth sensor data transmitted over SIP, DHCP/HELD/LOST-based location information retrieval, data, and location visualization for the automated generated call.

## CONCLUSIONS

In this article the EMYNOS platform is presented, focusing on managing complex emergency data in the form of voice, video, photos, real-time text messages, diverse local information visualization, social media information, eCall information, and additional medical data to process. The architecture of the EMYNOS framework, consisting of three main parts, is explained: the client side, able to provide access to people with disabilities; promoting the use of social media and at the same time using WebRTC technology that is natively supported by all modern web browsers, together with a native SIP client that can be the bridge among native SIP and IMS. The network part presents how the interconnection of the WebRTC, SIP, and traditional interconnection is achieved between the client side and the PSAP side. Next, on the PSAP side we present how a next generation PSAP will promote the use of open technologies like WebRTC and native SIP.

## ACKNOWLEDGMENT

The present work was undertaken in the context of the Next Generation Emergency Communications (EMYNOS) project with contract number 653762. The project has received research funding from the H2020-EU.3.7 European Framework Programme.

## REFERENCES

- [1] ITU-T Rec. Y.2001, "General Overview of NGN," 2004.
- [2] NENA 08-002, "NENA Functional and Interface Standards for Next Generation 9-1-1 (i3)," Dec. 18, 2007.
- [3] NG112, E.ENA, "Next Generation 112 Long Term Definition Standard for Emergency Services Document, v. 1.1," 2013.
- [4] Emergency Context Resolution with Internet Technologies (ECRIT), <http://datatracker.ietf.org/wg/ecrit/charter/>; accessed Sept. 2, 2016;
- [5] T. Hardie et al., "LoST: A Location-to-Service Translation Protocol," IETF RFC 5222, Aug. 2008

- [6] H. Schulzrinne, "Dynamic Host Configuration Protocol (DHCPv4 and DHCPv6) Option for Civic Addresses Configuration Information," RFC 4776, Network WG, 2006, <http://tools.ietf.org/html/rfc4776>; accessed Sept. 2, 2016.
- [7] J. Winterbottom, H. Tschofenig, and L. Liess, "A Routing Request Extension for the HTTP-Enabled Location Delivery (HELD) Protocol, IETF RFC 7840, 2016.
- [8] ANSI/TIA-1057-2006), Link Layer Discovery Protocol for Media Endpoint Devices, Apr. 2006.
- [9] G. Machado and T. O'Brien, "Advanced Mobile Location (AML) in the UK"; [http://www.eena.org/uploads/gallery/files/operations\\_documents/2015\\_02\\_18\\_AML\\_FINAL.pdf](http://www.eena.org/uploads/gallery/files/operations_documents/2015_02_18_AML_FINAL.pdf), accessed Sept. 2, 2016;
- [10] Y. Nikoloudakis et al., "A Fog-Based Emergency System for Smart Enhanced Living Environments," *IEEE Cloud Computing Mag.*, Nov./Dec. 2016.
- [11] R. Cabacas, et al., "Context-Aware Emergency Messaging System Framework Utilizing Social Relations as Services," *Int'l. J. Multimedia Ubiquitous Engineering*, vol. 9, no. 2, 2014, pp 77-86.

## 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 at TEI of Crete and 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 acts as Workshop Co-Chair for the IEEE SDN-NFV Conference.

ASIMAKIS LYKOURGIOTIS (asly@ece.upatras.gr) received his engineering diploma from the Electrical and Computer Engineering Department of the University of Patras, Greece, in 2008. Since 2008, he has been a Ph.D. candidate in 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.

ILIAS POLITIS received his Ph.D. in multimedia networking from the University of Patras in 2009. He is a postdoctoral research fellow at the School of Science and Technology of 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.

TASOS DAGIUKLAS is a leading researcher and expert in the fields of Internet and multimedia technologies for smart cities, ambient assisted living, healthcare, and smart agriculture. He is the leader of the Smart Internet Technologies (SuITE) research group at London South Bank University, United Kingdom, where he also acts as the head of the Division in Computer Science. His research interests include smart Internet technologies and cloud infrastructures and services.

YACINE 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 next generation emergency services and next generation network security, primarily DoS attacks detection, SPAM mitigation, and fraud and service misuse detection. He is the EMYNOS project coordinator

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 the 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 Distinguished Member of the Union of Regional Television in Greece.

The IoT emergency gateway must adopt the service oriented approach in order to orchestrate the devices according to their behavior and determine the order of the exchanged messages. A decision making engine processes these data and creates alerts as soon as certain threshold values are exceeded.

# Utilizing an NG 9-1-1 Test Lab to Validate Standards Compliance

Walter R. Magnussen, Ping Wang, and Yangyong Zhang

When implementing a state-wide NG 9-1-1 system, how do you ensure standards compliance? Which standards are relevant? How would you establish a lab environment to test NG 9-1-1 interoperability? The authors examine these questions and others. To answer these questions, they established an NG 9-1-1 test lab, developed test scenarios, and tested against a subset of NG 9-1-1 functional requirements.

## ABSTRACT

When implementing a state-wide NG 9-1-1 system, how do you ensure standards compliance? Which standards are relevant? How would you establish a lab environment to test NG 9-1-1 interoperability? The objective of this article is to examine these questions and others. To answer these questions, we established an NG 9-1-1 test lab, developed test scenarios and tested against a subset of NG 9-1-1 functional requirements.

## INTRODUCTION

Standards have long been used in the telecommunications industry when implementing service delivery architectures. The benefits of these standards are ease in implementation, reduced costs, and enhanced competition. The problem is that ensuring standards compliance is not always an easy thing to do. Often vendors add enhancements that are touted to be better than the standards, but the reality is that they actually violate the standards. Some vendors even claim that their product has the functionality of the standards while not complying with the standards at all. This has been the case in the NG 9-1-1 space for several years.

A recent study released by the Industry Council for Emergency Response Technologies (iCERT) [1] looks at the current status of NG 9-1-1 deployment across the United States. In summary, less than a handful of states have completed the NG 9-1-1 transition, a few more are early in the transition stage, several others are in the data collection mode, and the rest have yet to begin the transition. One of the issues that all of the responsible entities face is gaining an understanding of the level of standards compliance of the underlying functional elements that comprise the architecture. In the previous generation of emergency calling, or E911, standards compliance was part of the legacy telephone network. These systems all underwent rigorous testing in the labs of the service providers. The IP-centric next generation solution will not involve the service providers in all cases; therefore, the end users will either need to rely on market literature or complete their own testing. The State of Texas Commission on State Emergency Communications (CSEC) [2] chose to do the latter, and have established their own testing process.

In late 2015 CSEC initiated an NG 9-1-1 test lab project in support of their strategic plan for NG 9-1-1 transition adopted in May 2014 [3]. The strategic plan states that a full transition to NG

9-1-1 would be completed by 2019. The project is divided into three phases, each of which tests a segment of the NG 9-1-1 architecture. Phase I evaluates transition elements, such as the legacy network gateways (LNGs) and legacy selective router gateways (LSRGs). These elements precede the ESInet on the service provider side, and are critical until the last legacy network is decommissioned. The ESInet is the Emergency Services IP network as defined by the National Emergency Number Association (NENA) i3 specification. It is a walled garden network protected in all directions by the border control function (BCF). The BCF is a combination of a firewall to examine all non-Session Initiation Protocol (SIP) traffic and a session border controller to examine all SIP traffic. The NG 9-1-1 ecosystem will ultimately be made up of many ESInets interconnected in a hierarchical manner. Phase II evaluates the ESInet functional elements, such as the emergency communications routing functions (ECRFs), emergency service routing proxies (ESRP), and the BCF. Phase III examines the NG 9-1-1 equipment that would be housed in public safety answering points (PSAPs) or NG 9-1-1 call centers. This equipment is often referred to as customer premises equipment" (CPE).

The testbed project is led and funded by the CSEC with support contracts and agreements with four external entities. The first entity is Capgemini, a worldwide system integration and consulting firm with almost 180,000 employees in 40 countries [4]. Capgemini was responsible for developing test scenarios, providing technical support, and overseeing the testing process. The second entity is Mission Critical Partners (MCP), which is a leading emergency communications consulting firm; their role is to serve as subject matter experts. The third entity, the State of Texas Department of Information Resources (DIR), is the state agency that is tasked with supporting the underlying transport network and developing best practices in supporting partitioning network infrastructure as related to wall gardened networks, such as ESInets. Lastly, the fourth entity, the Texas A&M University Internet2 Technology Evaluation Center (ITEC), is tasked with hosting the test lab.

The NENA document that defines the NG 9-1-1 architecture is the "Detailed Functional and Interface Specification for the NENA i3 Solution - Stage 3 Version 08-003 V1" as approved June 14, 2011. It was updated in late 2016. This document defines all of the functional element requirements of an NG 9-1-1 network and describes the interface

specifications between the functional elements. Also referred to simply as i3, this document defines all of the requirements for our testlab.

The testing completed under this project is similar to that of NENA Industry Collaboration Events (ICEs) [5]. Both evaluate interoperability standards compliance, and are based on predefined test scripts. There are two significant differences between these two sets of tests. The first is that an ICE focuses on end-to-end testing in a multivendor environment but does not specifically target each and every NENA i3 requirement. It assumes that end-to-end success implies underlying compliance of all requirements. The second is that ICEs are held under a strict code of conduct. The code prohibits any dissemination of any test results to anyone outside of the ICE community. Since lessons learned cannot be shared with organizations, such as CSEC, which will ultimately have the responsibility of overseeing the NG 9-1-1 networks, it forces them to seek other means of becoming informed. This is not a criticism of ICEs since this confidentiality is paramount to the success of ICEs. Without the assured confidentiality vendors would not agree to expose their weaknesses.

There were three CSEC defined goals of the test lab project. The first desired outcome is to ensure standards compliance of all functional elements of the ESInet. The second outcome is the desire for CSEC to fully understand what is required to support the ongoing operation of the NG 9-1-1 network. The third outcome is the desire to inform the people involved in the procurement and subsequent contract process enough to ensure the best solution possible for the State of Texas.

## ESTABLISHING THE TEST LAB

The ITEC is a public safety research center at Texas A&M University that focuses on NG 9-1-1 and public safety broadband networks such as FirstNet. The ITEC was selected to host the test lab due to its history in supporting similar projects. In 2007 the ITEC was selected by the United States Department of Transportation to design, implement, and test the NG 9-1-1 proof of concept. In 2009 the ITEC was funded by the National Science Foundation to be part of an NG 9-1-1 security testbed project. Then, in 2010, the ITEC entered into an agreement with Harris County, Texas, to support their FirstNet public safety broadband LTE project. The ITEC has supported public safety communications initiatives through involvement in NENA, the Association of Public-Safety Communications Officials (APCO), National Public Safety Telecommunications Council (NPSTC), Public Safety Communications Research (PSCR), and FCC committee work.

Since the ITEC's inception in 2004, we have received donations of about \$6 million in leading edge technology from several industry firms. This investment serves as the basis of the CSEC test lab. Where additional elements were needed, we solicited donations. Where equipment was already in place, we completed any necessary updates. While donations were solicited and appreciated, we made clear that equipment donation was not a requirement to be considered in the final solution.

There were a few principles that guided the design of the test lab:

1. We would utilize actual systems and data to the extent possible. When not possible, we would simulate or develop the pieces necessary.
2. The architecture would be designed to allow the interoperability of layers of ESInets at the local, state, and national level.
3. We would only utilize components in the ITEC, DIR, and service providers' labs. By not connecting to any components that support live emergency traffic, we eliminated the risk of life safety issues.

Since Phase 1 focuses on the legacy transition devices, the first challenge was to secure access to a legacy network. Working with CenturyLink and the resources in their labs in Littleton, Colorado, we gained access to such a network. We were able to place a set of Cisco routers (one in Colorado and one at the ITEC) configured with T1/PRI interfaces, ear and mouth (E&M) interfaces, and foreign exchange office/subscriber (FXO/FXS) interfaces. To be specific, the primary rate interface (PRI) is a telecommunications interface standard used on an integrated services digital network (ISDN) for carrying multiple digital signal 0 (DS0) voice and data transmissions between the network and a user. E&M is a type of supervisory line signaling that is traditionally used in the telecommunication industry between telephone switches, and FXO and FXS interfaces are the names of the ports used by analog phone lines.

Then a Lucent 5ESS (a Class 5 telephone electronic switching system developed by Western Electric) in Colorado served as the host central office, and a Nortel DMS100 supported the selective router functionality. Twenty DS0s were configured in the 5ESS to point to a Cisco call manager acting as the enterprise private branch exchange (PBX) over the PRI. The two Cisco routers were logically connected to each other through an IPSec tunnel across the commodity Internet; IPSec is a protocol suite for secure IP communications that works by authenticating and encrypting each IP packet of a communication session. The connection between the TAMU ITEC lab and CenturyLink labs is depicted in Fig. 1.

In order to test NG 9-1-1 calls transferred to a legacy E911 PSAP, it was also necessary to secure a PSAP. Thanks to a donation from the Brazos Valley Council of Governments (BVCOG), we were given an Airbus PSAP. The network connection for the PSAP was over centralized automated message accounting (CAMA) trunks terminated in a set of FXO/FXS interfaces to the selective router in Colorado.

The next challenge was to be able to support both automatic location information (ALI) data for the legacy E911 systems and location information server (LIS) data for the NG 9-1-1 systems. These connections were made, again over IPSec tunnels, but to the Intrado test labs in Longmont, Colorado. Intrado (also known as West) is one of the largest providers of database services that are used to identify the location of emergency 911 callers in the United States. Since ALI is basically an RS-232 serial interface, the connection to the legacy PSAP was made over a terminal server using a Telnet connection to the server in Colorado. The LIS connection was made through an HTTPS connection using the HTTP Enabled Location Delivery (HELD) protocol. HELD is the standard specified in the NENA i3 specification for

The Internet2 Technology Evaluation Center (ITEC) is a public safety research center at Texas A&M University that focuses on NG 9-1-1 and public safety broadband networks such as FirstNet. The ITEC was selected to host the test lab due to its history in supporting similar projects.

Since this project involves the validation of standards compliance, the first step was to define the appropriate standards. This was fairly simple since the NENA i3 version 1 specification is the specification that has industry concurrence.

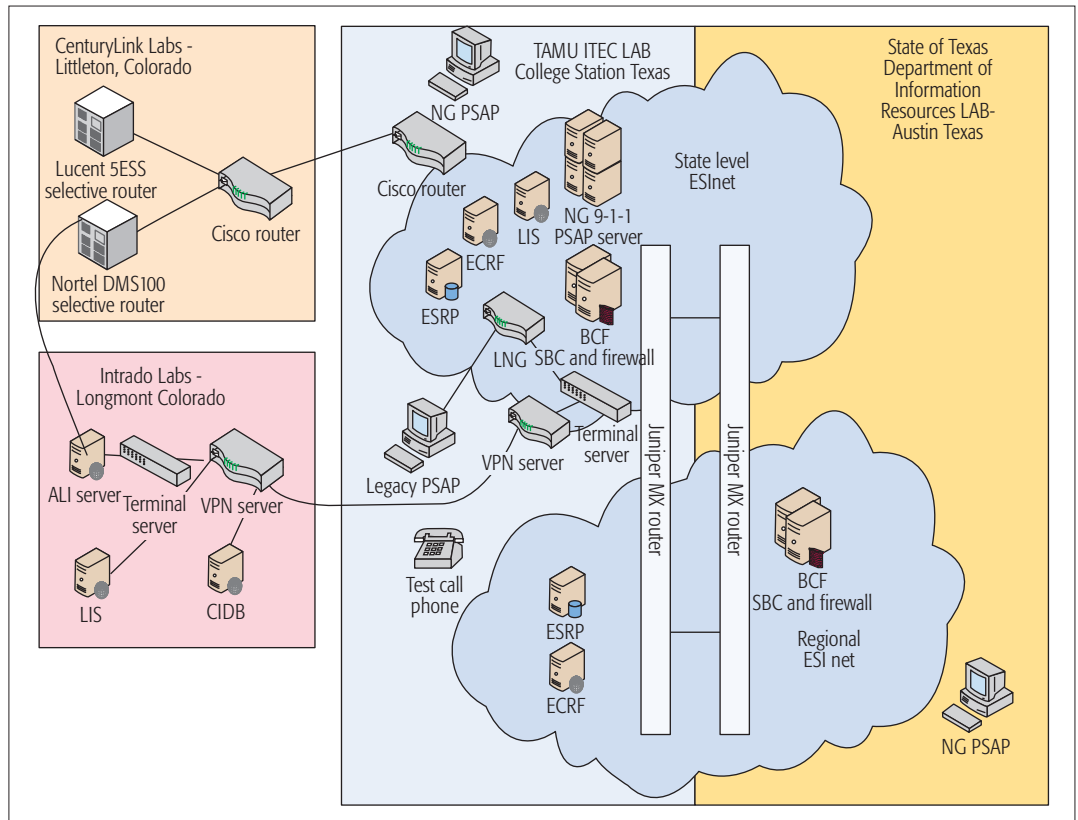


Figure 1. The connection diagram to CenturyLink labs.

NG 9-1-1 and defined in Internet Engineering Task Force (IETF) RFC 5985 [6], The IETF develops and promotes voluntary Internet standards, in particular the standards that comprise the IP suite (TCP/IP). While it was not specifically part of the Phase 1 requirements, we had to erect enough of an ESInet and PSAP to complete the testing. Since the Phase I tasks included connecting a call initiated in a legacy network to an NG 9-1-1 PSAP, connecting a call initiated in an NG 9-1-1 ESInet to a legacy PSAP, or transferring calls between NG 9-1-1 and legacy PSAPs, we needed to implement both an ESInet and an NG 9-1-1 PSAP. Functional elements from Oracle (BCF), Geocomm (ECRF), and Experient (ESRP) made up the ESInet. The NG 9-1-1 PSAP was provided by Experient. As of the writing of this article, we are in the process of installing additional ESRPs from Solacom and Oracle, and PSAPs from Solacom. Additional systems from other vendors are in negotiation. Figure 2 depicts the logical connections that make up the CSEC test lab.

### DEVELOPING TEST SCENARIOS AND TESTING

Since this project involves the validation of standards compliance, the first step was to define the appropriate standards. This was fairly simple since the NENA i3 version 1 specification is the specification [7] that has industry concurrence. This version was formally adopted on June 14, 2011, and is currently under review. NENA derives its authority as a standards development organization from the American National Standards Institute (ANSI) through their accreditation [8].

It is important to understand that the NENA i3 specification is not so much a standard as it is a reference architecture. Titled “Detailed Functional and Interface Specification for the NENA i3 Solu-

tion — Stage 3,” the specification relies heavily on other standards to accomplish its task. The document itself contains 147 references, most of which are IETF, the International Standards Organization (ISO), the Organization for the Advancement of Structured Information Standards (OASIS), ANSI, and other standards bodies.

For the purposes of the test lab project, the team lead by Capgemini went through the i3 specification and documented each functional requirement that related to the legacy gateways. This resulted in 61 test scripts that were divided into two phases.

Each test script resulted in a call scenario that was mapped against the i3 specification. For each test, a caller would initiate a 911 call, which would route through the test network. The typical call routing can be seen in Fig. 3.

The actual testing occurred in January–March 2016 with much of it happening in College Station. The following is the test team at work during the actual testing.

Test tools were installed to support troubleshooting and provide call flow documentation. These tools were Wireshark, used for deep packet inspection and documentation of all non-SIP transactions, and the Oracle Communications Operations Monitor tool, or OCOM. OCOM was used to document all SIP call flows. Figure 4 shows an example of one such flow.

### FINDINGS TO DATE

As of the publishing of this article, Phase 1 testing has been completed (Fig. 5). Overall, the results have been positive. With approximately 100 individual test scripts, only one has failed. Some more specific findings are:

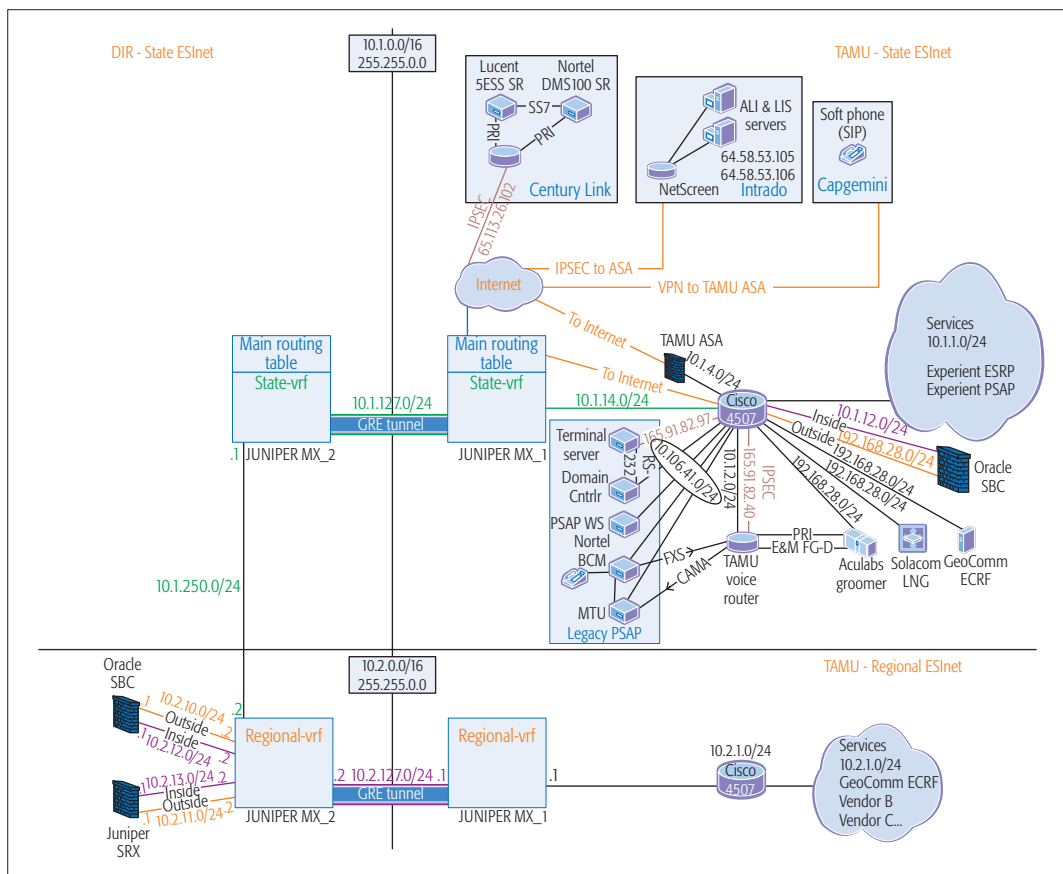


Figure 2. Connection diagram of the CSEC test lab.

•The standards appear to be comprehensive enough to allow states to proceed with their ESInet implementations, at least from the aspect of the legacy gateways. There is little reason to believe that Phase 2 will have any different results.

•The mapping of the ALI data (E911) to the LIS data (NG 9-1-1) is one of the most complex aspects of the transition. While there is standardization of what ALI fields are, there are no standards on which fields are required. In several cases we found fields required in the LIS (e.g., county) not populated in the ALI data. Another example is that the ALI address is now mapped into several fields in the LIS database. This resulted in the requirement for manual data manipulation in several cases. In defense of the data, we did select civic addresses that we knew would be challenging.

•There are several standards used in E911 when it comes to P-ANI format (P-ANI, or pseudo ANI, is used as a proxy for a number of wireless towers used in the location determination and routing of wireless E911 calls). This issued required significant header remapping in the LNG to deal with all cases. It is important to select legacy gateway systems that are flexible to support all formats.

•It appears that there could be legacy networks and gateways in place for the next 10 years. We found in the project that the skillsets and tools to configure and support the legacy networks are becoming more difficult to find as these old networks are shut down and staff retire. This is the case when an issue requires in-depth understanding of SS7 signaling, ISDN User Part (ISUP) signaling (used in PRI 911 trunks), and analog CAMA signaling. This problem will likely get worse as

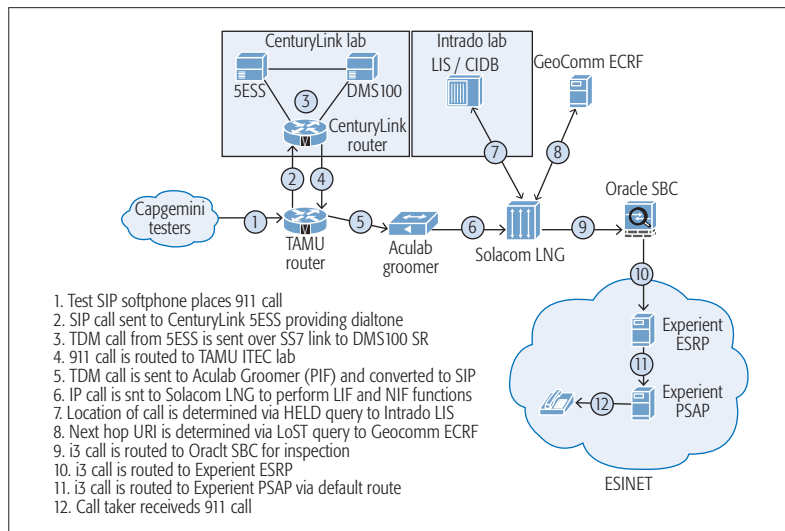


Figure 3. The typical call routing.

we get further along in the transition. In our current approach, we manually abstracted necessary information from SS7 messages, which share similar controlling processes to SIP messaging. We then carefully mapped different parameters so that calling processes would not be unintentionally affected by this customized gateway. We expect that a more delicate mechanism will be created in the future. Note that a document has been created for recording this matching process; it will be a valuable reference for future improvement and automation.

While this complex and comprehensive process is costly in terms of both investments of dollars and time, it will certainly pay off in terms of not requiring restarts once the transition begins, and will ensure an adequate, ongoing, support budget once the transition is complete.

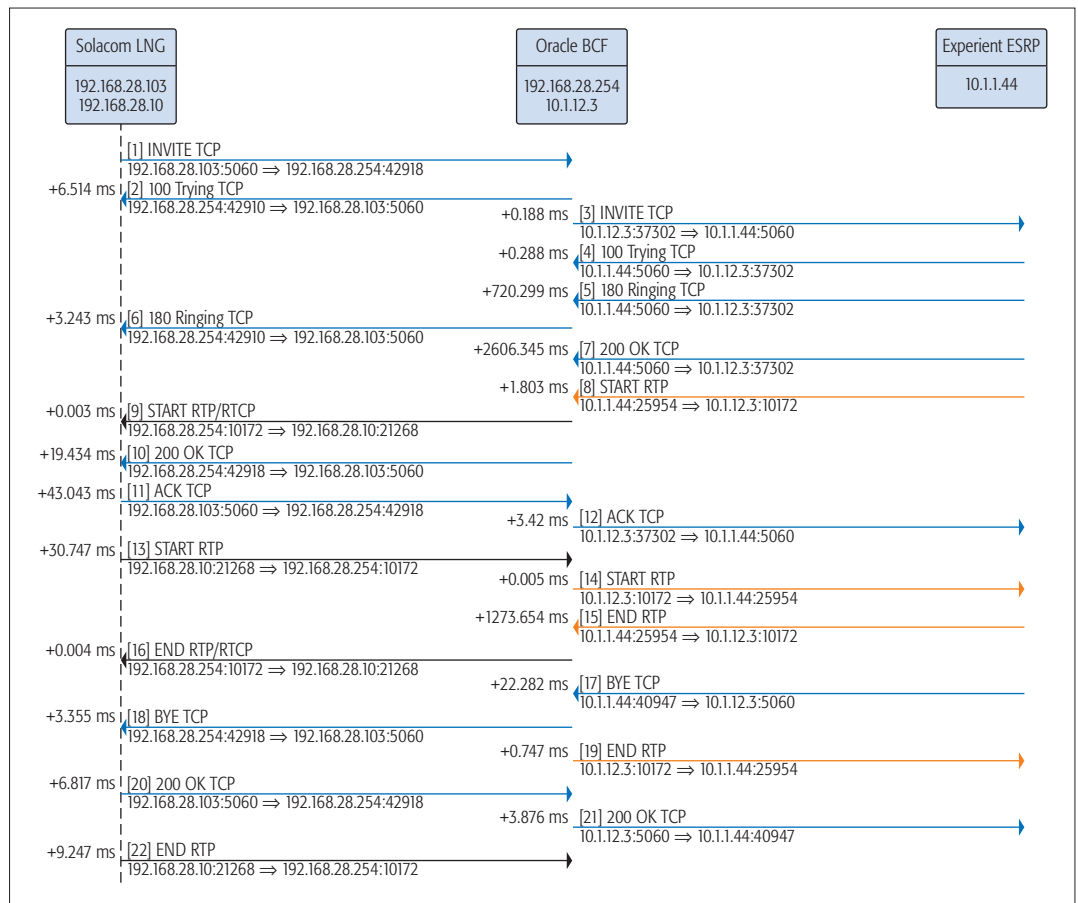


Figure 4. OCOM SIP call flow monitoring.

- We had to develop our own in-house server for use as a proxy for the mobile positioning center (MPC) and voice over IP (VoIP) positioning center (VPC). Getting access to live systems that would meet the requirements of the project proved to be impossible.

- This process can and should be replicated as we begin the integration of NG 9-1-1.

### CONCLUSION

In conclusion, to date this project has proven to be invaluable in terms of both further understanding the requirements of establishing procurement bids and supporting the network once it is established. While this complex and comprehensive process is costly in terms of investment of both dollars and time, it will certainly pay off in terms of not requiring restarts once the transition begins, and will ensure an adequate ongoing support budget once the transition is complete.

The ITEC lab (Fig. 6) established as a result of this project will continue to be of value once this project is complete. It can be used to support research for other states. Additionally, the lab is currently connected to the Defense Research-Development Canada (DR-DC) labs in Regina where the Canadian Homeland Security organization is mapping out a plan for Canada. There is also a Cooperative Research and Development Agreement (CRADA) in place with the Public Safety Communications Research Center in Colorado, which will support applications testing (including NG 9-1-1) over the FirstNet network.

This project became necessary since there is

no NG 9-1-1 certification process. While NENA does support the ICES, these events were created to allow industry manufacturers to be able to complete interoperability in a non-threatening environment. This requirement is inconsistent with a certification or approval process with results that could be made available to the public. It is not likely or efficient for this process to take place in every state, and the NENA Next Generation Partners Program is contemplating ways of resolving this gap.

During the establishment of the test lab we came across a few areas where we felt that the i3 specification was vague. Whenever we required clarification, we would reach out to a NENA member that we knew was also a member of an appropriate committee. Whenever this happened, we were provided clarification and assured that the issue in question would be resolved in the pending re-write. The testing resulted in the answering of two questions, the first being the extent to which the functional elements complied with the i3 specification. By mapping the i3 requirement to a test parameter, we were able to verify this compliance. The second question related to the ongoing operation of the ESNets. We wanted to document issues in the configuration of functional elements and maintenance of required databases. The testing process consisted of making test calls while using both OCOM and Wireshark to capture the call setup packets. This allowed us to ensure that not only were the outcomes of the call setup consistent, but the methods used complied as well. Per the research contract, we can publish the results, but the data is confidential.



A similar process should be repeated as we move into the NG 9-1-1 integration with FirstNet phases. The NPSTC has recently documented the high-level integration requirements, but there is much work to be done to ensure interoperability [9].

#### ACKNOWLEDGMENT

This project was made possible by the vision of the State of Texas CSEC staff who are driven by the desire to make the Texas NG 9-1-1 system the model for the rest of the country. The team is directed by Kelli Meriweather. Susan Seet is the NG 9-1-1 project lead, and she is supported by her team, which includes Kevin Rohrer and Monica Watt.

The TAMU ITEC support team includes Dr. Robert Arnold, Lauri Ditto, Ping Wang, Yangyong Zhang, Aaron Heald, Kevin Schmidgall, and Derek Ladd. This group managed all of the lab activities.

The Capgemini team included Tim Lindler, Jon Samuelson, Ryan Chandler, Justin Jensen, Michael Kyle, and Shara Tidwell. These members and others were responsible for developing test scripts, supporting network systems integration, and documenting the final results.

The DIR team was led by Wayne Egeler, and includes Sharon Blue, Steven Pyle, and Gerardo Lopez. This team made their labs in Austin, Texas, available to support the testing of transport requirements as they apply to multi-agency, multi-level ESNets.

The Mission Critical Partner (MCP) team served as subject matter experts.

A special thanks goes out to all of the vendors that provided lab access, equipment, and support resources to make this project possible. The vendors include CenturyLink, Intrado, Acculabs, Oracle, Juniper, Solacom, Experient, Geocomm, and RedSky

Technical advice was also provided by Roger Hixson of NENA and Brian Rosen of Neustar.

This project is a collaboration with multiple parties, which leads to an export control of research data disclosure. However, for readers who are interested in knowing test results and procedures, please contact the authors for further information.

#### REFERENCES

- [1] W. Magnussen, iCERT NG911 Research Report, Apr. 2015; <http://www.theindustryCouncil.org/publications/>.
- [2] CSEC, accessed Mar. 2016; <http://www.csec.texas.gov/>.
- [3] Texas NG9-1-1 Master Plan Recommended Updates, May 2016; [http://www.csec.texas.gov/Texas\\_CSEC\\_NG911\\_Master\\_Plan\\_Update\\_Recommendations\\_051514\\_Final.pdf](http://www.csec.texas.gov/Texas_CSEC_NG911_Master_Plan_Update_Recommendations_051514_Final.pdf)
- [4] Capgemini Consulting, accessed Mar. 2016; <https://www.capgemini-consulting.com/about-capgemini-group>.
- [5] NENA, accessed Mar. 2016; [http://www.nena.org/?page=NG911\\_ICE](http://www.nena.org/?page=NG911_ICE).
- [6] HTTP-Enabled Location Delivery (HELD), Sept. 2010; <https://tools.ietf.org/html/rfc5985>.
- [7] "Understanding NENA's i3 Architectural Standard for NG9-1-1," June 2011; [https://c.yimcdn.com/sites/www.nena.org/resource/resmgr/Standards/08-003\\_Detailed\\_Functional\\_a.pdf](https://c.yimcdn.com/sites/www.nena.org/resource/resmgr/Standards/08-003_Detailed_Functional_a.pdf).
- [8] NPSTC Library, accessed Mar. 2016; <https://www.nena.org/?ANSProcess>.
- [9] NENA ANS Process, accessed Mar. 2016; <http://www.npstc.org/>.

#### BIOGRAPHIES

WALTER R. MAGNUSSEN (w-magnussen@tam.u.edu) is currently the Director of the Texas A&M University (TAMU) ITEC. He has his Bachelor and Master degrees from the University of Minnesota and his Ph.D. from TAMU. He oversaw the implementation

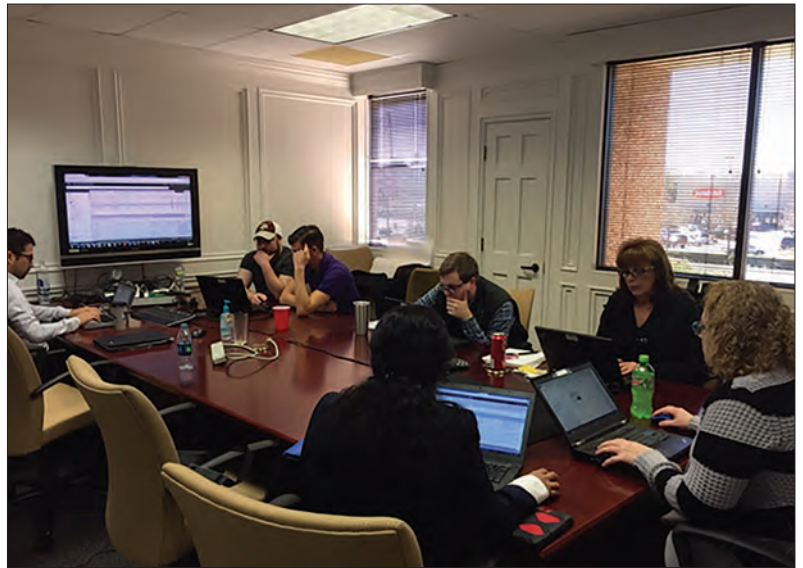


Figure 5. The actual testing in College Station.



Figure 6. TAMU ITEC laboratory.

of the U.S.DoT NG 9-1-1 proof of concept as well as several FirstNet application experiments. He currently serves on several NENA, APCO, and NPSTC committees, and he has served on Federal Communications Commission (FCC) committees. He has numerous presentations and publications to his credit.

PING WANG (pingwang.tamu@gmail.com) is currently a research assistant with the TAMU ITEC. She is currently a Ph.D. student studying in the Department of Electrical and Computer Engineering, TAMU. She has published nine papers in IEEE journals and conference proceedings. Her research interests are in NG 9-1-1 systems, LTE, QoS, wireless resource management, multicore and distributed computer architectures, and parallel computing.

YANGYONG ZHANG (yangyong@tam.u.edu) is currently a graduate assistant of the TAMU ITEC and a Ph.D. student studying in the Department of Computer Science and Engineering, TAMU. He has his Bachelor degree from the State University of New York, Buffalo, and is now a member of the SUCCESS lab directed by Dr. Guofei Gu, where his research focus is on SDN security. He has a broad interest in public safety networks, wireless SDN, and network security.

# Implementation of NG9-1-1 in Rural America—The Counties of Southern Illinois: Experience and Opportunities

Barbara Kemp

The author describes the deployment of a next generation 9-1-1 network in Southern Illinois. Thirteen counties and one municipality banded together to design, build, test and deploy this network, which provides voice, text, video and data services to emergency callers, call takers and first responders. She describes key challenges where contributions have been and will continue to be made through the collaboration of industry, academia, government, and standards bodies.

## ABSTRACT

This article describes the deployment of a next generation 9-1-1 network in Southern Illinois. Thirteen counties and one municipality banded together to design, build, test and deploy this network, which provides voice, text, video and data services to emergency callers, call takers, and first responders. The author describes key challenges where contributions have been and will continue to be made through the collaboration of industry, academia, government, and standards bodies. Lessons learned and potential next steps are examined.

## INTRODUCTION

The Counties of Southern Illinois (CSI) is a consortium of 13 counties and one municipality, clustered close to the Indiana border on the east, the Kentucky border on the south and Missouri on the west. Together they created a not-for-profit legal entity to build a next generation network to deliver emergency services with capabilities including video, text, and data from any device anywhere. The counties had a history of working together during disaster recovery efforts of various kinds. They are situated in the New Madrid Fault Earthquake zone, and are also frequently ravaged by severe storms, tornadoes, and derechos. They viewed the next generation 9-1-1 network (NG9-1-1) as a necessary next step. The work began in 2007, and by June 2015, the last of the 17 Public Safety Answering Points (PSAP) in the project's scope had been brought online. The project was honored February 23, 2016, by the NG9-1-1 Institute for its "leadership and efforts to improve the nation's 9-1-1 system." [1]

CSI faced many challenges as the project began. The counties lacked an adequate infrastructure for the delivery of traditional 9-1-1 services. The existing infrastructure had many single points of failure (SPOFs), and lacked redundancy and diversity. Elements of circuit-switched networks that are responsible for routing a 9-1-1 call to the correct PSAP, called selective routers (SRs) [2], were not redundant. Other equipment was aging and had been discontinued by the manufacturer. Four of the counties did not have E9-1-1 access for wireless callers, although 70 to 90 percent of all 9-1-1 calls originate from mobile phones [3]. In fact, wireless 9-1-1 access was first provided to

Alexander, Pope, Hardin, and Hamilton Counties through participation in the CSI initiative.

The National Emergency Numbers Association (NENA) specifies a set of requirements for next generation 9-1-1 networks, called the i3 standards [4]. The goal of CSI was to adhere to these requirements as closely as possible, given the state of the existing infrastructure and funding, and the relatively low density of the population in their area. The Illinois Communications Commission (ICC), which needed to review and approve the CSI plans, took these factors into consideration. The adaptations that were made by CSI will help inform and in some cases modulate the existing standards and best practices.

When CSI began its work in 2007, the NENA standards were in their infancy. There was as yet no transition plan for moving communities from legacy public switched telephone network (PSTN)-based 9-1-1 and E9-1-1 infrastructure to the IP-based emergency backbone specified in the NENA standards. NENA published a set of considerations for a transition plan in 2013 [5].

The remainder of this article is organized as follows. The next section provides an overview of the CSI network. We then present requirements that guided the work. Following that, we describe the testing that was done to ensure that 9-1-1 calls would be successfully terminated. The final section contains lessons learned and concluding remarks.

## OVERVIEW OF THE CSI WORK

Components of the CSI architecture include:

- A physical broadband data network with a fiber core
- A managed IP network and its associated switches and routers overlaying the broadband network
- Two data centers where the various networking and NG components are housed
- Multiple PSAPs where operators, called telecommunicators, answer emergency calls
- Interfaces to the various access carriers that serve the CSI area

Figure 1 illustrates key aspects of the architecture.

A 9-1-1 system service provider (SSP), *NG 911, Inc.*, was selected to have overall responsibility for the end-to-end emergency service. Another provider, *Clearwave Communications*, built an underlying broadband network funded through



In order to achieve load balancing, a set of engineering rules were implemented. Primary and alternate routes for phone calls from the access carriers' networks to the datacenters were established. Included with these were engineering rules for all participating access carriers.

and actual operational experience have proved this to be the case. In actual operation on the live network, backup PSAPs have been able to handle all calls in the event of outages.

#### LOAD BALANCING AND REDUNDANCY

Each data center was designed and built with the full capability of supporting all the 9-1-1 traffic originating in the CSI serving area. All the calls delivered to the ESInet are able to appear at either data center. Load balancing between the data centers when they are both up and running is built into the system. Both centers are designed to be up and running at all times so that either can take over the entire operation if necessary, and both are operational so that any defects in either one can be fixed before that one is required to handle the load of both. IP network service providers and equipment vendors have default configurations and best practices that may not conform to the particular requirements on NG9-1-1 systems, so care was needed to prevent these from overwriting the CSI requirements. The centers, for example, must not be in an active/standby mode, where one works and other is idle, but default configurations on some equipment implements just such an operating mode.

In order to achieve load balancing, a set of engineering rules were implemented. Primary and alternate routes for phone calls from the access carriers' networks to the data centers were established. Included with these were engineering rules for all participating access carriers.

The NG9-1-1 architecture conversion included a formal method of operation (MOP). As part of the MOP, plans were created to adjust the staffing of PSAPs and other functions in case of actual failures. In the case of a recent unplanned event, there were no calls completing to three PSAPs covering four counties, but the alternate PSAPs handled all the calls via the system reroutes while service was restored, and no 9-1-1 calls were lost. The event was observed by CSI's monitoring system, but a public formal FCC Report was not required since no calls were lost and the public was not endangered.

#### DATA CENTERS AND PSAPs

The data centers were housed in two PSAP equipment rooms located within the County Sheriff's Offices. These locations provided significant physical security, commercial and uninterrupted power, and generators with battery backup. The buildings are close to the central offices of the legacy carriers, and facilities were easily extended to connect to the data centers. Private line data circuits to the national database providers for mobile and VoIP, *West Safety Services (formerly Intrado)*, *Comtech Telecommunications Corp. (formerly TCS)*, and *Bandwidth.com*, were required to terminate at the new data centers rather than at the old PSAP locations.

Physical layer requirements for each data center and PSAP had to be verified. These include proper grounding, synchronization and timing, diversity and reliability of commercial power, uninterrupted power supplies, power generators, battery backup and associated maintenance logs and procedures, heating, air conditioning, physical security, cleanliness, space for the added NG9-1-1 equipment and racks and for terminating network facilities, and cabling adequate for the equipment.

#### DATABASES AND ROUTING

Geographic information systems (GISs) that capture, store, display, analyze, and manage spatially referenced data [2, p.89] are essential to the delivery of NG9-1-1 services. Much of the data required for these systems can be found in legacy automatic location information (ALI) databases and the master street address guide (MSAG), both of which are used by the local 9-1-1 authorities and are maintained for them by third-party vendors. Before the GIS data can be used for NG9-1-1 services, the records must be reconciled and synchronized with the legacy ALI and MSAG information. The creation of the GIS records, and their testing and maintenance is a time-consuming process and requires meticulous attention to accuracy and detail.

CSI creates and maintains the databases under the direction of the 9-1-1 SSP. The monitoring and intelligent alerting system used on the CSI network is capable of detecting exceptions that impact the GIS database and making these exceptions known to their third-party vendors *West Safety Services*, *Comtech Telecommunications Corp.*, and *Bandwidth.com*. Making sure this information is kept accurate in near real time and is shared with the database providers is not a trivial effort. It is the opinion of the CSI implementation team that top-down management of rules and responsibilities at the state level will be required to make NG9-1-1 work since eventually ESInets will need to interoperate across boundaries.

NG9-1-1 routing decisions are based on the geographic proximity of the caller to the PSAP and may also depend on other factors such as whether the PSAP is staffed, and whether it can provide special services such as foreign language support. These routing decisions are made, and may be changed, in near real time. Overflow and alternate routes in case of PSAP failures are determined by CSI. Routing rules and transfers to alternate PSAPs were defined and tested prior to cutover. This was especially challenging when the first PSAPs came online. Some counties had only a single PSAP. If that PSAP is brought online, it may be backed up by a legacy PSAP in a neighboring area that is not yet on the NG9-1-1 network. The delays introduced by the re-routing can cost lives. Since CSI was the first NG9-1-1 provider in Illinois, and no other adjoining states were ready to test interfaces, ESInet-to-ESInet testing was not possible. Building routing and transfer agreements and testing these between ESInets are both left to the future.

#### NETWORK MANAGEMENT AND MONITORING

Network monitoring has historically been the responsibility of the access carriers who own and manage the legacy SRs, and perform the physical routing and connection to the PSAPs. But access carriers are no longer solely responsible for the emergency function. As the responsibility has spread among various vendors and service providers, the state commissions and the PSAPs demand that one party be responsible for monitoring the end-to-end quality of the emergency service. Monitoring of the network is not a NENA requirement at this time. NENA has recently created a working group to develop recommendations on operational procedures associated with the transition to NG9-1-1. [6] CSI developed its requirements based on their experience and that of the organi-

zations with which they contracted. These requirements are described below and are summarized as follows. The network management system must monitor the ESInet and the data centers including their signaling, switching and data functions. It must also monitor the PSAPs, including their equipment, telecommunicator login status, and infrastructure sensors. Alerts based on the results of monitoring must be sent to the appropriate organization using text, email, trouble ticket generation, and/or network operations center (NOC) operations support system (OSS) and display. Such a monitoring system will be able to provide information to all participating organizations including wireline, wireless, and VoIP carriers, and database service providers. The monitoring system that meets these requirements and was selected by CSI is the product of Assure9-1-1, whose chief technical officer is the author of this article.

## REQUIREMENTS

Below are key requirements that were derived from many sources including the experience of the CSI team and the technical references and specifications mentioned above. The physical data network and its fiber core, together with the managed IP network and its data centers, must meet the same standards that the legacy PSTN-based networks did. Surveillance and reporting mechanisms that allow the 9-1-1 SSP to react in near real time to any network failures and anomalies are mandatory but as yet unspecified. These must enable the maintenance, recording, and reporting necessary to keep a lifeline service available and efficient at all times. Requirements for the PSAPs and data centers specific to the NG9-1-1 standard are considered below.

Figure 2 is a simplified illustration of a network that includes access to the ESInet as well as the ESInet and the PSAPs. It is included in this article to help the reader unfamiliar with the many specialized terms used here. A legacy network gateway (LNG) is shown providing access from the PSTN. Services on the ESInet include the emergency call routing function (ECRF), the emergency services routing proxy (ESRP), the border control function (BCF), and a legacy PSAP gateway. Descriptions of these elements can be found in [4].

## DATA CENTERS

The requirements for the data center and other components of the architecture are derived from the NENA i3 standard [4] and from FCC rulings such as the one described in [7]. Requirements for the data centers include the following:

- There must be a minimum of two geographically diverse data centers, each with a full complement of NG components.
- Each data center must be fully capable of serving the total network load.
- Data centers must provide a border control function for security.
- They must load share in real time.
- Each of them must provide full access to each PSAP.
- They must facilitate automatic failover to an alternate PSAP(s) should the primary PSAP fail.
- Finally, they must be secure and protected from denial of service attacks or general overload conditions, whether generated externally or from within the ESInet and PSAPs.

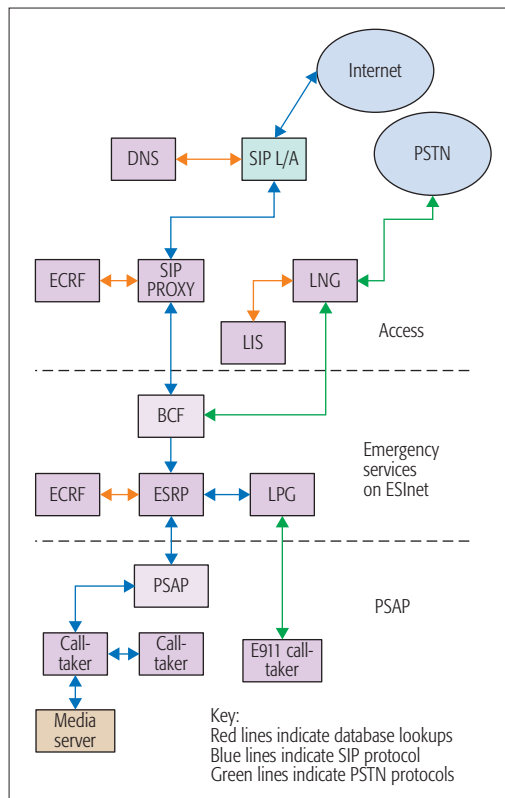


Figure 2. Access, ESInet, and NG components.

The full capability required of the redundant data centers includes the BCF, database, routing, and rules-based servers, and the logging/recording and monitoring systems. A major concern at the data center is the integration with the legacy PSTN infrastructure. Many challenges exist due to the presence of multiple access carriers, each with its own administrative methods, databases, and network implementations.

## MONITORING AND MAINTENANCE SYSTEMS

The system must be capable of monitoring the ESInet at an interface to its support system. The call processing core must be monitored for any exceptions including lack of heartbeat on a very short duty cycle. Other systems that need constant monitoring include the database infrastructure — the LIS, ECRF, and ALI Links to the various services providers and the sensors in the heating and air conditioning systems, power equipment, and in generators and battery alarms. Information from all these sources adds to the patterns of trouble that can be seen and resolved.

The system must look for any deviation from the norm. Failures are not the only things that require alerts. Overloads and abnormal conditions, which may indicate the need to change the service delivery method, must also be detected and logged, and alerts generated. Abnormal patterns may require operations changes such as opening backup centers, shifting staff, or bringing in added resources. Alerts must be shared with responsible entities as they occur.

The system must report the number of customers at risk for a given reported event. Reports that include this information are mandated by the ICC and FCC. Indicators must provide the reasons that

A major concern at the datacenter is the integration with the legacy PSTN infrastructure. Many challenges exist due to the presence of multiple access carriers each with their own administrative methods, databases and network implementations.

As part of its approval process, the State of Illinois required a test plan. Network testing took place over several months culminating in the network test report to the ICC. Several types of testing were done, including: network design, network operation, security, and performance.

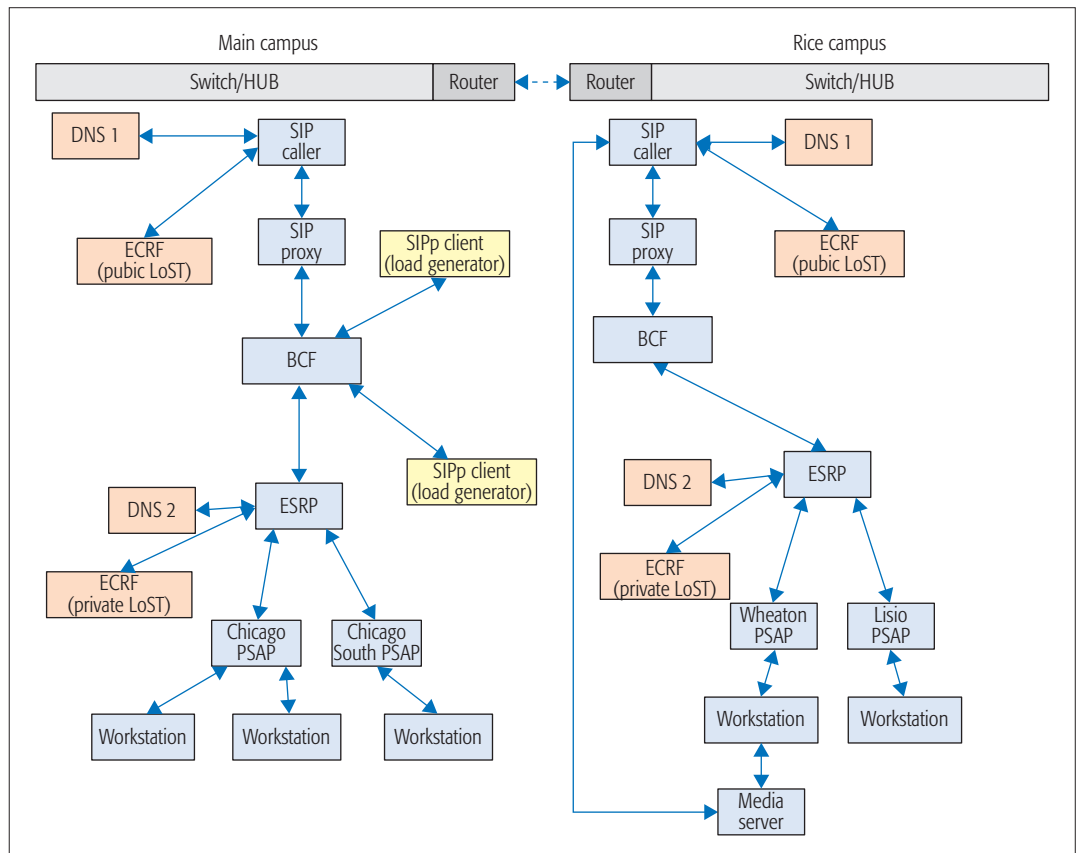


Figure 3. Lab test architecture for conformance and failover tests prior to deployment.

the report was issued. An expert system that learns the patterns of failure and typical root causes should enable the NOC, over time, to predict and avoid failures by indicating that corrective action be taken in advance of failures. It should be able to provide the most likely causes of the problem, for example, power outage, fiber cut, loss of logging and recording devices, routers down, and connections to carriers lost. Displays must be useful to several classes of users: database managers, PSAP managers, those who must make legal and regulatory reports, and the technical support personnel who must find and resolve the problem quickly. The information collected must be sorted and categorized for action. These types of failures are typical of what has been observed in CSI. No negative service impacts have been observed due to the diversity of the network and the ability to rapidly generate alerts.

### TESTING

As part of its approval process, the State of Illinois required a test plan. Network testing took place over several months, culminating in the network test report to the ICC. Several types of testing were done, including: network design, network operation, security, and performance. In all tests, whether done in the lab or on the actual network, an originating caller is attempting to reach a PSAP telecommunicator. Such calls are referred to here as “end-to-end” calls.

#### NETWORK DESIGN TESTS

The network design and operation were tested for conformance to CSI’s technical requirements. The architecture was tested first in the Real Time

Communications Lab (RTCL) at the Illinois Institute of Technology (IIT) [8] and then in the field after the installation of the network elements was complete. Figure 3 illustrates the RTCL test network. The transport consisted of two routers and two switches located on different campuses. Each switch represents a domain. Routers and a virtual private network (VPN) tunnel connect the two domains. Each switch is home to its own data center, ESInet, emergency services applications, and PSAPs. In Fig. 3, the logical communication paths are shown, while the physical connections to the switches and router are not. The network does not duplicate the products on the CSI architecture; rather, it serves to demonstrate that the CSI solution meets the NENA communications standards for signaling, routing, information retrieval, overload, and handoff. Later in the project, after the network and equipment were in place, the university lab remotely tested the security and overload capabilities of the actual network. The tests performed on the lab network were eventually performed on the CSI network in each data center and PSAP under the supervision of ICC staff and the participating vendors and service providers.

The lab testing included resiliency, security, and performance tests. Resiliency tests dealt with fail-over scenarios and expected outcomes such as the following:

- \* A function such as the ESRP or ECRF fails for one of several reasons. Test that a 9-1-1 call sent to its ESInet data center will fail over to the redundant element in the mirrored data center.

- The PSAP associated with one network fails for one of several reasons. Test that a 9-1-1 call will failover to the other network's PSAP.

Security tests dealt with scenarios in which unauthorized callers attempt to access management ports on the legacy network gateway (LNG) or the border control function (BCF), and ensure that the PSAPs' call takers can originate and terminate calls, filtering out unauthorized access. Performance test scenarios included the system's response to distributed denial of service (DDoS) attacks and characterizing the system's behavior when multiple test calls make all telecommunicator workstations busy. The complete set of tests is on file with the ICC.

### FIELD TESTING

Actual field testing on the CSI network began after the databases and networks were built and the PSAPs were ready. In the field, vendor products and IP-based call flow underwent the same tests that were performed in the university lab in addition to many more. Cooperation with the access-carriers was essential. Accurate information about the identity and configuration of their switches, SRs and facilities was required, not only when creating the connections to their networks, but also for testing these to ensure proper routing. Wireline, wireless, and VoIP access carriers have different methods for routing and identifying emergency callers so, for each type of access carrier, different tests were needed. Test numbers for each carrier were arranged, databases obtained, and end-to-end testing with each was undertaken.

For wireline carriers, each exchange was tested separately. For wireless carriers, each cell tower and each emergency service routing key (ESRK) range needed separate testing. VoIP carriers have their own methods for routing and identifying emergency callers, and each needed their own set of tests. End-to-end testing included validating split exchanges—number groups, some of whose calls are routed to one PSAP and others to a different PSAP.

The response of the network monitoring system was observed during this phase of testing. The monitoring system had an interface to a smartphone application configured to reflect the CSI network. As elements were removed from service by means of power interruptions or the removal of cables, the monitoring system's responses were tested. Exceptions are detected and analyzed, and alerts are texted to the designated user in a secure fashion. The user is able to see each exception, and the date and times of the events or anomalies. The access carriers and the 9-1-1 SSP used Wireshark, a freely available IP-based protocol analyzer, to verify correct operation of the network both before it was turned up and during the cutover. This tool was also used to verify the effectiveness of the 9-1-1 SSP's network monitoring system.

### LESSONS LEARNED AND CONCLUDING REMARKS

The implementation of NG9-1-1 networks is a large undertaking. Standards and best practices are evolving, and each cutover will yield new information that will contribute to the evolution-

ary process. In this work IT professionals and telecommunications engineers and administrators must learn each other's tools and terminology. Systems and processes that stood the test of time in the pre NG9-1-1 world still have value. Cooperation is necessary, and funding sources must be available. Below are some observations and lessons drawn from the CSI experience. It is hoped that these will inform future versions of the standards and best practices which will support the emergency services community as it moves toward next generation capabilities.

### CUTOVER AND TRANSITION ISSUES FOR NG9-1-1

After all tests were completed and problems resolved, the cutover of the network and the various PSAPs was scheduled. The initial two data centers and first pair of PSAPs, selected for their limited coverage area, were cut over in a single day. The least busy day for the PSAPs is the best choice, and for 9-1-1 that is usually a Wednesday or Thursday. The following operating guidelines should be helpful to organizations choosing to make such a transition:

- Identify a person at each access carrier who will test with the organization, and provide them with all necessary test numbers.
- Choose friendly customers and/or governmental agencies to make the calls that will be handled, transferred, and verified.
- Do not skip tests or go too fast in case tests go well.
- Identify a call leader and provide a conference bridge including all locations on the test.
- Vendors should be available to resolve issues with their products.
- Test tools including protocol analyzers and monitoring systems, together with technical support personnel need to be present during cutover.
- Database personnel must be on the call to resolve any issues with boundaries.
- The access carriers, the state regulatory staff, and the 9-1-1-SSP may have representatives on the calls. Legal and regulatory personnel should be notified and on call.
- Train the PSAP personnel shortly before the day of cutover. 9-1-1-SSP personnel should remain in each PSAP for several days after the cut to provide support.
- Leave the old network in place for an agreed upon timeframe following the cut. In case of unforeseen problems with the new network, a fallback network is needed.

### ACCESS SWITCHES AND SIGNALING

All call signaling on the ESInet uses SIP. Thus, an efficient technical solution for delivering calls that originate on an access carrier's network to the ESInet would be to connect the access carrier's network to the ESInet using a SIP trunk, a data facility that sends SIP messages between SIP proxy functions at each end. Most access carriers were reluctant to use SIP 9-1-1 direct trunking, however. Various reasons for this reluctance can include confusion about which entity must pay for the direct, diverse facilities to the two data centers. Additionally, if the carrier does not already have a SIP interface associated with its switch, it

In the field, vendor products and IP-based call flow underwent the same tests that were performed in the university lab in addition to many more. Cooperation with the access-carriers was essential. Accurate information about the identity and configuration of their switches, SRs and facilities was required, not only when creating the connections to their networks, but also for testing these to ensure proper routing.

The legal requirements and regulatory matters for NG9-1-1 are massive. The CSI Project team worked closely with federal and state commissions to change the language in the requirements for NG9-1-1 and create a baseline of documentation for others to follow. Attorneys helped provide guidance and common sense suggestions.

may not want to take on the additional expense. Thus, the access for 99 percent of the calls to CSI still goes through a daisy chain of up to three legacy SRs to reach the NG9-1-1 SR. This architectural choice has led to at least one routing failure and to the blocking of some landline calls. The 9-1-1 SSP's monitoring system also was not able to process the access carrier trunk and transport exception data that resulted from these failures.

The data in the local exchange routing guide (LERG), which provides numbering plan codes, central office types, and points of interface, was not valid in many cases: carrier test numbers were valid less than 10 percent of the time, requiring CSI to get the correct central office test numbers from each access carrier and add them to the NG9-1-1-SSP's databases prior to testing. The types of switches, and their locations and configurations were also unreliable: The switch type is important information to the monitoring system, which may be able to trace abnormal behaviors to certain switch types. Split exchanges within switches, situations in which some calls route to one PSAP and others to a different PSAP, were also not reliably identified.

Interworking was compromised in some cases by the fact that some legacy SRs were several generic releases behind those required for NG9-1-1 compatibility. The reasons were financial, with the owners of these SRs questioning why it was their responsibility to pay for the upgrades.

### PSAPs AND CONSOLIDATION

Smaller PSAPs feel economic pressure to consolidate into larger installations, and the risk of non-redundant copper or fiber failure is always present. If the resulting consolidated PSAP is large, and there are no other PSAPs capable of handling the loads in near real time at peak hours, a dual fiber facility or alternate path of copper, wireless, or satellite technology is highly recommended and should be included in the network design and in its test plan. The strength and knowledge of the existing PSAPs and their staff were very important to the success of the CSI cutover. Experienced telecommunicators know their area and geography. Rapid response is essential for emergency services, and experience is important when the delivery system is new.

### SMALLER PSAPs VS. LARGER PSAPs

There were many conditions in the CSI footprint that are not a match for the large PSAPs and high-density communities across America. When a small PSAP fails, the number of calls and customers impacted is not great. The extra load will seldom require a change in staffing. When larger PSAPs fail, staffing at the backup location must be adjusted. An alert should be sent to the parties who must act. Such near-real-time alerts are usually mandatory under state and/or federal requirements. These alerts are in addition to the technician alerts that go to the person fixing the problems. In most cases, the diversity of the NG9-1-1 network means the service quality is rarely impacted while technical issues are being resolved.

Testing directly with enterprise customers was not necessary in the CSI project since CSI had no private switch/ALI service customers with direct trunking to the PSAP. Projects with a larger, more

diverse customer base may need to perform such testing, however.

### PUBLIC POLICY AND FUNDING ISSUES WITH NG9-1-1

The legal requirements and regulatory matters for NG9-1-1 are massive. The CSI Project team worked closely with federal and state commissions to change the language in the requirements for NG9-1-1 and create a baseline of documentation for others to follow. Attorneys helped provide guidance and common sense suggestions to get cooperative agreement with the state and the access carriers to progress the work. A better financial model and funding method would help move the country toward NG9-1-1 more rapidly.

### ACKNOWLEDGMENTS

Thanks to Professor Carol Davids of the IIT RTCL for her editorial support, and to the many colleagues who provided technical advice and support for this work.

### REFERENCES

- [1] 13th Annual 9-1-1 Honor Awards (2016). ng911 Inst. N.p., 2016; accessed 22 Sept. 2016; <http://www.ng9-1-1institute.org/events/2015-honor-awards-gala/2016-9-1-1-honor-award-winners>.
- [2] NENA Master Glossary of 9-1-1 Terminology — NENA-ADM-000.18-2014, 07/29/2014 prepared by NENA Development Steering Council; [www.nena.org/resource/resmgr/Standards/NENA-ADM-000.18-2014\\_2014072.pdf](http://www.nena.org/resource/resmgr/Standards/NENA-ADM-000.18-2014_2014072.pdf).
- [3] FCC, "911 Wireless Services.," accessed 22 Sept. 2016; <https://www.fcc.gov/consumers/guides/9-1-1-wireless-servicesAccess>.
- [4] "NENA Detailed Functional and Interface Standards for the NENA 13 Solution," 1st ed., 2016; accessed 22 Sept. 2016; [https://www.nena.org/default.asp?page=i3\\_Stage3](https://www.nena.org/default.asp?page=i3_Stage3)
- [5] "NENA NG9-1-1 Transition Plan Considerations Information Document," 1st ed., , accessed 22 Sept. 2016; [http://c.yimcdn.com/sites/www.nena.org/resource/resmgr/Standards/NENA-INF-008.2.1-2013\\_NG9-1-1.pdf](http://c.yimcdn.com/sites/www.nena.org/resource/resmgr/Standards/NENA-INF-008.2.1-2013_NG9-1-1.pdf).
- [6] "NENA NG9-1-1 System and PSAP Operational Features and Capabilities Requirements Document 57-750," 1st ed., accessed 22 Sept. 2016; <https://www.nena.org/?page=Standards>
- [7] "FCC Imposes New Requirements For 911 System Service Providers | Commlawblog," *CommLawBlog*, accessed 22 Sept. 2016; <http://www.commlawblog.com/2013/12/articles/cellular/fcc-imposes-new-requirements-for-911-system-service-providers/>.
- [8] "RTC Lab | IIT School of Applied Technology.," accessed 22 Sept. 2016; <http://appliedtech.iit.edu/rtc-lab/>

### BIOGRAPHY

BARBARA KEMP (barb1kemp@aol.com) is a partner in Assure911 LLC. Her mission is to ensure that clients have the ability to monitor and manage their emerging 9-1-1 networks end-to-end. Her current focus is next generation emergency services planning, design, testing, and monitoring. She taught network management at the Illinois Institute of Technology (IIT) as an adjunct professor, and she mentors graduate student IIT RTC Laboratory projects. She is a member of the IIT 911 Board of Directors headed by Professor Carol Davids. She has been on the team doing design, engineering, test planning and cutover, and ongoing support for the Counties of Southern Illinois NG9-1-1 Project. She is actively engaged in the Northern Illinois Next Generation Alliance project. She was project manager for Cable and Wireless Caribbean operations in 2006 and 2007 consolidating NOCs and building NOCs for 13 Caribbean nations. She has managed a wide range of IP and TDM consulting projects for telcos and cable TV companies nationally and internationally. She has designed and built major NOCs, including AT&T's Midwest Reliability Center and the ICG Telecommunication's Service Reliability Center as senior VP, Service Reliability, Quality Assurance, and Security. The centers included 9-1-1 problem resolution, business processes, systems integration, and security. In 1997, she managed a 9-1-1 technical resolution project after the City of Chicago embarked on their new city-wide OEMC and 9-1-1 Center, heading a task force. Her background includes engineering, operations, and system requirements design and implementation. Barbara is a graduate of Indiana University, 1983, and the Wabash College Executive Management Program, 1992.



# In-Vehicle Emergency Call Services: eCall and Beyond

Risto Öörni and Ana Goulart

## ABSTRACT

What is the status of in-vehicle emergency call services? Which standards are being adopted? Does the NG-911/112 architecture support such services? The objective of this article is to address these questions. To do so we review the evolution of eCall, compare approaches developed in different parts of the world, and discuss interoperability between selected systems. This study shows that it is challenging to compare and classify in-vehicle emergency call systems because of different standards, terminologies, and proprietary specifications. We conclude that the NG-911/112 framework provides the building blocks to support next generation eCall, and can contribute to a common standard for the interface between private service centers and public safety answering points.

## INTRODUCTION

According to the U.S. National Highway Traffic Safety Administration, 32,675 people died in traffic accidents in the United States in 2014 [1]. In the same year, the European Commission (EC) reported 25,900 fatalities [2]. Reducing the number of fatalities and injuries has been identified as an important objective of transport policy.

The consequences of traffic accidents depend on the timely arrival of emergency services. It may be delayed when no one in a vehicle can make an emergency call, or vehicle occupants have difficulty determining their location. This is more common at night, on the interurban road network, or in places with no landmarks. The response time depends on the ability of victims or bystanders to make an emergency call, and of the public safety answering point (PSAP) to locate the accident.

Technology has been used as a tool to help victims of car accidents. In Europe, the Universal Service Directive has mandated that cellular operators provide the location of wireless callers. In the United States, since 1996 the Federal Communications Commission (FCC) has required mobile operators to add the location information as well as the phone number of callers in emergency calls. However, the accuracy of the location provided by the network is usually best effort.

In addition to location information, there are other types of information that may support the PSAP in making a risk assessment. This information includes, for example, the cause of the emergency call, details and conditions of the vehicle, its direction of travel, and number of passengers.

Another technology that can help victims of accidents is *in-vehicle emergency call services*, in which the car's in-vehicle system (IVS) can automatically make an emergency call after an accident. The vehicle's occupants can also manually trigger an emergency call. The calls are connected using the cellular network and routed to an appropriate PSAP. A voice channel is then established between the IVS and PSAP.

There are several in-vehicle emergency call services being standardized or available on the market:

- eCall and third-party services supported eCall (TPS-eCall) have been standardized in Europe.
- ERA-GLONASS has been implemented in Russia.
- HELPNET is operational in Japan.
- Proprietary systems are available (e.g., GM OnStar™).
- Next generation eCall (NG-eCall) standardization efforts are under way in the United States and Europe.

Figure 1 illustrates the concept of eCall. While the architectures and implementations of vehicle emergency call services are different, all of them provide a voice connection between the vehicle and a PSAP or a private call center. They provide the PSAP with an exact vehicle location and vehicle information, known as the minimum set of data (MSD).

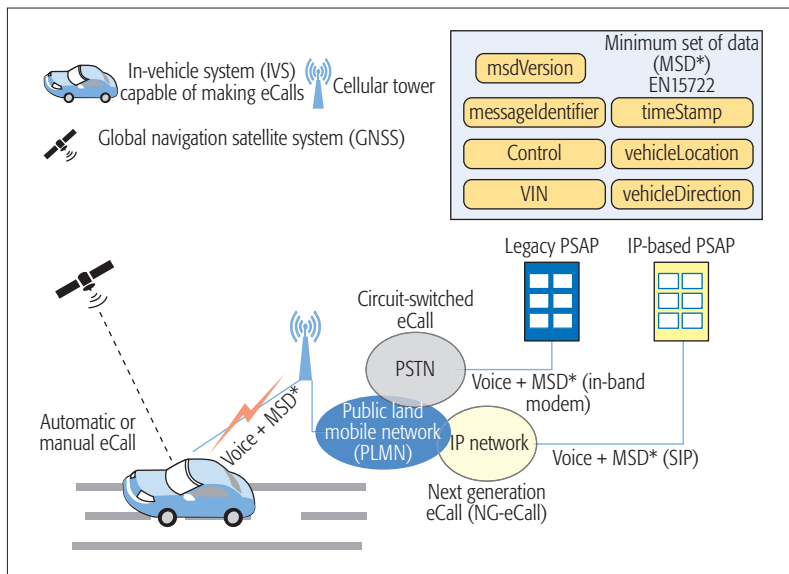
The Next Generation 9-1-1 (NG-911) proof of concept (POC) [3] in 2008 demonstrated an IP-based emergency call using the OnStar system. Now with vehicles that have cameras and sensors, and new generation cellular networks, how will the NG-911/112 framework support the evolution of eCall? To provide a big picture of in-vehicle emergency call services, this article offers these contributions:

- The history of eCall, and a description of in-vehicle emergency call services around the world
- A review of NG-eCall architecture
- Discussions on the interoperability of eCall and ERA-GLONASS, which share a regional boundary
- The evolution of eCall to NG-eCall

This article is organized as follows. In the next section we describe the requirements of in-vehicle emergency call systems. Standards in the circuit-switched domain are then reviewed, while next generation in-vehicle emergency call ser-

What is the status of in-vehicle emergency call services? Which standards are being adopted?

Does the NG-911/112 architecture support such services? The authors address these questions by reviewing the evolution of eCall, comparing approaches developed in different parts of the world, and discussing interoperability between selected systems.



**Figure 1.** eCall system concept. eCall standards are for the circuit-switched domain, but new standards for packet-switched NG-eCall are being developed.

VICES are discussed following that. Interoperability issues are then described, and conclusions are presented in final section.

## REQUIREMENTS

### REQUIREMENTS OF EMERGENCY CALLS

Requirements for citizen to authority emergency communications have been provided in European Telecommunications Standards Institute (ETSI) TR 102180 [4]. Emergency calls need to be free of charge, have high reliability and availability, be recognized and prioritized by the network, and be routed to an appropriate PSAP. The originating network must provide to the PSAP the caller's calling line identifier (CLI) and location. However, some requirements are subject to national regulations such as the availability of emergency call without a subscriber identity module (SIM) card and the possibility for the PSAP to call the mobile user back.

### REQUIREMENTS OF IN-VEHICLE EMERGENCY CALLS

The communication between IVS and PSAP is provided by public land mobile networks (PLMNs) and landline networks. These networks must provide the same reliability and security as a classic emergency call.

The eCall service has to offer cross-border interoperability, because vehicles cross national borders and because of the global and regional nature of the automotive market. In Europe, pan-European roaming capability was considered necessary. Also, there must be a defined PSAP that provides service for vehicle-originated calls in a regional area.

High-level functional and operational requirements for European eCall are summarized in EN16072, and the transmission of voice and MSD in 112 emergency calls are specified in ETSI TS 122.101. While the standards describe the eCall requirements, many of them are common to any other in-vehicle emergency call services, as follows:

- The call must be identified as an emergency call.

- The call must indicate if it is automatic or manual.
- The system must provide a voice connection between the PSAP and the vehicle.
- The MSD must be transmitted from the vehicle to the PSAP.
- The PSAP must acknowledge receipt of the MSD.
- The PSAP can request a new MSD from the IVS.
- PSAP can call the IVS back.
- The IVS can make test calls.

Next-generation eCall (NG-eCall) operates in the packet-switched domain; it offers elaborate new features [5] such as enhanced MSD with more than 140 bytes, calls with video and text, and commands to the IVS (e.g., a PSAP can send a request to unlock doors).

There are also requirements related to the in-vehicle environment and the usage context. First, the IVS must be physically robust to survive an accident. Second, the expected lifespan of vehicles is longer than phones'. Therefore, backward compatibility with previous generations of cellular networks and IVSs is required.

## IN-VEHICLE EMERGENCY CALL SERVICES IN CIRCUIT-SWITCHED NETWORKS

### eCALL

eCall is the European in-vehicle emergency call system. The functionality of eCall has been described in European Committee for Standardization (CEN) standards EN16062 and EN16072. An overview of the core standards of eCall was provided in [6].

Current standards specify eCall as follows. When the IVS is triggered manually, or the sensors in the vehicle detect an accident, the IVS registers to a second generation (2G) or 3G mobile network and makes an emergency call. The network uses a special emergency call type indicator – eCall discriminator – defined for manual or automatic eCalls to route the call to the most appropriate PSAP.

Once the call is connected, the IVS transmits the MSD to the PSAP. Transmission of the MSD takes place in the voice channel of a circuit-switched connection between the IVS and the PSAP using an in-band modem. A sample MSD is shown in Fig. 1. The MSD control information indicates the type of eCall and vehicle. The MSD also includes timestamp, location, direction as a number of two-degree steps, vehicle identification number (VIN), and propulsion sources present in the vehicle.

When the call is completed, the PSAP can hang up or send a *clear-down* to the IVS using the in-band modem. Then the IVS stays registered in the mobile network for some time. This allows the PSAP to call back the IVS.

Typical eCall behavior in error cases is explained in EN16062. If the MSD transmission fails, the call will continue as a voice call, and the PSAP operator may request an MSD retransmission by using the in-band modem. If the network registration fails, the IVS shall attempt to register with another mobile network and attempt an emergency call in limited service state. If the call setup fails or the call is disconnected before

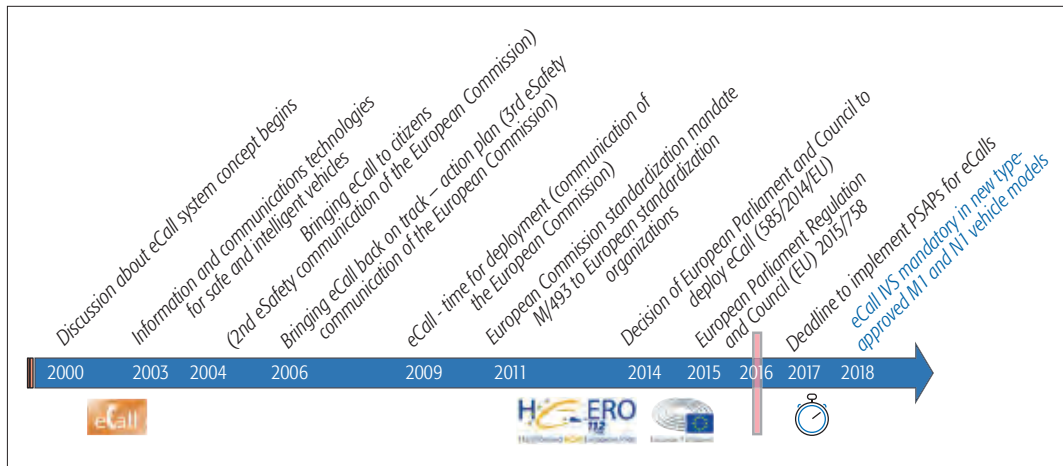


Figure 2. History of pan-European eCall – a timeline.

the MSD transmission is concluded, the IVS shall attempt to redial.

Security aspects related to the eCall service are discussed briefly in EN16072. While eCall shares many of the security aspects of a mobile emergency call, it seems unlikely that eCall would increase risks for the PSAP. First, an eCall IVS is required to have a SIM card that allows authentication of the subscriber. Second, IVSs that make repeated false calls can be blacklisted. The IVS registers with the mobile network only after automatic or manual activation. Therefore, it does not allow tracking of the vehicle.

Current eCall standards do not specify how automatic eCalls should be triggered, because this allows equipment manufacturers to provide innovative solutions. Instead, EN16072 requires that the trigger shall be “safe, robust and reliable” and provide as many positive detections and as few false positive detections as possible.

The history of eCall is illustrated as a timeline in Fig. 2. Talks about eCall began around 2000. Discussions about the service continued during the eSafety and Intelligent Car Initiatives of the EC. After 2009, European Commission shifted from a voluntary to a regulatory approach to achieve the deployment of the service to improve road safety. This was followed by a standardization mandate to European standardization organizations (2011) and decisions on the mandatory deployment of eCall in PSAPs (2017) and in new type-approved M1 and N1 vehicle models (2018) for all EU member states.

Before the decision on mandatory deployment in 2014, an impact assessment [7] was carried out on the European level and in several individual countries. For example, in Finland an analysis of accident records indicated that an automatic emergency call system could prevent 3.6 percent of fatalities [8]. It was assessed that the safety impact could be about 4–8 percent if possibly preventable fatalities were also included. Preparations for the deployment of eCall have been made in the Harmonized eCall European Pilot projects: HeERO and HeERO2.

### TPS-eCALL

Third-party services supported eCall (TPS-eCall) (Fig. 3) is a private service in which the voice call from the IVS and the related data set are received

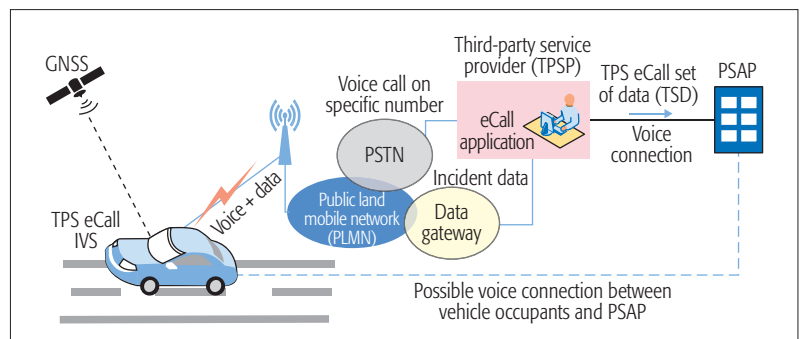


Figure 3. TPS-eCall – a high-level description (adapted from EN16102).

by a third-party service provider (TPSP). The TPSP talks with the vehicle occupants, receives the incident data from the vehicle, and determines whether the accident requires emergency services. If so, TPSP sends the TPS-eCall set of data (TSD) to an appropriate PSAP and provides information about the incident. The TPSP also makes best efforts to establish a voice connection between vehicle occupants and the PSAP if this is required by the PSAP.

TPS-eCall and TPSP-PSAP interfaces are defined in EN16102. In practice, private in-vehicle emergency call services that do not conform with EN16102 are also called TPS-eCall. According to European regulation, the car owner can opt for a private in-vehicle emergency call instead of eCall. However, the 112 eCall has to be available in all cars. The decision to support TPS-eCall in PSAPs to allow service providers to connect to PSAPs belongs to individual EU members and is subject to national regulations

### ERA-GLONASS

ERA-GLONASS [9] (Fig. 4) is the Russian in-vehicle emergency call system standardized by Rosstandart. The functionality of ERA-GLONASS is similar to eCall, although their service architectures are different.

The IVS initiates a TS12 emergency call with the eCall discriminator. The call is routed to the ERA-GLONASS regional switching node (RCC) to which the IVS transmits an emergency message (MSD) with the in-band modem. In case of

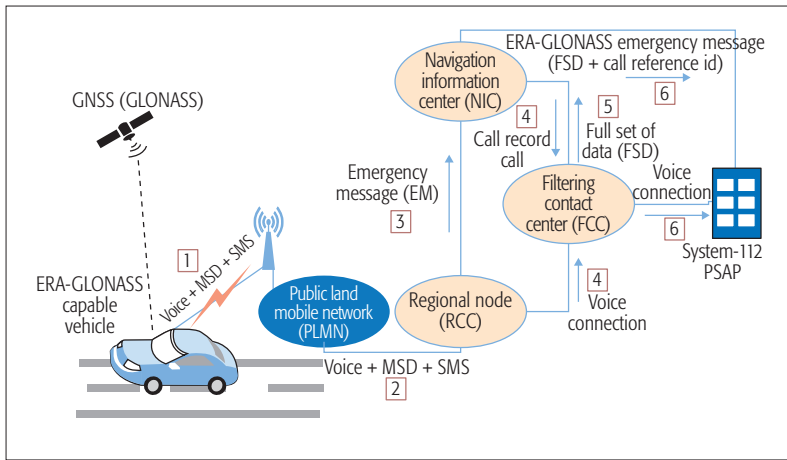


Figure 4. ERA-GLONASS architecture.

in-band transmission failure, the IVS transmits an MSD via short message service (SMS).

The RCC then forwards the MSD to the navigation information center (NIC), which decodes and reformats the emergency message, forms the full set of data (FSD), stores it in a database, assigns a reference identifier, and creates a call record card. Next, this call record card is forwarded to the filtering contact center (FCC), and the voice call is connected to the FCC at the same time. The FCC agent speaks to the occupants of the vehicle and verifies the emergency situation.

If rescue is needed, the FCC agent completes information in the emergency card and initiates a three-party conference with the appropriate PSAP. At the same time, the NIC forwards the ERA-GLONASS emergency card to the PSAP.

Once the call is cleared down, the IVS stays registered. This allows the FCC or PSAP to call the IVS back. The FCC may request an MSD retransmission via in-band modem or SMS.

The ERA-GLONASS system is already operational in Russia. The in-vehicle system has been mandatory in all new M and N class vehicles (i.e., passenger vehicles, buses, and trucks) beginning on January 1, 2017 and in new type-approved M and N class vehicle models since January 2015. Automatic or manual eCall initiation is mandated for M1 and N1 class vehicles under 2.5 tons. Manual eCall is mandated for all M and N class vehicles. Also, starting January 1, 2017, it is mandatory to trigger eCall automatically by rollover detection in all M and N class vehicles.

## HELPNET

HELPNET is an in-vehicle emergency call service that is operational in Japan. HELPNET is provided by a private company, Japan Mayday Service Co. Ltd., and is available for Toyota and Honda. HELPNET uses a 3G mobile network for voice and data transmission [10]. The service has similar functionality as TPS-eCall; it uses both circuit-switched and packet-switched networks. HELPNET employs two stages of answering service: a members center, and a HELPNET center. First, the IVS initiates a circuit-switched call to the members center and sends location, recent positions, sensor data, and subscriber ID using HTTP. Then the members center supplements the data with subscriber and vehicle information, and

sends it to the HELPNET center. If needed, it contacts the emergency services or the police via fax or HTTP.

In summary, from the business, standardization, and governance points of view, HELPNET has common characteristics with TPS-eCall.

## IN-VEHICLE EMERGENCY CALL SERVICES INTEGRATED WITH 911

In the United States, the National Emergency Number Association (NENA) provides guidelines for emergency calls. One of NENA's technical reports [11] describes the interface between third-party call centers and PSAPs, similar to TPS-eCall. It requires that the call center provide the network with the vehicle's location (not the call center location) so that the emergency call can be forwarded to a PSAP in the same jurisdictional area of the incident.

Additionally, PSAPs can accept an emergency call directly from an IVS. When it is answered, an audio message says: "this is an emergency call from a vehicle." The PSAP call taker can then talk with the vehicle occupants, or receive a computer-generated message with the vehicle's location with no additional vehicle data. An example of such a system is Chrysler's UConnect 911.

## NG-911/112 ARCHITECTURE AND eCALL

Standardization efforts for NG-eCall are in progress. The NG-911/112 framework's best practices are specified in the Internet Engineering Task Force (IETF) RFC 6881, and there are two IETF drafts that address NG-eCall and MSD delivery [12, 13]. Additionally, ETSI TR 103 140 [5] makes recommendations for eCall over voice over IP (VoIP).

One motivation for NG-eCall is that it provides a richer set of services and media types. Another motivation is that 4G cellular networks (e.g., LTE) operate in the packet-switched domain. NG-eCall uses Session Initiation Protocol (SIP) over TCP to set up the call. Once the SIP session is established, the IVS and PSAP can send voice, video, and other media packets over Real-Time Transport Protocol (RTP) and UDP. This is similar to regular IP-based emergency calls, or NG-911.

The NG-911 proof of concept [3] adopted the Emergency Services IP Network (ESInet) [12] as the core IP network to route emergency calls. It is managed by public safety authorities; for example, in Texas, the Commission on State Emergency Communications (CSEC) manages the ESInet. The PLMN sends all emergency calls to the ESInet. The ESInet's advantage is that it can set its own policies to route calls. Another approach is to use the IP Multimedia Subsystem (IMS) to route emergency calls, which is what ETSI proposes [5, 14].

An overview of NG-eCall and ESInet is shown in Fig. 5, where the PLMN directs the eCall SIP Invite request to the ESInet. There are two ESInets in Fig. 5: country-wide and regional. Inside the ESInet, there are several components needed for the routing process, as described next.

## NG-eCALL BUILDING BLOCKS

NG-eCall can be described in terms of five building blocks.

**Call Identification:** When an eCall is triggered, the IVS sends a SIP Invite message. In its header there is a service uniform resource name (URN)

to signal that it is an emergency call. The URN tells the PLMN that this call must receive priority and be routed to the ESInet. It must be unique for each type of eCall, that is, automatic, manual, or test eCalls. The following service URNs are proposed for NG-eCall [12]:

- urn:service:sos.ecall.automatic
- urn:service:sos.ecall.manual
- urn:service:test.sos.ecall

**Location Determination:** Location is needed for routing the call to the PSAP and dispatching emergency services. The SIP Invite message may have location information or a reference to a server that has the location. In the IMS emergency services infrastructure, it is assumed that the location is provided by the end user (which is the IVS in eCall) or by the network operator. This is also recommended by ETSI. IETF assumes that the end user or the network may also obtain the location from a location information server (LIS). Thus, when an eCall is triggered, the SIP request will have two sources of location information: location by value or reference in the SIP message and the location information in the eCall vehicle data (i.e., MSD).

**Location Conveyance:** The location information is conveyed using the geolocation header field of the SIP Invite. It indicates if the location is by value or by reference. Location by value means the location information is in the body of the SIP message, in an Extensible Markup Language (XML) format known as Presence Information Data Format Location Object (PIDF-LO). The IETF recommends using PIDF-LO for sending location information from IVS to PSAP.

**Call Routing:** The service URN and the location are inputs for routing the call to the PSAP. As shown in Fig. 5, when the SIP proxy receives the eCall SIP Invite, it sends the location of the caller to the Location-to-Service Translation Protocol (LoST) server. A LoST server contains a geo-database of all PSAPs' jurisdictional boundaries, and maps the caller's location to the PSAP's Uniform Resource Identifier (URI). This URI resolves to the IP address of a specific PSAP or to the IP address of an emergency services routing proxy (ESRP). In this case, the ESRP represents a group of PSAPs. The ESRP may then do another LoST query to find the appropriate PSAP.

**Additional Data Delivery:** This is a new component that complements the NG-911/112 architecture. For eCall, additional data means MSD or control data, and can be sent by both vehicle and PSAP. Different approaches to send the additional data are being discussed by IETF and ETSI:

- IETF uses the term *eCall-specific control/meta-data*. It is exchanged between IVS and PSAP in both directions. In one direction, the IVS needs to send the MSD and other information that the PSAP may request. In the other direction, the PSAP may send commands to the IVS to retransmit the MSD, perform an action, or send other data (e.g., video). These data can be sent in the header of the SIP message using the Call-Info header field, or in the body of the message as part of the PIDF-LO, by either value or reference [12, 13]. Consequently, the SIP message may contain multiple body objects. The SIP messages involved are SIP Invite and SIP Info, where SIP Info messages carry control requests and meta-

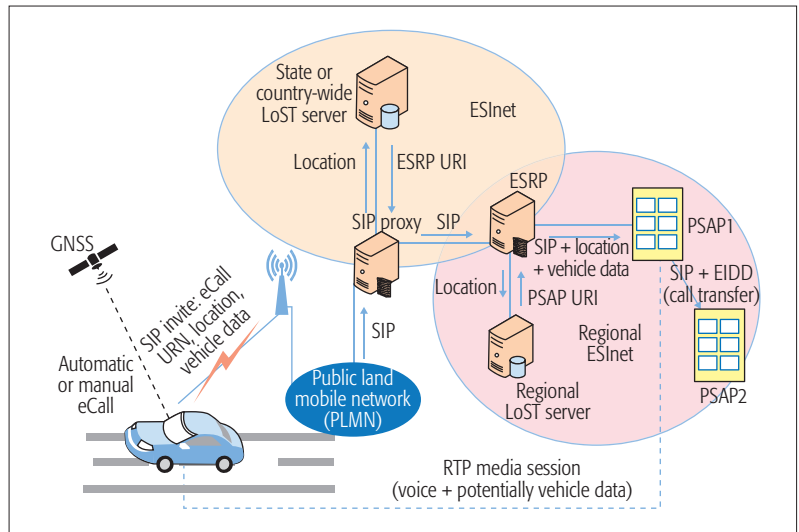


Figure 5. NG-eCall operation in the ESInet.

data during an ongoing eCall (i.e., a PSAP may request MSD retransmission or send a command to the vehicle in the middle of the call).

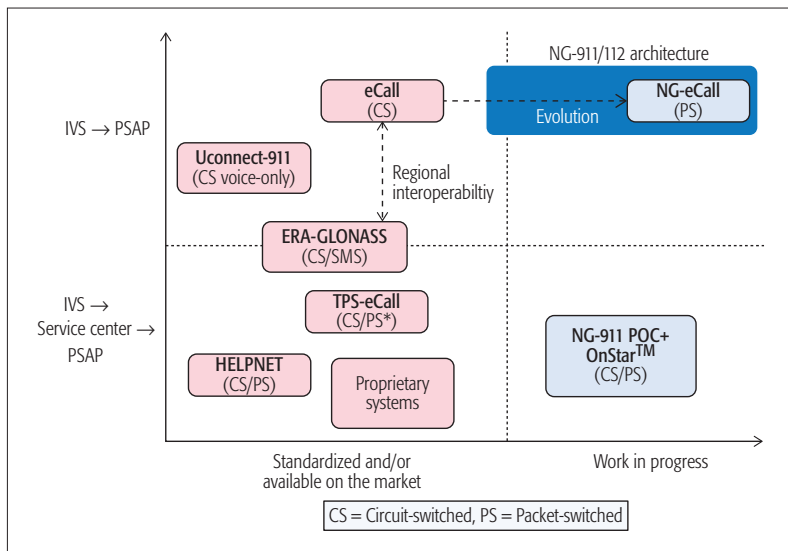
- ETSI also recommends that SIP messages carry MSD. Because this approach uses the control plane to send data, the ETSI report also considers the reliable transmission of text using an RTP media session, with Message Session Relay Protocol (MSRP) in the application layer.

Additionally, for IP-based emergency calls the NG-911/112 framework supports the transmission of Emergency Incident Data Documents (EIDDs) [15] in SIP messages. The EIDD can be transmitted between PSAPs, for instance, when a call is transferred to another PSAP (Fig. 5). An EIDD is an XML data structure. A component of an EIDD is called a *vehicle information data component*, which includes vehicle registration, time-stamp, and the relationship of the vehicle with the emergency situation.

### NG-eCALL SECURITY

The security of NG-eCall depends on the security of the mobile network and ESInet. Although ETSI assumes NG-eCall "inherits the security of the cellular network" [5], it is possible that NG-eCall uses less secure wireless networks, such as WiFi. In the ESInet, SIP proxies and ESRPs have access to SIP headers and must safely transfer these messages. Location and MSD privacy issues, such as secondary use and misattribution, are compelling [14]. Misattribution means that data related to one call is attributed to somebody else. Thus, how do the standards address security threats such as eavesdropping, denial of service, misrouting, PSAP or LoST server impersonation, or corrupted DNS responses? There is no single solution to prevent such attacks, but per NG-911/112 architecture, emergency call signaling should be protected using Transport Layer Security (TLS).

The use of TLS depends if data is sent by value or by reference. If by value, TLS should be enabled between SIP servers to encrypt the SIP messages' headers and bodies. While this requires a TLS handshake, persistent TLS sessions may be applied to reduce delays. If the handshake fails, the call must go through unprotected. On the



**Figure 6.** In-vehicle emergency call systems organized in terms of standardization efforts and architecture.

other hand, when SIP carries data by reference, the information will be accessed by PSAPs (and first responders) via HTTPS. Mutual authentication between servers and PSAPs is necessary; however, choosing or designing the right public key infrastructure (PKI) that works across service areas is a challenge. Another open issue is how the system should behave when the authentication fails.

### INTEROPERABILITY

As a big picture of in-vehicle emergency systems, Fig. 6 organizes the systems in terms of standards and call types (i.e., if the IVS calls the PSAP or if the call is mediated by a service center). Note that eCall is the only in-vehicle emergency call system that can send data and voice directly to a PSAP, but it is based on circuit-switched technology, using an in-band modem to send data. As it evolves to NG-eCall, it will adopt the building blocks of the NG-911/112 architecture.

In this section, we discuss the interoperability of eCall and NG-eCall and the role of the NG-911/112 architecture. We also discuss details of the interoperability of eCall and ERA-GLONASS, which operate in the same continent.

#### INTEROPERABILITY OF eCALL AND NG-eCALL

In Europe, eCall standards are stable. NG-eCall may use IMS as defined by ETSI in 2014. When NG-eCall standards are finalized, vehicles will likely support both systems. One recommendation by the European Emergency Number Association (EENA) is that the cellular network has to advertise if it supports NG-eCall (i.e., IMS eCall indication). Also, in-band modems should not send MSD over VoIP because the receiving modem is sensitive to timing.

The NG-911/112 architecture has the necessary elements to support NG-eCall. It has introduced a new service URN to identify NG-eCall, and IETF has been working to define data structures to carry additional data from IVS to PSAP and vice versa [12,14]. Sending additional data by reference in SIP signaling enables the use of web services so that PSAPs and first responders can retrieve the vehicle incident information using

HTTP or HTTPS. HELPNET in Japan uses HTTP to communicate data from vehicle to service center. This was also the approach used in the NG-911 proof of concept with OnStar back in 2008. When OnStar initiated the SIP call, crash data URI was carried in the SIP header, which also contained a passcode to allow the PSAP to access that web link using HTTPS. This URI was part of the call and could be retrieved later by first responders. OnStar also sent the location of the vehicle in the emergency call, acting as a LIS. One of the SIP proxies at the entrance of the ESInet converted it to PIDF-LO format to query the LoST server. Then the call was routed to an appropriate PSAP.

In Europe, a specification based on a web service interface (EN16102) has already been published. We believe the NG-911/112 framework provides a good framework for the standardization of the interface between private service centers and PSAPs. As for the future landscape in the United States and Europe, it will likely be a combination of systems operating in the circuit- and packet-switched domains.

#### INTEROPERABILITY OF eCALL AND ERA-GLONASS

The interoperability of pan-European eCall and the Russian ERA-GLONASS system has been analyzed in [6], which has interworking use cases. The use cases involved an eCall IVS interacting with the ERA-GLONASS back office system, and an ERA-GLONASS IVS contacting a PSAP supporting eCall. The results in [6] indicated that the core functions of both systems — MSD transmission and voice connection — are available in the interworking use cases. Callback from the PSAP to the IVS will probably not be possible.

Both systems support MSD retransmission and retransmission request using eCall in-band modem. While an MSD retransmission mechanism based on SMS is included in the specifications of ERA-GLONASS, it is not supported by eCall.

Additionally, the specifications of eCall allow the PSAP to send a *clear-down* instruction with the in-band modem, but call *clear-down* is not included in the specifications of ERA-GLONASS.

A new version of the eCall MSD was published as a CEN standard in 2015. The current version (EN15722) states that MSD v. 1 is not supported. The latest ERA-GLONASS standards describe MSD v. 1 and 2, and use of MSD v. 1 is allowed until January 2018. Thus, ERA-GLONASS IVSs using the older version of MSD will not be interoperable with eCall PSAPs based on existing eCall specifications. In conclusion, the specifications of eCall and ERA-GLONASS allow interoperable solutions, but there is no interoperability between all versions of each system. The analysis in [6] needs to be verified in future interoperability tests.

### CONCLUSION

This article has reviewed in-vehicle emergency call systems being implemented and developed. Two systems (eCall and ERA-GLONASS) have been standardized and deployed with a regulatory approach. They have both been developed for circuit-switched networks, while ERA-GLONASS supports SMS as a backup for an in-band modem.

Systems that use a third-party service center usually combine both circuit and packet switching. The packet-switched NG-911/112 architec-

ture has introduced eCall service identification, location, and vehicle data transmission, and control data from PSAP to vehicle, using SIP signaling. It provides a foundation for future interfaces between proprietary systems and PSAPs, and for the evolution from eCall to NG-eCall.

#### ACKNOWLEDGMENTS

The authors wish to express their gratitude to Dr. Evgeni Meilikhov, Dr. Raimo Kantola, and Dr. Walt Magnussen. This work was supported by the Digile IoT Programme, Finnish Agency for Technology and Innovation (Tekes), VTT Technical Research Centre of Finland. Thanks also to Texas A&M's Association of Former Students, which funded Ana Goulart's faculty development leave in Finland.

#### REFERENCES

- [1] NHTSA, Press release NHTSA 47-15, Nov. 2015; <http://www.nhtsa.gov/About+NHTSA/Press+Releases/2015/2014-traffic-deaths-drop-but-2015-trending-higher>, accessed July 20, 2016.
- [2] EC, "Mobility and Transport, Road Safety, Statistics – Accidents Data," Mar. 2016; [http://ec.europa.eu/transport/road\\_safety/specialist/statistics/index\\_en.htm](http://ec.europa.eu/transport/road_safety/specialist/statistics/index_en.htm), accessed July 20th, 2016.
- [3] J. Y. Kim, W. Song, and H. Schulzrinne, An Enhanced VoIP Emergency Services Prototype, *Proc. 3rd Int'l. ISCRAM Conf.*, Newark, NJ, May 2006.
- [4] ETSI TR 102180, "Emergency Communications (EMTEL); Basis of Requirements for Communication of Individuals with Authorities/Organizations in Case of Distress (Emergency Call Handling)," July 2015; [https://www.etsi.org/deliver/etsi\\_tr/102100\\_102199/102180/01.05.01\\_60/tr\\_102180v010501p.pdf](https://www.etsi.org/deliver/etsi_tr/102100_102199/102180/01.05.01_60/tr_102180v010501p.pdf), accessed July 20, 2016.
- [5] ETSI TR 103140, "Mobile Standards Group (MSG): eCall for VoIP," Apr. 2014; [https://www.etsi.org/deliver/etsi\\_tr/103100\\_103199/103140/01.01.01\\_60/tr\\_103140v010101p.pdf](https://www.etsi.org/deliver/etsi_tr/103100_103199/103140/01.01.01_60/tr_103140v010101p.pdf), accessed July 20, 2016.
- [6] R. Öörni, E. Meilikhov, and T. O. Korhonen, "Interoperability of eCall and ERA-GLONASS In-Vehicle Emergency Call Systems," *IET Intelligent Transport Systems*, vol. 9, no. 6, 2015, pp. 582–90.
- [7] EC, Impact Assessment, accompanying the document, Commission Recommendation, on support for an EU-wide eCall service in electronic communication networks for the transmission of in-vehicle emergency calls based on 112 ("eCalls"), 2011; [http://ec.europa.eu/newsroom/dae/document.cfm?doc\\_id=2252](http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=2252), accessed July 20th, 2016.

- [8] N. Sihvola *et al.*, "In-Depth Evaluation of the Effects of an Automatic Emergency Call System on Road Fatalities," *European Transport Research Review*, vol. 1, no. 3, Oct. 2009, pp. 99-105; [http://rd.springer.com/content/pdf/10.1007\\_percent2Fs12544-009-0016-3.pdf](http://rd.springer.com/content/pdf/10.1007_percent2Fs12544-009-0016-3.pdf), accessed July 20, 2016.
- [9] GLONASS Union, "Development of the ERA-GLONASS Program," 2015; <http://en.glonassunion.ru/era-glonass>, accessed July 20, 2016.
- [10] HELPNET System Specifications, presentation at UNECE Informal Group on Accident Emergency Call Systems, Session 3, Moscow, Feb. 2014; <https://www2.unece.org/wiki/download/attachments/16449877/AECS-03-05%20%28J%29%20HELPNET.pdf?api=v2>, accessed July 20, 2016.
- [11] NENA, Automatic Collision Notification and Vehicle Telematics Technical Information Document (NENA 07-504), June 2007.
- [12] R. Gellens and H. Tschofenig, "Next-Generation Pan-European eCall," July 2016, work in progress; <https://tools.ietf.org/html/draft-ietf-ecrit-ecall-08>, Accessed on July 26th, 2016.
- [13] R. Gellens *et al.*, "Additional Data Related to an Emergency Call," Apr. 2016, work in progress; <https://tools.ietf.org/html/draft-ietf-ecrit-additional-data-38>, accessed July 26, 2016.
- [14] G. Camarillo, *The 3G IP Multimedia Subsystem (IMS) : Merging the Internet and the Cellular Worlds*, 3rd ed., Wiley, 2008.
- [15] NENA, NENA/APCO Emergency Incident Data Document (EIDD) Information Document (NENA/APCO-INF-005), Feb. 2014; [https://c.ymcdn.com/sites/www.nena.org/resource/resmgr/Standards/EIDD\\_INF-005\\_FINAL\\_20140221.pdf](https://c.ymcdn.com/sites/www.nena.org/resource/resmgr/Standards/EIDD_INF-005_FINAL_20140221.pdf), accessed July 30, 2016.

#### BIOGRAPHIES

RISTO ÖÖRNI (risto.oorni@vtt.fi) received his M.Sc. in electrical and communications engineering from Helsinki University of Technology in 2004. Since 2004, he has been working as a research scientist and later as a senior scientist at VTT Technical Research Centre of Finland Ltd. His research interests include in-vehicle emergency call systems, data and service quality in ITS applications, and analysis of impacts of ITS.

ANA GOULART (goulart@tamu.edu) received her M.Sc. from North Carolina State University and her Ph.D. in electrical and computer engineering from Georgia Tech in 2005. She is an associate professor in the Electronics Systems Engineering Technology program at Texas A&M University. In 2015–2016 she was a visiting professor at Aalto University. Her research interests are IP-based emergency communications, IoT, smart grid, and cyber-security.

The packet-switched NG-911/112 architecture has introduced eCall service identification, location and vehicle data transmission, and control data from PSAP to vehicle, using SIP signaling. It provides a foundation for future interfaces between proprietary systems and PSAPs, and for the evolution from eCall to NG-eCall.

## AD HOC AND SENSOR NETWORKS



Edoardo Biagioni



Silvia Giordano



Ciprian Dobre

**W**e live in exciting times! As we develop into a society increasingly becoming more technology-dependent, much closer to Mark Weiser's vision of the computer disappearing into the fabric of everyday life [1], we develop technologies to support novel ways to communicate. With the miniaturization of computer technology, we expect today technology-driven ordinary objects to behave "intelligently," to communicate between themselves and with large data centers privately holding our "digital life," and running smart algorithms against our data, all designed to improve our life. This large-scale ubiquitous computing vision embraces a model in which users, services, and resources discover other users, services, and resources, and integrate them into a useful experience [2]. This ubiquitous society is similar to what Manuel Castells defines as the "network society" [3], where, similar to how the Internet has become a pervasive utility, we reach a phase when networking logic becomes applicable in every realm of daily activity, in every location and every context. In such a ubiquitous society, billions of miniature, ubiquitous inter-communication devices will be spread worldwide, "like pigment in the wall paint" [3]. This is the time when computing is expected to be available for us anywhere anytime, at the reach of a button. And many times that button is made available by a smartphone, a portable or wearable device, or a miniaturized wireless sensor. And, of course, communication has to cope with the high demand of the mobile world, to exchange more data much more reliably and faster than before!

This issue explores recent advances in ad hoc communication to support this vision. It includes technology efforts toward scaling communication by leveraging everyday mobile devices, using social information to make smart relaying decisions. It includes advances in the field of wireless sensor networks. And it includes a study of the realization of a concrete video-related application. The issue presents communication paradigms that contribute to the emergence of the pervasive and ubiquitous computing vision, based on the proliferation of sensor-rich portable devices. Until now, such sensor-rich devices were used mostly standalone, but when combined or as a complement to an infrastructure-based computation substrate, such as the cloud, they leverage the mobility of end users, and

the processing power of these end devices, to enhance users' ability to communicate, sense, and compute in the absence of reliable end-to-end connectivity.

Toward this vision, one development in the field of ad hoc networks is that of opportunistic networks (OppNets). OppNets are related to delay-tolerant networks (DTNs) in that connectivity is expected to be intermittent. One main feature of OppNets is that they are networks among devices carried by individuals for a variety of purposes, but mostly to support mobile communication for the individual. OppNets use short-range communications such as WiFi or Bluetooth, and are infrastructure-free by design. The purpose of OppNets is to disseminate content rather than (as in DTNs) to connect devices. As in DTNs, communication may be enabled by the physical motion of the devices.

In the first article, "A Decade of Research in Opportunistic Networks: Challenges, Relevance, and Future Directions," Trifunovic *et al.* give a good overview of this field before analyzing the challenges and future research directions possible with OppNets. Unlike a typical survey of the up-to-date results obtained by authors working in this field, the authors make an analysis of today's challenges that still forbid us from seeing real-world implementations of such technology, from the lack of support at the mobile operating system level to the existence of alternative technologies (Google's Project Loon, Internet.org by Facebook) that can lead to the realization of the typical use cases advertised as the killer applications for OppNets. Finally, the authors review the oft-stated motivations for having OppNets become reality, identifying the technical areas in which they fit applications and examining the economic drivers for their development.

In the second article, Zhang *et al.* discuss similar scenarios, but refer to the network as smartphone ad hoc networks, or SPANs. Rather than providing an overview, the focus of this article is to leverage smartphone users' social connections to improve routing in the lower layers. As in the previous article, the focus is on content dissemination rather than node connectivity. As an example, content sharing is more likely between friends than between people who are not socially connected. Protocols that take advantage of this are found to provide higher delivery ratios under otherwise comparable circumstances.



The third article, by Mate and Curcio, also focuses on content, but this time on automatic remixing of video content from several sources. The crucial insight of this article is that information from sensors available in many modern handheld devices, particularly accelerometers, GPS, compass, and gyroscope, can be used to automatically choose among several available video streams of the same event uploaded over high-speed wireless networks by different users using different devices.

The final article, by Iova, Theoleyre, and Noel, returns to the topic of ad hoc networks, in particular networks serving the Internet of Things (IoT). Many devices in the IoT are power-constrained and use potentially lossy network links, placing them in the world of low-power and lossy networks (LLNs). IEEE 802.15.4-2006 has defined a protocol, IPv6 Routing Protocol for LLNs (RPL), which effectively ignores information from the medium access control (MAC) layer such as the number of MAC-layer packet retransmissions. Ignoring MAC-layer information also leads to redundant routes that may have undesirable sharing of wireless link characteristics. This article provides a survey of interesting questions and conclusions on this general topic.

In short, these articles provide a range of answers to questions about the continuing evolution of the field of wireless ad hoc networks, in supporting both different applications and different technologies used to solve specific issues.

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

## REFERENCES

- [1] M. Weisser, "The Computer for the Twenty-First Century." *Scientific American*, vol. 265, no. 3, 1991, pp. 94–104.
- [2] J. Coutaz et al., "Context Is Key," *Commun. ACM*, 48.3 (2005): 49-53.
- [3] M. Castells, *The Rise of the Network Society*, Vol. 1 of *The Information Age: Economy, Society and Culture*, Blackwell, 1996.

## BIOGRAPHIES

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

SILVIA GIORDANO [M] (silvia.giordano@supsi.ch) received her Ph.D. from EPFL, and she is a full professor at SUPSI, Switzerland, and an associate researcher at CNR. She directs the NetworkingLab. She has published extensively in the areas of QoS, traffic control, and wireless and mobile networking. She is a co-editor of the books *Mobile Ad Hoc Networking* (IEEE-Wiley, 2004) and *Mobile Ad Hoc Networking: The Cutting Edge Directions* (Wiley, 2013). She is an ACM Distinguished Scientist, ACM Distinguished Speaker, and on the Board of ACM N2Women. She has been a Series Co-Editor for *IEEE Communications Magazine* since 2004.

CIPRIAN DOBRE (ciprian.dobre@cs.pub.ro) is a professor at the Polytechnic University of Bucharest (Habil. in 2014, Dr. in 2008 with cum laude). He leads the activities within the Laboratory on Pervasive Products and Services, and MobyLab. His current research interests involve topics related to mobile wireless networks and computing applications, pervasive services, context awareness, and people-centric sensing. He has contributions in mobile applications and smart technologies to reduce urban congestion and air pollution, context-aware applications, opportunistic networks, and mobile data offloading, monitoring, and high-speed networking. He is Director or Principal Investigator of 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.

# A Decade of Research in Opportunistic Networks: Challenges, Relevance, and Future Directions

Sacha Trifunovic, Sylvia T. Kouyoumdjieva, Bernhard Distl, Ljubica Pajevic, Gunnar Karlsson, and Bernhard Plattner

After a decade of research, opportunistic networks have not yet been ubiquitously deployed. The authors explore the reasons for their absence. They take a step back, and first question whether the use-cases that are traditionally conjured to motivate opportunistic networking research are still relevant. They also discuss emerging applications that leverage the presence of opportunistic connectivity.

## ABSTRACT

Opportunistic networks are envisioned to complement traditional infrastructure-based communication by allowing mobile devices to communicate directly with each other when in communication range instead of via the cellular network. Due to their design, opportunistic networks are considered to be an appropriate communication means in both urban scenarios where the cellular network is overloaded, as well as in scenarios where infrastructure is not available, such as in sparsely populated areas and during disasters. However, after a decade of research, opportunistic networks have not yet been ubiquitously deployed. In this article we explore the reasons for their absence. We take a step back, and first question whether the use cases that are traditionally conjured to motivate opportunistic networking research are still relevant. We also discuss emerging applications that leverage the presence of opportunistic connectivity. Further, we look at past and current technical issues, and we investigate how upcoming technologies would influence the opportunistic networking paradigm. Finally, we outline some future directions for researchers in the field of opportunistic networking.

## INTRODUCTION

In recent years we have witnessed the spectacular success of the mobile Internet, driven by the rise of smart mobile devices. The demand for data is exponentially increasing as more and more services are based on a cloud infrastructure with a prediction of 24 EB of monthly mobile data traffic by 2019 according to Cisco's Visual Networking Index. At this pace, the mobile Internet is about to become a victim of its own success. On one hand, improving the infrastructure is becoming increasingly costly, and coping with the demand during large gatherings such as sports events is already hardly feasible. Furthermore, in some places, even when communication is technically possible, it might be restricted by censorship, thus blocking information dissemination. On the other hand, in sparsely populated areas the deployment of communication infrastructure might not be economically beneficial for operators. Finally, infrastructure may break during natural or man-made disasters, leaving rescue services and people in need unable to communicate.

Opportunistic networks, or *OppNets* (sometimes referred to as pocket-switched networks [1] and people-centric networks [2]), are a special type of mobile ad hoc networks (MANETs) in which human-carried mobile devices (often referred to as nodes) communicate directly via some short-range wireless technology such as Wi-Fi or Bluetooth whenever they are in transmission range. By design, *OppNets* are infrastructure-free: nodes *store* data, and *carry* it according to the underlying user mobility until a new communication opportunity arises to *forward* the data. This store-carry-forward paradigm was first introduced in the general field of delay-tolerant networking (DTN) [3]. While DTN embraced the idea of leveraging mobility as a means of transporting information, it still kept the traditional Internet-inspired user-centric approach for delivering data between particular source-destination pairs. To facilitate data dissemination, various routing algorithms were introduced [4]. Contrary to DTNs, in *OppNets* the focus shifts from user-centric to *content-centric* data dissemination. This reduces network complexity, as choosing appropriate intermediate nodes for forwarding information is no longer a priority. Instead, data dissemination depends on the mobility patterns of humans as well as some shared content interests. Due to these characteristics, *OppNets* have been considered as a potential solution to complement the infrastructure and mitigate the aforementioned shortcomings that network operators are experiencing. However, after a decade of research efforts, *OppNets* have not yet been widely deployed. It may thus be time to take a step back and pose the question: *Why are OppNets not used to solve these problems?*

There are two main reasons *OppNets* have never been deployed beyond small-scale testbeds. First, *OppNets* did not present companies with a clear business case. Instead, the infrastructure-free design has been perceived as a threat by mobile operators. Second, even if a particular business case were available, the prohibitive battery consumption of the mobile devices to maintain the network would still prevent the deployment of *OppNets*. To discover communication opportunities without the aid of an infrastructure, the mobile devices need to continuously advertise their presence in the network. With the available technologies, this operation is too power-hungry

for the limited battery capacity of modern smartphones [5].

In this article we look beyond the current showstoppers and first ask ourselves the question: *Is opportunistic communication still a relevant concept in today's highly connected world?* We revisit well established use cases, and discuss their applicability and positioning with respect to other upcoming technologies. Then we take a look into the future, examining what emerging applications and technologies are on the horizon and how they might impact the paradigm of opportunistic communication. Finally, we outline the next steps that could lead to eventual deployment of OppNets.

## ARE OPPNETS STILL RELEVANT?

In this section we evaluate the relevance of the motivational scenarios used to justify research in the field of opportunistic networks during the past decade, and examine how well suited OppNets currently are for these scenarios in comparison to newly emerging technologies.

### CLASSICAL USE CASES

For a decade researchers have been searching for the “killer application” that will boost the global deployment of OppNets. Below are four distinct application areas that have been promoted in the community.

**Cellular Network Offloading:** Operators struggle to cope with the traffic demands of large crowds, especially if they are sporadic in nature, such as festivals and street fairs. They deploy ever smaller cells and greatly overprovision the supply, but this is costly and is still unable to deal with unforeseen traffic peaks. Mobile operators could utilize OppNets to offload their infrastructure by seeding popular content to a few devices in a crowded space which then opportunistically disseminate it to others in their vicinity.

However, as operators do not like to relinquish control and as users still expect to have their requests for data answered with minimal delay, the type of OppNets that could succeed in this scenario might be operator controlled.

**Communication in Challenged Areas:** A challenged area is often defined as an area in which infrastructure is partially or fully unavailable. Reasons for such unavailability may be due to:

1. A natural or man-made disaster that has destroyed available infrastructure
2. A lack of economic motivation for deploying infrastructure, for instance, in sparsely inhabited regions
3. The inaccessibility of certain areas, for instance, in mines

Due to their infrastructure-free design, OppNets enable local communication, and could even serve as a bridge between the challenged areas and the infrastructure (wherever available). During disasters, this could be of great importance for supporting the operation of rescue teams. In sparsely populated areas, both above or underground (e.g., in mines), OppNets could provide an alternative means of communication.

**Censorship Circumvention:** OppNets may become an appropriate tool for enabling freedom of speech in regions governed by oppressive institutions that are inclined to censor traditional

communication via the Internet. Participants in opportunistic communication benefit from the fact that links established in an opportunistic manner are hard to intercept or jam, and individual users are not easy to track down, especially in crowded scenarios. However, simply promoting OppNets as a censorship circumvention technology may not appeal to governmental bodies. Therefore, this application might only be seen as an added value instead of a primary solution.

**Proximity-Based Applications:** A promising use case for OppNets are proximity-based applications. Proximity-based applications take advantage of the co-location of nodes to provide add-on services on top of available infrastructure.

However, employing OppNets in the proximity-based applications domain for providing services such as proximal social networking has failed due to the following two limitations:

1. The lack of an explicit business model
2. The possibility to provide similar functionality via traditional centralized communication

A special use case of proximity-based applications for OppNets might be seen in applications that target people in the creative sector (e.g., artists or musicians); applications have been tailor-made for these industries and have been met with interest.

### NEW RESEARCH DIRECTIONS

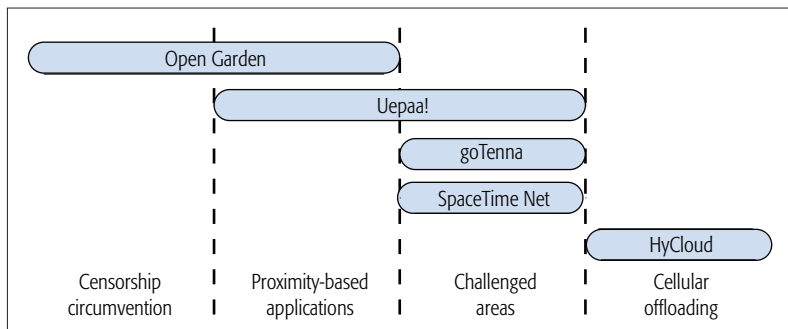
In addition to the classical use cases, in recent years the research community has been investigating other promising application areas that exploit the characteristics of opportunistic communication.

**Opportunistic Mobile Sensing:** Today's mobile devices (both smartphones and wearables) are equipped with a rich set of embedded sensors including accelerometers, cameras, microphones, GPS, and more. Opportunistic mobile sensing exploits all of these sensing devices available in an environment to collect data in a fully automated way [6]. It is expected that larger populations may engage in the data collection process. The objective of opportunistic mobile sensing is to investigate human behaviors and socio-economic relationships by analyzing the digital footprint of people in the surrounding physical world.

**Opportunistic Mobile Computing:** OppNets make use of contact opportunities among mobile devices purely in the context of data dissemination. However, when two (or more) devices are in direct communication range, they could potentially share more than just data; for example, they could exploit each other's software and hardware resources, and even execute tasks remotely. This lays the foundation for the newly emerging concept of opportunistic computing [7]. The objective of opportunistic computing is to enrich the functionality of a single device by allowing nodes to utilize resources on other devices in proximity in a trustable and secure way. Opportunistic computing is expected to find application in pervasive healthcare, intelligent transportation systems, and crisis management, among other fields.

However, as these use cases are not directly targeted at providing connectivity, their direct competition is the same or a similar service provided over a centralized communication infrastructure. If these use cases perform better when using opportunistic communication, most proba-

As a special use-case of proximity-based applications for OppNets might be seen applications that target people in the creative sector, such as artists or musicians; applications have been tailor-made for these industries and have been met with interest.



**Figure 1.** Distribution of companies utilizing the opportunistic networking paradigm across potential use case scenarios. It is interesting to notice that most solutions cater to providing connectivity in challenged areas. Data offloading, which may be the most economically beneficial application, has not yet seen actual industrial deployments, only early-stage prototyping.

bly stemming from reasons linked to the classical use case of data offloading, such novel services can indeed be seen as an additional motivator.

#### ALTERNATIVE SOLUTIONS FOR FUTURE CONNECTIVITY

OppNets are not the only suggested paradigm for overcoming the limitations in connectivity listed above. In 2014, four companies publicly announced their goal of providing or facilitating mobile connectivity and Internet access on a global scale. We can divide them into two broad categories based on the way they mitigate infrastructure: floating and orbital.

The two main representatives of floating infrastructure are Google's Project Loon and Internet.org by Facebook. Project Loon aims to provide air-floating cellular infrastructure in the form of LTE-equipped balloons. Initial trials show that the balloons can be kept in the air for months, and public trials began in 2016. In contrast, the Internet.org initiative intends to use solar-powered drones to provide a backbone to scattered cellular base stations that provide connectivity in remote areas.

Orbital infrastructure is suggested by SpaceX and OneWeb, which intend to provide connectivity through a swarm of satellites in low Earth orbit. In contrast to current geostationary satellite technologies, supporting an infrastructure of satellites at lower orbits would offer users shorter end-to-end communication delays, but at a cost of a larger amount of equipment. Expected initial trials are scheduled for 2020.

#### OPPNETS VS. FUTURE CONNECTIVITY SOLUTIONS

In 2009 researchers in the field of OppNets and DTN stated that "delay-tolerant systems will progress to become the mainstream default networking paradigm" [8]. Nowadays however, with emerging paradigms for providing global connectivity such as floating and orbital infrastructures, a valid question is whether some of the classical OppNet use cases could be better addressed by these infrastructures instead. Table 1 summarizes the applicability of different approaches to the scenarios introduced earlier.

As expected, the emerging paradigms are best suited for the scenarios for which they were initially designed, that is, providing connectivity in challenged areas, such as sparsely inhabited areas as well as during disasters. Floating infrastruc-

tures, especially Project Loon, are perfectly suited to provide Internet access to underdeveloped regions. Depending on the speed with which the network can be rearranged, communication might be provided during or after disasters. In contrast, orbital infrastructure is likely to be always present, and thus immediately available during disasters. Enabling communication via a satellite is also well suited for underdeveloped regions, but the cost of putting infrastructure in the orbit might make the solution expensive.

However, both floating and orbital infrastructures are not appropriate when targeting communication in inaccessible areas, as in the case of providing connectivity in mines. Furthermore, due to the larger cell sizes, they are ill suited for supporting proximity-based services. Offloading mobile data is also not a potential application since floating and orbital infrastructures are facing the same issues with traffic volumes as current terrestrial deployments of base stations. Finally, in the case of censorship circumvention, access to these emerging technologies may be blocked by oppressive governmental bodies as is done with current infrastructure.

We can thus conclude that the concept of OppNets is relevant to this day. While some of the classical use case scenarios, such as communication in sparsely populated areas, may be better served by emerging communication paradigms, there is still a strong case for the usage of opportunistic communication, most notably in the context of mobile data offloading and proximity-based applications. The latter is further strengthened by the increasing interest in the Internet of Things (IoT) domain where direct communication between devices is dominant. In fact, as of Release 12, the Third Generation Partnership Project (3GPP) is focusing on utilizing device-to-device communications for providing proximity-based services on top of current cellular infrastructure [9], which is an indication of the potential deployment of OppNets. Finally, the emerging application paradigms that make use of opportunistic communication, such as opportunistic mobile sensing and opportunistic mobile computing, can be construed as a positive sign for the future development of OppNets.

#### OPPNETS TODAY

*The Research View* — Most research on OppNets addresses issues in the area of content dissemination, with the focus being on routing and mobility modeling as enablers of data sharing. Due to the absence of centralized control, security and privacy have also been investigated. The high battery consumption of nodes in OppNets has led to designing energy-efficient discovery protocols. However, not all research topics are fully exhausted, as we show later.

*The Industry View* — Few industrial applications have been developed on top of the opportunistic networking paradigm, as shown in Fig. 1. *Space-time networks* base their business model on an opportunistic router developed in the SCAMPI [10] research project and aim to deploy OppNets as a communication tool in challenged environments such as mines and underground tunnels. *Uepaa!* has developed an alpine safety application to be used in areas with no cellular coverage.

*Open Garden* offered FireChat, an off-the-grid application that gained popularity during the Hong Kong protests. Both Uepaa! and Open Garden aim to release their platform with an open application programming interface (API) for the convenience of third-party developers. To circumvent the prohibitive energy costs of establishing OppNets, *goTenna* takes an entirely different approach to providing ad hoc communication capabilities. They provide an add-on device linked to the smartphone, with communication ranges of 500 m in urban areas for an operational duration of over 30 hours.

It is interesting that although mobile data offloading would be the most economically beneficial application, currently there are no real industrial applications developed. HyCloud [11] is the only academic project to prototype opportunistic networking for data offloading.

## EVOLUTION OF OPPNET TECHNOLOGY

While few companies attempt to create business models on top of the opportunistic networking concept, they all face similar technical limitations. The functional support for opportunistic communication provided by the mobile operating system is currently nonexistent. Furthermore, the lack of radio technology tailored to providing efficient device discovery at low energy cost still presents a challenge.

### BRIEF HISTORIC OVERVIEW

At the dawn of opportunistic networks, researchers only had access to two widely deployed technologies: Wi-Fi in ad hoc mode and Bluetooth. Wi-Fi in ad hoc mode was often the preferred radio technology for early-stage proof-of-concept implementations due to its higher data rates, longer communication ranges, and lack of manual pairing. However, researchers quickly encountered a number of limitations. First, Wi-Fi in ad hoc mode is extremely energy-hungry due to the fact that the energy spent in idle state (while trying to catch a signal) is on the same order of magnitude as that spent on actual transmission and reception of data. Due to the implicit requirement of continuous device discovery, a device can only operate for a few hours before it completely drains its battery [12]. Moreover, support for Wi-Fi in ad hoc mode is also restricted, requiring users to operate their devices in privileged mode if they are to participate in any opportunistic content sharing. This has naturally limited users' interest in OppNets.

To combat the aforementioned issues, the research community created WLAN-Opp [13], an 802.11-based technology that leverages the tethering mode of devices. However, due to the lack of standardization, WLAN-Opp has not been widely adopted in current devices, and its usage is limited solely to research activities.

Both Bluetooth and Wi-Fi have evolved since; however, neither of these technologies has become more suitable for opportunistic communication. Bluetooth Low Energy (BLE) was the first technology on the market to tackle the problem of energy-efficient device discovery. However, the required manual pairing makes it inappropriate for opportunistic networking. Furthermore, scanning intervals are on the order of minutes, which makes discovery slow, with a potential of

Use-case scenario	OppNets	Floating infrastructure	Orbital infrastructure
Network data offloading	★★★	★★	★
Proximity-based apps	★★★	—	—
Censorship circumvention	★★★	★	★★
Inaccessible areas	★★	★	★
Disaster scenarios	★★	★★	★★★
Sparsely populated areas	★	★★★	★★★

**Table 1.** Comparison: OppNets vs. future connectivity paradigms. One star denotes that a paradigm is ill suited; three stars denote a good fit.

skipping a lot of contact opportunities in dynamically changing environments such as urban areas. When Wi-Fi Direct gained momentum in 2012, it brought a new wave of excitement to the research community. However, Wi-Fi Direct was originally created as a competitor of BLE, and as such it is ill suited for performing opportunistic device discovery and communication: not only are its energy consumption profiles unbalanced, but discovery is time consuming and requires manual pairing.

### FUTURE TECHNOLOGIES

The 3GPP is currently discussing the introduction of device-to-device communication as a complement to traditional communication via the cellular infrastructure. As a result, two new technologies have been proposed to allow energy-efficient proximity-based service discovery and communication for users on the go, catering to the whole potential of OppNets: unlicensed spectrum Wi-Fi Aware and in-band LTE-Direct. While there are no products available yet using these new technologies, LTE-Direct has already been implemented and tested, making it currently the only radio technology designed *specifically* for opportunistic device discovery. Due to its synchronous duty-cycling scheme, it is expected to significantly reduce the energy consumption in devices.

The fact that technologies are developed entirely for the specifics of opportunistic device discovery is partially linked to the rise of the IoT, and can be seen as a strong indication of the uprise of OppNets. It is still unclear whether OppNets would operate in unlicensed spectrum as envisioned by researchers a decade ago, whether they would be entirely under the control of cellular network operators, or if a hybrid approach would prevail. However, once a stable technological foundation is built, one that decreases the energy consumption in the devices while simultaneously allowing them to discover nodes in a quick and efficient manner, it would be technically possible for OppNets to see mass deployment.

### FUTURE DIRECTIONS IN OPPNET RESEARCH

As the concept of OppNets remains relevant and more timely than ever, as the industry expresses interest in its potential, and as promising technological enablers are emerging on the horizon, the natural question for researchers to ask is: *What is to be done next?* In this section we first outline a three-step action plan for future research toward

Current research efforts have only evaluated the performance under the assumption of a single available service in the opportunistic domain. Thus, it is unclear how many services would constitute a bottleneck, as well as in which scenarios this may be an actual performance issue.

ubiquitous deployment of OppNets, and then discuss open research questions.

#### ACTION PLAN TOWARD DEPLOYMENT OF OPPNETS

**First Large-Scale Experiments:** While waiting for the technological progress to happen, researchers should take an active role in setting up large-scale experiments. This can be done in three possible ways:

1. By using the most energy-efficient method to establish OppNets, and recruiting people to participate in support of research with the explicit warning that energy consumption may be higher
2. By using a controlled testbed of mobile devices with a rooted or modified OS that integrates OppNet functionality in an energy-efficient way, maybe even using prototypes of future protocols such as Wi-Fi Aware
3. By using external devices such as goTenna<sup>1</sup> for performing long-distance experiments

Each approach has its own advantages and limitations. Implementation on smartphones provides a few options in terms of radio technology used (either Bluetooth or WLAN-Opp; using a pure ad hoc mode may also be possible if paired with additional energy saving schemes [14]). The benefit of integration in the OS is better control of duty cycling and background operation without interfering with the user. If researchers decide to use external devices such as goTenna, energy consumption on the mobile device during neighbor discovery would only depend on the energy spent on communicating with the goTenna. However, it is unclear how traffic will impact battery consumption, especially if data is also relayed for other devices.

**Exploring Scalability:** Large-scale deployments will result in exploring a feature of OppNets that has not previously been addressed, that is, their scalability. It is thus important to perform extensive scalability tests and determine the bounds, in terms of density of participants, below which the performance of OppNets is acceptable, also taking into account application requirements. A good way to measure scalability would be to provide OppNets as an alternative communication means during large gatherings such as outdoor festivals.

Another aspect of scalability researchers should consider is related to the abundance of services competing to use the communication opportunities. Current research efforts have only evaluated the performance under the assumption of a single available service in the opportunistic domain. Thus, it is unclear how many services would comprise a bottleneck, as well as in which scenarios this may be an actual performance issue.

**Economical Validation:** Finally, researchers should address the economic benefits of ubiquitous deployment of OppNets. In this context, economic validation should be understood in a broader sense than simply monetizing the OppNet concept. Instead, it should evaluate the potential benefits of OppNet deployment for all involved market players. Emerging use cases should also be considered: offloading network traffic and providing proximity-based services

are a good starting point, but as deployments advance, other use cases, especially in the IoT domain, are worth investigating.

#### OPEN RESEARCH QUESTIONS

Not all research questions have been fully addressed in the domain of opportunistic communication. A number of issues still need the attention of the research community to make OppNets a reliable and trustworthy communication paradigm.

**Privacy vs. Security:** It is crucial to provide good privacy and security in OppNets, not only for the classical use cases but also when considering emerging application paradigms such as mobile sensing and opportunistic computing.

Currently, all proposed privacy-enabling schemes are based on changeable identifiers, to change the medium access control (MAC) address to be changed, which limits the applicability of these approaches. Furthermore, current privacy schemes are difficult to implement alongside certain security and routing schemes that make use of social information [15]. It is still not clear whether it is possible to combine privacy and security in a single scheme that satisfies all requirements. If not, the trade-off between these aspects should be thoroughly evaluated, possibly with respect to the application at hand. For example, privacy and security may be handled by the cellular infrastructure in the case of network offloading, while ensuring privacy should be a priority of the OppNet itself in the case of providing freedom of speech.

**Short-Range vs. Long-Range Communication:** Until now, researchers have always assumed that opportunistic communication would occur over short-range radios and be characterized by short contact durations. Thus, a general goal when designing protocols for neighbor discovery has been to provide quick and efficient discovery mechanisms. However, with the advances of 3GPP's LTE-Direct as well as emerging products like goTenna, which promise to operate at ranges of up to 500 m in urban environments, it may be necessary to re-evaluate the assumptions for opportunistic communication, and investigate the implication of long-range communication links on both protocol design and performance. It is possible that long-range communication links are better suited for implementing the well studied MANET paradigm where mobility of nodes is obscured instead of explicitly utilized. A longer range might, however, increase interference and thus result in lower capacity for a covered area (less spectral reuse).

#### CONCLUDING REMARKS

After a decade of research in the field of opportunistic networking, are we about to witness the age of OppNets? The research area is mature, as most research questions have been addressed. However, implementations have been scarce, thus making large-scale evaluations impossible. As of now, only a few start-up companies have ventured into creating products based on the opportunistic networking paradigm.

Meanwhile, 3GPP coined the term *device-to-device* (D2D) communication to define a concept similar to opportunistic networking. In D2D,

<sup>1</sup> www.gotenna.com

devices are allowed to establish a direct communication link and exchange information when in range, but under the supervision of the network operator. In other words, the cellular network partially or fully assists with one or more procedures during the connection establishment phase, such as authentication and radio resource allocation. Although OppNets are designed to be entirely infrastructure-free, the fundamental principles of opportunistic networking are really not dependent on the involvement of the cellular network in the connection establishment process. Thus, it may be valuable for researchers in the OppNet community to transfer the knowledge they have cultivated over the past decade toward the D2D domain.

Although employing OppNets is advantageous in scenarios where the network is unavailable or inaccessible (Table 1), opportunistic communication is best suited for providing proximal services such as data offloading, proximal social networking and proximal entertainment. However, such applications would require cellular operators to relinquish some of the network control. On the contrary, network-assisted D2D as defined by 3GPP allows operators to preserve their control over the network. However, it raises privacy concerns as communicating devices are expected to reveal their identity as well as periodically report their location. Thus, it is still an open question how D2D and OppNets will coexist as proximity-based networks of the future.

## REFERENCES

- [1] P. Hui *et al.*, "Pocket Switched Networks and Human Mobility in Conference Environments," *Proc. ACM WDTN*, 2005, pp. 244–51.
- [2] M. Conti *et al.*, "From MANET to People-Centric Networking: Milestones and Open Research Challenges," *Computer Commun.*, vol. 71, 2015, pp. 1–21.
- [3] K. Fall, "A Delay-Tolerant Network Architecture for Challenged Internets," *Proc. ACM SIGCOMM*, 2003, pp. 27–34.
- [4] S. Basagni *et al.*, Eds., *Mobile Ad Hoc Networking: The Cutting Edge Directions*, 2nd ed., Wiley-IEEE Press, 2013.
- [5] S. Trifunovic *et al.*, "Adaptive Role Switching for Fair and Efficient Battery Usage in Device-to-Device Communication," *ACM SIGMOBILE Mobile Computing and Commun. Rev.*, vol. 18, no. 1, 2014, pp. 25–36.
- [6] N. Lane *et al.*, "A Survey of Mobile Phone Sensing," *IEEE Commun. Mag.*, vol. 48, no. 9, Sept. 2010, pp. 140–50.
- [7] M. Conti *et al.*, "From Opportunistic Networks to Opportunistic Computing," *IEEE Commun. Mag.*, vol. 48, no. 9, Sept. 2010, pp. 126–39.
- [8] A. Lindgren and P. Hui, "The Quest for a Killer App for Opportunistic and Delay Tolerant Networks," invited paper, *Proc. ACM CHANTS*, 2009, pp. 59–66.
- [9] 3GPP Tech. Spec. Group Services and System Aspects, "Feasibility study for Proximity Services (ProSe)," Release 12, TR 22.803 v. 12.2.0, June 2013.
- [10] M. Pitkänen *et al.*, "Scampi: Service Platform for Social Aware Mobile and Pervasive Computing," *SIGCOMM Comp. Commun. Rev.*, vol. 42, Sept. 2012, pp. 503–08.
- [11] J. Danhelka, D. Giustiniano, and T. Hossmann, "Hycloud: A System for Device-to-Device Content Distribution Controlled by the Cloud," *Proc. IEEE WoWMoM*, June 2014, pp. 1–5.
- [12] O. Helgason *et al.*, "A Mobile Peer-to-Peer System for Opportunistic Content-Centric Networking," *Proc. ACM SIGCOMM MobiHeld Wksp.*, 2010.
- [13] S. Trifunovic *et al.*, "Wlan-Opp: Ad-Hoc-Less Opportunistic Networking on Smartphones," *Ad Hoc Networks*, vol. 25, Part B, no. 0, 2015, pp. 346–58.
- [14] S. T. Kouyoumdjieva and G. Karlsson, "Energy-Aware Opportunistic Mobile Data Offloading for Users in Urban Environments," *Proc. IFIP/TC6 Networking*, 2015.
- [15] M. Li *et al.*, "Swing & Swap: User-Centric Approaches towards Maximizing Location Privacy," *Proc. WPES, ACM*, 2006, pp. 19–28.

## BIOGRAPHIES

SACHA TRIFUNOVIC (sascha.trifunovic@gmail.com) received his Ph.D. degree in electrical engineering from ETH Zurich in 2015. He holds an M.Sc. degree in electrical engineering received from ETH Zurich in 2010. His research interests lie in opportunistic networking, mainly in efficient resource consumption, fairness, and security.

SYLVIA T. KOUYOUMDJIEVA is a post-doctoral researcher at the Laboratory for Communication Networks at KTH Royal Institute of Technology, Stockholm, Sweden. She received her Ph.D. degree in electrical engineering from KTH in 2015. She holds an M.Sc. degree in telecommunications from KTH (2009) and a B.Sc. degree in telecommunications from the Technical University of Sofia, Bulgaria (2006). Her research focuses on challenges in system design for opportunistic networks.

BERNHARD DISTL received his Ph.D. degree in electrical engineering from ETH Zurich in 2015. He previously received his M.Sc. degree in electrical engineering in 2006, also from ETH Zurich. His research interests are in the area of security and privacy of opportunistic networks.

LJUBICA PAJEVIC is a doctoral student at the Laboratory for Communication Networks, KTH. She received her M.Sc. degree in system engineering and radio communications in 2009 and her B.Sc. degree in telecommunications in 2007, both from the School of Electrical Engineering, University of Belgrade, Serbia. Her research interests are in the area of opportunistic networking and mobility modeling.

GUNNAR KARLSSON [SM] is a professor at KTH and director of the Laboratory for Communication Networks. His current research relates to opportunistic networking, mobility modeling, and online education. He is a member of ACM and a Senior Editor of the *IEEE Journal on Selected Areas in Communications*. He is the 2015 recipient of the KTH Pedagogic Prize.

BERNHARD PLATTNER is an Emeritus Professor of Computer Engineering at ETH Zurich, where he led the Communication Systems Research Group for 30 years. His research focuses on self-organizing and opportunistic networks, systems-oriented aspects of information security, and future Internet research. He holds the position of an adjunct professor with the Institut Eurecom at Sophia Antipolis, France, and served as the head of the Department of Information Technology and Electrical Engineering at ETH Zurich.

Network-assisted D2D as defined by 3GPP allows operators to preserve their control over the network. However, it raises privacy concerns as communicating devices are expected to reveal their identity as well as periodically report their location. Thus, it is still an open question how D2D and OppNets will co-exist as proximity-based networks of the future.

# A Social-Aware Framework for Efficient Information Dissemination in Wireless Ad Hoc Networks

Yanru Zhang, Lingyang Song, Chunxiao Jiang, Nguyen H. Tran, Zaher Dawy, and Zhu Han

The authors present a social-aware framework for optimizing SPANs by exploiting two layers: users' relationships in the online social network layer and users' offline connections and interactions in the physical wireless network layer. The online content popularity distribution is also studied as a result of the users' online interaction profiles.

## ABSTRACT

In wireless ad hoc networks, each node participates in routing by forwarding data to other nodes without a pre-existing infrastructure. Particularly, with the wide adoption of smart devices, the concept of smartphone ad hoc networks (SPANs) has evolved to enable alternate means for information sharing. Using unlicensed frequency spectrum and short-range wireless technologies, a SPAN enables a new paradigm of applications and thus is seen as an attractive component in future wireless networks. In a SPAN, smartphones form local peer-to-peer networks to cooperate and share information efficiently. Recent studies have shown that if the users' social relations are considered while designing cooperation schemes and protocols in SPANs, the cooperation initialization and content dissemination can be notably improved to increase the overall network efficiency and communications reliability. In this article, we present a social-aware framework for optimizing SPANs by exploiting two layers: users' relationships in the online social network layer and users' offline connections and interactions in the physical wireless network layer. The online content popularity distribution is also studied as a result of the users' online interaction profiles. In the end, we integrate both online and offline layers, and discuss possible applications to further enhance the network performance.

## INTRODUCTION

The concept of wireless ad hoc networks (WANETs) has been introduced as devices directly transmit data signals to each other while bypassing the wireless infrastructure, that is, the cellular network's base stations (BSs). In WANETs, communications usually happen over the unlicensed spectrum by using existing short-range wireless technologies such as Bluetooth and Wi-Fi-Direct, and thus the interference between cellular and WANETs can be avoided. This can significantly improve the performance of cellular networks by offloading traffic, extending coverage, increasing throughput, and enhancing reliability. Given these nice properties, WANETs are promising to serve as a good backup for the cellular network with a variety of applications, which include public and

military areas. Furthermore, smartphone ad hoc networks (SPANs) particularly focus on smartphone users, which is a special case of WANETs that comes closest to our daily life. Given the wide adoption of smartphones all over the world, SPANs are thus regarded as a promising technology for introducing new applications such as proximity services and public safety applications.

In such an ad hoc network made up of mobile phones, movement and changes are constant factors, leading to the burden of maintaining reliable routing information. The existing literature on SPANs mainly focused on technical challenges such as communication issues, routing protocol, and resource allocation, while smartphone users' social connections are another major issue that can be further explored. To integrate the social networks and SPANs, it is necessary to study users not only as friends online, but also who would like to meet offline. Thus, we define the online social network (SN) as the online platform that reflects users' friendships and influence on one another. On the other hand, the offline social network refers to the physical wireless network in which users' mobility and proximity are considered, and where the contents data is transmitted.

The benefits of integrating SNs into SPANs are based on recent studies on information/content spreading behaviors. Researchers have realized that SN websites such as Facebook and Twitter are playing significant roles in the propagation of information over the Internet [1]. Observing these online SNs, people have found that in delay-tolerant networks (DTNs), with a small number of initial seeds in a local area, an efficient content dissemination mechanism through opportunistic sharing can guarantee content delivery while satisfying target delay requirements. Therefore, we see that if social awareness in both online SNs and offline SNs can be exploited to design efficient data forwarding mechanisms in SPANs, we can achieve efficient information dissemination with the lowest transmission cost. One brief illustration of content dissemination by integrating online SN and offline SN in SPANs is shown in Fig. 1.

The work in [2] has already shown that by grouping mobile nodes with similar mobility patterns into a cluster, the proposed cluster-based routing protocol can achieve higher delivery ratio,

This work was made possible by NPRP grant # 4-347-2-127 from the Qatar National Research Fund (a member of the Qatar Foundation). The statements made herein are solely the responsibility of the authors.

Digital Object Identifier: 10.1109/MCOM.2017.1600029CM

Yanru Zhang and Z. Han are with the University of Houston; Lingyang Song is with Peking University; Chunxiao Jiang is with Tsinghua University; Nguyen H. Tran is with Kyung Hee University; Zaher Dawy is with the American University of Beirut.



and lower overhead and delay. The authors in [3] also proposed the utilization of network-wide clustering to facilitate data propagation. The previous two works provided the original form of content sharing in SPANs. However, none of them focused on the social structure within each cluster, which can play a key role in further optimizing the efficiency of SPANs. To facilitate the design of social-aware SPANs, there are three keys that need to be fully exploited:

- **Offline SNs:** These constitute the physical layer where users locate within transmission distance and data is transmitted. Users' mobility patterns and encounter histories can be studied to derive the social structure in the physical wireless network, which can be utilized to establish reliable SPANs.
- **Online SNs:** These contain information such as virtual connections and friendships. Users' social online influence has a strong impact on the spreading breadth of the posted content. Studying the social connections among people in an online SN can assist in designing an efficient data dissemination mechanism in the offline SN.
- **Content popularity distribution:** If the popularity of online content can be accurately modeled, we can analyze the probability with which content will be requested, and thus design the content transmission mechanism more efficiently.

The future wireless technology needs to provide efficient means to bridge online and offline communities, that is, using online social structures and iterations to enhance data transmission in physical offline SNs. Therefore, we study both offline SNs and online SNs, and their combination in facilitating SPANs for data delivery. We first introduce existing models for user social connections in offline wireless networks. Then we talk about different structures of online SNs. In the end, we give examples of this emerging category of applications, and analyze the potential and benefits of combining offline and online SNs to facilitate data transmission.

## OFFLINE SOCIAL NETWORK

Similar to the content delivery in DTNs, due to the mobility of users, contents are shared among SPAN users through opportunistic encounters. Studying the social properties of SPAN users by considering their locations, mobility patterns, and interests will further improve the security and delivery delay. In the following subsections, we introduce three different aspects from which we can derive the offline social structures.

### LOCAL COMMUNITY

Communities are the SNs that mirror our daily lives. People such as neighbors, co-workers, and classmates who we meet regularly are the perfect candidates for initialing a cooperative SPAN, since they are not only trustful content providers, but also easily meet the data transmission requirements of short-range communication such as Bluetooth. In those scenarios, slow mobility patterns and close transmission distances are critical factors to guarantee the high quality of short-range communication. Therefore, it is worthwhile to derive offline SNs from local communi-

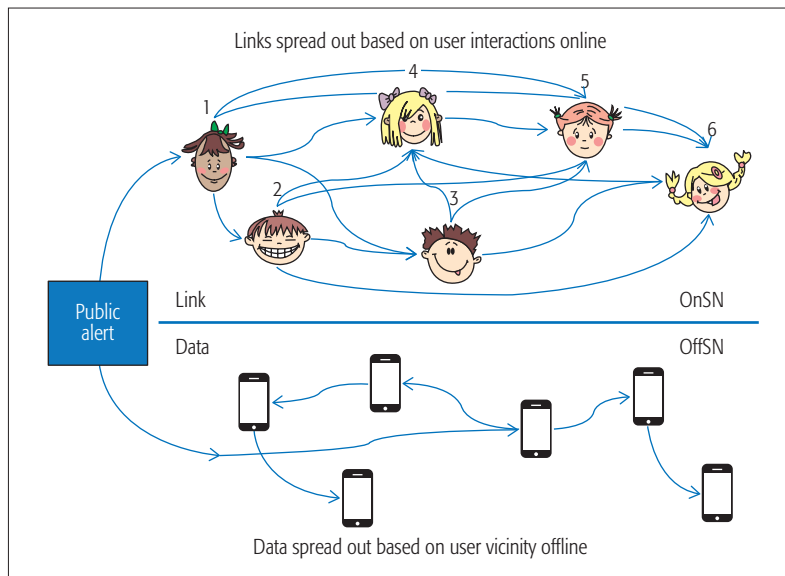


Figure 1. An integration of online and offline social networks in SPANs.

ties. There are extensive studies that exploit these friendship-based SN characteristics to assist content delivery in DTNs. Furthermore, there are also algorithms that can be used to detect communities such as  $k$ -clique [4].

### OPPORTUNISTIC ENCOUNTER

In the SNs of local communities, users are known to each other as families, friends, co-workers, classmates, and so on. However, there are many people who work or study in the same building but are strangers to each other. Users who are stably within proximity of each other are also ideal candidates for SPANs to transfer or relay data. Achieving cooperative communication from nearby users or strangers will be feasible if users are well motivated and secured, and it is a mutually beneficial action to expand the SPAN coverage for multicast [5]. This kind of SN can be developed from peer-to-peer encounter histories. Metrics such as frequency and length of the encounters are the key parameters that determine the social connections of users. One way is to find the probability of two users having an effective encounter. An effective encounter means that the two users are within the maximum transmission distance, and their encounter duration is long enough to finish the data transmission. Based on the Bayesian parameter estimation method, deciding the distribution that fits the data best is essential; known works include Beta distribution [6].

### COMMON INTEREST GROUPS

Despite the locality and mobility of users, interest is another critical social aspect of users. Common interest groups differ from local communities in the sense that people may live, work, and study in various locations and do not necessarily go to the same place every day. However, they love to access similar contents, and may like to go to activities and events arranged by themselves or others related to games, movies, books, pets, careers, or hobbies. Forming the offline SN through users' interests can improve the effectiveness of SPANs in the following three aspects.

In recent years there has been a dramatic rise in the number of mobile users who are connected to social network websites, such as Facebook and Twitter.

In particular, social network websites have been playing a significant role in the propagation of information online, and content data has become the main source of traffic in wireless networks offline.

First, users with common interests are more likely to have opportunistic encounters than others. For example, people who are interested in comic books will possibly attend a comics show. The mobile app Meetup brings people the convenience of arranging group meetings offline for people who share the same interests. Indeed, SPANs have drawn lots of attention for assisting temporary large-scale events where a huge scale of communication is needed for a short period of time. Second, within such an SN, the probability of successfully acquiring desirable content will be much higher than that in a local community opportunistic encounter, since people with the same interests are more likely to have mutually interesting contents. Third, it has been shown in [7] that there is a high locality of interest for data within a certain region (i.e., certain contents are usually popular within a certain local region). For example, users who want to stream a live football game are mostly to come from regions of the home and guest teams.

### ONLINE SOCIAL NETWORKS

In recent years there has been a dramatic rise in the number of mobile users who are connected to SN websites, such as Facebook and Twitter. In particular, SN websites have been playing a significant role in the propagation of information online, and content data has become the main source of traffic in wireless networks offline. Given that users actively exchange information over those SNs, it will be helpful to extract the social structure among mobile users by exploring the interacting patterns online. This social structure derived from user online iterations can be regarded as the online SN, corresponding to the offline SN extracted from users' wireless connections offline. Next, we discuss online SN structures from three aspects.

#### BIDIRECTIONAL

The SN applications such as Facebook and Wechat belong to the bidirectional online SNs, since in those online SNs, connections are confirmed by both sides of users, and contents posted by either side are viewable by the other. Thus, a user's activity will influence anyone in a bidirectional online SN. A bidirectional online SN is similar to the local community in the offline SN, where the connection is formed between mutually known users. A bidirectional online SN can be utilized to assist a cooperative SPAN, since users are trustworthy friends to each other.

#### UNIDIRECTIONAL

In popular SN websites such as Twitter and Weibo, users can follow anyone they are interested in, such as movie stars, singers, and celebrities in other areas. It is unnecessary for two users to know each other, and a social connection can be established. An online SN extracted from a unidirectional social connection shows great potential in designing efficient data forwarding mechanisms. For example, daily posts, news, and advertisements from accounts with huge followers are able to make more people see them than other accounts. Thus, the information achieves the maximum propagation within the shortest time.

### CENTRALIZED CONTENT

The centralized content social structure can be found in the popular SN website called Douban, where users can create contents related to film, books, music, and recent events and activities. Users are also allowed to form groups based on the topics in which they are interested. Inside the groups, users do not have to know each other, but they can access the same contents inside the group, and may contribute to enrich the content if they are willing to do so. This online SN is similar to offline common interest groups.

### ONLINE CONTENT POPULARITY

Users' online requests for content can be influenced by both internal (users' own interests) and external (influence from outside) reasons. While a user's internal factors are difficult to predict a priori, the external influence that comes from users' interactions in an online SN can be estimated. Such an analysis of network externality (external influence from media, friends, etc.) constitutes one of the major topics studied in online SNs, but it focuses more on social influences than on the structures discussed in the previous sections.

With the information circulating over the online SN and the data transmitted over the offline SN, the popularity distribution of online content is worthwhile studying to assist in the design of a content propagation mechanism, such as different forwarding schemes for popular and regular contents, and thus enhances the efficiency of SPANs. Fortunately, based on practical measurement, spreading impact modeling, and user profiling, it has been proven that it is not difficult to predict popular trends and access patterns [8].

#### PARAMETRIC METHOD

One simple way is the parametric method by assuming that the popularity of online content follows a certain distribution. Then the problem becomes among those known distributions, which of them will fit the online content popularity? There have been many studies in this area, and we select Zipf's law, Pareto's law, and power-law distributions for a brief introduction.

**Zipf's Law:** Zipf's law was first introduced in language study to model the frequency of used words, where the frequency of any word is inversely proportional to its rank in the frequency table. Later, people found that a similar relation occurs in many other areas, including the population ranks of cities' corporation sizes, ranks of the number of people watching the same TV channel, and so on. Based on numerous studies, Zipf's law has been established as a good model of the measured popularity of online videos. In [4], the authors adopted Zipf's law to model video resources online for BS-assisted device-to-device (D2D) communication, which is similar to data transmission in SPANs but over licensed spectrum.

**Pareto's Law:** Pareto's law was first used to describe the distribution of income. Different from Zipf's law, instead of asking what the largest income in a specific rank is, Pareto's law asks how many people have an income greater than a given number. Indeed, there is a tricky connection between Zipf's law and Pareto's law in the way the cumulative distribution is plotted. While

the Zipf rank distribution is plotted with ranking on the horizontal x axis and number on the vertical y axis, in the Pareto distribution, the x and y axes are flipped. Therefore, we can see that Pareto's law is also able to fit the online content popularity as Zipf's law does. In [1], the authors did a test on whether Pareto's law applies to user generated content video popularity, and the simulation results based on real data show a good fit on the Pareto's law.

**Power Law Distribution:** A power law distribution does not tell us how many people have an income greater than a specific number as in Pareto's Law, but the number of people whose income is exactly the given number. Indeed, the power-law distribution is a direct derivative of Pareto's law, which covers a wide range of varieties including the frequencies of words in most languages, frequencies of family names, and so on. The authors in [9] used power law distribution to fit the online content popularity, and compared the power law fit against other alternative distributions such as exponential distribution and log-normal distribution. The results indicated that the power-law is a better fit than the alternative distributions.

### NON-PARAMETRIC METHOD

Finding a closed-form expression of online contents' popularity distribution can be challenging. In addition, the distribution of online contents is highly time-varying as users continue to access them. Therefore, many studies argue that assuming content is drawn from a given probability distribution may not be appropriate. Fortunately, the nonparametric method provides another way of estimating popularity instead of fitting any parameterized distributions. The use of this model can automatically derive a distribution model from the observed data and learn the network structure.

**Non-Parametric Regression:** As one form of regression analysis, non-parametric regression does not refer to a pre-specified functional form but a flexible functional form constructed by information extracted from the observed data. The authors of [10] used a Cox proportional hazard regression model to infer the popularity of content with publicly observable metrics, such as the threads lifetime, views, and comments number. The Cox proportional hazard regression does not assume any parametric structure for the baseline hazard.

**Stochastic Process:** In probability theory, there are some discrete-time stochastic processes that are similar to the users' content selection process online. One example is the Bayesian non-parametric methods' extension, Indian Buffet Process (IBP), which models an Indian buffet where each diner chooses several samples from infinite dishes. The first customer selects its preferred dishes according to a Poisson distribution without any external influence, whereas the following customers make their selection with the prior information based on the first customer's feedback. Therefore, the decisions of subsequent customers are influenced by previous customers. One simple illustration of IBP is shown in Fig. 2.

The online content popularity spreading process in an online SN is analogous to the stochastic processes of IBP if regarding the online SN as an Indian buffet, the online content as dishes, and

		Dish										
		1	2	3	4	5	6	7	8	9	10	...
Customer	1	1	1	1	1	0	0	0	0	0	0	
	2	1	0	0	1	0	1	0	0	0	0	
	3	0	1	0	1	1	1	0	0	0	0	
	4	1	0	1	0	0	1	1	1	0	0	...
	5	1	0	0	1	0	1	1	1	0	0	
	6	1	1	0	0	1	1	0	0	0	1	
	...					...						...

Figure 2. One realization of the Indian buffet process.

the users as customers. Users enter the online SN sequentially to request their desired content, which changes the popularity distribution of content, and thus affects the probability of this content being requested by others. Popular content is requested more frequently, whereas content that is only favored by a few people or newly produced content is requested less frequently. For comparison, there are three real datasets of YouTube video viewing count, which are sampled from top ranked videos, random videos, and copyrighted videos. We observe the effectiveness of using the IBP to describe the online content popularity in Fig. 3.

## INTEGRATION OF ONLINE AND OFFLINE SOCIAL NETWORKS

From the previous discussions, we learn the characteristics of online and offline SNs. Future wireless technologies need to provide efficient means to bridge online and offline communities, that is, using online social structures and iterations to enhance data transmission in physical offline SNs. In this section, we give examples of this emerging class of applications, and analyze the potential and benefits of combining offline and online SNs to facilitate efficient data transmission in WANETS.

### CONTENT SHARING BETWEEN FRIENDS

Given a local-community-based offline SN and a bidirectional online SN, one application is the pre-loading of content from social networking applications, which has been shown to consume a significant portion of the overall wireless network traffic. Real trace experiments conducted by the CRAWDAD team at the University of St. Andrews [11] demonstrated the effectiveness of online SN and offline SN integration. The researchers first explored the online SN and offline SN from participants' Facebook friendships and sensor mote encounter records on campus. Then they used the detected communities and cliques to see whether they helped determine routing paths. Their results show that when the content dissemination uses the information from either an online SN or an offline SN, the delivery ratios do not show significant difference between the two. However, integrating the two leads to a higher delivery ratio and significantly lower communication cost.

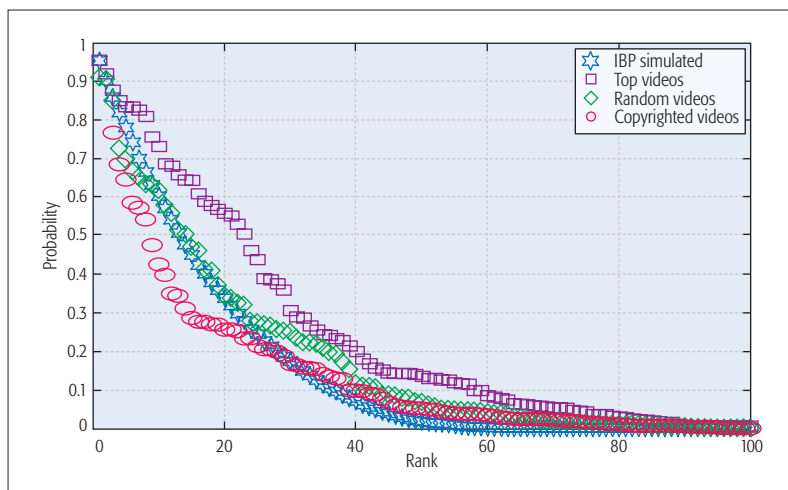


Figure 3. Video rank and selection probability by real trace and IBP.

This result is not surprising, since the social structures obtained from online and offline SNs help better targeting of content data, relays, and destinations. We can elaborate on this idea by the following example. User A and user B frequently interact online. If A posts or shares content online, it is highly likely that B will access it. Thus, the content should be pre-loaded and forwarded to B to avoid increasing traffic during peak hours. While A is located far away from B, the content shared by A is accessible by B's two other offline SN friends C and D. Then the data can be forwarded to B from C during the day since they work in the same office building. D can forward the data to B at night since they are neighbors.

In summary, integration of online and offline SNs increases delivery ratio and reduces communication cost in the following ways. First, with the users' friendship structures and content popularity in online SNs, one can decide which data to forward and to which user to forward. Second, the transit communities that are formed by users' mobility patterns in offline SNs determine when to forward the data, and which user to forward it to as an ad hoc relay [12].

### INFORMATION PROPAGATION

In the case of pure information broadcasting, friendship is not necessary as in the previous application. In online SNs such as Twitter and Instagram, tweets and pictures are pushed to all followers once posted. In offline SNs, data can be broadcast among strangers as long as they are within a certain transmission range. The TOSS algorithm proposed by [13] integrates online and offline SNs together to assist information broadcasting. TOSS first models the content access probability online using a Weibull distribution to decide which content has the higher priority to be forwarded. Then it exploits users' spreading impact and user-dependent access delay between the content generation time and each user's access time from the SN; it also extracts users' mobility impact from the offline SN. The authors use these inputs to derive the probability of accessing content. The objective function is to maximize the access probability without infrastructure-assisted communication and subject to the access delay constraint. Finally, TOSS

can decide which seed users are responsible for pushing the content to other users they opportunistically meet. The trace-driven evaluation demonstrates that TOSS can reduce up to 86.5 percent of the cellular traffic while satisfying the access delay requirements of all users.

When the objective is to maximize information propagation, it is critical to target certain initial seeds who have large influence on the others [14]. Most works on offline SNs select seeds with high mobility patterns in or between communities, which makes them have larger probability to encounter the others and thus increase delivery ratio. However, as we all know, a celebrity on Twitter can have his/her tweets pushed to millions of followers, which is far more influential than any active nodes offline. Thus, taking one's online influence into consideration when disseminating information offline will have great potential to facilitate data dissemination and reduce communication cost. This system model is promising in areas such as local advertisement pushing and public emergency alarm systems.

### INTEREST GROUPS

People grouped in small networks, such as by regions and interests, are more likely to access the same content and meet in certain places. The work done in [15] implies that to achieve better diffusion performance, each node should diffuse data similar to their common interests when it meets a friend, and diffuse data different from their common interests when it meets a stranger. This work tried to take strangers as bridges and relays to forward one's disinterested data to interested communities. Such an algorithm reaches a high delivery ratio; however, a limited amount of communication cost can be reduced, since many disinterested nodes are also infected.

The social patterns of interest groups online include useful information such as users' interests and the groups to which they belong. In addition, people tend to have multiple interests, and some of them are active members in different groups. Thus, if each node can access that information before they forward data in an encounter, they can better target the bridge and relay nodes in different interest groups. On one hand, with the information extracted from the online SN, we know which interests groups the users belong to, and thus which data should be forwarded to them, to ensure that most of the infected users are those who are interested in the content. On the other hand, with the information extracted from the offline SN, we know which users do actively move among the different groups. Therefore, one can design algorithms to better target the carrier nodes to achieve minimal delay and reduce communication cost.

### CONCLUSIONS

In this article, we study how to use a social-aware framework to optimize information dissemination in SPANs. We exploit users' social connections in two layers: the online SN and offline wireless network. We also provide further discussion on online content popularity, which is another representative of user interactions online. In the end, we propose three applications to integrate offline SNs and online SNs to enhance content delivery

in SPANs from three different aspects: bidirectional friendship-based content sharing, unidirectional information broadcasting, and interests-oriented content sharing. Overall, there is great potential to combine offline and online SNs to facilitate content spreading SPANs.

## REFERENCES

- [1] M. Cha et al., "Analyzing the Video Popularity Characteristics of Large-Scale User Generated Content Systems," *IEEE/ACM Trans. Net.*, vol. 17, no. 5, Oct. 2009, pp. 1357–70.
- [2] H. Dang and H. Wu, "Clustering and Cluster-Based Routing Protocol for Delay-Tolerant Mobile Networks," *IEEE Trans. Wireless Commun.*, vol. 9, no. 6, June 2010, pp. 1874–81.
- [3] N. Golrezaei, A. F. Molisch, and A. G. Dimakis, "Base-Station Assisted Device-to-Device Communication for High-Throughput Wireless Video Networks," *IEEE ICC*, Ottawa, Canada, Jun. 2012.
- [4] F. Li and J. Wu, "LocalCom: A Community-Based Epidemic Forwarding Scheme in Disruption-Tolerant Networks," *6th Annual IEEE Commun. Society Conf. Sensor, Mesh and Ad Hoc Communications and Networks*, Rome, Italy, June 2009.
- [5] W. Gao et al., "Social-Aware Multicast in Disruption-Tolerant Networks," *IEEE/ACM Trans. Net.*, vol. 20, no. 5, Oct. 2012, pp. 1553–66.
- [6] Y. Zhang et al., "Social Network Aware Device-to-Device Communication in Wireless Networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, Jan. 2015, pp. 177–90.
- [7] E. Jaho and I. Stavrakakis, "Joint Interest- and Locality-Aware Content Dissemination in Social Networks," *6th Int'l. Conf. Wireless On-Demand Network Systems and Services*, Snowbird, UT, Feb. 2009.
- [8] M. P. Wittie et al., "Exploiting Locality of Interest in Online Social Networks," *Proc. Co-NEXT '10*, Philadelphia, PA, Nov. 2010.
- [9] A. Tatar et al., "From Popularity Prediction to Ranking Online News," *Social Network Analysis and Mining*, vol. 4, no. 1, Feb. 2014, pp. 1869–5450.
- [10] J. G. Lee, S. Moon, and K. Salamatian, "An Approach to Model and Predict the Popularity of Online Contents with Explanatory Factors," *2010 IEEE/WIC/ACM Int'l. Conf. Web Intelligence and Intelligent Agent Technology*, Toronto, Canada, Aug. 2010.
- [11] G. Bigwood et al., "Exploiting Self-Reported Social Networks for Routing in Ubiquitous Computing Environments," *IEEE Int'l. Conf. Wireless and Mobile Computing Networking and Communications*, Avignon, France, Oct. 2008.
- [12] W. Gao et al., "On Exploiting Transient Social Contact Patterns for Data Forwarding in Delay-Tolerant Networks," *IEEE Trans. Mobile Computing*, vol. 12, no. 1, Jan 2013, pp. 151–65.
- [13] X. Wang et al., "TOSS: Traffic Offloading by Social Network Service-Based Opportunistic Sharing in Mobile Social Networks," *Proc. IEEE INFOCOM*, Toronto, Canada, Apr. 2014.
- [14] Z. Lu et al., "Towards information diffusion in mobile social networks," *IEEE Trans. Mobile Computing*, vol. 15, no. 5, May 2016, pp. 1292–304.
- [15] Y. Zhang et al., "Social-Aware Data Diffusion in Delay Tolerant MANETs," in *Handbook of Optimization in Complex Networks: Communication and Social Networks*, Springer, 2012, pp. 457–81.

## BIOGRAPHIES

YANRU ZHANG [S'13-M'16] received the B.S. degree in electronic engineering from the University of Electronic Science and Technology of China (UESTC) in 2012, and the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Houston (UH) in 2016. She is now working as a research associate at the Wireless Networking, Signal Processing and Security Lab, UH. Her current research involves the contract theory and matching theory in network economics, Internet and applications, wireless communications and networking. She received the best paper award at IEEE ICCS 2016.

LINGYANG SONG [S'03, M'06, SM'12] received his Ph.D. from the University of York, United Kingdom, in 2007, where he received the K. M. Stott Prize for excellent research. He worked as a research fellow at the University of Oslo, Norway, until joining Philips Research UK in March 2008. In May 2009, he joined the School of Electronics Engineering and Computer Science, Peking University, China, as a full professor. His main research interests include MIMO, cognitive and cooperative communications, security, and big data. He wrote two textbooks, *Wireless*

*Device-to-Device Communications and Networks* and *Full-Duplex Communications and Networks* (Cambridge University Press). He was the recipient of the IEEE Leonard G. Abraham Prize in 2016 and IEEE Asia Pacific (AP) Young Researcher Award in 2012. He is currently on the Editorial Board of *IEEE Transactions on Wireless Communications*. He has been an IEEE Distinguished Lecturer since 2015.

CHUNXIAO JIANG [S'09, M'13, SM'15] received his B.S. degree in information engineering from Beijing University of Aeronautics and Astronautics (Beihang University) in 2008 and his Ph.D. degree from Tsinghua University, Beijing, in 2013, both with highest honors. During 2011–2012, he visited the Signals and Information Group at the Department of Electrical & Computer Engineering, University of Maryland. During 2013–2016, he was a postdoctoral researcher at the Department of Electronic Engineering, Tsinghua University. He is currently an assistant research fellow at Tsinghua Space Center. His research interests include the applications of game theory and queuing theory in wireless communication and networking. He received the Best Paper Award at IEEE GLOBECOM '13, Best Student Paper Award from IEEE GlobSIP '15, Tsinghua Outstanding Postdoc Award in 2015, Beijing Distinguished Graduated Student Award, Chinese National Fellowship, and Tsinghua Outstanding Distinguished Doctoral Dissertation in 2013.

NGUYEN H. TRAN [S'10, M'11] received the B.S. degree from Hochiminh City University of Technology, and Ph.D. degree from Kyung Hee University, in electrical and computer engineering, in 2005 and 2011, respectively. Since 2012, he has been an assistant professor with the Department of Computer Science and Engineering, Kyung Hee University. His research interest is in applying the analytic techniques of optimization, game theory, and stochastic modelling to cutting-edge applications such as cloud and mobile-edge computing, datacenters, heterogeneous wireless networks, and big data for networks. He received the best KHU thesis award in engineering in 2011, and several best paper awards, including at IEEE ICC 2016, APNOMS 2016, and IEEE ICCS 2016. He is an editor of *IEEE Transactions on Green Communications and Networking*.

ZAHER DAWY [SM'09] received his B.E. degree in computer and communications engineering from the American University of Beirut (AUB), Lebanon, in 1998 and his M.E. and Dr.-Ing. degrees in communications engineering from Munich University of Technology (TUM), Germany, in 2000 and 2004, respectively. Since 2004, he has been with the Department of Electrical and Computer Engineering, AUB, where he is currently a professor. His research and teaching interests include wireless communications, cellular technologies, context-aware mobile computing, mobile solutions for smart cities, computational biology, and biomedical engineering. He is an Editor of *IEEE Communications Surveys & Tutorials*, *IEEE Transactions on Communications*, *IEEE Transactions on Wireless Communications*, and *Elsevier Physical Communications*. He also served as Executive Editor for *Wiley Transactions on Emerging Telecommunications Technologies* from 2011 to 2014. He received the Abdul Hameed Shoman Award for Young Arab Researchers in 2012, the IEEE Communications Society 2011 Outstanding Young Researcher Award in Europe, Middle East, and Africa Region, the AUB Teaching Excellence Award in 2008, the Best Graduate Award from TUM in 2000, the Youth and Knowledge Siemens Scholarship for Distinguished Students in 1999, and the Distinguished Graduate Medal of Excellence from Hariri Foundation in 1998.

ZHU HAN [S'01, M'04, SM'09, F'14] received his B.S. degree in electronic engineering from Tsinghua University in 1997, and his M.S. and Ph.D. degrees in electrical and computer engineering from the University of Maryland, College Park, in 1999 and 2003, respectively. From 2000 to 2002, he was an R&D engineer of JDSU, Germantown, Maryland. From 2003 to 2006, he was a research associate at the University of Maryland. From 2006 to 2008, he was an assistant professor at Boise State University, Idaho. Currently, he is a professor in the Electrical and Computer Engineering Department as well as in the Computer Science Department at the University of Houston. His research interests include wireless resource allocation and management, wireless communications and networking, game theory, big data analysis, security, and smart grid. He received an NSF Career Award in 2010, the Fred W. Ellersick Prize of the IEEE Communication Society in 2011, the EURASIP Best Paper Award for the *Journal on Advances in Signal Processing* in 2015, the IEEE Leonard G. Abraham Prize in the field of Communications Systems (best paper award in *IEEE JSAC*) in 2016, and several best paper awards at IEEE conferences. Currently, he is an IEEE Communications Society Distinguished Lecturer.

We proposed three applications to integrate the offline SN and online SN to enhance content delivery in SPANs from three different aspects: bidirectional friendship based content sharing, unidirectional information broadcasting, and interests oriented content sharing. Overall, there exists great potential to combine the offline SN and online SN in facilitating content spreading SPANs.

# Automatic Video Remixing Systems

Sujeet Mate and Igor D. D. Curcio

There has been a tremendous increase in the amount of user generated content (UGC), and many mobile devices are also equipped with sensors (magnetic compass, accelerometer, gyroscope, etc.). We present an automatic video remixing system (AVRS), which intelligently processes UGC in combination with sensor information. The system aims to generate a video remix with minimal user effort.

## ABSTRACT

There has been a tremendous increase in the amount of user generated content (UGC), and many mobile devices are also equipped with sensors (magnetic compass, accelerometer, gyroscope, etc.). We present an automatic video remixing system (AVRS), which intelligently processes UGC in combination with sensor information. The system aims to generate a video remix with minimal user effort. We present sensor-based as well as sensor-less architectures of such a system. The sensor-based AVRS system involves certain architectural choices that meet the key system requirements (leverage user generated content, use sensor information, reduce end-user burden) and user experience requirements. Architecture adaptations are required if any of the operating parameters need to be constrained for real world deployment feasibility. We present sensor-less architecture adaptations that enable the usage of the automatic remixing logic in different operating scenarios. The challenge for these system adaptations is to improve the benefits for certain key performance parameters, while reducing the compromise for other parameters. Subsequently, two key sensor-less AVRS architecture adaptations (Cloud Remix System and Smartview system) are presented. We show that significant reduction in system complexity can be achieved when a smaller reduction in the user experience is allowed. Similarly, the system can also be optimized to reduce the need for other key requirement parameters such as storage and user density.

## INTRODUCTION

Due to the ubiquitous availability of high-quality content capture, we often find someone recording video at an event or even multiple individuals recording videos. Most of today's mobile devices have in-built sensors including accelerometer, magnetic compass, gyroscope, GPS, and so on. The rapid growth in the available network bandwidth coupled with the social consumption of content has led to an explosion in the amount of video that is recorded by individual users and shared socially using various methods. The coming together of high-quality content capture and sensors, connected to the Internet via high-speed networks, provides unprecedented opportunities for various multimedia applications. It is becoming difficult for a user to manage content manually. For example, even though there might be many users recording videos during a concert, the following limitations may limit its proper utilization.

First, there is lack of quality assurance (in terms of objective as well as subjective quality parameters) of the individual videos. Second, there may be significant redundancy in the captured content. These two factors together reduce the usability of multiple video clips from the same event, due to difficulty in finding the best parts in terms of viewing value, as well as the objective media quality.

The increasing gap in the amount of content generated related to an event vs. the content consumers' ability to utilize it for various purposes can be addressed by automatically suggesting the best-quality content version from multiple similar versions of it (e.g., to find the best video clip from multiple clips of the same song recorded by different users at a concert). A manual approach can quickly become unmanageable for most users; hence, the need for automation. A vast majority of the users who capture video content do not perform any post-processing of the videos to improve the viewing experience. An even smaller fraction bothers to get involved in making multi-camera video edits with content captured by multiple users.

An automated system that leverages the high-quality content capture from multiple users in combination with sensor data can provide two important benefits. First, it can significantly reduce the threshold for a large demography to get involved into creating value added content like video remixes with their own content or that from multiple users. Second, the use of widely available in-built sensors in mobile devices can help produce a high-quality remix with high efficiency in terms of resource usage.

In this article we focus on the following. We present the state of the art in automatic remixing with crowd-sourced user generated content (UGC) and compare it to our approach. We give an overview of a sensor-based automatic video remixing system (AVRS), which creates multi-camera remix videos of live events (e.g., music concerts) fully automatically. We also present the AVRS requirements and describe their implications. Subsequently, we present sensor-less AVRS optimized for different operating scenarios and key performance parameters. We compare the sensor-based and sensor-less approaches in terms of benefits and compromises. Finally, we conclude the article.

## RELATED WORK

In this section we present related work in the area of automatic video remixing, which uses UGC. In [7], the proposed system utilizes audio-visual con-

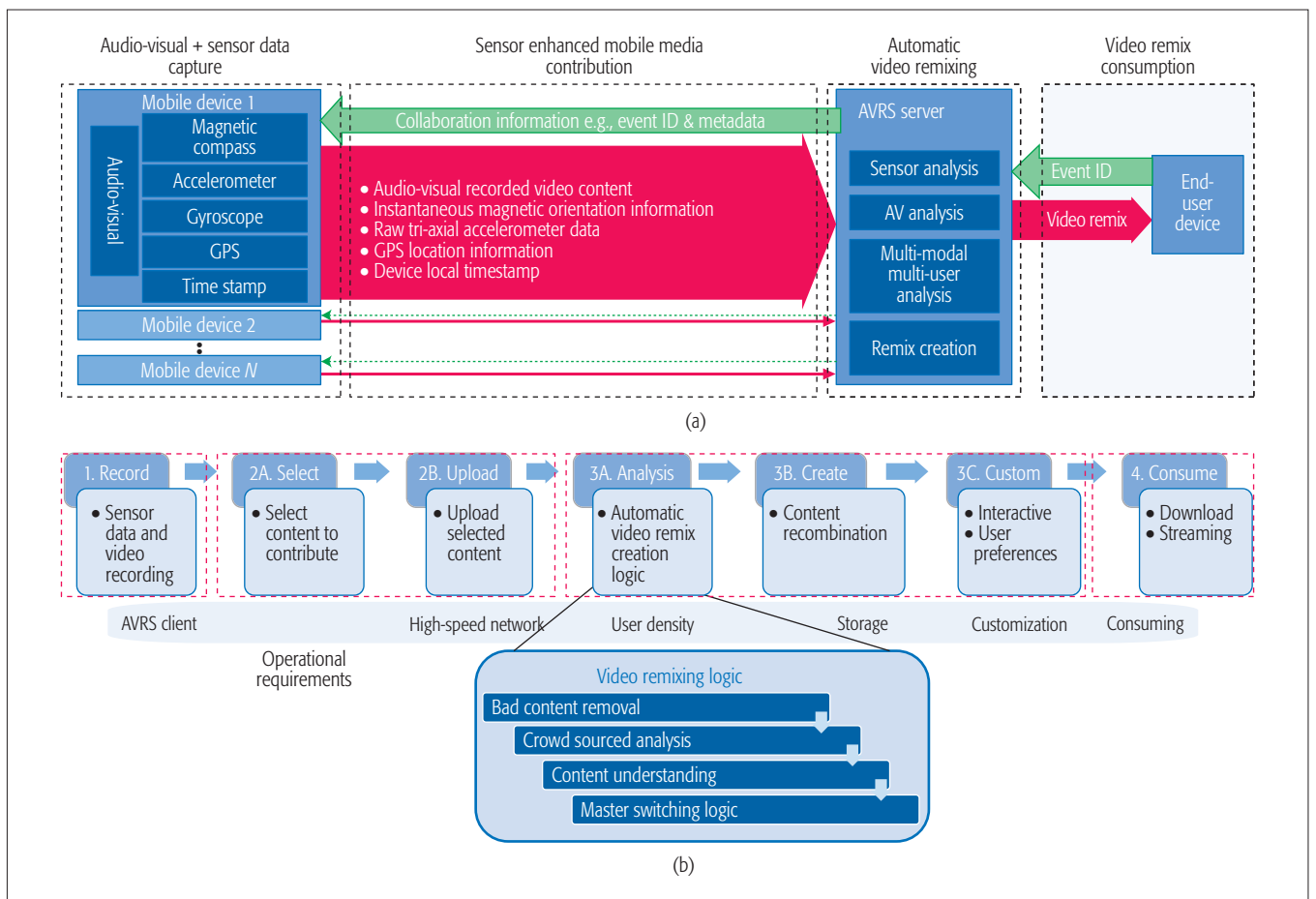


Figure 1. a) AVRS end-to-end system; b) functional overview.

tent analysis in combination with a pre-defined criteria as a measure of interest for generating the mash-up. This approach does not leverage the sensor information to determine semantic information. The system proposed in [8] utilizes video quality, tilt of the camera, diversity of views and learning from professional edits. In comparison, our system utilizes multimodal analysis involving sensor and content data where higher-level semantic information is used in combination with cinematic rules to drive the switching instance and view selection. The work [9] presents a collaborative sensor and content-based approach for detecting interesting events from an occasion like a birthday party. The system consists of grouping related content, followed by determining which view might be interesting and finally the interesting segment of that view. Our approach takes the sensor analysis as well as content analysis into account to generate semantically more significant information from the recorded sensor data (region of interest) as well as video data (audio quality, audio rhythm, etc.). The approach in [10] uses the concept of focus of multiple users to determine the value of a particular part of the event. The focus is determined by estimating camera pose of the devices using content analysis. This approach also utilizes cinematic rules as well as the 180° rule for content editing. Compared to this approach, ours is significantly less computationally intensive, since we utilize audio-based alignment of content and also sensor-based semantic information. A narrative description-based approach

for generating video edits is presented in [11]. This approach utilizes end-user programming for generating remixes corresponding to different scenarios. None of the approaches presented above address the issues related to architectural choices that improve particular performance parameters for certain operating scenarios while reducing the compromise on other parameters as the present article does.

## SENSOR-BASED AUTOMATIC VIDEO REMIXING SYSTEM OVERVIEW

The AVRS was conceived with three key requirements. The first requirement was to leverage the increasing amounts of high-quality UGC. The second motivation was to reduce the burden for the end user to create value added content, like a multi-camera video remix. The third key requirement was to leverage the trend of having different types of sensors as part of the mobile devices. These sensors provide a means of generating semantic and objective media content information with significantly lower computational resources, compared to the conventional media content analysis only approach. The sensor-based AVRS uses crowd-sourced UGC and sensors in the video recording device to create multi-camera video remixes [5, 6]. In the following, we first present the sensor-based AVRS end-to-end system overview. This is followed by further elaboration about the automatic video remix creation methodology.

These sensors provide a means of generating semantic and objective media content information with significantly lower computational resources compared to the conventional media content analysis only approach. The sensor-based AVRS uses crowd-sourced UGC and sensors in the video recording device to create multi-camera video remixes.

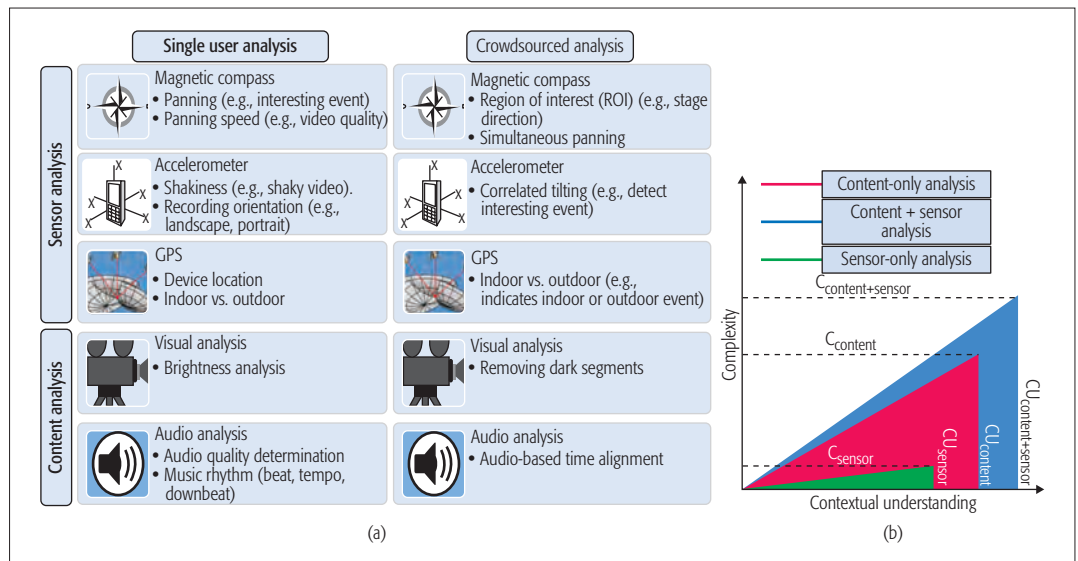


Figure 2. a) Sensor and content analysis methods; b) their comparison.

### AVRS END-TO-END SYSTEM OVERVIEW

Logically, the AVRS automatic multi-camera video remix creation can be divided into four main steps (Fig. 1a).

- The first step consists of capturing media and associated time-aligned sensor information from the recording device, which are data from the magnetic compass, accelerometer, GPS, and gyroscope. The sensor data provide motion and location information of the recording device. It is encrypted and stored in the same file container as the video file.

- The second step involves having a service set up that facilitates multiple users attending an event to effectively co-create a video remix. An “event,” created in the system by one of the participants of the event itself or the organizers of the event, acts as a nodal point for content contribution. Based on the user’s selection, media items (along with the associated sensor data) are uploaded to the server. In order to ensure robustness over an unreliable network, upload with small chunks of data over HTTP is used.

- The third step starts with processing of all the contributed content (also referred to as source media), which consists of sensor data in addition to the audio-visual data, to understand the semantic value and objective media quality of the received media from multiple users. The sensor data from heterogeneous devices is normalized to a common baseline and utilize vendor-specific sensor data quality parameters to filter data. The AVRS needs to support iterative and incremental remix creation, since the source media is not received in one go, but over an extended period of time. Successive remixes can include portions of the newly contributed content if they offer new and better views compared to the previous version of the remix.

- The fourth and final step involves downloading or streaming the video remix generated as a video file. This also includes additional metadata to acknowledge the contributing users and is done by overlaying a contributing user’s information when her contributed source is being used. This ensures transparency and due accreditation to all the contributors.

### VIDEO REMIXING

The sensor-based AVRS analysis and automatic video remix creation logic consist of four main steps (Fig. 1b): bad content removal, crowd-sourced media analysis, content understanding, and master switching logic. The use of sensor data, in addition to the traditional content analysis only approach, provides significant advantages. Figure 2a presents in brief the sensor and content analysis methods utilized in this system. It can be seen from Fig. 2b that high efficiency for contextual understanding can be achieved by using sensor data, whereas better contextual understanding can be obtained by combining sensor and content analysis [14][15]. Thus, sensors can play a significant role in improving efficiency as well as expanding the envelope of semantic understanding. With reference to Fig. 1b, step 3A:

**Bad content removal** primarily involves removing content segments with poor objective quality. Sensor-based methods (using accelerometer and magnetic compass data) can be applied on each video file to remove shaky or blurred video segments, segments recorded with incorrect orientation (portrait vs. landscape), and also those that may be recording irrelevant parts, such as feet. Dark segments are removed with content analysis. This process is significantly optimized due to the combination of content analysis and motion sensor data analysis, compared to the traditional content analysis only approach (Fig. 2).

**Crowd-sourced media analysis** consists of analyzing source media and the corresponding sensor data contribution by multiple users in the event. This information, which may be insignificant for one user, when combined with the same information from multiple users in the same event can provide valuable semantic information about the salient features of the event. By utilizing magnetic compass data from all the contributing users, information regarding the significant direction of interest (e.g., a stage) in the event can be determined. Simultaneous panning/tiltings can indicate occurrence of an interesting event. Precise time alignment of all the contributed videos is done by analyzing the source media audio



content envelope. Some methods to understand the semantic information and event type with the help of multimodal analysis have been described in [13, 14]. This is an essential requirement for seamless recombination of different source videos. The power of the crowd and the sensor information add significant value without requiring heavy computational requirements.

**Content understanding** starts with determining the characteristics of the source media. Sensor data corresponding to each source media item can efficiently provide orientation (w.r.t. the magnetic North as well as the horizontal plane), and fast or slow panning/tilting information about the recorded content. Other information consists of rhythm information in case of music and face information, which is determined with content analysis. This data is used to find the appropriate instance for changing a view, and to select the appropriate view segment from the multiple available views.

**Master switching logic** embodies the use of all the information generated in the previous steps in combination with cinematic rules to create a multi-camera video remix. Some content type specific methods like sports have been studied in [12]. The master switching logic determines the appropriate switching times of views for a multi-camera experience, and uses a method for ranking the views based on the interest derived from the previous steps. Bad quality content is penalized. A seamless audio experience is obtained by selecting the best quality audio track from the source content and switching to a different track only when the currently selected track ends. The video remix can be personalized by providing user-specific preferences to the master switching logic parameters: for example, users can indicate whether they prefer more frequent view switches or would like to have more of their own content as part of the video remix. In addition, the switching algorithm can adopt a switching pattern based on a model derived from a reference video (e.g., a multi-camera video clip of a pop music video) [16].

The sensor-based AVRS has been validated with users in multiple trials involving different scenarios [1, 3]. The findings indicate that although manual remixes outperform automatic remixes, for certain specific applications such as event memorabilia, automatic remixes performed as well as manual remixes [2]. Taking into account the effort required for producing a manual remix, user studies indicated a clear preference for using the AVRS, since it requires negligible effort. The AVRS was considered to work well, especially at music concerts, followed by party events and then sports events (based on a study of 30 users).

## SENSOR-BASED

### AVRS USAGE REQUIREMENTS AND IMPLICATIONS

The AVRS implementation was optimized for high subjective viewing quality experience without constraining system complexity, bandwidth, and storage requirements. Following are the operating requirement implications of the AVRS (Fig. 1b):

**1. AVRS recording client** equipped mobile devices. This client is required to record video and sensor information in parallel, and consequently, embed them in a format that is understandable by the AVRS server.

**2. High-speed uplink** bandwidth connectivity for end users to enable source media contribution. Uploading content was found to be the single biggest component in terms of early video remix availability. With the rapid increase in video resolutions from VGA to 720p to 1080p to 4K (and beyond), the amount of data that needs to be uploaded for every minute of video content is growing rapidly.

**3. User density** is an essential but uncertain variable of a system that depends on UGC contribution by the “crowd” present at any event, and hence cannot be known a priori. Even though there may be many people at the event, a fraction of those are likely to record media, and even fewer are likely to contribute their content. For events with a large number of participants, user density is not a problem; however, for events with fewer people, this can become a bottleneck.

**4. Storage** is a significant infrastructure requirement for user contributed source media and the subsequent iteratively generated video remixes. This requirement is expected to be an increasing component from the maintenance perspective, as the amount of events and the corresponding media (source as well as remixes) continue to grow over time.

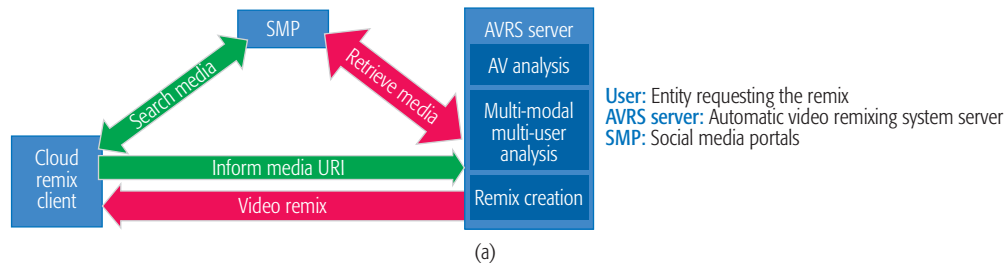
**5. Customization** of the automatic video remix is an important feature to allow personalization based on the user’s own preferences. The customization can be done before a video remix is generated, and further customization requires creation of a new video remix file.

**6. Consuming** the video remix involves streaming or downloading the video file. Viewing a newer version of the remix or a customized version of the remix requires downloading or streaming the complete remix again. This implies the need for reliable and high-speed connectivity for a pleasant user experience.

## SENSOR-LESS AUTOMATIC VIDEO REMIXING

The sensor-based AVRS was developed with the key requirements listed above. Assumptions regarding the operating scenario for the sensor-based AVRS have driven the architectural choices, and their implications have been described in the above section. In summary, the operating scenario generally expects availability of a sensor data capture in parallel with video recording on the participating users’ mobile device, and also that the infrastructure on the service side is capable of accepting and holding the sensor data together with the audio-visual data. In addition, availability of high-speed upload capability, minimum critical density of sensor data enriched video contributors, is assumed. Overall, the above choices aim for high-quality user experience without constraints on resource requirements. Real world deployment scenarios limit the support for devices with sensor data annotated capture of videos, as well as support for handling sensor data in the mainstream social media portals. These limitations directly affect the achievement of minimum critical density of users who can participate, and also the business model, as such a system would require proprietary support for end-to-end system realization. To overcome these limitations, an architecture adaptation of the video remixing system is required, which is optimized for a differ-

Real world deployment scenarios limit the support for devices with sensor data annotated capture of videos, as well as support for handling sensor data in the mainstream social media portals. These limitations affect directly the achieving of minimum critical density of users who can participate, and also it affects the business model.



User: Entity requesting the remix  
 AVRS server: Automatic video remixing system server  
 SMP: Social media portals

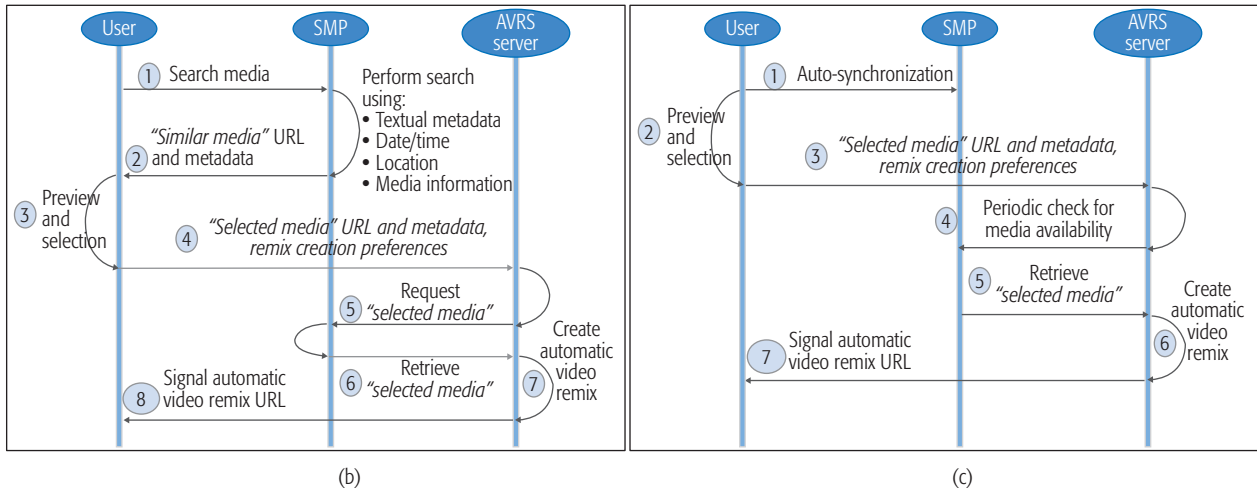


Figure 3. a) Cloud remix system overview; b) sequence diagrams without auto-synchronization; c) sequence diagrams with auto-synchronization.

ent set of operating scenario parameters. In the following, two sensor-less architecture adaptations of the AVRS are presented.

### CLOUD REMIXING SYSTEM

From the sensor-based AVRS described above, it was found that operating requirement 1 needs wide availability of devices equipped with a non-standard video recording client. In its absence, devices that do not have such a client would not be able to contribute. Consequently, operating requirement 3 might also be compromised. In addition, operating requirement 2 would be difficult to satisfy for users in regions having low network bandwidth, unreliable connectivity, or high data usage costs. The problem is more pronounced in terms of user experience when a user explicitly uploads videos to get a video remix, because she has limited patience to wait before seeing any result. Consequently, this architecture adaptation of the video remixing system envisages removing the need for uploading videos with the sole purpose of generating a video remix. Contributing content to the sensor-based AVRS by uploading videos was identified as a pain point by the users.

The cloud remixing system retrieves source media directly from social media portals (e.g., YouTube). This approach leverages the content uploaded by other users from the same event and also allows the users to leverage the uploaded content for sharing it with friends, in addition to creating remixes. All content available in the social media portals can be used for video remix creation. In practice, the content retrieval directly from the cloud can be done in two ways.

The first method (Fig. 3b) consists of the user

querying one or more social media portals (SMPs) for content of interest using the search parameters supported by the respective SMPs (step 1). The SMPs return the results based on the search parameters (step 2). The user previews the media and selects the source media to be used for generating the video remix (step 3). The selected media URLs are signaled to the AVRS server (step 4). This retrieves the source media using the signaled URLs directly from the SMPs (steps 5 and 6). The automatic video remix video is generated in the AVRS server (step 7). Finally, the video remix URL is signaled to the user (step 8). The video remix file is stored on the AVRS server for a limited period, during which the user is notified to view and download or stream the video.

In the second method (Fig. 3c), the cloud remix system leverages the auto-synchronization of media on the device and the cloud (e.g., DropBox, Microsoft OneDrive, Google Drive), which is available on an increasing number of mobile devices. This feature can be used by the cloud remixing client on users' mobile devices to contribute their content to the AVRS server, and it significantly mitigates the perceived delay in the upload, since the content contribution is explicit but does not require an explicit upload. The contributed source media URLs or media identifiers are signaled from the cloud remix client to the AVRS server (step 3). The AVRS server periodically checks for the availability of the contributed source media on the user's SMP (step 4). When the source media is available on the user's SMP, the AVRS server retrieves the content directly from the SMP (step 5). The AVRS server creates the video remix (step 6), and subsequently stores it for a limited duration (as described in the above paragraph).

## SMARTVIEW REMIXING SYSTEM

This architecture adaptation envisages a system that can work completely on a mobile device, without the need for any network connectivity for generating the video remix [4]. In addition, it is envisaged that this architecture adaptation of the video remixing system should enable creation of value added content from even a single user recording a single video clip from an event. Consequently, the operating parameters are drastically different from the sensor-based AVRS. This requires a radically different architecture compared to the sensor-based AVRS, while retaining the essential methodology of generating a fully automatic video remix. This implies that the core cinematic rules, content understanding, and low footprint are essential for such a system. Thus, this architecture removes the need to comply with operation parameter implications 2, 3, and 4. In the following sections, the terms SV and MTSV will refer to single video and multiple video SmartView, respectively.

### Single Video and Multiple Video SmartView:

There is currently a trend of many high-end devices being equipped with a display having a native resolution that is lower than the device's maximum supported video recording resolution. For example, it is not uncommon to see devices equipped with 4K video recording capability having a 1080p resolution display. The resolution difference between the native video capture resolution and the display resolution is used to generate sub-resolution rendering (e.g., close-ups) without compromising the viewing experience. The SV creates a multi-camera video remix viewing experience from a single video. The MTSV extends the SV concept to incorporate multiple videos. The MTSV creation involves analyzing the multiple videos to generate rendering metadata, which is used by a metadata-aware player.

**SmartView System Overview:** The SV/MTSV creation is initiated (Fig. 4) for single or multiple videos by a user first selecting the videos to be used for remix creation (step 1). The SV application (SVA) extracts and analyzes the one or more audio tracks, and time aligns the multiple videos using their audio track information (step 2). In step 3, audio characteristics like music rhythm and downbeat information are determined to semantically derive coherent switching points for rendering different views. This information is used to analyze the video frames corresponding to the switching instances. Such analysis can consist of detecting faces in the video frames from one (SV) or more source videos (MTSV) to rank the inclusion of different views for each temporal segment (step 4). This information is used in combination with cinematic rules to generate rendering metadata (step 5). The rendering metadata consists of source media identifier(s) for audio and visual track rendering for each temporal segment (step 6). The spatio-temporal rendering coordinate information is stored as SV or MTSV rendering metadata. A SmartView rendering is performed with the help of a player application on the same device that is able to scale the video rendering and/or render the different source videos to deliver the desired multi-camera remix experience (step 7). The remix creation is limited to generating metadata and does not involve video editing

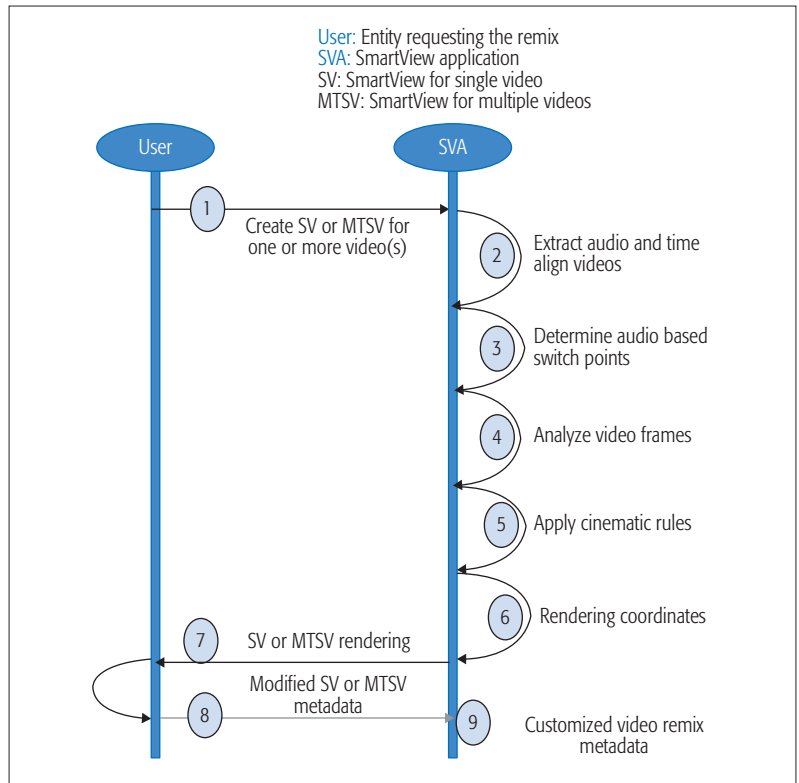


Figure 4. Single and multiple video SmartView architecture adaptation.

or re-encoding. The overall footprint of such a system is minimized to enable remix creation of videos completely on the device. This approach also opens the possibility of instantaneous video remix manual customization to the user with very lightweight processing (step 8). The modified SV metadata is stored within the original video file in a suitable format in case of a single source video input. For multiple source videos, the MTSV metadata is stored either in the source videos or separately (step 9).

MTSV can use multiple videos for video remix experience on a device without the need for network connectivity. The content from multiple users is collected via side loading to remove any dependency on the cloud. The SV metadata creation process for the multiple videos scenario is similar to the earlier scenario, except for the addition of time alignment of the multiple videos. In case of multiple source videos, this step can be repeated to rank different source videos or analyze objective visual quality to avoid bad quality views (step 4).

For multiple source videos, the rendering coordinates consist of a source video identifier for video and audio track for each temporal segment. The audio track switches are minimized to provide a seamless experience. Thus, it is possible to leverage an audio track from a different source video than the one from which the video track is rendered. In the case of multiple source videos, the multi-camera remix experience is generated by switching the rendering between multiple source videos for the corresponding temporal segments.

Based on our user studies, most users did not notice any visual quality difference between the native resolution video played with a conventional

	Sensor-less		Sensor-based
	SmartView (SV as well as MTSV)	Cloud remix system	AVRS
Min. number of videos	1	> 1	> 1
Min. number of people	1	1	> 1
Source videos from	Any standard-compliant MP4 video	You Tube or other portals (no capture required) or mobile platforms with auto-synchronization	Sensor data enriched video recording client
Social media portal support	Not needed	Yes	No
Explicit upload required	No	No (or autosync services)	Yes
Final output downloading required	No	Optional	Optional (streaming is preferred)
Manual customization capability	Yes (new video file not created)	Change remix parameters (new video file created)	Change remix parameters (new video file created)

**Table 1.** Comparison of video remixing system adaptations for different operating scenarios.

video player vs. the SmartView rendering. The instantaneous customization option was a major hit with almost all the users [4].

### IMPLICATIONS OF ARCHITECTURE ADAPTATIONS

Adapting the AVRS system from the sensor-based to the sensor-less approach has impact on different aspects. In this section we discuss the implications in terms of three factors: first, video remix quality and richness of re-live experience; second, system complexity and infrastructure requirements; and finally, user density requirement. The effect on different parameters is presented in Table 1.

#### SENSOR-BASED AVRS APPROACH

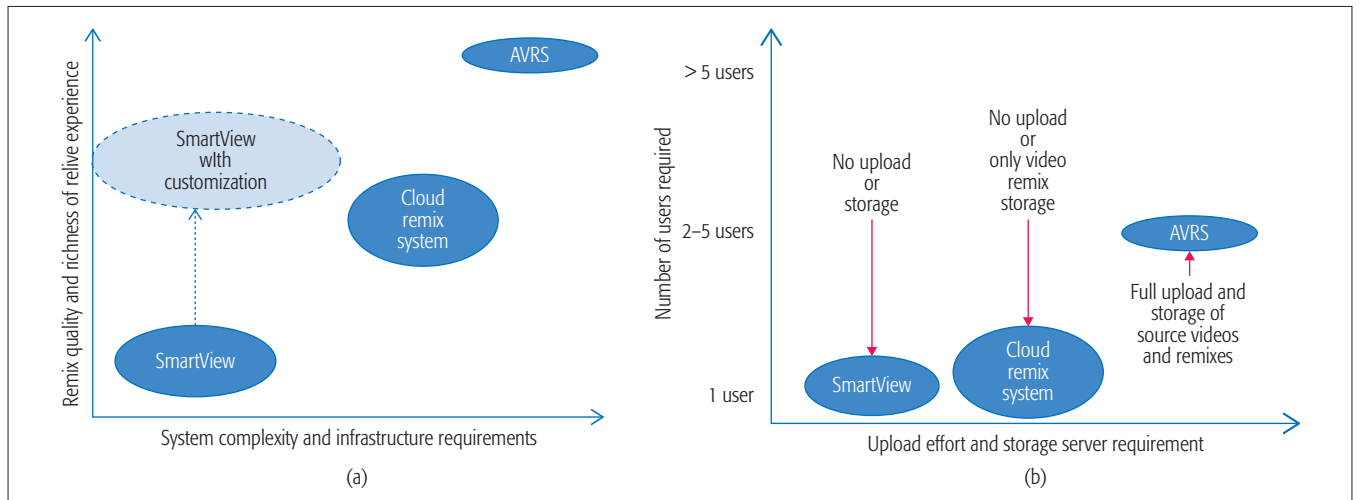
The **sensor-based AVRS** utilizes sensor augmented source media from a large number of users. This enables the video remixing process to have a higher amount of information to generate a high-quality video remix. The need for wide penetration of sensor equipped multimedia capture clients adversely affects the user density requirement. The lack of inherent support for sensor data enriched UGC media from popular SMPs inhibits widespread use due to increased system complexity and infrastructure requirement.

#### SENSOR-LESS AVRS APPROACH

The **cloud remix system** architecture, since it relies on the content from SMPs, may or may not have sensor augmented source media. This reduces the amount of semantic information available for choosing the views in the remix (e.g., device landscape/portrait orientation during recording). Such an approach removes the need for users to upload content for a specific purpose (video remix creation) and also allows use of various SMPs. The user density requirement is down to one person, since it allows leveraging the content available on various SMPs. Consequently, it is of great advantage in terms of managing costs and reducing system complexity.

The **SmartView (SV as well as MTSV)** architecture is the leanest since there are no infrastructure requirements. It achieves good user experience in focused operating scenarios (e.g., music dominated situations). It is ideal for a single user or a small group of users, since the user density requirement threshold is just one.

A comparison of remix quality and overall complexity for the sensor-based and sensor-less approaches is presented in Fig. 5a. Figure 5b illustrates the comparison between the user density requirements and the upload effort as well as the storage server requirement.



**Figure 5.** a) Comparison in terms of remix quality and overall complexity; b) comparison in terms of required number of users vs. uploading and storage.

## CONCLUSIONS

An automatic video remixing system delivers the best quality video remix when it can leverage sensor data enriched video content, although the need for recording sensor data simultaneously with audio-visual content means that it is difficult to have a minimum critical mass of persons in an event who can contribute such content. Also, there is an absence of such sensor data aware social media services. This drives the need for adaptation of the system architecture such that it can improve the desired performance parameters while limiting the reduction in the overall user experience. As described in this article, the first video remixing system, referred to as sensor-based AVRS, provides a high-quality overall user experience without imposing operating parameter constraints. The sensor-less cloud remix system removes the need to have multiple users as well as the need to upload videos specifically for making remixes, but compromises on computational efficiency as well as semantic information due to the absence of sensor augmented information. SmartView reduces the overall system complexity to an extent where no back-end infrastructure is required, making it feasible to use it on a standalone mobile device and operated by a single user. The sensor-based and sensor-less video remixing approaches presented in this article exemplify the need for prioritizing certain performance parameters in the end-to-end system design to make the system suitable for the chosen operating parameters.

## REFERENCES

- [1] S. Vihavainen *et al.*, "We Want More: Human-Computer Collaboration in Mobile Social Video Remixing of Music Concerts," *ACM SIGCHI Conf. Human Factors in Computing Systems*, Vancouver, B.C., Canada, 7–12 May 2011, pp. 287–96.
- [2] S. Vihavainen *et al.*, "Video as Memorabilia: User Needs for Collaborative Automatic Mobile Video Production," *ACM SIGCHI Conf. Human Factors in Computing Systems*, Austin, TX, 5–10 May 2012, pp. 651–54.
- [3] J. Ojala *et al.*, "Automated Creation of Mobile Video Remixes: User Trial in Three Event Contexts," *13th Int'l. Conf. Mobile and Ubiquitous Multimedia*, Melbourne, Australia, 25–28 Nov. 2014, pp. 170–79.
- [4] S. Mate *et al.*, "Automatic Multi-Camera Remix from Single Video," *30th ACM Symp. Applied Computing*, Salamanca, Spain, 13–17 Apr. 2015, pp. 1270–77.
- [5] F. Cricri *et al.*, "Sensor-Based Analysis of User Generated Video for Multi-Camera Video Remixing," *18th Int'l. Conf. Advances in Multimedia Modeling*, 2012, pp. 255–65.
- [6] F. Cricri *et al.*, "Multimodal Extraction of Events and of Information About the Recording Activity in User Generated Videos," *Multimedia Tools and Applications*, vol. 70, no. 1, 2014, pp. 119–58.
- [7] P. Shrestha *et al.*, "Automatic Mashup Generation from Multiple-Camera Concert Recordings," *18th ACM Int'l. Conf. Multimedia*, Firenze, Italy, 25–29 Oct. 2010, pp. 541–50.

- [8] M. K. Saini *et al.*, "Movimash: Online Mobile Video Mashup," *20th ACM Int'l. Conf. Multimedia*, Nara, Japan, 2012, pp. 139–48.
- [9] X. Bao and R. Choudhury, "Movi: Mobile Phone Based Video Highlights Via Collaborative Sensing," *ACM Int'l. Conf. Mobile Systems, Applications, and Services*, San Francisco, CA, 15–18 June 2010, pp. 357–70.
- [10] I. Arev *et al.*, "Automatic Editing of Footage from Multiple Social Cameras," *ACM Trans. Graphics*, vol. 33, no. 4, article 81, July 2014, pp. 1–11.
- [11] V. Zsombori *et al.*, "Automatic Generation of Video Narratives from Shared UGC," *22nd ACM Conf. Hypertext and Hypermedia*, Eindhoven, The Netherlands, 6–9 June 2011, pp. 325–34.
- [12] F. Chen and C. De Vleeschouwer, "Personalized Production of Basketball Videos from Multi-Sensored Data Under Limited Display Resolution," *Computer Vision and Image Understanding*, vol. 114, no. 6, 2010, pp. 667–80.
- [13] X. Chen, A. O. Hero, and S. Savarese, "Multimodal Video Indexing and Retrieval Using Directed Information," *IEEE Trans. Multimedia*, vol. 14, no. 1, 2012, pp. 3–16.
- [14] F. Cricri *et al.*, "Sport Type Classification of Mobile Videos," *IEEE Trans. Multimedia*, vol. 16, no. 4, Feb. 2014, pp. 917–32.
- [15] F. Cricri *et al.*, "Salient Event Detection in Basketball Mobile Videos," *IEEE Int'l. Symp. Multimedia*, Taichung, Taiwan, 10–12 Dec., 2014, pp. 63–70.
- [16] M. Roininen *et al.*, "Modeling the Timing of Cuts in Automatic Editing of Concert Videos," *Multimedia Tools and Applications*, Feb. 2016, pp. 1–25.

## BIOGRAPHIES

SUJEET MATE (sujeet.mate@nokia.com) is a senior researcher at Nokia Technologies, Tampere, Finland. He received his B.E. degree in electrical engineering from SVNIT, Surat, India, and M.S. degree in electrical engineering from the University of Texas at Dallas. He joined Nokia in 2004, where he has been active in developing applications that leverage multimodal sensing for automatic media remixing, interactive video services, and real-time applications like videoconferencing. His interests include VR audio-visual systems, multimedia applications and services architectures, end to end system prototyping, and multimodal context sensing.

IGOR D. D. CURCIO [S'91, M'03, SM'04] (igor.curcio@nokia.com) received an M.Sc. degree in computer science from the University of Catania, Italy, and a Ph.D. degree in signal processing from Tampere University of Technology, Finland. After 11 years working as a freelance consultant and trainer, in 1998 he joined Nokia Corporation, where he covered several research and management positions in areas related to mobile media. He is now a principal scientist at Nokia Technologies, and also a lecturer in video compression at Tampere University of Technology, 2012–2016. He has been active for over 10 years in several standardization organizations (e.g., 3GPP, MPEG, DLNA, IETF, ARIB) where he has also covered sub-working group and task force chair positions, and (co-)authored over 200 standardization contributions in the areas of adaptive media, VoIP, QoE, and transport protocols. He holds more than 40 granted patent families, and several patents are currently pending. He has been an ACM member since 1990 and an ACM professional member since 2012. He has published over 70 papers in the areas of mobile media applications and services. He has served on the organizing committees and TPCs of IEEE CCNC, IEEE PerCom, IEEE WoWMoM, IEEE ICCNC, IEEE PIMRC, IEEE ISCC, and ACM MUM. His current interest areas include mobile video applications and services, multimodal sensing applications, contextual media systems, crowd-sourced cloud media services, and virtual reality media services.

The sensor-based and sensor-less video remixing approaches presented in this article exemplify the need for prioritizing certain performance parameters in the end-to-end system design, to make the system suitable for the chosen operating parameters.

# The Love-Hate Relationship between IEEE 802.15.4 and RPL

Oana Iova, Fabrice Theoleyre, Thomas Watteyne, and Thomas Noel

RPL creates a routing topology without a priori knowledge about the topology created at the MAC layer. This negatively impacts the number of redundant paths, their quality, and the overall performance of the routing protocol. The authors highlight the need for an intermediate layer between MAC and network layers to solve these problems. They describe the protocols to be used in future Internet of Things, emphasize their weaknesses when deployed together, and propose areas of improvement.

## ABSTRACT

Low-power and lossy networks (LLNs) are at the core of many Internet of Things solutions. Significant standardization effort has been put in creating a protocol stack suited for LLNs. Among these standards, IEEE 802.15.4-2011 and RPL allow LLN devices to form a multi-hop mesh network. Today, RPL creates a routing topology without a priori knowledge about the topology created at the MAC layer. This negatively impacts the number of redundant paths, their quality, and the overall performance of the routing protocol. In this article, we highlight the need for an intermediate layer between the MAC and network layers to solve these problems. We describe the protocols to be used in future Internet of Things, emphasize their weaknesses when deployed together, and propose areas of improvement.

## INTRODUCTION

Miniaturization of computation and communication solutions has enabled the creation of small, durable, and inexpensive wireless devices often called “motes.” Motes can be programmed to interconnect wirelessly, and form a multi-hop low-power wireless network, known as a “low-power and lossy network” (LLN). LLNs are one of the core technologies in the Internet of Things (IoT). LLN protocols and standards need to take into account their specific constraints in terms of energy, memory, and processing power.

The IEEE and IETF, two major standards development organizations (SDOs) in the telecommunication arena, have published several standards that contribute to the creation of a fully standards-based protocol stack for LLNs. IEEE 802.15.4 [1] is arguably the standard with the most impact on low-power wireless technology. It defines both the physical layer (i.e., modulation scheme, data rate) and the medium access control (MAC) layer for low-rate wireless personal area networks (WPANs). In 2012, the IETF ROLL working group published the “IPv6 Routing Protocol for Low-Power and Lossy Networks” (RPL) [2], which enables low-power devices to form a multi-hop topology. Because of energy constraints, the focus was on single interface nodes.

While blind layer separation allows modularity, it also comes with some limitations, especially in constrained environments. An LLN is a canonical example of a constrained network: a large num-

ber of low-end and energy-constrained devices form a multi-hop mesh network using unreliable links over which small packets can be transmitted at a low data rate. In such an environment, there is great potential for cross-layer optimization, where different (theoretically independent) layers could exchange information to coordinate their actions. In some cases, a sublayer might be introduced to perform adaptation between two layers otherwise unaware of each another. One such example is 6LoWPAN. Situated above the MAC layer, it compacts (long) IPv6 headers so they fit in (short) IEEE 802.15.4 frames.

In this article, we show the shortcomings of blind layer separation in the current protocol stack for LLNs, focusing on the MAC and network layers. In short, the contributions of this article are:

- We propose to use the same topology control at the MAC and routing layers.
- We highlight the instability problem of RPL when using the current routing metrics.
- We propose to estimate link quality by exploiting all the parents at the topology created by RPL.

## A STANDARDS-BASED PROTOCOL STACK FOR LLNs

This section discusses the standards with the most impact on LLN technology.

### MAC LAYER: IEEE802.15.4-2011

The IEEE 802.15.4 standard was introduced in 2003 to be used in WPANs. Two revisions later (2006, 2011), and with one upcoming revision (2015), IEEE 802.15.4 is arguably the standard with the highest impact on low-power wireless in general, and on IoT in particular.

**Link Layer Topology:** In an IEEE 802.15.4-2011 network, the devices are managed by a controller known as the “PAN coordinator” (or “sink,” two terms that we use interchangeably in this article). The standard defines two types of network topologies, star and peer-to-peer, both illustrated in Fig. 1a.

In a star topology, all devices communicate only with the PAN coordinator, over a single hop. While devices can run on batteries, the PAN coordinator is usually mains powered, as it needs to keep its radio on at all times.

In a peer-to-peer topology, communication

is not restricted to the PAN coordinator. In contrast to a star topology, devices communicate with one another, enabling multi-hop connectivity. Multi-hop is a key feature in many IoT applications where not all nodes are deployed sufficiently close to the PAN coordinator to be in its radio range. The drawback of the peer-to-peer topology is that a node must always stay awake as it can receive a packet from a neighbor at any time. A third topology, called “cluster-tree”, can be used to overcome this. In a cluster-tree topology, a tree rooted at the PAN coordinator organizes the sleeping periods of the different router nodes to enable multi-hop communication with energy savings.

**Medium Access and Energy Efficiency:** In IEEE802.15.4-2011, accessing the medium can be done either in an asynchronous (beacon-less) or synchronous (with beacons) mode.

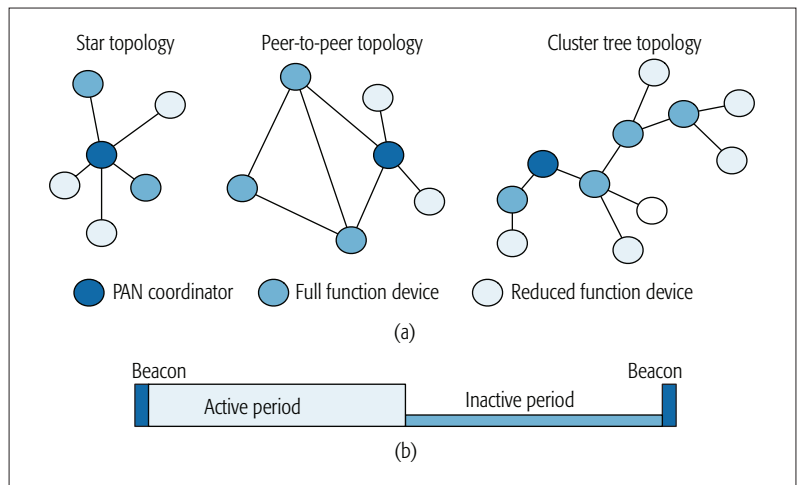
In beacon-less mode, nodes use unslotted carrier sense multiple access with collision avoidance (CSMA/CA), without exchanging request-to-send and clear-to-send (RTS/CTS) messages. In a multi-hop topology, all routing nodes must stay awake to be able to receive packets, which can be sent at any time. Preamble-sampling can reduce energy consumption: the transmitter node pre-sends a long preamble (a series of well known bytes) to its packet; a receiver periodically samples the medium and stays on when it hears an ongoing preamble. Unfortunately (besides breaking compliance with IEEE 802.15.4, which does not include it), preamble sampling puts the energy burden on the transmitter, and significantly lowers the throughput of the network.

In beacon mode, IEEE 802.15.4 cuts the time into superframes. Each superframe starts when the coordinator (possibly a router) sends a beacon. As we can see in Fig. 1b, this is followed by an active period (in which all transmitters compete using slotted CSMA/CA), and an inactive period (in which nodes sleep until the next beacon). The beacon mode saves energy in multi-hop topologies only when using the cluster-tree topology. As previously stated, in the peer-to-peer topology, nodes have to always keep their radio on.

### ROUTING OVER LLNs WITH RPL

RPL is a distance-vector routing protocol designed to scale to thousands of devices in an LLN. It organizes the topology in a destination oriented directed acyclic graph (DODAG), a directed graph with no cycle. This DODAG is rooted at the sink (or at each sink when multiple sinks are present). To build the DODAG, RPL assigns a *rank* to each mote, i.e., a virtual distance to the sink. An *objective function* defines how routing metrics (e.g., link quality, hop count) are used to compute a node’s rank. For example, if the objective is to create shortest paths, a node computes its rank by adding a scalar value to the rank of its preferred parent.

**DODAG Construction:** DODAG construction starts when the sink is switched on. It periodically broadcasts a DODAG information object (DIO), a control packet containing its rank, as well as configuration parameters. When a joining node receives a DIO, it inserts the transmitter’s address in its list of possible parents. From that list, it chooses its preferred parent as the node that advertises the smallest rank<sup>1</sup>. Once this par-



**Figure 1.** IEEE802.15.4-2011 concepts: a) topology examples; b) superframe structure of IEEE802.15.4.

ent-child relationship is established, a node forwards all packets for the sink through its preferred parent. After a node has computed its own rank (usually using the rank of its parent and link and node metrics), it starts to periodically broadcast its own DIOs.

Figure 2 illustrates this DODAG construction routine. For simplicity, we use *hop count* (the number of hops to the sink) as routing metric. The rank of a node is computed as the rank of its parent plus a constant *step value* of 1. The sink R starts broadcasting DIO messages (Fig. 2a). The neighbors of R choose it as their preferred parent, compute their rank, and start broadcasting their own DIOs (Fig. 2b). The network is fully formed when all the nodes have chosen their preferred parent (Fig. 2d).

**The Trickle Algorithm:** Even after the RPL DODAG has formed, nodes keep transmitting DIOs to update the DODAG to topological changes. Unlike IEEE 802.15.4, which sends beacons at a fixed rate, the rate at which the DIOs are being sent is tuned using the Trickle algorithm [3]. The idea is for nodes to send fewer DIOs when the topology is stable, leading to less energy consumption.

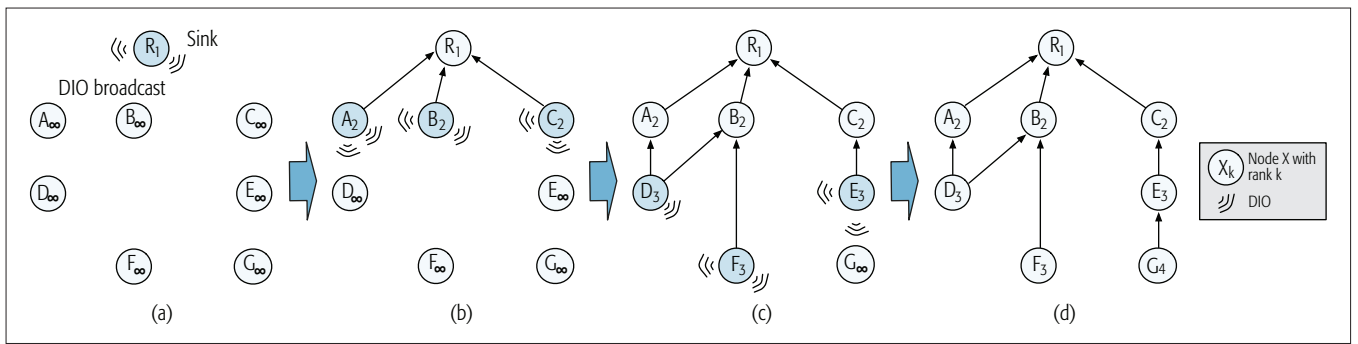
When a node receives DIO messages that contain the same information as the last ones, it doubles its own period for sending DIOs. When an inconsistency is detected (e.g., the rank of the preferred parent has changed), the Trickle algorithm resets this period to an initial value. This causes the nodes to send DIOs more frequently, and the DODAG to adapt more quickly to the change.

### GOTCHAS WHEN USING LOSSY LINKS

Wireless phenomena such as external interference and multi-path fading cause links to be unreliable. It is therefore crucial for a mote to continuously estimate the quality of the links to its neighbors, in order to choose the subset of “good” links to forward packets on. Routing metrics such as hop count are not enough, as nodes might elect a preferred parent that is close to the sink, but with which has poor connectivity.

De Cuoto *et al.* propose to use the expected transmission count (ETX) [4] as a link metric, and

<sup>1</sup> Alternatively, a node can choose its parent as the neighbor that gives it the smallest *Rank*; this takes into account the neighbor’s *Rank*, and the cost of the link between the current node and the neighbor.



**Figure 2.** DODAG construction with hop count as a routing metric. a) bootstrap: the sink starts broadcasting DIO messages; b) the neighbors of the sink choose it as the preferred parent; c) DIO propagation continues until reaching all the nodes; and d) after all the nodes chose their preferred parent, a DODAG is formed.

use only “good” links. ETX estimates the number of required transmissions needed before the neighbor correctly receives the frame. It can also be used to estimate the energy cost associated with communicating over that link.

Yet, as highlighted by Liu *et al.* [5] and Passos *et al.* [6], using ETX causes network churn (i.e., nodes changing routing parent) because of its greedy approach. That is, a node always searches for the link with the best (instantaneous) quality.

The IETF has defined several routing metrics [7] that can be used by RPL to construct the DODAG:

- Node metrics: node characteristics, hop count to the sink, and residual energy of the node.

Parameter	Value
Simulation duration	3600 s (1 hour)
Number of nodes	50 or 100 nodes deployed uniformly on a disk
Simulated area	400m <sup>2</sup> (50 nodes), 800m <sup>2</sup> (100 nodes)
Traffic model	CBR, 1 pkt/min, convergecast
Data packet size	127 bytes
RPL parameters	MinHopRankIncrease = 256
RPL Objective Function	MRHOF (for ETX) and OF0 (for hop count)
MRHOF parameter	PARENT_SWITCH_THRESHOLD = 0.5
Trickle parameters	$I_{min} = 27ms, I_{max} = 16, k = 1$
MAC protocol	IEEE802.15.4-2011
Beacon mode parameters	BO = 7, SO = 2
PHY model	Path-loss shadowing
PHY parameters	Path loss = 1.97, standard deviation = 2.0, $Pr(2m) = -61.4$ dBm
Simulation runs	10 (results are average over 10 random topologies)

**Table 1.** Simulation parameters.

- Link metrics: throughput, latency, link reliability, and link color (a semantic constraint).

Unlike RPL, IEEE 802.15.4 does not specify any metric for the construction of its cluster-tree. Cuomo *et al.* propose to select the routing nodes based on the LQI (link quality indicator) from the physical layer, or a combination of LQI and hop count [8]. It is very common for each layer (network, MAC) to use its own (routing, link) quality metric. We will show in the next section the limits of such an approach.

## EVALUATION, LIMITS AND RECOMMENDATIONS

In this section we simulate a LLN and highlight its poor performance when the MAC and routing protocols are used independently, while offering guidelines for improvement.

### METHODOLOGY

We simulate the behavior of RPL and IEEE802.15.4 on multi-hop networks in WSN, a well known network simulator for LLNs [9]. We use either the peer-to-peer topology of IEEE 802.15.4 operating in beacon-less mode, or the cluster-DAG topology from [10] for the beacon-enabled mode. As stated earlier, the peer-to-peer topology cannot be used together with the beacon mode. Table 1 lists the simulation parameters.

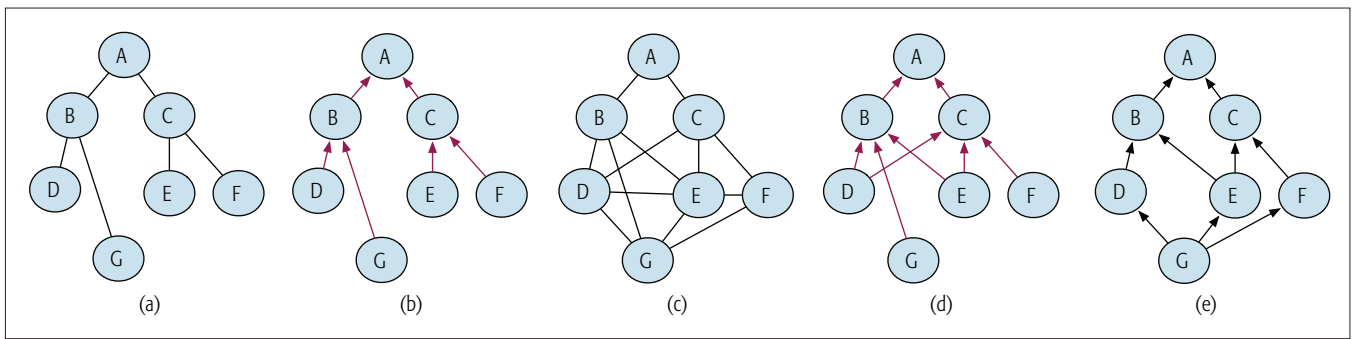
### TOPOLOGY CONTROL

We are interested in how a node chooses the neighbors to communicate with, at both the MAC layer and the routing layer. The problem is that under blind layer separation, decisions made by these layers might conflict. In this section we evaluate how this affects network performance.

**Context:** Usually, the MAC layer only filters the neighbors a node may use. However, IEEE 802.15.4 (the MAC layer) imposes a topology on the network: star, peer-to-peer, or cluster-tree. If the MAC layer structures the network as a cluster-tree (Fig. 3a), the routing layer is presented with a topology without redundancy, and has to stick to the neighbors selected by the MAC layer (Fig. 3b).

A better solution is for the MAC layer to structure the network as a peer-to-peer topology (Fig. 3c). RPL then creates a redundant DODAG (see Fig. 3d, where D and E have redundant paths to the sink). The peer-to-peer mode does not, however, implement low radio duty-cycle, so the network’s energy consumption is high.





**Figure 3.** Topology control using hop count as a metric: a) cluster-tree MAC topology; b) running RPL on top of a cluster-tree MAC topology; c) peer-to-peer MAC topology; d) running RPL on top of a peer-to-peer MAC topology; e) using a cluster-DAG topology.

To introduce redundancy at the MAC layer, while remaining energy efficient, Pavkovic *et al.* [10] propose to use a cluster-DAG in IEEE 802.15.4-2011. This allows RPL to select multiple parents (Fig. 3e), at no extra costs: the same amount of DIOs are sent regardless of the number of parents selected.

To avoid loops in the cluster-tree or the cluster-DAG, a path metric is required at the MAC layer, but none is defined in IEEE 802.15.4. Here again, the MAC layer and routing layer can be in conflict: if the MAC layer uses hop count, it creates a cluster-DAG with long and potentially bad links. Even if the routing protocol uses a different metric, it can only choose from MAC links, negatively impacting the network's reliability and energy consumption.

Having several applications run on the same network imposes further requirements. RPL can implement a DODAG instance per application, each DODAG potentially using a different routing metric. This requires the MAC layer to offer sufficient neighbor choices.

**Evaluation:** We quantify the impact of blind layer separation on topology control, through simulation.

Figure 4a shows the complementary cumulative distribution function (CCDF) of the number of routing neighbors a node has, when using both cluster-DAG and peer-to-peer MAC topologies. The peer-to-peer MAC topology gives RPL a larger choice of neighbors, hence more diversity.

Figure 4b illustrates the CCDF of the end-to-end packet delivery ratio, when using either hop count or ETX at both the MAC layer and routing layer. The network performs best when both the MAC layer and network layer use ETX, offering the largest end-to-end reliability. In this case, the MAC layer and the routing layer make consistent decisions and use the links with the smallest ETX, improving the end-to-end reliability.

**Recommendations:** We recommend that the MAC protocol does not impose a topology on the network, but only filters out bad links (e.g., links with quality below a certain threshold). It is up to the routing protocol to use the set of good links presented by the MAC layer and construct a multi-hop redundant routing topology.

### ROUTING TOPOLOGY DYNAMICS

**Context:** When a node changes its preferred parent, it resets its trickle timer, which generates more DIOs and higher energy consumption.

Changing a parent too often is not efficient. One option is to limit parent changes by reducing the number of MAC neighbors. However, this also comes with the price of limiting routing diversity.

**Evaluation:** We quantify the impact of the number of neighbors on network dynamics by simulation. We implement the minimum rank with hysteresis objective function (MRHOF) [11], in which a node changes its preferred parent only when its new rank differs significantly from the old one.

Figure 4c shows the CCDF of the average number of parent changes for a node, over a simulated hour, when both RPL and IEEE 802.15.4 use ETX. A node changes its preferred parent more frequently when using a peer-to-peer MAC topology. It offers more choices, and a small variation in the link quality estimation can result in changing the preferred parent. Figure 4c also confirms the conclusions of [12] that parent changes are more frequent in larger networks.

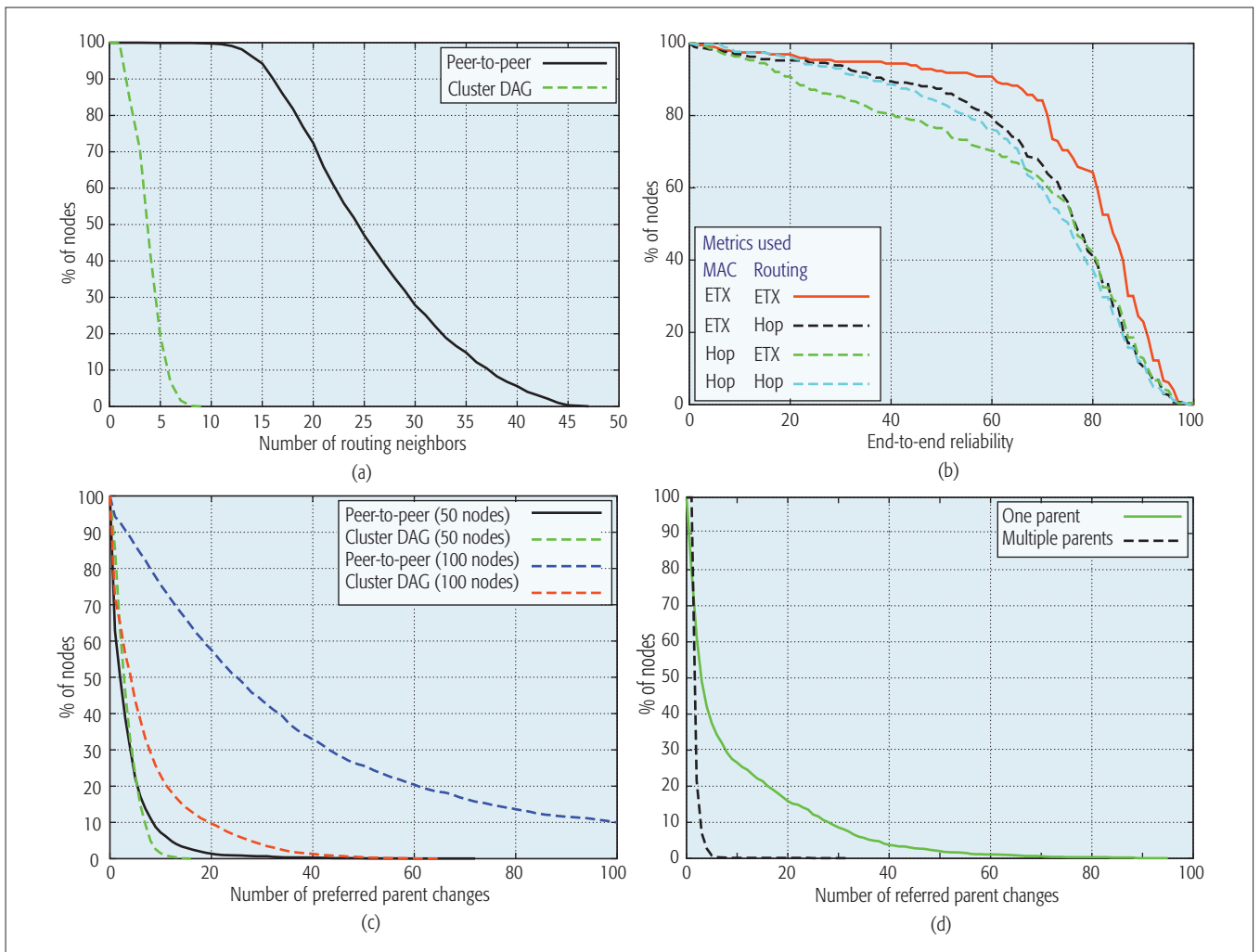
**Recommendations.** To reduce the number of RPL parent changes, we recommend the use of hysteresis when estimating link quality, such as the window mean with exponentially weighted moving average (WMEWMA). Several other techniques have been proposed in the literature, but WMEWMA offers the highest performance [13].

### ESTIMATING LINK QUALITY

**Context.** To estimate the quality of a link to a neighbor, a node can use statistics of data packets exchanged in the network. The main problem with this passive approach is that the estimation is done only for the neighbors the node communicates with. In an active approach, the node can send probe packets, at the cost of extra overhead. OpenWSN 1.9.0, Contiki 2.6, and TinyOS 2.1.2 all use the passive approach.

In RPL, a node communicates only with the preferred parent, so there is no way to estimate link quality to the other neighbors. The preferred parent can be a set of parents if those parents are equally preferred. Still, RPL does not specify the forwarding rule. The authors in [14] propose that a node send outgoing traffic to all its neighbors simultaneously. One parent will be selected opportunistically to forward the packets.

**Evaluation:** Figure 4d plots the CCDF of the number of parent changes, when using a single parent or a set of parents. Not only does the latter approach allow a node to estimate link quality to several neighbors, but it also increases the stability of the routing topology.



**Figure 4.** Simulation results: (a) A node has more routing neighbors when using a peer-to-peer rather than a cluster-DAG MAC topology (the network has 50 nodes); (b) The network has the highest end-to-end reliability when both MAC and routing layers use the same metric ETX (the topology is a cluster-DAG and has 50 nodes); (c) The difference in the number of preferred parent changes between the beacon and the non beacon mode increases with the size of the network; and (d) The number of preferred parent changes decreases when using a multipath technique for RPL (the topology is a cluster-DAG and has 50 nodes).

**Recommendations:** We propose for a node to use all parents in its parent set to route packets to the root (not just the preferred one), allowing the quality to the links to all parents to be passively monitored.

#### COMPUTING ETX

**Context:** The ETX of a link is defined as  $1/PDR$ , with PDR the packet delivery ratio of that link. The PDR is computed as the ratio between the number of acknowledgments received and the number of packets sent. Both Contiki and TinyOS compute ETX by simply counting the number of retransmissions, without taking into account packets dropped by the MAC layer (because of successive CCA failures or buffer timeout). Still, these dropped packets reflect the quality of those links.

**Evaluation:** Using the simulation setting from earlier, we observed that while only 0.05 percent of packets are dropped when using a peer-to-peer MAC topology, this ratio shockingly increases to 12 percent with the cluster-DAG. The latter can be attributed to additional contention because of the inactive periods of the beacon mode, and clearly should be taken into consideration when evaluating the quality of a link.

**Recommendations:** We recommend that the computation of ETX accounts for the packets dropped by the MAC layer, including because of successive CCA failures or buffer timeout.

#### TOMORROW'S LLN TECHNOLOGY: IEEE802.15.4E TSCH AND IETF 6TiSCH

The IEEE802.15.4e amendment was published in 2012, and introduces a radically new medium access control technique: time slotted channel hopping (TSCH). In a TSCH network, nodes are tightly synchronized, and time is cut into time slots. Slots are grouped in a slotframe, which continuously repeats over time. As depicted in Fig. 5, a slot is long enough (typically 10ms) for a node to send a packet to its neighbor, and for that neighbor to indicate successful reception through a link-layer acknowledgment (ACK).

Communication is orchestrated by a schedule that indicates to each node what to do in each slot: transmit, receive, or sleep. "Scheduling" a network corresponds to populating the slotframe with communication slots. Figure 5 shows a canonical example schedule for the depicted

topology. When E needs to communicate data to A, it sends the data packet to C at slot offset 1, on channel offset 2. C acknowledges successful reception (causing E to clear the data from its transmit buffer), after which C sends to A at slot offset 4, channel offset 3. The slotframe repeats continuously, giving the nodes repeated opportunities to communicate. Figure 5 is simplified to be easily explained: real-world slotframes are 10's to 1000's slots long, with typically 16 channel offsets (when using the IEEE802.15.4 physical layer at 2.4GHz).

There is a subtle but important difference between channel offset and frequency. The schedule indicates the former; the channel offset is translated on-the-fly into a frequency through a pseudo-random hopping pattern each time the device turns its radio on. This means that in successive slotframe iterations, the same channel offset translates into different frequencies. The result is "channel hopping": when two nodes communicate, successive retransmissions happen at different frequencies, thereby combating external interference and multi-path fading. The authors in [15] highlight the effectiveness of channel hopping in IEEE802.15.4 networks.

TSCH allows the network to be abstracted by its communication schedule. This schedule must be built to match link-layer resources (the cells) to the requirements of the applications running on the network. This allows a clean trade-off between throughput, latency, and energy consumption. IEEE 802.15.4e does not define how to build or maintain the TSCH schedule. Hence, a "standardization gap" exists between the IEEE 802.15.4e link-layer standard and upper layer standards such as 6LoWPAN, as neither define the entity responsible for building and managing the TSCH schedule.

The IETF 6TiSCH standardization working group was created in 2013 to fill this standardization gap by defining mechanisms to manage the TSCH communication schedule. 6TiSCH defines the 6top sublayer, which operates at layer 2.5, between IEEE 802.15.4e and 6LoWPAN. 6top offers a management interface (detailed below), and gathers statistics about each communication cell. Statistics include the number of transmitted frames in that cell, and the portion of those frames that were acknowledged.

6top supports centralized and distributed scheduling. In a centralized approach, the 6top sublayer of each node implements a CoAP-based management interface. This allows a central scheduling entity (called a path computation element (PCE)) sitting outside of the network to gather information about the topology of the network, compute an appropriate schedule, and configure each node with the cells of the schedule it participates in (using the CoAP application-level protocol). When using the distributed approach, no PCE is present in the network, and nodes need to agree on the schedule to use in a distributed fashion. The 6top sublayer implements a management interface, allowing two neighbor nodes to negotiate adding/removing cells to one another. Communication happens through "information elements" in the IEEE 802.15.4e header, fields that can serve as containers for a layer 2.5 protocol. In this distributed approach, a node monitors

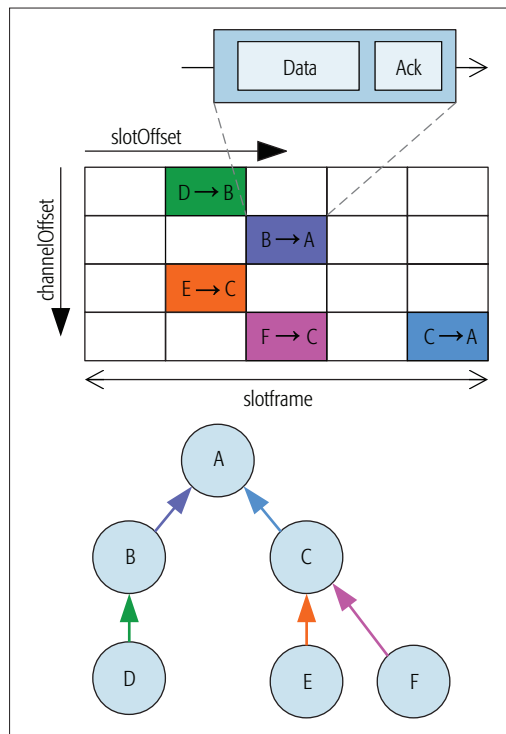


Figure 5. IEEE802.15.4e TSCH network example.

the transmission queue to each of its neighbors: if the queue overflows (resp. underflows), the node contacts its neighbor to negotiate to add (resp. remove) cells.

6TiSCH defines the mechanisms that support both centralized and distributed schedule management (packet formats and typical interaction). As a standardization entity, IETF 6TiSCH does not define the policy (when to use centralized/distributed, and the scheduling algorithm). Identifying the scheduling approach to adopt, and the associated limits, is an open research problem. Intuitively, a centralized approach can compute near-optimal schedules, provided it has up-to-date information about the network topology and the needs of the applications running in the network. A distributed approach might be preferred when the topology is highly dynamic (e.g., a swarm of mobile robots), or when a PCE cannot be installed (e.g., a very simple home network where a PCE is not cost-efficient).

This article highlights several issues that are closely related to the current standardization work at IETF 6TiSCH.

**Topology Control:** The choice of the topology at the MAC layer impacts the diversity of the routes. 6TiSCH proposes to use shared slots for exchanging broadcast control packets such as DIOs. This allows RPL to use any neighbor as a parent; 6top is then in charge of negotiating dedicated slots between the node and its parent. This is directly in line with the recommendations made earlier. What is missing in the 6TiSCH solution is a mechanism to modify the schedule on-the-fly in a distributed fashion.

**Using Different Metrics:** The 6top sublayer gathers cell statistics, which RPL uses to select the best routes. However, link metrics are needed to aggregate these statistics, some of which are TSCH-specific (e.g., per-frequency transmis-

Intuitively, a centralized approach can compute near-optimal schedules, provided it has up-to-date information about the network topology and the needs of the applications running in the network. A distributed approach might be preferred when the topology is highly dynamic or when a PCE cannot be installed.

The resulting architecture, which combines the performance of IEEE802.15.4e TSCH with the IPv6-based upper stack, has the potential to revolutionize LLN technology. With the work already done, IETF 6TiSCH is starting to do so, but several challenges remain open.

sion counters). One option is to develop a unified link metric that encompasses both MAC and routing metrics. This remains an open research problem, especially when several applications run on the same network and RPL has to implement a DODAG instance per application.

**Routing Topology Dynamics:** TSCH uses channel hopping to make links more reliable, and the connectivity in the network more stable. However, capturing the variations over time, while not overreacting to inaccurate estimators, remains an open challenge.

**Link Quality Estimation (ETX) with 6TiSCH:** 6top maintains per-cell statistics (including the number of packets sent and acknowledged). The ETX between two neighbor nodes is calculated by aggregating the statistics from the different cells scheduled between those nodes. What is missing is a way to discover neighbors that a node is not communicating with, and estimate the quality of the link to that neighbor.

## CONCLUSION

Network performance depends not only on the MAC and routing protocols used, but also on their interaction. This article highlights the interaction between the topology defined at the MAC layer and the decision made by the routing protocol. Through simulation, we show that a peer-to-peer MAC topology yields higher performance compared to a cluster-DAG topology, as it presents more neighbors to the RPL routing protocol. However, having more neighbors does mean that RPL might change a node's preferred parent more often (especially with large networks). Still, this article shows how using a set of parents and passive link quality estimation reduces network churn.

These observations are being addressed by the new IETF 6TiSCH working group, which defines 6top, a sublayer between the link-layer (IEEE 802.15.4e) and networking/routing layers (RPL). The resulting architecture, which combines the performance of IEEE 802.15.4e TSCH with the IPv6-based upper stack, has the potential to revolutionize LLN technology. With the work already done, IETF 6TiSCH is starting to do so, but several challenges remain open, including those highlighted in this article, i.e., an on-the-fly distributed reservation mechanism, and a link metric that combines MAC and routing statistics.

## REFERENCES

- [1] 802.15.4-2011: IEEE Standard for Local and Metropolitan Area Networks. Part 15.4: Low-Rate Wireless Personal Area Networks (LR-WPANs), June 2011.
- [2] T. Winter *et al.*, "RPL: IPv6 Routing Protocol for Low-Power and Lossy Networks," IETF RFC 6550, March 2012.
- [3] P. Levis *et al.*, "The Trickle Algorithm," IETF RFC 6206, March 2011.

- [4] D. De Couto *et al.*, "A High-Throughput Path Metric for Multi-hop Wireless Routing," *ACM MobiCom*, Sept. 2003.
- [5] T. Liu *et al.*, "Performance Evaluation of Link Quality Estimation Metrics for Static Multihop Wireless Sensor Networks," *IEEE SECON*, June 2009.
- [6] D. Passos *et al.*, "Mesh Network Performance Measurements," *Int'l. Information and Telecommunications Technologies Symp.*, 2006.
- [7] J. P. Vasseur *et al.*, "Routing Metrics Used for Path Calculation in Low-Power and Lossy Networks," IETF RFC 6551, March 2012.
- [8] F. Cuomo, E. Cipollone, and A. Abbagnale, "Performance Analysis of IEEE 802.15.4 Wireless Sensor Networks: An Insight into the Topology Formation Process," *Elsevier Computer Networks*, vol. 53, no. 18, Dec. 2009, pp. 3057–75.
- [9] E. Ben Hamida, G. Chelius, and J.-M. Gorce, "On the Complexity of an Accurate and Precise Performance Evaluation of Wireless Networks using Simulations," *ACM-IEEE MSWiM*, Oct. 2008.
- [10] B. Pavkovic, F. Theoleyre, and A. Duda, "IEEE 802.15.4 and RPL Cross-optimization for Reliable Opportunistic Routing in WSN," *ACM MSWiM*, Oct. 2011.
- [11] O. Gnawali and P. Levis, "The Minimum Rank with Hysteresis Objective Function," IETF RFC 6719, Sept. 2012.
- [12] O. Iova, F. Theoleyre, and T. Noel, "Stability and Efficiency of RPL under Realistic Conditions in Wireless Sensor Networks," *IEEE PIMRC*, Sept. 2013.
- [13] A. Woo, T. Tong, and D. Culler, "Taming the Underlying Challenges of Reliable Multihop Routing in Sensor Networks," *ACM SenSys*, Nov. 2003.
- [14] Q. Lampin *et al.*, "Exploiting Long-Range Opportunistic Links to Improve Delivery, Delay and Energy Consumption in Wireless Sensor Networks," *IEEE MASS*, Oct. 2012.
- [15] T. Watteyne, A. Mehta, and K. Pister, "Reliability Through Frequency Diversity: Why Channel Hopping Makes Sense," *ACM PE-WASUN*, Oct. 2009.

## BIOGRAPHIES

OANA IOVA (oanatedora.iova@unitn.it) is a post-doctoral researcher at the University of Trento, where she works in the Department of Information Engineering and Computer Science (DIS). Her research interests include routing solutions for low power and lossy networks, and MAC protocols for multihop wireless networks. She received her Ph.D. from the University of Strasbourg, France in 2014, and her M.Sc. in computer science from Ecole Normale Supérieure de Lyon, France in 2011.

FABRICE THEOLEYRE (theoleyre@unistra.fr) is a senior research scientist in the CNRS, and is now affiliated with ICUBE/University of Strasbourg. He received his Ph.D. in computer science from INSA, Lyon (France) in 2006. He was a visiting researcher at INRIA Sophia-Antipolis in 2005 and a visiting scholar at the University of Waterloo in 2006. He has been an associate editor for *IEEE Communications Letters* and a guest editor for *Computer Communications* and *Eurasip JWCN*.

THOMAS WATTEYNE (thomas.watteyne@inria.fr) is a researcher at Inria-Paris, on the EVA team, where he builds the Industrial Internet of Things. He is a senior networking design engineer at Linear Technology/Dust Networks. He co-chairs the IETF 6TiSCH WG. He completed his postdoctoral research at UC Berkeley, with Prof. Pister. He co-leads the OpenWSN project. From 2005 to 2008 he was research engineer at Orange Labs. He holds a Ph.D. (2008), M.Sc. (2005), and M.Eng. (2005) in telecommunications, from INSA Lyon, France.

THOMAS NOEL (noel@unistra.fr) is a professor at the University of Strasbourg, France. His research activities include several aspects of wireless communications networks and telecommunications systems. He is particularly interested in network mobility, self-organized mobile networks, mobile network architecture and protocols, wireless sensor networks, ubiquitous computing, and multicast and group communications.

CALL FOR PAPERS  
*IEEE COMMUNICATIONS MAGAZINE*

## BEHAVIOR RECOGNITION BASED ON WI-FI CHANNEL STATE INFORMATION (CSI)

### BACKGROUND

Human behavior recognition is the core technology that enables a wide variety of human-machine systems and applications, e.g., health care, smart homes, and fitness tracking. Traditional approaches mainly use cameras, radars, or wearable sensors. However, all these approaches have certain disadvantages. For example, camera-based approaches have the limitations of requiring line of sight with enough lighting and potentially breaching human privacy. Low-cost radar-based solutions have limited operation range of just tens of centimeters. Wearable sensor-based approaches are attracting increasing attention. The rationale is that different human behaviors introduce different multi-path distortions in Wi-Fi CSI. Compared with traditional approaches, the key advantages of Wi-Fi CSI-based approaches are that they do not require lighting, provide better coverage as they can operate through walls, preserve user privacy, and do not require users to carry any devices as they rely on the Wi-Fi signals reflected by humans. As a result, the recognition of quite a number of behaviors that are difficult based on traditional approaches have now become possible, e.g., fine-grained movements (e.g., gesture and lip language), keystrokes, drawings, gait patterns, vital signals (e.g., breathing rate and heart rate), etc. However, Wi-Fi CSI-based behavior recognition still faces a number of challenges: How to build the CSI-behavior model and algorithms that are robust for different humans? How to overcome the impact of noise and ensure the performance of CSI-enabled systems? How to simultaneously recognize the behavior of multiple users? How the CSI-enabled system can adapt and evolve according to the environment change?

This FT provides the opportunity for researchers and product developers to review and discuss the state-of-the-art and trends of Wi-Fi CSI-based behavior recognition techniques and systems.

In the light of the above, the main goals of this FT are threefold:

- To promote unparalleled / first-time approaches and techniques in signal processing, feature extraction, data mining and model construction for behavior recognition based on Wi-Fi CSI.
- To identify open issues which remain a challenge towards the convergence of computation theories and technologies for behavior recognition based on Wi-Fi CSI.
- To exploit novel application areas and demonstrate the benefits of Wi-Fi CSI in contrast with more traditional sensing approaches.

Topics may include (but are not limited to):

- Behavior Recognition Model/Theory based on Wi-Fi CSI
- Behavior Recognition Algorithms based on Wi-Fi CSI
- Wi-Fi CSI Signal Processing for Behavior Recognition
- Wi-Fi CSI Data Mining for Behavior Recognition
- Novel Behavior Recognition Applications/Systems Supported by Wi-Fi CSI
- Evaluation Metrics and Empirical Studies of Wi-Fi CSI enabled Systems
- Quality-enhanced and adaptive sensing models with Wi-Fi CSI

### SUBMISSIONS

Papers must be tailored to the problems of Wi-Fi CSI-enabled behavior recognition and explicitly consider the above issues. The Guest Editors reserve the right to reject papers they deem to be out of scope of this FT. Only originally unpublished contributions and invited articles will be considered for this FT.

Articles should be tutorial in nature, with the intended audience being all members of the communications technology community. They should be written in a style comprehensible to readers outside the specialty of the article. Mathematical equations should not be used (in justified cases up to three simple equations are allowed). Articles should not exceed 4500 words. Figures and tables should be limited to a combined total of six. The number of references is recommended to not 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 manuscript 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 "October 2017/Behavior Recognition Based on Wi-Fi CSI" as the Feature Topic category for your submission.

### IMPORTANT DATES

- Manuscript Submission: February 1, 2017
- Decision Notification: June 1, 2017
- Final Manuscript Due Date: July 15, 2017
- Publication Date: October 2017

### GUEST EDITORS

Bin Guo (Corresponding Guest Editor)  
Northwestern Polytechnical Univ., China  
[guobin.keio@gmail.com](mailto:guobin.keio@gmail.com)

Jennifer Chen  
Stevens Institute of Technology, USA  
[yingying.chen@stevens.edu](mailto:yingying.chen@stevens.edu)

Nic Lane  
Bell Labs and University College London, UK  
[niclane@acm.org](mailto:niclane@acm.org)

Yunxin Liu  
Microsoft Research Asia, China  
[yunxin.liu@microsoft.com](mailto:yunxin.liu@microsoft.com)

Zhiwen Yu  
Northwestern Polytechnical University, China  
[zhiweny@gmail.com](mailto:zhiweny@gmail.com)

## NETWORK AND SERVICE MANAGEMENT



George Pavlou



Jürgen Schönwälder

This is the 22nd issue of the Series on Network and Service Management, which is typically published twice a year, in January and July. The Series provides articles on the latest developments, highlighting recent research achievements and providing insight into both theoretical and practical issues related to the evolution of the network and service management discipline from different perspectives. The Series provides a forum for the publication of both academic and industrial research, addressing the state of the art, theory, and practice in network and service management.

The most recent notable event of the network and service management community was the Conference on Network and Service Management (CNSM 2016), which took place October 31–November 4 in Montreal, Canada. During CNSM 2016, the technical program of the first key event in 2017, the International Symposium on Integrated Network Management (IM 2017), was selected. IM 2017 will take place in Lisbon, Portugal, May 8–12, 2017, and focus on integrated management in the cloud and 5G era. Another important forthcoming event is the Conference on Network Softwarization (NetSoft 2017), which will take place in Bologna, Italy, on July 3–7.

The network and service management community has been working on a taxonomy for the network and service management field. Such a taxonomy can, for example, help tools such as conference management systems to quickly find experts on certain topics. The taxonomy discussions have taken place in the Technical Committee on Network Operation and Management (CNOM) of the IEEE, the IFIP Management of Networks and Distributed Systems Working Group (WG 6.6), and the Network Management Research Group (NMRG) and the Internet Research Task Force (IRTF). The result has been published as an article on Wikipedia: Network and service management taxonomy

We again experienced excellent interest for the 22nd issue with 18 submissions in total. For all submissions in the scope of our Series, we obtained at least three indepen-

dent reviews. We finally selected two articles, resulting in an acceptance rate of 11.1 percent. It should be noted that some additional submissions are currently being revised and may appear in the July issue. The acceptance rate of all previous issues has ranged between 14 and 25 percent, making this series a highly competitive place to publish.

The first article, “Increasing DNS Security and Stability through a Control Plane for Top-level Domain Operators” by Hesselman, Moura, Schmidt, and Toet, presents the authors’ efforts to build a control plane for DNS operators, which increases the security and stability of the DNS services provided.

The second article “Service Provider DevOps” by John, Marchetto, Németh, Sköldström, Steinert, Meirosu, Papafili, and Pentikousis, investigates how development practices that aim to bridge the gap between development and operations, often called DevOps, can be applied to the operational complexity of carrier-grade software-defined service provider infrastructures.

We hope that readers of this issue find the articles informative, and we will endeavor to continue with similar issues in the future. We would finally like to thank all the authors who submitted articles to this Series and the reviewers for their valuable feedback and comments on the articles.

## BIOGRAPHIES

GEORGE PAVLOU (g.pavlou@ucl.ac.uk) is a professor of communication networks in the Department of Electronic Engineering, University College London, United Kingdom, where he coordinates networks and services research activities. His research interests focus on networking and network management, including traffic engineering, autonomic networking, information-centric networking, and software-defined networks. He has been instrumental in a number of research projects that produced significant results with real-world uptake, and has contributed to standardization activities in ISO, ITU-T, and the IETF.

JÜRGEN SCHÖNWÄLDER (j.schoenwaelder@jacobs-universiy.de) is a professor of computer science at Jacobs University Bremen, Germany. His research interests include network management and measurement, network security, embedded systems, and distributed data processing. He is an active member of the IETF, where he has edited more than 30 network management related specifications and standards. He has contributed in various roles to the organization of IEEE and IFIP sponsored academic conferences and journals.

# Increasing DNS Security and Stability through a Control Plane for Top-Level Domain Operators

Cristian Hesselman, Giovane C. M. Moura, Ricardo de Oliveira Schmidt, and Cees Toet

## ABSTRACT

We present a control plane for operators of top-level domains (TLDs) in the DNS, such as “.org” and “.nl,” that enables them to increase the security and stability of their TLD by taking on the role of a threat intelligence provider. Our control plane is a novel system that extends a TLD operator’s traditional services and detects potential threats in the TLD by continuously analyzing the TLD operator’s two key datasets: the typically large amounts of DNS traffic that it handles and its database of registered domain names. The control plane shares information on discovered threats with other players in the TLD’s ecosystem and can also use it to dynamically scale the TLD operator’s DNS infrastructure. The control plane builds on a set of open source modules that we have developed on top of a Hadoop-based data storage cluster. These enable, for example, TLD operators to run and develop threat detectors and to easily import their DNS traffic into the control plane. Our control plane uses policies to protect the privacy of TLD users and is based on our operational experience of running .nl TLD (Netherlands), which we are also using as the use case for our implementation.

## INTRODUCTION

Since their inception, domain names have been used as a simple identification label for hosts, services, applications, and networks on the Internet (RFC 1034). Until the mid-1980s, the mappings from domain names to IP addresses were distributed as text files (HOSTS.TXT) via ftp to the relatively small number of hosts that were connected to the Internet at that time. The Domain Name System (DNS) (RFC 1034) replaced this mechanism to provide domain name to IP address mappings in a scalable way and has become a critical part of the Internet infrastructure.

The DNS uses a hierarchical namespace and a tree-like structure in which each level uses so-called authoritative name servers to provide pointers to the next lower level. As an example, consider a user trying to reach the website www.example.nl (Fig. 1). The user’s computer first connects to a resolver, which is a recursive name server that interacts with authoritative name servers on behalf of the user and is usually located in

the network of the user’s Internet access provider. The resolver obtains a reference to the “.nl” namespace from the root name servers, then a reference to “example.nl” from the .nl name servers, and finally the reference to “www.example.nl” from the name server of example.nl. This last name server knows the requested IP address, which the resolver returns to the user, allowing its browser to reach www.example.nl.

The second level of the DNS namespace currently contains over 1300 top-level domains (TLDs), classified into country code TLDs (e.g., “.nl” and “.br”), generic TLDs (e.g., “.com” and “.org”), and new generic TLDs such as “.amsterdam” and “.shop.” The operators of these TLDs manage the TLD’s authoritative name servers and the database of all registered second-level domain names (usually of the form [domain].[tld]). They regularly export the database contents to a so-called zone file, which is the input for the TLD’s authoritative DNS servers. The other levels in the DNS tree follow this same principle, as Fig. 1 illustrates.

A recent development is that some TLD operators have extended their traditional role as DNS operator to also take on the role of threat intelligence provider. They leverage the updates of their domain name database and the DNS traffic they handle on their name servers to detect potential threats in their TLD, such as phishing sites [1], distributed denial of service (DDoS) attacks on the DNS [2, 3], and sites that distribute malware. The underlying rationale is to protect the TLD’s users by making this threat information available to other players in the TLD, such as hosting and access providers, thus helping them to better fight these threats (collaborative security).

The contribution of our work is that we have developed and implemented a so-called control plane that enables TLD operators to become threat intelligence providers. The control plane is a novel system that extends a TLD operator’s traditional services (registration and DNS) to automatically derive potential threats from DNS traffic, database updates, and potentially other sources. Our control plane makes this threat information available to other players in the TLD and can also use it to dynamically scale the TLD operator’s DNS services. Together, these two functions increase the level of automation of operating a

The authors present a control plane for operators of top-level domains (TLDs) in the DNS, such as “.org” and “.nl,” that enables them to increase the security and stability of their TLD by taking on the role of a threat intelligence provider.

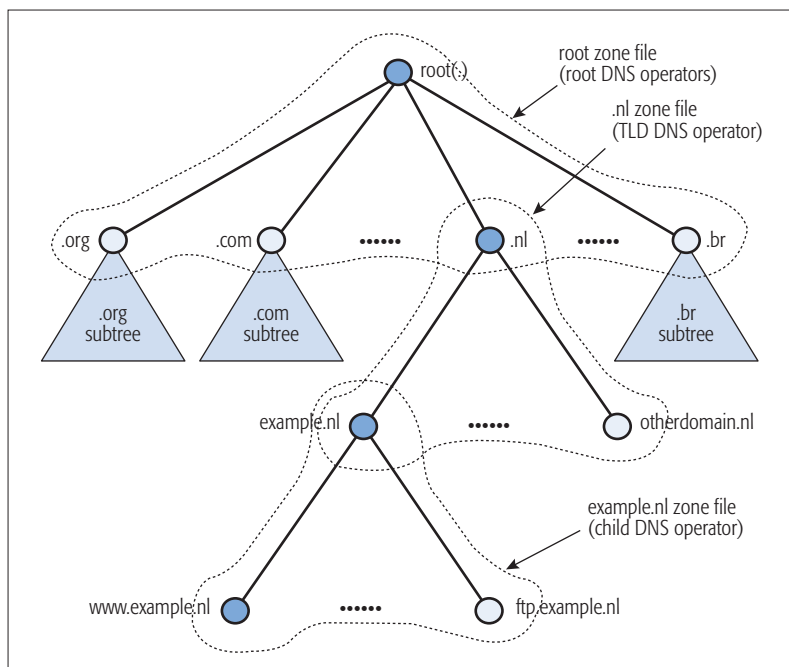


Figure 1. DNS naming hierarchy and DNS operators.

TLD because threat detection and DNS reconfiguration are mostly manual and ad hoc tasks today.

Our control plane builds on several open source modules we have developed on top of a Hadoop-based data storage cluster. For instance, they enable TLD operators to detect phishing sites and to easily import their DNS traffic into the control plane. Our modules are currently being used by at least six TLD operators, including .ca (Canada) and .at (Austria). Our control plane uses policies to protect the privacy of TLD users and is based on our operational experience of running the .nl TLD (7th largest TLD, 5.6 million domain names). We are also using .nl as the use case for our implementation.

In this article, we focus on the design and principles of our control plane and refer the interested reader to our previous work for more technical details and extensive analysis.

We first provide an overview of infrastructure that a TLD operator typically manages. Next, we discuss the threats to which TLDs are exposed, the functions our control plane needs to mitigate them, and how we realized the control plane. We end with a discussion on related work, conclusions, and future work.

## TLD OPERATOR INFRASTRUCTURE

A TLD operator traditionally manages the set of authoritative name servers for the TLD and the TLD's registration database.

### AUTHORITATIVE NAME SERVERS

Because TLD operators form the second highest level in the DNS naming hierarchy (Fig. 1), they typically use multiple layers of redundancy to provide their DNS services in a fault-tolerant way. For example, they replicate their name servers across multiple DNS services (e.g., ns1.dns.nl and ns2.dns.nl for the .nl TLD), use multiple types of name server software, and use IP anycast [3] to replicate their DNS services across sites. The

advantage of IP anycast is that it also enables TLD operators to scale their DNS capacity to deal with an increasing DNS load and to reduce response times by placing machines closer to end users. IP anycast relies on the Internet's inter-domain routing protocol (Border Gateway Protocol, BGP, RFC 4271) to route clients to the closest name server and is heavily used by the DNS root (11 of its 13 "letters" use anycast across more than 500 different locations [3]).

As an example, the DNS infrastructure for the .nl TLD consists of six unicast name servers and two anycast services. The anycast service is distributed across several dozens of sites, with one anycast service mostly co-located with large Dutch access providers ("local anycast") and the others worldwide (through third parties). We use several different types of name server software for reasons of diversity, and changes to our infrastructure go through a tightly controlled change management process.

Four of our six unicast name servers together handle around 850 million DNS queries a day coming from approximately 1.3 million resolvers.<sup>1</sup> This is a subset of the total amount of queries because resolvers use local caches to avoid having to completely walk the DNS tree for every lookup. This increases performance and DNS scalability, but implies that authoritative servers only receive part of the queries that a resolver receives from clients.

### REGISTRATION DATABASE

A TLD operator's registration database usually contains all the second-level domain names in a TLD, which are of the form [domain].[tld] (some TLD operators also allow for thirdlevel registrations, e.g., under .com.br). The TLD operator typically enables so-called registrars to register a domain name (or update or delete it) in the database on behalf of Internet users, which are called registrants. A registration corresponds to adding a leaf under a TLD in the DNS tree (Fig. 1).

Different registrars provide different registration interfaces, but the registrar-registry interface is often based on the Extensible Provisioning Protocol (EPP, RFC 5730). Registrars typically sell domain names in combination with hosting services.

As an example, the .nl registration database is synchronized across multiple sites, contains 5.6 million domains, and serves around 1500 domestic and international registrars. We offer both an EPP and a web-based interface, and generate and export the .nl zone file to our name servers every hour.

## THREATS

The DNS and the domain names in a TLD are exposed to various threats. Some affect the services of a TLD operator, others those of other players within the TLD. We distinguish four types of threats in this article and refer to RFC 3833 for a more detailed description of DNS-related threats.

Zone file integrity violation: These threats involve compromising the TLD zone file (cf. [16]), for instance, by stealing users' or registrar credentials, allowing the attacker to change certain records in the zone file. This leads the authorita-

<sup>1</sup> <http://stats.sidnlabs.nl>



tive server to respond to queries with fraudulent answers, ultimately pointing the user to a malicious domain name.

**Name server unavailability:** This type of threat purposely reduces the availability of name servers in the DNS, for instance, through a DDoS attack [2–4]. This results in name servers becoming unavailable or unstable (partial availability), which means that clients do not receive a response to their DNS request (in time) and are unable to reach the intended server.

**DNS response integrity violation:** Bad actors tamper with DNS responses, for instance, through man-in-the-middle attacks, DNS hijacking, or cache poisoning (RFC 3833). This results in a user being redirected to a malicious or unsolicited server. DNSSEC (RFC4035) detects this type of attack at the resolver.

**Abuse:** The DNS is being exposed to various sorts of abuse, such as phishing, malware distribution, and command-and-control botnet channels. While the malicious content is hosted outside the DNS, the DNS is misused to direct victims to such sites.

## DATA AND FUNCTIONS

The goal of our control plane is to leverage the data that a TLD operator handles to detect potential threats in the TLD and to automatically reconfigure the TLD operator's name servers. The analysis of the TLD operator's data requires a third function, which is privacy protection.

### TLD OPERATOR DATA

A TLD operator has two key datasets that it can use to detect threats: DNS authoritative traffic (incoming DNS queries for domains in the TLD's zone) and the TLD's domain registration database. The latter furthermore gives a TLD operator a real-time view of domain registration changes (creates, deletes, updates) across different registrars.

TLD operators can use these datasets to automatically detect patterns and suspicious behaviors in their zone. For example, the TLD operator would be able to detect spam campaigns based on bulk registrations, which has been reported on in [14]. It would also be able to detect phishing attacks based on unusual DNS traffic patterns for a domain that has just been registered (discussed later). TLD operators could furthermore cautiously carry out active measurements on all domain names in their zones and use this information to augment the threat detection logic.

While resolvers and DNS operators at lower levels in the DNS hierarchy would be able to carry out a similar analysis, they miss the real-time centralized view that a TLD operator has as a result of its position at the second-highest level in the DNS (Fig. 1). This makes it difficult for them to detect and correlate malicious domain names created through different registrars, such as the automatically generated domain names that botnets use.

The limitation of a TLD operator's data is that it provides a "sampled" view of the DNS because resolvers cache queries [15]. Also, TLD operators are likely to gradually receive less DNS information because of QNAME minimization (RFC 7816), which is a recent DNS extension that reduces the amount of data in DNS queries to protect the privacy of users. QNAME minimiza-

tion resolvers only put example.nl in the queries they send to TLD operators instead of www.example.nl, which is the fully qualified domain name (FQDN). The uptake of QNAME minimization is currently limited.

### THREAT DETECTION

The purpose of threat detection is to automatically detect potential threats in a TLD, such as phishing domains and unavailability of DNS name servers. To accomplish this, the control plane needs to be able to quickly analyze large datasets covering a year or more of relatively high-volume DNS data. Speed is crucial to quickly detect and mitigate threats such as the appearance of phishing sites, which will affect fewer victims the sooner they are removed.

To accomplish this, the control plane needs to provide near-real-time response times when analyzing a TLD operator's datasets, and needs to continuously store large volumes of DNS and other data. Data streaming warehouses (DSWs) [5] are designed with this in mind: they continuously digest incoming data, and use optimized file formats (columnar storage) and parallel processing to achieve near-real-time response times. DSWs can also easily be extended with extra nodes, enabling the control plane to increase its capacity when the TLD operator's datasets grow. DSWs typically also provide an easy interface for data analysis, which eases application development and interaction with a human operator.

Our control plane's DSW needs to be able to obtain the transport and IP-level information in DNS packets, which might be relevant, for example, to detect reflection attacks based on ICMP messages. The DSW should also introduce limited changes on the TLD operator's name servers. This is essential because TLD-level name servers are high availability resources that are typically tightly managed. The format for importing DNS packets from name servers into the control plane should furthermore be widely used so that different TLD operators can easily implement our control plane irrespective of their particular name server setup (as discussed earlier).

We discuss our DSW later, and our threat detection modules and their performance after that.

### ON-DEMAND DNS RECONFIGURATION

The purpose of on-demand DNS reconfiguration is to dynamically adapt the DNS anycast infrastructure of a TLD operator, for instance, to handle a DDoS threat (name server unavailability) as it occurs. TLD operators frequently use IP anycast because of its ability to handle stress situations [3] and because it allows them to easily scale their authoritative name server infrastructure.

By adapting we mean starting and stopping anycasted and virtualized DNS name servers at specific (external) hosting platforms [6]. The result is that our control plane manages a potentially large set of DNS name servers that grows and shrinks dynamically over time, which is unlike today's static and relatively small DNS anycast networks. A precondition is that the control plane is able to interface with the TLD operator's name servers so that it can send reconfiguration commands to them.

A TLD operator has two key datasets that it can use to detect threats: DNS authoritative traffic (incoming DNS queries for domains in the TLD's zone) and the TLD's domain registration database. The latter furthermore gives a TLD operator a real-time view on domain registration changes (creates, deletes, updates) across different registrars.

Privacy protection is an important function because the DNS traffic that the control plane analyzes for threat detection and DNS reconfiguration contains IP addresses of resolvers and domain names being looked up, which may constitute Personally Identifiable Information (PII), depending on the jurisdiction.

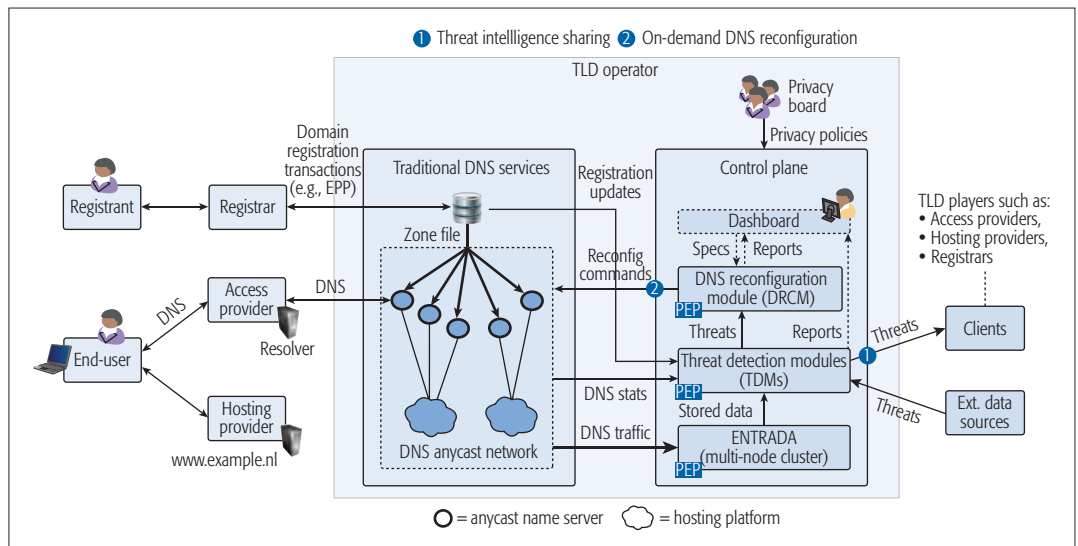


Figure 2. A TLD operator's traditional DNS services (left) and its control plane (right).

Automatic reconfiguration requires the control plane to collect a rich set of statistics on every DNS anycast node it manages. This includes basic statistics such as processing and storage resources usage, which may be collected using tools such as Nagios.<sup>2</sup> More extended statistics include EDNS Client-Subnet (ECS) extensions (RFC 7871). ECS contains crude geographic information on the location of clients, which the control plane may use to map query demands to the geographical location of end users (i.e., queries' origin) rather than of resolvers.

Our ultimate goal is that the control plane raises the abstraction level of operating a DNS name server infrastructure, allowing human operators to focus on handling rare incidents because the control plane handles the "regular" ones automatically. We expect this will require advanced visualizations through a TLD operator-wide dashboard [7], but that is outside the scope of this article.

We discuss our reconfiguration module and its performance later.

### PRIVACY PROTECTION

Privacy protection is an important function because the DNS traffic that the control plane analyzes for threat detection and DNS reconfiguration contains IP addresses of resolvers and domain names being looked up, which may constitute personally identifiable information (PII), depending on the jurisdiction. For example, under Dutch law this type of information is regarded as PII [8].

Privacy protection requires a mechanism that allows a TLD operator to systematically balance the privacy of Internet users on one hand and the targeted increase in the security and stability of the TLD by the control plane on the other. This mechanism needs to be flexible so that it can work with applicable privacy regulations, and it needs to be easy to use for engineers who need to develop new software to detect a new type of threat. It also needs to protect privacy through technical means within the control plane. We refer to [8] for more details on privacy requirements.

We discuss our privacy protection mechanism below.

## REALIZATION

Figure 2 provides an overview of our control plane, which consists of a high-speed data streaming warehouse called "ENTRADA," threat detection modules, a DNS reconfiguration module, and a privacy framework.

### ENTRADA

ENTRADA<sup>3</sup> (Enhanced Top-Level Domain Resilience through Advanced Data Analysis) [7, 9] is our open source DSW for the TLD control plane. ENTRADA consists of a set of modules that run on top of Apache Hadoop,<sup>4</sup> which is open source as well.

Figure 3 provides an overview of the ENTRADA DSW and how it stores DNS authoritative traffic. Steps I—III refer to domain name resolution. We export the incoming DNS traffic from the .nl authoritative servers to a staging server (step IV), in which the raw PCAP format is converted to an optimized open source column storage format (Parquet, step V), and later imported into the Hadoop File System (HDFS, VI). Impala<sup>5</sup> provides a massively parallel processing query engine with a standard SQL interface (VII). Applications and services use this interface to connect to ENTRADA.

We choose PCAP as our format for importing DNS traffic from name servers because it includes transport and IP-level headers in addition to their DNS payloads, because it requires few to no changes on name servers (a mirror port on the network or a PCAP process on the name servers), and it is widely used.

ENTRADA delivers the performance we need to build threat detection modules and perform hypothesis tests. For example, we showed in [9] that ENTRADA is able to analyze the equivalent of 52 TB of PCAP data in less than 3.5 min in a four-data-node cluster, using Impala and SQL syntax, which would be infeasible using PCAP format.

Our ENTRADA instance for .nl currently receives DNS traffic from four of our six unicast authoritative name servers and has been operational on our research network uninterruptedly as of March 2014. It currently stores more than

<sup>2</sup> <http://www.nagios.org>

<sup>3</sup> <http://entrada.sidnlab.nl>

<sup>4</sup> <http://hadoop.apache.org>

<sup>5</sup> <http://impala.io>

320 billion DNS query-response pairs in 15 TB of Parquet-compressed format.

### THREAT DETECTION MODULES

A TDM is an ENTRADA application that discovers potential threats, possibly in combination with other data feeds such as domain name database transactions, logs, and feeds from external threat information providers such as ShadowServer.<sup>6</sup>

An example of a TDM we developed is the New Domains Early-Warning System (nDEWS) [10], which leverages the known fact that newly registered malicious domains receive a much higher number of DNS queries immediately after their registration than normal domains.

Figure 4 shows this based on the daily number of queries for 20,000 randomly chosen normal domains (purple line) and phishing domains (green line). nDEWS thus enables a TLD operator to monitor all new domains added to its zone on a daily basis. It uses the *k*-means clustering algorithm to classify them based on their DNS query patterns.

We evaluated nDEWS using historical 8-month-data collected from one of the .nl authoritative servers. nDEWS yielded almost 3000 suspicious domains, which we had to validate using several techniques because we did not have a ground truth for them, since the contents of their websites might have changed during this period.

We are also evaluating nDEWS using current data and performing web content analysis if a domain is classified as suspicious. Besides phishing, nDEWS is able to detect other types of suspicious sites, such as allegedly counterfeit drugs and shoes. We automatically share the information coming out of nDEWS with 32 .nl registrars as part of a pilot.

Another TDM we have developed detects the DNS traffic pattern of a specific botnet. We identified what this pattern looks like, and our TDM uses ENTRADA to continuously scan for it. When our TDM detects a resolver that exhibits this behavior, it sends the resolver's IP address to the Abuse Information Exchange (AbuseHUB).<sup>7</sup> Members of this platform include large Dutch access providers, who use the information to clean up the botnet infections located within their network. With this TDM we are thus able to actively disrupt the distribution of spam mail and other malicious activity.

We refer to [7] for other TDMs we have developed.

### DNS RECONFIGURATION MODULE

The DNS reconfiguration module (DRCM) dynamically decides which name servers to start or stop at which locations. The DRCM is a logical entity that may be fully distributed across the name servers of a DNS anycast service [6].

Our current DRCM focuses on minimizing the latency between resolvers and the TLD operator's authoritative name servers. To develop the DRCM, we studied the impact of the number of anycast instances and their physical locations on the latency of the anycast service and reported on this study in [11]. By measuring real-world anycast deployments from C-, F-, K-, and L-Root DNS name servers using the RIPE Atlas framework, we were able to show that a handful of well placed anycast instances provide better and more stable

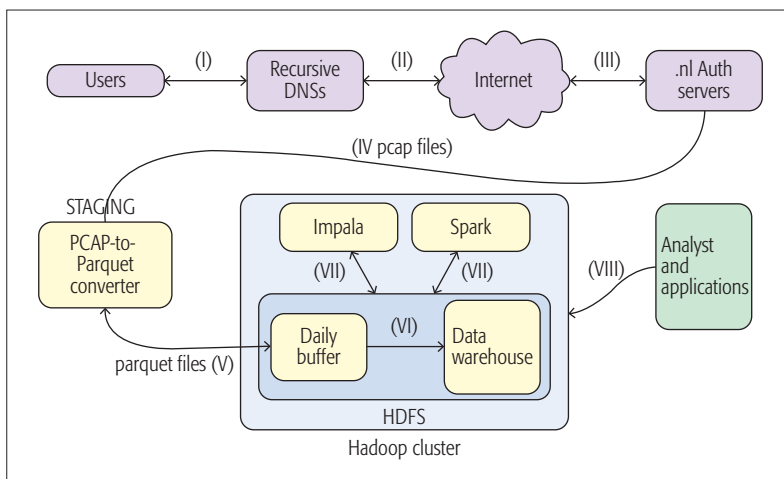


Figure 3. ENTRADA overview.

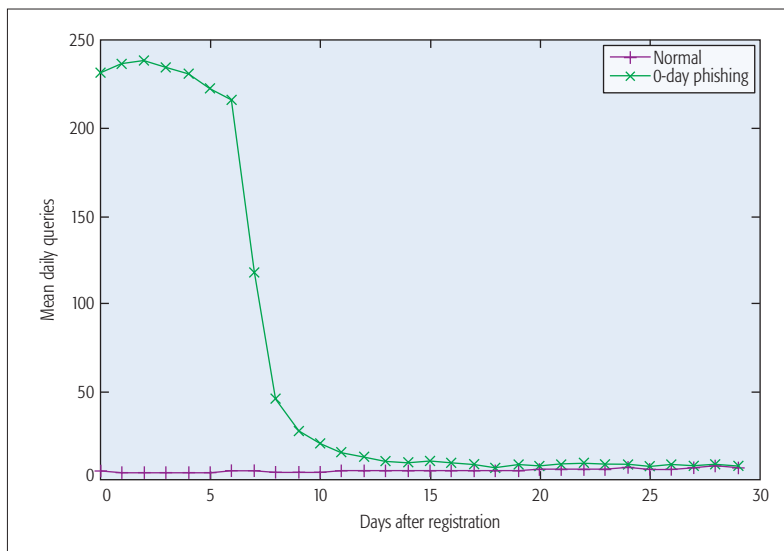


Figure 4. Queries to normal and phishing domains.

latency than a large-scale infrastructure consisting of several dozens of nodes. For example, C-Root with 8 anycast sites (4 in Europe and 4 in North America) achieved a worldwide median latency of 32 ms, while L-Root with 144 sites (18× more than C) all over the globe achieved a median latency of 30 ms.

Figure 5 shows the distribution of latency for C-Root and L-Root as seen from around 7900 vantage points around the globe. Note that the larger deployment of L-Root did not result in a shorter distribution tail as well: the 75th percentile of the latency distribution is 76 ms for C-Root and 73 ms for L-Root. These results suggest that connectivity of the anycast site is far more important for the performance of the anycast service than the number of deployed sites, which is an important finding for our DRCM.

We have also set up a worldwide anycast test-bed,<sup>8</sup> which we are using to further investigate the relationship between the number of anycast sites and their respective connectivity to service latency, in particular, to understand the efficiency and impact of traffic engineering through anycast for the mitigation of DDoS attacks. We are actively probing the anycast infrastructure to understand

<sup>6</sup> <http://shadowserver.org>

<sup>7</sup> <http://www.abuseinformationexchange.nl> (in Dutch)

<sup>8</sup> <http://www.anycast-testbed.nl>

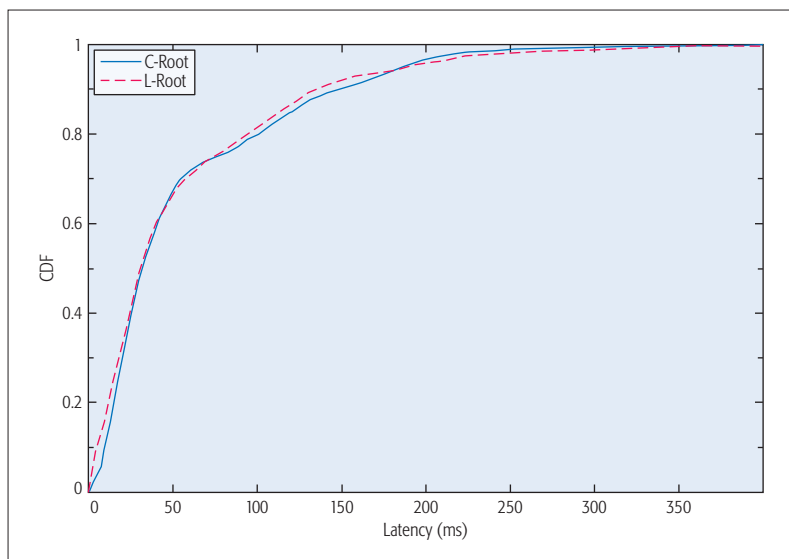


Figure 5. Distribution of latency for C- and L-Root.

the effects of runtime reconfigurations. We also evaluated the use of the ECS extension based on real data measured at two name servers that are authoritative for popular second-level domains such as `apache.com`, and we have modified them to receive and process queries with the ECS extension.

#### PRIVACY FRAMEWORK

Our Privacy Framework protects the privacy of the users of a TLD [8]. Its key concept is a privacy policy, which is a structured document in natural language that defines what data ENTRADA and its applications process for a particular purpose using which data filters. A filter is an operation that ENTRADA or an application applies to personal data. Examples are pseudonymization and aggregation. Filters form an essential element in the Privacy Framework, because they ensure that the privacy policies are verifiably enforced by technical means.

ENTRADA application developers and researchers formally submit privacy policies to the privacy board, which is a body within the TLD operator's organization that reviews policies. The privacy board approves or rejects the policy and informs the author through a policy evaluation report.

After policy approval, the author implements it as part of a policy enforcement point (PEP), which is the technical component within ENTRADA or one of its applications that realizes an approved privacy policy and actually applies the policy's filters at runtime.

Our implementation of the framework for `.nl` conforms to both EU and Dutch laws, and we reported it to the Netherlands Data Protection Authority. Our privacy board consists of a technical expert, a legal expert, and a member of our management team. They approved several privacy policies as of mid-2015, which we activated through PEPs.

#### RELATED WORK

To the best of our knowledge, we are the first to propose a system that enables TLD operators to become threat intelligence providers and increase

the robustness of their DNS services. However, there is scattered prior work on individual components.

The operator of the `.uk` TLD (United Kingdom) developed Turing,<sup>9</sup> a system that appears to be similar to ENTRADA. Turing, however, is a commercial closed source solution, and there is little publicly available information about its technical implementation. As far as we know, they did not extend their platform with functions to dynamically reconfigure name servers; nor did they include privacy protection mechanisms. We are also unaware of deployments of Turing at TLD operators other than at the `.uk` operator.

There have been several research works that use DNS TLD data for detection of malicious domains, but not as part of a larger modular system such as our control plane. Hao *et al.* [12] analyzed the initial lookup behavior of malicious domains under `.com` and `.org` using a spam trap. Also, there are different methods to classify malicious websites. For example, Abbasi and Chen [13] present a comparison of tools to detect fake websites and perform content analysis to classify the websites.

The dynamic reconfiguration of DNS anycast networks is a technique that has been used to guide clients to the best node of a content distribution network (CDN) in terms of network latency [6]. But the topic has not been explored before in the context of TLD operators, which need to support all networked applications that use the DNS and cannot assume that the roles of DNS operator and content provider are collocated as in the case of [6].

#### CONCLUSIONS AND FUTURE WORK

We present a control plane for operators of top-level domains in the DNS that enables them to increase the security and stability of their TLD by becoming a threat intelligence provider. Our control plane is a system that extends a TLD operator's traditional services and leverages the DNS traffic and the domain registration transactions that a TLD operator handles. The control plane continuously stores and analyzes this data to automatically detect potential threats in the TLD and shares this information with other players in the TLD, such as hosting and access providers. It can also use the information to dynamically scale the TLD operator's DNS infrastructure, which increases the robustness of the TLD operator's DNS services.

Our control plane builds on the ENTRADA open source software, which we have developed on top of a Hadoop-based data storage cluster. ENTRADA enables TLD operators to easily feed their authoritative DNS traffic into the control plane, to run our threat detection modules, and to add their own. ENTRADA is currently being used by at least six operators of country code TLDs. It comes with a Privacy Framework that enables TLD operators to manage the personally identifiable information of TLD users.

Our future work consists of further refining and implementing the control plane, for instance, in terms of modeling the DNS ecosystem using a variety of data sources, extending the control plane to other types of DNS operators, the interfaces a TLD operator needs to provide toward its DNS services, the impact of adding and remov-

<sup>9</sup> <http://nominet.uk/turing>

ing nodes from a DNS anycast network, and new threat detection modules such as for the detection of booter sites.

### ACKNOWLEDGMENTS

We thank Kees Neggens (SIDN Supervisory Board), Aiko Pras (University of Twente), Marc Groeneweg (SIDN), Moritz Müller (SIDN), and the anonymous reviewers, who provided valuable feedback at various stages of this article.

Ricardo de O. Schmidt's work is sponsored by the SAND project (<http://www.sand-project.nl>).

### REFERENCES

- [1] "Phishing Activity Trends Report, 1st Quarter 2016," Anti-Phishing Working Group (APWG), May 2016; [https://docs.apwg.org/reports/apwg\\_trends\\_report\\_q1\\_2016.pdf](https://docs.apwg.org/reports/apwg_trends_report_q1_2016.pdf).
- [2] A. Ozgit, "DDoS Attack on .tr," ICANN55 Tech Day, Marrakech, Morocco, Mar. 2016; <https://meetings.icann.org/en/marrakech55/schedule/mon-tech/presentation-ddos-07mar16-en.pdf>.
- [3] G. C. M. Moura *et al.*, "Anycast vs. DDoS: Evaluating the November 2015 Root DNS Event," *Proc. ACM Internet Measurement Conf.*, Santa Monica, CA, Nov. 2016.
- [4] J. J. Cardoso de Santanna *et al.*, "Booters — An Analysis of DDoS-as-a-Service Attacks," *IFIP/IEEE Int'l. Symp. Integrated Network Management*, 2015, pp. 243–51.
- [5] A. Bar *et al.*, "Large-scale Network Traffic Monitoring with DBStream, a System for Rolling Big Data Analysis," *2014 IEEE Int'l. Conf. Big Data*, Oct. 2014, pp. 165–70.
- [6] A. Flavel *et al.*, "FastRoute: A Scalable Load-Aware Anycast Routing Architecture for Modern CDNs," *12th USENIX Symp. Networked Systems Design and Implementation*, Oakland, CA, May 2015.
- [7] M. Wullink *et al.*, "ENTRADA: Enabling DNS Big Data Applications," *APWG Symp. Electronic Crime Research*, Toronto, Ontario, Canada, June 2016.
- [8] C. Hesselman *et al.*, "A Privacy Framework for DNS Big Data Applications," tech. rep., Nov 2014; [https://www.sidnlabs.nl/downloads/whitepapers/SIDN\\_Labs\\_Privacyraamwerk\\_Position\\_Paper\\_V1.4\\_ENG.pdf](https://www.sidnlabs.nl/downloads/whitepapers/SIDN_Labs_Privacyraamwerk_Position_Paper_V1.4_ENG.pdf).
- [9] M. Wullink, G. Moura, and C. Hesselman, "ENTRADA: A High Performance Network Traffic Data Streaming Warehouse," *IEEE/IFIP NOMS '16*, Istanbul, Turkey, Apr. 2016.
- [10] G. Moura *et al.*, "nDEWS: A New Domains Early Warning System for TLDs," *IEEE/IFIP Int'l. Wksp. Analytics for Network and Service Management*, Istanbul, Turkey, Apr. 2016.
- [11] R. Schmidt, J. Heidemann, and J.H. Kuipers, "Anycast Latency: How Many Sites Are Enough?," tech. rep. ISI-TR-2016-708, May 2016; <http://wwwhome.cs.utwente.nl/~schmidtr/docs/ISI-TR-2016-708.pdf>.

- [12] S. Hao, N. Feamster, and R. Pandrangi, "Monitoring the Initial DNS Behavior of Malicious Domains," *Proc. 2011 ACM SIGCOMM Conf. Internet Measurement*, ser. IMC '11, 2011, pp. 269–78.
- [13] A. Abbasi and H. Chen, "A Comparison of Tools for Detecting Fake Websites," *Computer*, no. 10, 2009, pp. 78–86.
- [14] S. Hao *et al.*, "PREDATOR: Proactive Recognition and Elimination of Domain Abuse at Time-Of-Registration," *Proc. 2016 ACM CCS*, Oct. 2016.
- [15] Y. Yu *et al.*, "Authority Server Selection in DNS Caching Resolvers," *SIGCOMM Comp. Commun. Rev.*, vol. 42, no. 2, Mar. 2012, pp. 80–86.
- [16] M. Korczynski, M. Krol, and M. van Eeten, "Zone Poisoning: The How and Where of Non-Secure DNS Dynamic Updates," *ACM Internet Measurement Conf. 2016*, Nov. 2016.

### BIOGRAPHIES

CRISTIAN HESSELMAN ([cristian.hesselman@sidn.nl](mailto:cristian.hesselman@sidn.nl)) directs SIDN Labs, the research team of the .nl operator, SIDN. His work focuses on mechanisms and systems that further enhance the security, stability, and interoperability of .nl, the Domain Name System, and the wider Internet. He was previously a senior researcher at Telematica Instituut and a software engineer at Lucent Technologies. He holds a Ph.D. and an M.Sc. in computer science, both from the University of Twente, the Netherlands.

GIOVANE C. M. MOURA ([giovane.moura@sidn.nl](mailto:giovane.moura@sidn.nl)) is a data scientist with SIDN Labs, the research arm of the top-level domain registry of the Netherlands (.nl). He works in research projects that involve performance and security of computer networks. Prior to SIDN, he worked as a postdoctoral researcher at TU Delft, and obtained his Ph.D. from the University of Twente, both in the Netherlands. He has also a Master's degree in computer science from the Federal University of Rio Grande do Sul, Brazil.

RICARDO DE OLIVEIRA SCHMIDT ([r.schmidt@utwente.nl](mailto:r.schmidt@utwente.nl)) is a postdoctoral researcher within the Design and Analysis of Communication Systems group at the University of Twente. He obtained his Ph.D. from the same university in 2014. His research interests are in Internet security and management, with research approaches strongly based on Internet measurements and monitoring. He has been involved in various Dutch and European projects, and is currently responsible for running two national projects in management and security of the DNS system.

CEES TOET ([cees.toet@sidn.nl](mailto:cees.toet@sidn.nl)) is SIDN's director of IT, responsible for the operation, security, and stability of the authoritative name servers and the domain registration database of the .nl TLD (The Netherlands). Following his IT education at the Dutch Ministry of Defense, he has worked at several governmental and commercial companies as IT Manager.

Our implementation of the framework for .nl conforms to both EU and Dutch laws and we reported it to the Netherlands Data Protection Authority. Our privacy board consists of a technical expert, a legal expert, and a member of our management team. They approved several privacy policies in mid 2015, which we activated through PEPs.

# Service Provider DevOps

Wolfgang John, Guido Marchetto, Felicián Németh, Pontus Sköldström, Rebecca Steinert, Catalin Meirosu, Ioanna Papafili, and Kostas Pentikousis

The authors present what lies beyond the first evolutionary steps in network management, identify the challenges in service verification, observability, and troubleshooting, and explain how to address them using their Service Provider DevOps (SP-DevOps) framework.

## ABSTRACT

Although there is consensus that software defined networking and network functions virtualization overhaul service provisioning and deployment, the community still lacks a definite answer on how carrier-grade operations praxis needs to evolve. This article presents what lies beyond the first evolutionary steps in network management, identifies the challenges in service verification, observability, and troubleshooting, and explains how to address them using our Service Provider DevOps (SP-DevOps) framework. We compendiously cover the entire process from design goals to tool realization and employ an elastic version of an industry-standard use case to show how on-the-fly verification, software-defined monitoring, and automated troubleshooting of services reduce the cost of fault management actions. We assess SP-DevOps with respect to key attributes of software-defined telecommunication infrastructures both qualitatively and quantitatively, and demonstrate that SP-DevOps paves the way toward carrier-grade operations and management in the network virtualization era.

## INTRODUCTION

Software-defined networking (SDN) and network functions virtualization (NFV) enable operators to use network service function chains that are no longer static and embedded into special-purpose physical network elements or deployed at pre-planned and fixed points in the infrastructure. This change has a profound effect on network operations. As advocated in [1], virtualized networks based on network functions (NFs) and endpoints chained together through network function forwarding graphs (NF-FGs) can be highly dynamic and programmable in terms of service definition and execution. In this context, open application programming interfaces (APIs) will take precedence over, for instance, vendor-specific command line interfaces (CLIs) as currently used by expert administrators in the field. The availability of NF APIs combined with SDN programmability calls for handling carrier infrastructure and resources using techniques common in the software engineering realm, thus changing the practice of network management significantly.

The first contribution of this article is a compendious tutorial addressing the wider network research and practitioner communities about how to handle the operational complexity of carrier-grade software-defined infrastructures (SDIs)

through Service Provider DevOps (SP-DevOps). We explain how SP-DevOps eases verification and activation of complex services using novel network and service observability, diagnostics, and troubleshooting methods ready to be integrated into developer and operations workflows. The second contribution of this article is the assessment of SP-DevOps and a qualitative comparison with earlier proposals and publicly available toolsets. Finally, we summarize and provide pointers to the publicly available open source contributions that implement SP-DevOps in practice. Interested readers can delve into the full details of SP-DevOps in our publicly available technical report [2].

Next, we review the requirements, objectives, and related initiatives for network operations in SDIs. We then present the SP-DevOps paradigm followed by an illustrative use case. We conclude this article with an assessment and summary of our contributions.

## REQUIREMENTS, OBJECTIVES, AND RELATED INITIATIVES

Traditional telecom and IT infrastructure operations are governed by extensive processes typically following eTOM [3] and ITIL [4]. Originally designed for preplanned, hardware-oriented, physical infrastructure, recent work in the TM Forum ZOOM group as well as the Internet Engineering Task Force (IETF) has initiated the adaptation of these processes for SDIs. Requirements for **carrier-grade** telecom infrastructures include high availability (“five 9s”); scalability to hundreds of thousands of nodes covering large geographical areas; and the ability to monitor performance parameters for service level agreements (SLAs). As a first objective, SDIs must conform to said requirements, but also meet new challenges [5].

The key characteristics of SDIs are **agility** to introduce new services in the market in minutes rather than in months or years, and **elasticity** to dynamically optimize demand-responsive resource allocation in accordance with policy. SDIs are fueled by programmability and automation, which reduce manual interaction with equipment and management systems. We use the term **orchestrated assurance** to refer to the integration between fulfillment (i.e., programmable orchestration) and assurance systems, which can generate actionable insights based on huge quantities of data.

Wolfgang John and Catalin Meirosu are with Ericsson AB; Guido Marchetto is with Politecnico di Torino;

Felicián Németh is with Budapest University of Technology and Economics; Pontus Sköldström is with ACREO Swedish ICT AB;

Rebecca Steinert is with SICS Swedish ICT AB; Ioanna Papafili is with Hellenic Telecommunications Organization; Kostas Pentikousis is with Traveling.

Digital Object Identifier:  
10.1109/MCOM.2017.1500803CM

Project or framework	Carrier-grade	Agility	Elasticity	Infrastructure validation	Service verification	Orchestrated assurance
eTOM [3]	Yes	Partial <sup>1</sup>	Partial <sup>1</sup>	Manual	No	No
ITIL [4]	Partial <sup>1</sup>	No	Partial <sup>1</sup>	Manual	No	No
MANO – OPNFV	Yes	Ongoing <sup>2</sup>	Ongoing <sup>2</sup>	Partial <sup>1</sup>	No	Ongoing <sup>2</sup>
CloudWave [7]	No	Partial <sup>1</sup>	Yes	No	No	Yes
T-NOVA [8]	Partial <sup>1</sup>	Yes	Partial <sup>1</sup>	Partial <sup>1</sup>	No	No
IBM BlueMix with DevOps services	No	Yes	Yes	Partial <sup>1</sup>	No	Yes
UNIFY SP-DevOps [2]	Yes	Yes	Yes	Ongoing <sup>2</sup>	Ongoing <sup>2</sup>	Yes

<sup>1</sup> Partial implies that some requirement areas are addressed, but significant open issues remain with little or no activity to tackle them as of June 2016.

<sup>2</sup> Ongoing means that we consider the requirement as fulfilled from a conceptual point of view, but stable version documents, technical descriptions, or full implementations have yet to be released as of July 2016.

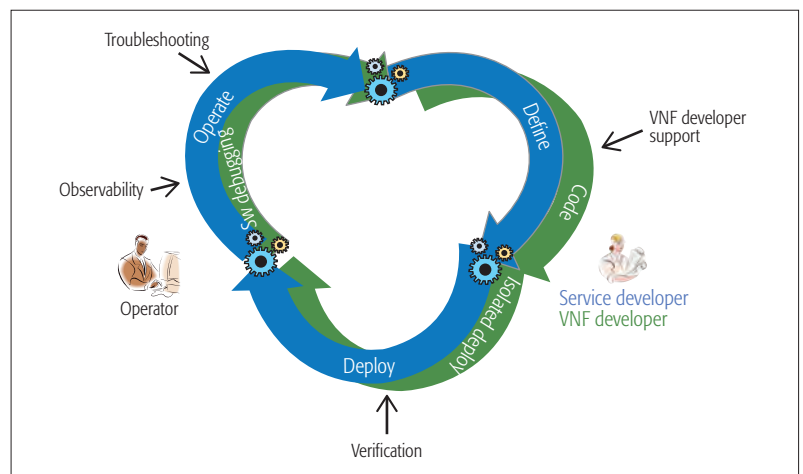
**Table 1.** Carrier network management frameworks.

Agility, elasticity, and programmability introduce new objectives in two additional areas, **infrastructure validation** and verification of all components. In the context of SDIs, we regard validation as an evolution of the traditional performance checks prior to service activation. In contrast, **service verification** relates to the formal correctness of the code executed at different SDI levels. Specifically, correctness against a set of rules and policies must be verified, whether the code represents a single network element or a complex service chain.

A comprehensive summary of industry and research activities related to NFV and SDI can be found in [6]. Table 1 reflects our assessment of selected management frameworks with respect to the above-mentioned SDI characteristics. For instance, CloudWave [7] developed a platform-as-a-service that presents detailed views of the enterprise cloud infrastructure to application developers and enables fast development cycles. T-NOVA [8] provides a Network Function Store as part of a self-service portal where customers can select virtual appliances to add to their services, complemented by automatic deployment and monitoring. The implementation of the European Telecommunications Standards Institute (ETSI) MANO framework through components selected and integrated through the OPNFV project focused initially on the virtual infrastructure layer and its management. Finally, IBM BlueMix addressed the needs of enterprise mobile and web developers with DevOps services that simplify application development and deployment. However, none of them is complete with respect to the SDI characteristics outlined in this section. We examine how SP-DevOps addresses these objectives later.

### SERVICE PROVIDER DEVOPS

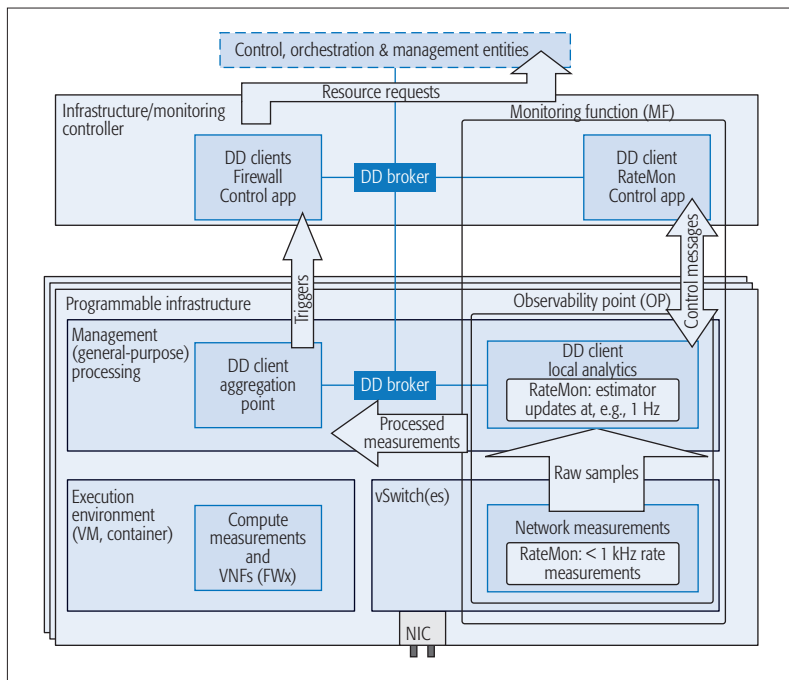
IT DevOps tackles objectives comparable to the ones listed above but for data center (DC) environments. DevOps, however, is not a single method that can be directly applied to telecommunication SDIs. Telecommunication infrastructure exhibits several orders of magnitude more distribution than DC infrastructure, spreading



**Figure 1.** SP-DevOps processes and roles in the telecom service life cycle.

over very large geographical areas, making carrier-grade requirements such as high availability or strict latency bounds harder to meet. Furthermore, telecom resources range over multiple network and DC domains, typically consisting of heterogeneous hardware and software, which contrasts starkly with the homogeneity of DC resources. Finally, basic DC operational assumptions regarding network latency and capacity do not hold.

Adopting DevOps involves an organizational shift as well: developer, operations, and quality assurance teams must work closely together. Typical IT companies do not have significant business boundaries between different teams, whereas such boundaries are not uncommon in telecommunication service providers (SPs). Today, network elements and functions are developed by vendors, and deployed and operated by SPs, with some parts of the network operated by sub-contractors. Figure 1 illustrates the SP-DevOps service life cycle and highlights technical processes shared by different roles – **Verification**, **Observability**, and **Troubleshooting**, which address this shift and further challenges [2, 5, 9]. We define the following SP-DevOps roles for this service life cycle:



**Figure 2.** Software defined monitoring components in a programmable infrastructure. RateMon serves as a local observability point, and DoubleDecker (DD) provides a hierarchical messaging architecture.

- The **service developer** assembles the service graph for a particular service category, similar to the traditional operator role.
- The **VNF developer** implements virtual NFs (VNFs) and would be associated with the traditional equipment vendor role in today's terms.
- The **operator** ensures that a set of performance indicators associated with a service are met when the service is deployed on the SDI.

#### PRE-DEPLOYMENT SERVICE VERIFICATION

SP-DevOps relies on repeatable and reliable processes, calling for automated service verification as an integral part of the deployment processes. Identifying problems early in the service/product life cycle increases availability, and significantly reduces time and cost spent on debugging and troubleshooting tasks. For telecommunication service definitions and configurations this is especially true due to the high spatial distribution and the lower levels of infrastructure redundancy in an operator environment compared to centralized DCs.

The first SP-DevOps process we introduce is pre-deployment service verification based on formal methods, which can prove that the involved functions fulfill certain properties, thus addressing the objectives of agility, elasticity, and programmability for the envisioned service model. Service verification employs VNF models that can be combined to build a formal service description, ensuring generality and supporting dynamic service definitions. In essence, network services described by NF-FGs involving multiple VNFs are translated into sets of formulas that can be analyzed and verified.

Our realization benefits from Z3 [10] and leverages an earlier VNF verification engine [11]

to create an engine compliant with the SDI objectives. We focus on model scalability to guarantee fast (on-the-fly) verification, which is still detailed enough to completely capture VNF behavior. For this, we complement the VNF model catalog with more complex, previously unsupported VNFs (i.e., active VNFs that alter packets). Specifically, we developed models for active VNFs including Network Address Translation (NAT), virtual private network gateway, and web cache, as well as additional models of currently unsupported passive VNFs, like antispam filter. As a result, SP-DevOps service verification applies to a wide range of dynamic service graphs.

#### OBSERVABILITY THROUGH SCALABLE MONITORING

In SP-DevOps we employ software-defined monitoring (SDM) designed to meet a number of goals. First, the design should provide accurate and scalable monitoring both in large-scale geographically distributed WAN and centralized DC scenarios. Second, it should be able to quickly trigger reactions on monitoring results locally for increased scalability. Third, network dynamics, such as migration of VNFs and associated monitoring functions (MFs), should be supported. Finally, it should allow for distributed, programmable data processing following SDN principles. Altogether, we devise SDM to effectively meet the elasticity, orchestrated assurance, and infrastructure validation objectives.

Figure 2 illustrates the main SDM components:

- MFs that can aggregate and process data at high rates and produce reliable monitoring results
- Multi-level aggregation with distributed and programmable aggregation points able to combine multiple metrics and forward results and/or trigger reactions
- A flexible, distributed, and carrier-grade messaging system that routes monitoring results and other messages between entities

By combining these three components we can reduce load on the control/management planes and react locally while avoiding transmitting high-rate monitoring results over WAN connections.

An MF performs lightweight node-local aggregation, processing, and analytics to fulfill the monitoring goals for the service component (e.g., measurement intensity and duration). An MF is implemented by one or several observability points (OPs) and an MF control app. An MF control app is the SDM equivalent of an SDN controller, that is, a logically centralized measurement control plane configuring OPs and performing parts of the processing. OPs run locally on the infrastructure nodes and implement functionality for performing measurements and lightweight data processing. MF monitoring results are sent to the closest aggregation point, typically on the same node. Aggregation points expose a simple API used by the management layers to configure the desired aggregation method and triggering thresholds. Multiple metrics can be combined, evaluated, and forwarded to higher layers or local control components when certain thresholds are met.

In line with the SDM design goals, we implemented RateMon [12], an MF that probabilistically models link utilization for assessing the risk of con-



gestion at various timescales. RateMon can quickly detect symptoms of persistent micro-congestion episodes with no communication overhead as it does not require forwarding raw measurements for further processing. Moreover, DoubleDecker [13], a multi-tenant distributed messaging system, provides connectivity between MFs, aggregation points, and higher-layer entities. DoubleDecker keeps messages local when possible and provides a simple messaging API with a publish/subscribe mechanism for distributing monitoring results and a notification mechanism for targeted messages such as alarms.

### AUTOMATED TROUBLESHOOTING

Troubleshooting involves a series of hypothesis tests in which the troubleshooter repeatedly analyzes the results of one test and decides whether another hypothesis needs to be tested, leading to a consecutive test by possibly a different debugging tool. Today such steps are performed manually by support teams and in practice are time-consuming, costly, and error-prone. Automation can reduce the time spent on troubleshooting incidents, but in SDIs must be combined with the NF-FG and its mapping to the underlying virtualized infrastructure, which is often not fully exposed to service or VNF developers.

SP-DevOps automated troubleshooting addresses this by facilitating fault management and service chain debugging on a large scale. Automated troubleshooting invocation includes the specification of a troubleshooting template, which states the troubleshooting steps and rules, the type of tools used along with their respective configurations, and specifications on how to report troubleshooting results. Troubleshooting is controlled by a function that executes the template instructions using available system functions and interfaces. High-level troubleshooting processes of varying complexity can therefore be implemented for different purposes without knowing the particular details of the underlying SDI. Such processes are easier to maintain and develop compared to complex functions that fully integrate multiple traditional and SDN-specific troubleshooting tools.

SP-DevOps automated troubleshooting is exemplified by EPOXIDE [14], a lightweight framework for testing troubleshooting hypotheses that enables ad hoc creation of tailor-made testing methods from predefined building blocks. Troubleshooting personnel employ EPOXIDE to write and execute troubleshooting graphs (TSGs) that define the interconnectivity of individual debugging/troubleshooting tools. Writing a TSG is faster than typing similar CLI commands, but more importantly, the EPOXIDE high-level language hides particularized technical details from the personnel (e.g., actual IP addresses) and provides reusable troubleshooting recipes. Moreover, EPOXIDE allows decision logic to be inserted into TSG nodes instead of just connecting different low-level tools by piping the output of one tool into the input of another. Decision nodes can analyze outputs and decide where to forward them. As a result, a TSG can test more than just one troubleshooting hypothesis, and can further automate the troubleshooting process by executing decision trees, as we see in the following section.

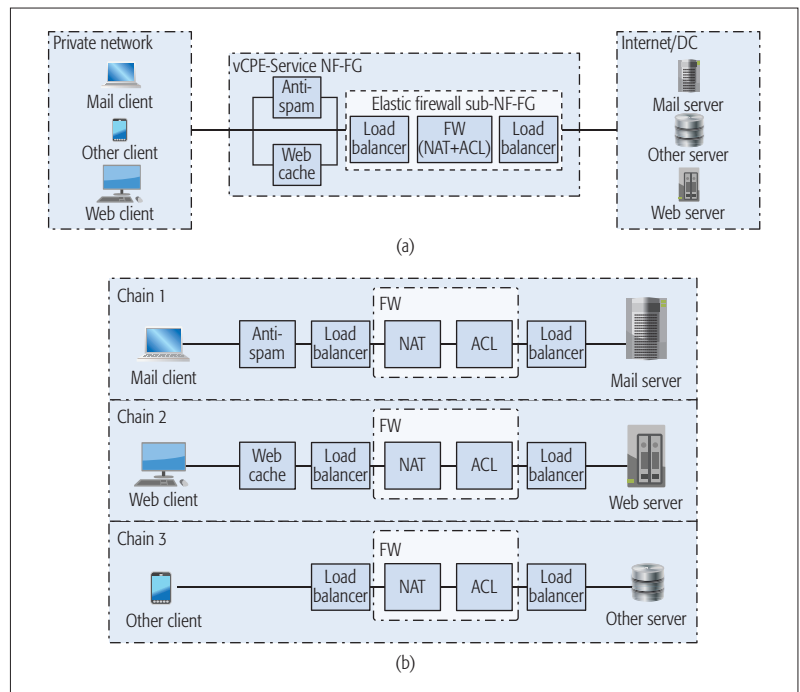


Figure 3. vCPE service chain with Elastic Firewall: a) Service NF-FG; and b) Chain extraction for service verification.

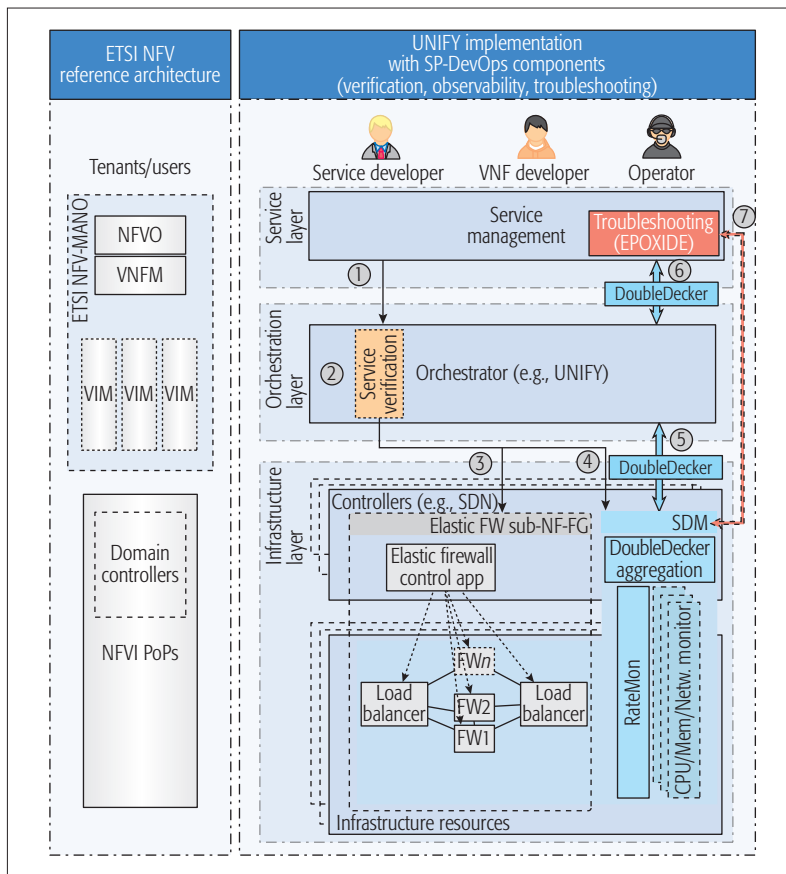
### SP-DEVOPS IN PRACTICE

We take virtual customer premises equipment (vCPE) as an illustrative SDI deployment example. The vCPE service is specified as a graph (NF-FG) of security and performance acceleration VNFs chained together with a firewall providing NAT and access control list (ACL) functionality. Figure 3a is a high-level system overview, with hosts in a private network communicating via vCPE service components to servers located in the Internet. Forwarding rules are configured such that email and web traffic is forwarded to an anti-spam function and a web cache, respectively, before reaching its destination server through the firewall.

We consider an elastic firewall as a dynamically scalable network element. It is “elastic” as it supports different scaling approaches triggered by continuous monitoring of certain conditions, including infrastructure resource utilization, changes in user patterns, and so on. In this article we consider horizontal scaling, that is, scaling by adding further virtual resource instances on scale-out.

We model the elastic firewall as a sub-NF-FG comprising load balancing functions, dynamically instantiated firewall data plane elements, and an elastic firewall control app (Fig. 4). The elastic firewall scales out/in by an action of the control app adding or removing firewall elements to the service chain, realized on resources requested from the orchestrator. Moreover, the control app instruments the load balancer elements that precede the firewall data plane elements to forward traffic according to dynamically configured flow rules.

The three SP-DevOps processes described earlier support the roles of operators as well as VNF and service developers throughout the life cycle of an elastic firewall during both service deployment (i.e., fulfillment) and assurance phases, as explained next.



**Figure 4.** NFV use-case of an elastic firewall and SP-DevOps processes embedded in the UNIFY NFV architecture — an architecture that is in conformance with ETSI NFV, as indicated in the left part of the figure.

### SERVICE DEPLOYMENT

Figure 4 illustrates the SP-DevOps processes applied to the vCPE case. Service deployment starts with a tenant/user employing service management interfaces to request a service graph as per step 1, which describes the ordered interconnection of abstract NFs and their corresponding key performance indicators (KPIs). In the orchestrator, the abstract NFs are translated into concrete VNF components including implementation version, placement definitions, and connectivity configurations.

In step 2, SP-DevOps service verification is invoked. In case of dynamic service chains including elastic functions, verification takes place before the initial service deployment as well as before scale in/out updates of the NF-FG by the orchestrator. Specific requests are verified before actually being deployed let alone executed in the production environment. If service verification fails, the request is rejected and sent back to service management, where it can be refined or cancelled. If the service and VNF definitions with their corresponding configurations are valid, the request is forwarded to the infrastructure layer in step 3, where VNF control and data plane components are deployed on the assigned network and compute resources. Pre-deployment verification is a quality assurance mechanism for the operator in the deployment phase. In the vCPE case, verification ensures that firewall and NAT configuration are correct, and that correctness is maintained throughout scaling operations.

In order to evaluate SP-DevOps verification, we consider the vCPE NF-FG shown in Fig. 3a. As a first step, three separate VNF chains are extracted (Fig. 3b):

- Chain 1 employs an anti-spam function, NAT, and ACL.
- Chain 2 is composed by a web cache, NAT, and ACL.
- Chain 3 uses only the NAT and ACL firewall functionalities.

Our service verification tool currently offers verification of reachability and isolation properties, that is, whether a network configuration can ensure that a given node is reachable, or whether specific traffic never reaches a given node, respectively. The tool internals and the specific steps it performs with respect to reachability verification are described in [15]. Concerning isolation, it is worth noting that this can be seen as the logical complement of reachability, which is actually the property that our tool currently uses to also verify isolation policies.

For the vCPE use case, we consider both reachability and isolation properties during pre-deployment verification to evaluate the impact of verification on the overall deployment time. In particular, given two different ACL configurations that should either allow or block traffic flows, we verified whether the three servers are actually reachable or isolated, respectively. In chains 1 and 2, the verification of the isolation property must be followed by a further verification step ensuring that server unreachability is indeed due to the ACL configuration rather than the anti-spam or web cache function. This is done by iterating verification of a reachability property between the clients and the firewalls on the paths. Our results indicate that the average time to verify reachability properties is less than 50 ms, with a maximum verification time of 200 ms. The average time to verify isolation properties (including the extra reachability tests) is less than 80 ms, with a maximum time of 310 ms. This is in line with SP-DevOps goals to support agile and elastic service deployment with on-the-fly verification of service requests/updates. With only negligible overhead, service verification can significantly reduce the number of trouble incidents, as it prevents erroneous and untrustworthy behaviors of the system ahead of deployment.

### SERVICE ASSURANCE

**Continuous Monitoring:** Highly dynamic firewall elasticity requires very frequent status updates about service components such as load balancing and firewall data plane instances. To support this requirement, we employ SDM for continuous monitoring. Monitoring components are deployed and configured automatically alongside the NF-FG components (step 4 in Fig. 4) based on first, monitoring intents derived from the KPIs, and second, requirements specified in the service graph definition. In the vCPE case, RateMon monitors network resource utilization.

MFs continuously collect status information about the service VNFs, and transfer the results using DoubleDecker. In case of performance degradation, the firewall control application decides on suitable elasticity operations, and SDM notifies the orchestrator to trigger resource scaling

in step 5. An updated NF-FG with new firewall instances resulting from a scale-out operation is automatically considered by the monitoring system, which instantiates further RateMon instances. The DoubleDecker pub/sub interface provides the operator with fast triggers for service elasticity mechanisms, and the service developer with status metrics for logging and SLA reporting purposes, as well as for triggering troubleshooting processes.

Besides dynamicity and programmability, SDM offers significant scalability and resiliency benefits. Following the distributed nature of service provider networks, SDM distributes the monitoring functionality (transport and storage of results, processing, and alarm generation) to reduce network overhead as well as the dependence and load on centralized components. Comparing the flow of information in SDM to centralized monitoring with equivalent functionality, such as OpenStack Telemetry, highlights the difference: In the vCPE case, traffic rate sampling at all the ports of up to 20 firewall instances at 100 Hz results in 4000 samples/s. In a centralized solution, this would translate into 4000 events/s that have to be stored, processed, and reacted to by central components. With SDM, samples are first reduced by a factor of 100 by RateMon's probabilistic parameter estimation and then processed by the local analytics engine. The local analytics engine decides whether other components need to be informed, for example, by sending an alarm to the firewall control application. The control application in turn decides what action should be taken locally or whether to request additional resources from the orchestrator. Centralized components are invoked only once per scale-out/in leading to several orders of magnitude fewer events that need to be handled centrally. These savings are crucial when considering large operator networks providing a large number of clients with many services in parallel, all of which are continuously monitored for multiple performance metrics.

**On-Demand Troubleshooting:** A VNF developer or operator may decide to debug a specific VNF or troubleshoot the complete service NF-FG once continuous monitoring results raise a troubleshooting incident (step 6 in Fig. 4). Automated troubleshooting employing EPOXIDE will instrument a set of monitoring and debugging tools (step 7). TSGs may include legacy networking tools, such as ping or iperf, next to complex SP-DevOps tools [2]. For instance, the elastic router control app can be verified using black box testing to analyze the behavior of the entire NF-FG or white box testing by logging into the control app container and debugging the app itself. Both approaches are supported.

Assume that SDM reports increased response time in web requests. After an initial investigation, a service developer suspects that automatic resource scaling is the culprit. The developer writes a TSG (Fig. 5) that uses the traffic generator node to overload the firewall. While running the traffic generator, the developer can observe key network characteristics, for example, how many VNF instances are deployed for selected VNF types, or the load of the WebCache-NAT link, which helps to determine whether the hypothesis is true. This method is not only faster compared

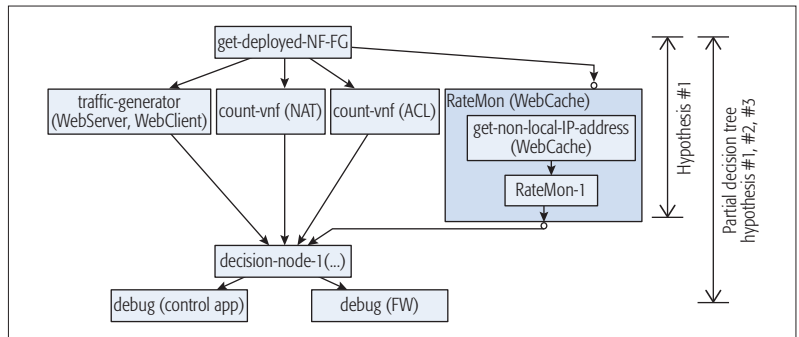


Figure 5. Example troubleshooting graph for the elastic firewall scenario.

to testing the hypothesis using traditional command line tools, but also less error-prone, because the process is described without particular details such as exact locations of the web server and web client.

Further measurement metrics (CPU load, memory use, etc.) may be additionally necessary to decide about hypotheses, so a TSG can be extended to collect information from many tools to decide on further hypotheses. For example, if the CPU loads of the firewall VNFs are unbalanced, the load balancer might require debugging; if the link load is high, but the number of VNFs is not increasing, the control app might be debugged instead. If the decision logic is expressed, say, by a numerical formula, a troubleshooter can configure a decision node and form a (partial) decision tree from individual hypothesis tests. As a result, one manual troubleshooting session can be turned into an automated test written especially for this service graph, thereby decreasing the execution time from hours to minutes. Moreover, subsequent occurrences of similar issues can be addressed by re-invocation of the TSG by less trained personnel without any particular knowledge of special-purpose tools used.

#### SP-DEVOPS ASSESSMENT

We assess the overall value of SP-DevOps in two ways: quantitatively, by estimating the potential of the described SP-DevOps processes in terms of OPEX savings; and qualitatively, by contrasting SP-DevOps and related frameworks with the objectives introduced earlier.

**OPEX Savings:** Admittedly, it is difficult to estimate the direct operational expenditure (OPEX) savings resulting from widescale adoption of SP-DevOps since an openly available OPEX model for SDIs is not available. Instead, we discuss OPEX savings based on how SP-DevOps components could address large categories of incidents that today cause service interruptions. According to the 2014 ENISA Annual Incident Reports, 44 percent of significant telecom network incidents were caused by software bugs, 19 percent by network overload, 10 percent by faulty software changes or updates, and 10 percent by faulty policies or procedures. Based on the ENISA data we derived a model, including the average duration of incidents, to determine the impact of SP-DevOps processes. Notably, integrated service verification during deployment reduces the number of incidents caused by faulty policies and software bugs. Scalable, continuous SDM shortens the duration of incidents via fast

On-the-fly verification of service definitions and configurations before actual deployment is practically feasible and performant as it reduces the number of incidents while introducing negligible overhead. Software-defined monitoring supports dynamic and elastic observability of deployed services and offers carrier-grade scalability. A novel framework enables automatic instrumentation of monitoring, verification, and debugging tools, thereby decreasing troubleshooting times from hours to minutes.

error discovery, and our specific example of Rate-Mon can significantly decrease operational costs and lost revenues related to network overload incidents. Finally, automated troubleshooting can decrease repair times for most incident types.

In a conservative scenario only a low percentage (about 30 percent) of addressable incidents could be avoided. In an optimistic scenario, a high percentage (around 80 percent for some relevant categories, e.g., faulty software changes and procedure faults) could be addressed. For incidents related to both fixed and mobile Internet connectivity, the optimistic scenario would record about 70–80 percent savings in terms of reduced incident occurrence and length, which translate into significant OPEX savings for operators, as well as additional benefits due to reduced customer churn and increased network uptime. Due to space considerations, interested readers are referred to the publicly available technical report [2] that details the model and the scenarios.

**Comparison with Related Initiatives:** We revisit the frameworks presented earlier and evaluate how SP-DevOps fulfills the following objectives:

- Address carrier network management
- Satisfy key characteristics associated with programmability (i.e., agility and elasticity)
- Fulfill infrastructure validation
- Perform service verification
- Provide orchestrated assurance

Table 1 summarizes our analysis of SP-DevOps in the context of the vCPE service use case. We conclude that SP-DevOps supports service agility and elasticity well. We acknowledge that processes such as billing and charging are not directly addressed by SP-DevOps, but consider them out of scope for this work since they can be adapted from traditional, standardized processes [3, 4]. SP-DevOps supports infrastructure validation, although at the time of this writing with a limited number of tools. Integration of further monitoring and diagnostic SP-DevOps tools is ongoing, as discussed in [2]. Further, service verification is currently updated with capabilities to verify additional properties and policies. Finally, SP-DevOps provides orchestrated assurance of network services due to the integration of deployment and assurance processes.

## SUMMARY

Carrier networks will evolve at a faster pace due to new networking paradigms such as SDN and NFV, which enable telecom operators to use programmable network service function chains. In this context, new network management challenges arise, which cannot be addressed by today's common practice and employed techniques. By combining network programmability with NF APIs, we can leverage techniques from the software engineering realm to define carrier-grade network management in the virtualization era.

Service Provider DevOps addresses said challenges via three integrated technical processes: verification, observability, and troubleshooting. We present a set of tools that each represents an advancement in the state of the art in its area while at the same time serving as a building block for the three integrated SP-DevOps processes. We consider an industry-standard use case,

a modern vCPE service with an elastic firewall, as an illustrative example that confirms the feasibility of our integrated approach. We show that on-the-fly verification of service definitions and configurations before actual deployment is practically feasible and performant as it reduces the number of incidents while introducing negligible overhead. Software-defined monitoring supports dynamic and elastic observability of deployed services and offers carrier-grade scalability. A novel framework enables automatic instrumentation of monitoring, verification, and debugging tools, thereby decreasing troubleshooting times from hours to minutes.

SP-DevOps processes are executed by different actors/roles during a service life cycle, thus establishing a common vocabulary and work routines that foster a DevOps-like approach for managing telecom infrastructure. Our simplified quantitative model points to savings of up to 80 percent in terms of OPEX costs with respect to the number of incidents and repair times. Our qualitative analysis confirms high compliance of SP-DevOps with respect to key objectives, enabling a carrier to address many network management challenges in the emerging network virtualization era. Moving forward, our ongoing efforts include the definition of metrics to evaluate service performance SP-DevOps tools. Moreover, we are enhancing SP-DevOps with further observability, diagnostic and verification tools, and capabilities targeting the application of SP-DevOps in real carrier networks and large-scale deployments. Finally, we are contributing to ongoing efforts at IETF [5] addressing DevOps challenges in telecommunication SDIs.

## ACKNOWLEDGMENT

This work is supported by FP7 UNIFY, a research project partially funded by the European Community under the Seventh Framework Programme (grant agreement no. 619609). The views expressed here are those of the authors only. The European Commission is not liable for any use that may be made of the information in this document. The authors would like to thank all anonymous reviewers for their constructive comments

## REFERENCES

- [1] ETSI GS NFV 002 V1.1.1 (2013-10), "Network Functions Virtualization (NFV); Architectural Framework"; [http://www.etsi.org/deliver/etsi\\_gs/nfv/001\\_099/002/01.01.01\\_60/gs\\_nfv002v010101p.pdf](http://www.etsi.org/deliver/etsi_gs/nfv/001_099/002/01.01.01_60/gs_nfv002v010101p.pdf) (accessed Oct. 12, 2016).
- [2] G. Marchetto *et al.*, "Final Service Provider DevOps Concept and Evaluation," CoRR, vol. abs/1610.02387, 2016; <http://arxiv.org/abs/1610.02387> (accessed Oct. 12, 2016).
- [3] Telemanagement-Forum (2016-09). "GB921 Business Process Framework (eTOM) R16.0.1"; <https://www.tmforum.org/resources/suite/gb921-business-process-framework-etom-r16-0-1/> (accessed Oct. 12, 2016).
- [4] AXELOS (2011). "Information Technology Infrastructure Library (ITILv3)"; <https://www.axelos.com/best-practice-solutions/itil> (accessed Oct. 12, 2016).
- [5] C. Meirosu *et al.*, "DevOps for Software-Defined Telecom Infrastructures," Internet Draft, draft-unify-nfvrg-devops (work in progress), July 2016; <https://datatracker.ietf.org/doc/draft-unify-nfvrg-devops/> (accessed Oct. 12, 2016).
- [6] R. Mijumbi *et al.*, "Network Function Virtualization: State-of-the-Art and Research Challenges," *IEEE Commun. Surveys & Tutorials*, vol. 18, no. 1, 2016, pp. 236–62, doi: 10.1109/COMST.2015.2477041.
- [7] F. Longo *et al.*, "Towards the Future Internet: The RESERVOIR, VISION Cloud, and CloudWave Experiences," *Int'l. J. High Performance Computing and Networking*, vol. 8, no. 3, 2015, pp. 235–47, doi: 10.1504/IJHPCN.2015.071260.

- [8] G. Xilouris *et al.*, "T-NOVA: A Marketplace for Virtualized Network Functions," *Euro. Conf. Networks and Commun.*, Bologna, Italy, 2014, pp. 1–5, doi: 10.1109/EuCNC.2014.6882687.
- [9] W. John *et al.*, "Research Directions in Network Service Chaining," *IEEE SDN for Future Networks and Services*, Trento, Italy, 2013, pp. 1–7, doi: 10.1109/SDN-4FNS.2013.6702549.
- [10] L. De Moura and N. Björner, "Z3: An Efficient SMT Solver," *Int'l. Conf. Tools and Algorithms for the Construction and Analysis of Systems*, Budapest, Hungary, 2008, pp. 337–40, doi: 10.1007/978-3-540-78800-3\_24.
- [11] A. Panda *et al.*, "Verifying Isolation Properties in the Presence of Middleboxes," *CoRR*, vol. abs/1409.7687, 2014; <https://arxiv.org/abs/1409.7687> (accessed Oct. 12, 2016).
- [12] P. Kreuger and R. Steinert, "Scalable In-Network Rate Monitoring," *IFIP/IEEE Int'l. Symp. Integrated Network Management*, Ottawa, Canada, 2015, pp. 866–69, doi: 10.1109/INM.2015.7140396; source code, <https://github.com/nig-sics/ramon> (accessed Oct. 12, 2016).
- [13] W. John *et al.*, "Scalable Software Defined Monitoring for Service Provider DevOps," *4th Euro. Wksp. Software Defined Networks*, Bilbao, Spain, 2015, pp. 61–66, doi: 10.1109/EWSN.2015.62; source code, <https://github.com/acreo/doubledecker> (accessed Oct. 12, 2016).
- [14] I. Pelle *et al.*, "One Tool to Rule Them All: A Modular Troubleshooting Framework for SDN (and Other) Networks," *ACM Sigcomm Symp. SDN Research*, Santa Clara, CA, 2015, pp. 24:1–24:7, doi: 10.1145/2774993.2775014; source code, <https://github.com/nemethf/epoxide> (Oct. 12, 2016).
- [15] S. Spinoso *et al.*, "Formal Verification of Virtual Network Function Graphs in an SP-DevOps Context," *Euro. Conf. Service-Oriented and Cloud Computing*, Taormina, Messina, Italy, 2015, pp. 253–62, doi: 10.1007/978-3-319-24072-5\_18; source code, <https://gitlab.com/mettiu/verigraph/tree/unify> (accessed Oct. 12, 2016).

## BIOGRAPHIES

WOLFGANG JOHN ([wolfgang.john@ericsson.com](mailto:wolfgang.john@ericsson.com)) is a senior research engineer at Ericsson Research in Sweden, working on novel management approaches for SDN, NFV, and cloud environments. He holds M.Sc. degrees from both Salzburg University of Applied Sciences (2001) and Halmstad University (2005), as well as a Ph.D. (2010) in computer engineering from Chalmers University of Technology. He was Technical Work-package Leader in EU FP7 projects SPARC and UNIFY, and has co-authored over 40 scientific papers and patent applications.

GUIDO MARCHETTO ([guido.marchetto@polito.it](mailto:guido.marchetto@polito.it)) is an assistant professor at the Department of Control and Computer Engineering of Politecnico di Torino. He got his Ph.D. in computer engineering in April 2008 from Politecnico di Torino. His research topics cover distributed systems and formal verification of systems and protocols. His interests also include network protocols and network architectures.

FELICIÁN NÉMETH ([nemethf@tmit.bme.hu](mailto:nemethf@tmit.bme.hu)) received his M.Sc. degree in computer science from BME in 2000. He is a research fellow at the Department of Telecommunications and Media Informatics of the same university. He was a member of several national research projects and FP7 EU projects (EFIPANS, OPENLAB, and UNIFY). His current research interests focus on software defined networking, congestion control methods, and autonomic computing.

PONTUS SKÖLDSTRÖM ([ponsko@acreo.se](mailto:ponsko@acreo.se)) holds an M.Sc. in communication systems from KTH Royal Institute of Technology (2008) and has since been a network researcher at Acreo Swedish ICT, in parallel with his graduate studies. At Acreo his research has been focused on network control, virtualization, and monitoring, particularly in the areas of GMPLS and SDN/NFV. He prefers implementation over speculation, and can often be found writing and debugging open source projects.

REBECCA STEINERT ([rebste@sics.se](mailto:rebste@sics.se)) is a senior research scientist and leader of telecom research at SICS Swedish ICT (employed since 2006). Her research group offers expertise in system-oriented solutions based on applied machine learning and data analytics for management of software-defined and virtualized networking infrastructures. From KTH Stockholm she has a B.Sc. in real-time systems (2002), an M.Sc. in autonomous systems and machine learning (2008), and a Ph.D. (2014) in distributed and probabilistic network management.

CATALIN MEIROSU ([catalin.meirosu@ericsson.com](mailto:catalin.meirosu@ericsson.com)) is a master researcher at Ericsson Research in Stockholm, Sweden, working on autonomic management of software-defined infrastructure. He holds a B.Sc. (1999) from Transilvania University in Brasov, Romania, and an M.Sc. (2000) and a Ph.D. in telecommunications (2005) from Politehnica University, Bucharest, Romania. He was a project associate at CERN, Geneva, Switzerland, working on the ATLAS experiment at the Large Hadron Collider. He has 10 granted patents and has co-authored over 50 scientific papers.

IOANNA PAPAFLI ([iopapafi@oterresearch.gr](mailto:iopapafi@oterresearch.gr)) received her B.Eng. in computer and telecommunications from the Polytechnic School of the University of Thessaly in 2006, and her M.Sc. and Ph.D. degrees in computer science from the Department of Informatics of Athens University of Economics and Business (AUEB) in 2008 and 2013, respectively. As a telecommunication engineer in the Hellenic Telecommunications Organization, and previously as a research associate at AUEB, she has participated in several European research projects (FP7, H2020).

KOSTAS PENTIKOUSIS ([k.pentikousis@traveling.com](mailto:k.pentikousis@traveling.com)) has 20 years of experience in the computer networks area. In the past, he has held development, research, and management positions in the United States, Finland, and Germany. As business development manager at Traveling GmbH in Berlin, Germany, he focuses on carrier-grade network functions virtualization and software-defined telecom infrastructures. He holds a Ph.D. in computer science from Stony Brook University.

Moving forward, our ongoing efforts include the definition of metrics to evaluate service performance SP-DevOps tools. Moreover, we are enhancing SP-DevOps with further observability, diagnostic and verification tools and capabilities targeting the application of SP-DevOps in real carrier networks and large-scale deployments.

# IEEE 802.15.7r1 Reference Channel Models for Visible Light Communications

Murat Uysal, Farshad Miramirkhani, Omer Narmanlioglu, Tuncer Baykas, and Erdal Panayirci

The authors present the reference channel models that were endorsed by the IEEE 802.15.7r1 Task Group for evaluation of VLC system proposals. These were developed for typical indoor environments, including home, office, and manufacturing cells. While highlighting the channel models, they further discuss physical layer techniques potentially considered for IEEE 802.15.7r1.

## ABSTRACT

The IEEE has established the standardization group 802.15.7r1 “Short Range Optical Wireless Communications”, which is currently in the process of developing a standard for visible light communication (VLC). As with any other communication system, realistic channel models are of critical importance for VLC system design, performance evaluation, and testing. This article presents the reference channel models that were endorsed by the IEEE 802.15.7r1 Task Group for evaluation of VLC system proposals. These were developed for typical indoor environments, including home, office, and manufacturing cells. While highlighting the channel models, we further discuss physical layer techniques potentially considered for IEEE 802.15.7r1.

## INTRODUCTION

In line with governmental plans worldwide to phase out incandescent and fluorescent lights in favor of more energy-efficient lighting technologies, it is predicted that light emitting diodes (LEDs) will be the ultimate light source in the near future. The expected widescale availability of LEDs in the near future opens the door for so called visible light communication (VLC). VLC involves the dual use of LEDs for illumination and communication purposes. White LEDs can be pulsed at very high speeds without any adverse effect on lighting output and the human eye. Therefore, the existing LED-based illumination infrastructure can be used as wireless access points. Such a ubiquitous wireless access technology is a powerful alternative or complement to radio-frequency counterparts.

VLC, also referred to as LiFi, has been receiving increasing attention from academia and industry (e.g., [1, 2] and the references therein). Recognizing these developments, the IEEE has established the Task Group 802.15.7r1 “Short Range Optical Wireless Communications”, which is currently in the process of developing a standard for LiFi. This standard aims to deliver peak data rates of 10 Gbits per second. It is expected that advanced physical layer techniques such as optical orthogonal frequency division multiplexing (OFDM), multi-input multi-output (MIMO) communications, link adaptation, and relay-assisted transmission will be employed to reach this ambitious goal [3].

The ultimate performance limits of a commu-

nication system are determined by the channel in which it operates. Therefore, realistic channel models are of critical importance for VLC system design, performance evaluation, and testing. In the past, many works were reported on infrared (IR) channel modeling. However, those results cannot be applied to VLC channel modeling in a straightforward manner due to significant differences between IR and visible light (VL) wavelengths. For example, an IR source is typically treated as a monochromatic emitter, while a white LED source is inherently wideband. This necessitates factoring in the wavelength-dependency characteristics of the light source in VLC channel modeling. Furthermore, at IR wavelengths, the reflectance values of surface materials in indoor environments (such as walls, floors, ceilings, etc.) are typically modeled as constant. At VL wavelengths, these are not constant any longer and significantly vary with wavelength.

Initial works on indoor VLC channel modeling have built on some simplifying yet idealistic assumptions, such as ideal Lambertian sources and purely diffuse reflections (see Table 1 of [4] for a comparison of these works). Furthermore, wavelength dependency at VL wavelengths is typically ignored in most works. The most realistic indoor VLC channel modeling at the time of this writing is presented in [4], where accelerated ray tracing features of Zemax® are used to obtain channel impulse responses (CIRs) for various indoor environments. This approach is able to obtain CIRs for any non-ideal source types as well as specular and mixed specular-diffuse reflections. Furthermore, a large number of reflections (more than 10) can be easily handled for better accuracy. Based on the methodology in [4], VLC channel models were developed for typical indoor environments including home, office, and manufacturing cells. These were accepted by the IEEE 802.15.7r1 Task Group as reference models [5, 6] for evaluation of VLC system proposals.

The rest of this article is organized as follows. We provide a brief overview of the VLC channel modeling approach. We describe the reference scenarios under consideration and present the associated CIRs along with fundamental channel parameters such as delay spread and channel DC gain. We discuss potential physical layer techniques suitable for such channels. Finally, we conclude the article.

This work is carried out as an activity of the “Optical Wireless Communication Technologies Excellence Centre” funded by the Istanbul Development Agency (ISTKA) under the Innovative Istanbul Financial Support Program 2015 (TR10/15/YNK-72 OKATEM). The statements made herein are solely the responsibility of the authors and do not reflect the views of ISTKA and/or the T.R. Ministry of Development. The works of Farshad Miramirkhani and Erdal Panayirci are supported by TUBITAK Research Grant No. 113E307.

## CHANNEL MODELING APPROACH

The channel modeling approach is built on three major steps. In the first step, a three dimensional simulation platform is created in Zemax® where the geometry of the indoor environment (i.e., size, shape, etc.), the reflection characteristics of the surface materials (i.e., floor, ceiling, walls, etc.) and the specifications of the light sources and detectors (i.e., field of view, lighting pattern, etc.) are precisely defined. Any objects within the environment such as furniture, human beings, etc. are modeled as CAD objects and imported to the simulation platform.

In the second step, the non-sequential ray tracing features of Zemax® are used to calculate the detected optical power and path lengths from source to detector. In ray tracing, the source emits the rays following a given statistical distribution (the distribution type depends on the source). Rays are then traced along a physically realizable path until they intercept an object. In addition to the line-of-sight (LOS) components, there are a large number of reflections among floor, ceiling, and walls, as well as any other objects within the environment. The Zemax® ray tracing tool generates an output file that includes the detected power and path length for each ray. This data is imported to MATLAB®, and using this information, the CIR is expressed as

$$h(t) = \sum_{i=1}^{N_r} P_i \delta(t - \tau_i) \quad (1)$$

where  $P_i$  is the optical power of the  $i$ th ray,  $\tau_i$  is the propagation time of the  $i$ th ray,  $\delta(t)$  is the Dirac delta function, and  $N_r$  is the number of rays received at the detector.

## REFERENCE SCENARIOS AND ASSOCIATED CIRs

Based on the IEEE 802.15.7r1 Technical Requirements Document [3], four reference scenarios were selected for channel modeling: workplace (open office floor and cubicles); office room with secondary light; living room; and manufacturing cell. In the following, we describe each of these scenarios, present associated CIRs, and discuss the relevant channel parameters.

### SCENARIO 1: WORKPLACE

In the first reference scenario, two workplaces are considered where six office desks with working personnel are located. Both workplaces have identical dimensions: 14 m × 14 m × 3 m. The first workplace has an open office layout (Fig. 1a), while the second workplace (Fig. 1b) has cubicles. Human bodies are modeled as CAD objects with different coating materials for body parts. Specifically, absorptive coating is assumed for their heads and hands along with cotton clothes and black gloss shoes. There are 32 luminaries uniformly located in a rectangular grid in the ceiling. The luminaries used in simulations are commercially available from Cree (LR24-38SKA35). They have a non-ideal Lambertian pattern, a half viewing angle of 40 degrees, and 73 lumens per watt efficacy. The average illumination level is 533 lx, satisfying typical illumination requirements for

<b>Scenario 1</b> Workplace	<b>Walls:</b> plaster; <b>ceiling:</b> plaster; <b>floor:</b> pinewood, <b>cubicles:</b> plaster; <b>desk:</b> pinewood; <b>chair:</b> pinewood; <b>laptop:</b> black gloss paint
<b>Scenario 2</b> Office room with secondary light	<b>Walls:</b> plaster; <b>ceiling:</b> plaster; <b>floor:</b> pinewood; <b>desk:</b> pinewood; <b>chair:</b> black gloss paint; <b>laptop:</b> black gloss paint; <b>desk light:</b> black gloss paint; <b>library:</b> pinewood; <b>window:</b> glass; <b>couch:</b> cotton; <b>coffee table:</b> pinewood
<b>Scenario 3</b> Living room	<b>Walls:</b> plaster; <b>ceiling:</b> plaster; <b>floor:</b> pinewood; <b>tables:</b> wooden; <b>chairs:</b> wooden; <b>couch:</b> cotton; <b>coffee table:</b> glass
<b>Scenario 4</b> Manufacturing cell	<b>Retaining walls:</b> concrete; <b>manufacturing gates:</b> aluminum metal; <b>cell boundaries:</b> plexiglas (PMMA); <b>ceiling:</b> aluminum metal, <b>floor:</b> concrete; <b>robot arm:</b> galvanized steel metal

Table 1. Coating materials for objects within different scenarios.

workplaces [7]. The specific materials for floor, ceiling, walls, and objects within the environment can be found in Table 1.

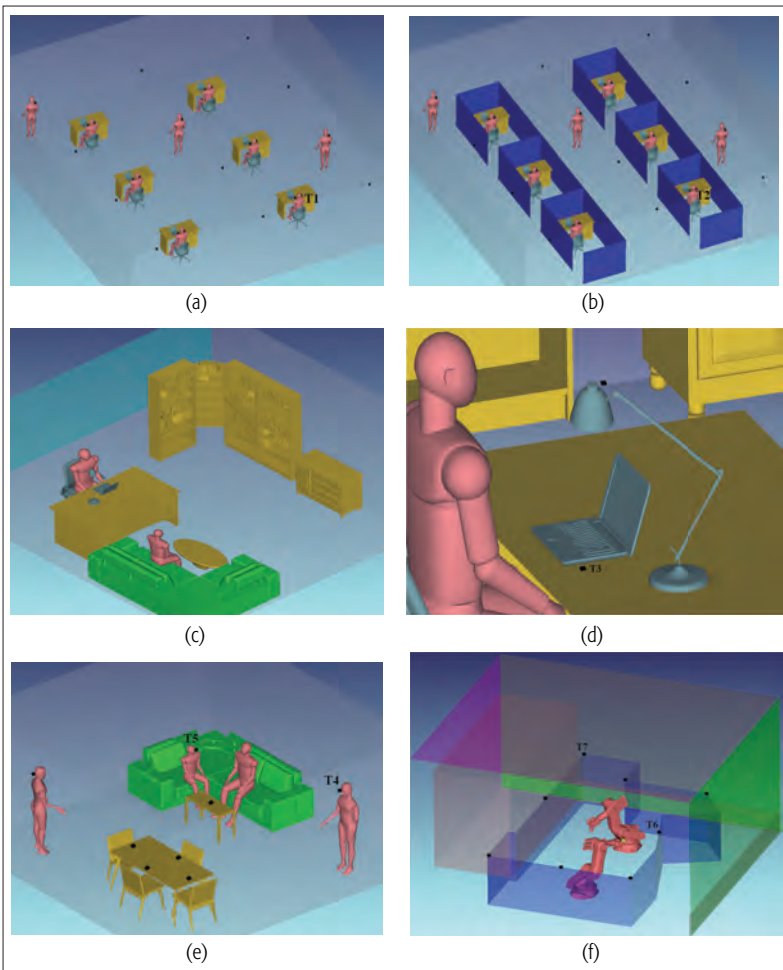
Different locations for the receivers are considered. For example, in one case, a standing person holds a cell phone in their hand next to their ear and the detector is located on the phone (i.e., the detector is at a height of 1.7 m with 45° rotation). In another case, a person works at their desk and holds a cell phone in their hand over their stomach (i.e., the detector is at a height of 0.95 m with 45° rotation). In the third case, the person sits with a cell phone in their hand to their ear (i.e., the detector is at a height of 1.1 m with 45° rotation). The field of view (FOV) and the area of the detector are 85° and 1 cm<sup>2</sup>, respectively.

CIRs for 24 different test points (see “black points” in Fig. 1a and Fig. 1b) can be found in [6]. As an example, two CIRs are presented in Fig. 2a and Fig. 2b. The associated test points are indicated by T1 in Fig. 1.a and T2 in Fig. 1b. These correspond to the person who sits with a cell phone in their hand to their ear. The average DC gain and root mean square (RMS) delay spread (averaged over 24 test points) are  $8.13 \times 10^{-4}$  and 15.27 ns, respectively for workplace with open office. In workplace with cubicles, those values decrease to  $7.51 \times 10^{-4}$  and 12.68 ns, respectively. The decrease in channel DC gain is a result of the presence of cubicle walls. The rays hit the cubicle walls and decay more rapidly than rays in an open office. Similarly, since the rays cannot pass through cubicle walls, delay spread decreases with respect to those in an open office.

### SCENARIO 2: OFFICE ROOM WITH SECONDARY LIGHT

In the second scenario, an office room with two light sources is considered. One of them is the main light source in the ceiling, the other is a desk light. Such a scenario is particularly useful to evaluate the performance of relay-assisted (cooperative) VLC systems [8] where the ceiling light acts as the source and the desk light serves as the relay. The destination receiver is on the desk next to the laptop (T3 in Fig. 1d). This might, for example, take the form of a USB device connected to the laptop. The relay receiver is on top of the desk light with 45° rotation toward the source in the ceiling. The room has dimensions 5 m × 5 m × 3 m. The specifications of luminaries and detectors are the same as those in the first scenario. Other material specifications can be found in Table 1.

The CIRs associated with links from source to



**Figure 1.** Reference scenarios [6]: a) workplace with open office concept; b) workplace with cubicles; c) office room with secondary light; d) enlarged version of (c) showing secondary light, i.e., desklight; e) living room; f) manufacturing cell.

destination ( $S \rightarrow D$ ) and relay to destination ( $R \rightarrow D$ ) are presented in Fig. 2c and Fig. 2d. As expected, the  $R \rightarrow D$  channel is stronger than the  $S \rightarrow D$  channel since the relay transmitter is closer to the destination and therefore experiences smaller path loss. It is also observed that the  $S \rightarrow D$  channel has more scattering components inducing a larger delay spread. Since the distance between source and destination is larger than the distance between relay and destination, the rays coming out from the source hit more surfaces (i.e., wall, floor, and objects inside the room) and result in more scattering.

### SCENARIO 3: LIVING ROOM

In the third scenario, a living room with dimensions  $6 \text{ m} \times 6 \text{ m} \times 3 \text{ m}$  is considered. Four persons are present in the room, two sitting on the couch and two standing. Nine luminaries (Cree CR6-800L) are uniformly located in a rectangular grid in the ceiling. They have a half viewing angle of 40 degrees and 67 lumens per watt efficacy. The average illumination level is calculated as 153 lx.

Similar to Scenario 1, various locations are considered for the receivers. In one case, the person is in a standing position and holds a cell phone in their hand next to their ear. The detector is located on the phone (i.e., the detector is at

a height of 1.7 m with  $45^\circ$  rotation). In another case, a person sits on the couch and holds a cell phone in their hand next to their ear. The detector is located on the phone at a height of 1.1 m with  $45^\circ$  rotation. The detectors on the dinner table are at a height of 0.9 m; the detector on the coffee table is at a height of 0.6 m with  $45^\circ$  rotation toward the person sitting on the couch.

CIRs obtained for eight test points (Fig. 1e) can be found in [6]. As an example, two CIRs are presented in Fig. 2e and Fig. 2f. The associated test points are indicated by T4 and T5 in Fig. 1e. The average DC gain and RMS delay spread (averaged over eight test points) are calculated as  $2.61 \times 10^{-4}$  and 9.24 ns, respectively. It is observed that these are smaller than those observed in the workplace since the room dimensions are now smaller.

### SCENARIO 4 – MANUFACTURING CELL

In the fourth scenario, a manufacturing cell with dimensions  $8.03 \text{ m} \times 9.45 \text{ m} \times 6.8 \text{ m}$  is considered. Six LED transmitters are located at the head of the robotic arm that has the shape of a cube. Each face of the cube is equipped with one transmitter, ensuring  $360^\circ$  coverage. The LEDs are commercially available from Cree (MC-E) with non-ideal Lambertian pattern and a half viewing angle of 60 degrees. The FOV and the area of the detector are  $35^\circ$  and  $1 \text{ cm}^2$ , respectively.

Eight test points are considered on top of the cell boundaries (Fig. 1f). As an example, two CIRs are presented in Fig. 2g and Fig. 2h. The associated test points are indicated by T6 and T7 in Fig. 1f. These are detectors rotated toward the robot arms, which are respectively placed in the middle and at the corner side of the cell boundary. It is observed that the amplitude of T6 is much larger than that of T7 because this detector is closer to the set of transmitters. Since T7 is located at the corner of the cell boundary, it receives more scattering from boundaries. On the other hand, the RMS delay spread of T7 is much larger than that of T6.

### COMPARISONS BETWEEN THE IR AND VL CHANNELS

For the above four scenarios, we have also obtained CIRs assuming the deployment of the IR source. The IR LED is from OSRAM®, operates at 880 nm, and has a non-ideal Lambertian pattern with a half viewing angle of 60 degrees. The average DC gain and RMS delay spread (averaged over test points in each scenario) are provided in Table 2. A comparison of results obtained for the IR and VL channels reveals that DC gains and RMS delay spreads in the VL channels are smaller than those in the IR channels for the same scenarios. This is mainly because reflectivity values in the IR band are larger than those in the VL band, as discussed in [4].

### THE EFFECT OF LED CHARACTERISTICS

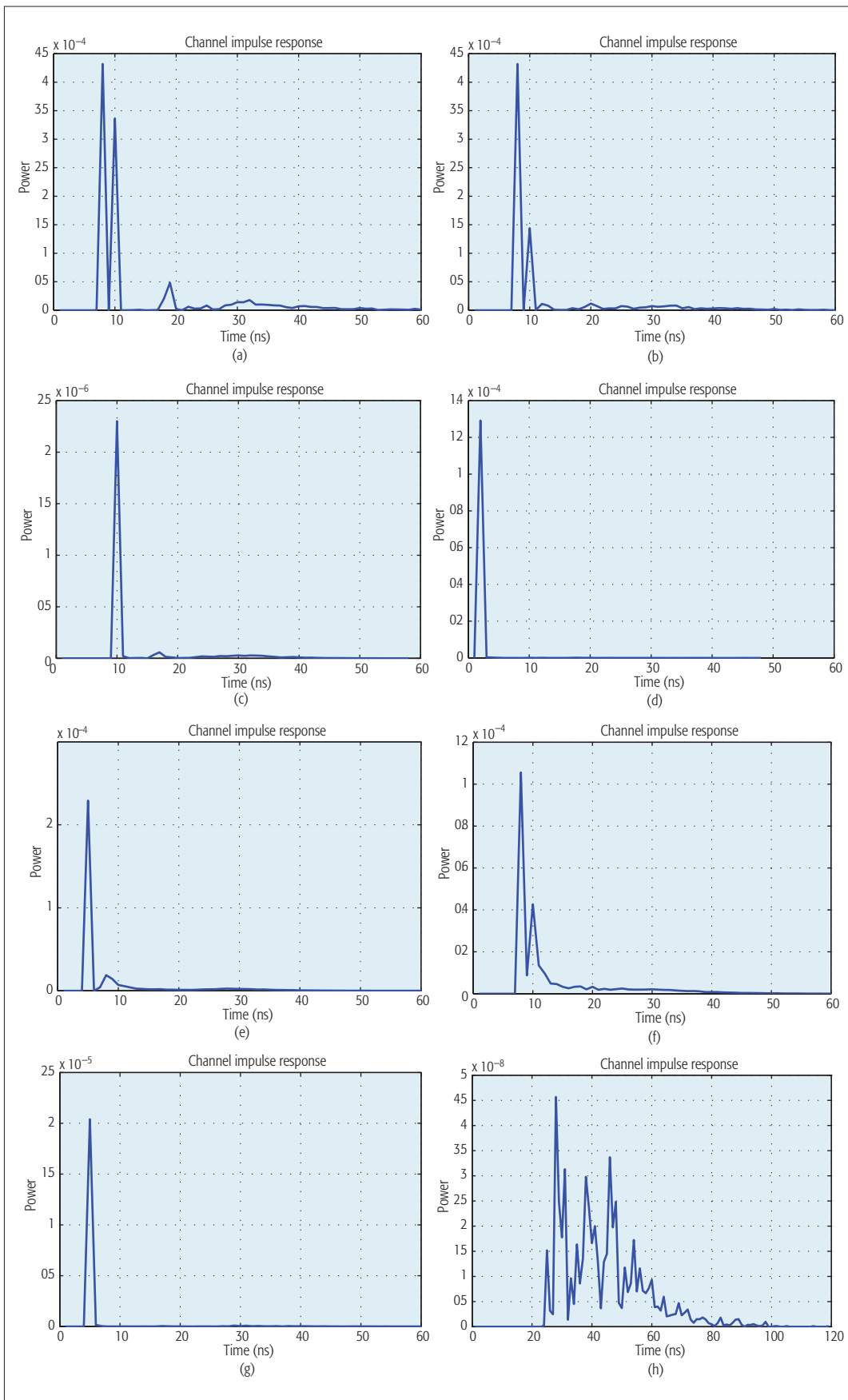
In addition to the multipath propagation environment, the effects of LED sources should be further taken into account in channel modelling. The frequency response of the LED is assumed to be [9]

$$H_{\text{LED}}(f) = \frac{1}{1 + jf/f_{\text{cut-off}}} \quad (2)$$

where  $f_{\text{cut-off}}$  is the LED cut-off frequency. The effective channel frequency response (taking



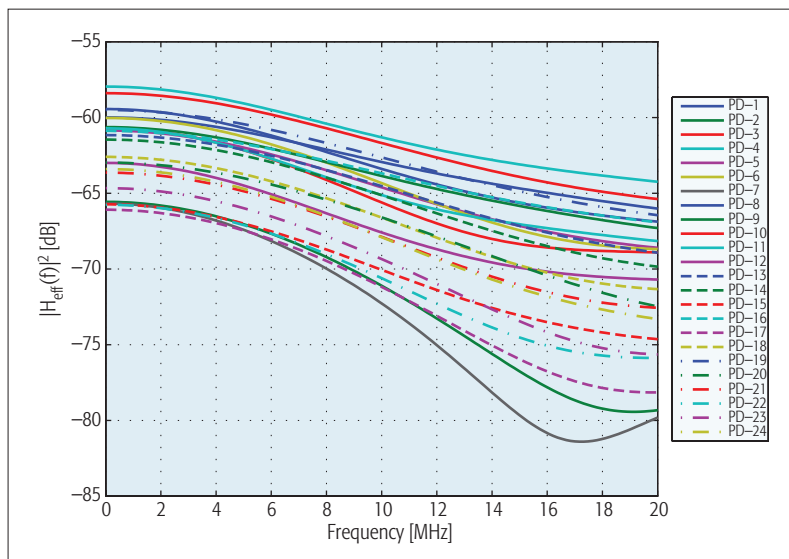
A comparison of results obtained for IR and VL channels reveals that DC gains and RMS delay spreads in VL channels are smaller than those in IR channels for the same scenarios. This is mainly due to the reason that reflectivity values in the IR band are larger than those in the VL band.



**Figure 2.** Channel impulse responses for: a) T1 in workplace with open office concept; b) T2 in workplace with cubicles; c) source to destination in office with secondary light; d) relay to destination in office with secondary light; e) T4 in living room; f) T5 in living room; g) T6 in manufacturing cell; h) T7 in manufacturing cell [6].

Optical bands	VL		IR	
	Delay spread (ns)	DC gain	Delay spread (ns)	DC gain
Workplace with open office concept	15.27	$8.13 \times 10^{-4}$	18.30	$9.04 \times 10^{-4}$
Workplace with cubicles	12.68	$7.51 \times 10^{-4}$	15.15	$8.07 \times 10^{-4}$
Office room with secondary light (R-D)	1.37	$1.30 \times 10^{-4}$	1.65	$1.35 \times 10^{-4}$
Office room with secondary light (S-D)	7.76	$2.81 \times 10^{-6}$	9.50	$2.89 \times 10^{-6}$
Living room	9.24	$2.61 \times 10^{-4}$	11.16	$2.71 \times 10^{-4}$
Manufacturing cell	12.10	$4.81 \times 10^{-6}$	12.58	$5.01 \times 10^{-6}$

**Table 2.** Comparison of VL and IR channels under identical scenarios.



**Figure 3.** Effective channel frequency responses for 24 test points in Scenario 1 taking into account the low-pass filter nature of LED.

into account the LED characteristics) can then be expressed as  $H_{\text{eff}}(f) = H_{\text{LED}}(f) H_{\text{VLC}}(f)$ , where  $H_{\text{VLC}}(f)$  denotes the frequency response of the VLC channel. As an example, the effective frequency channel responses associated with the 24 test points in Scenario 1, denoted as PD1-PD24 [6], are illustrated in Fig. 3, assuming a cut-off frequency of  $f_{\text{cut-off}} = 20$  MHz. It is observed that the low-pass characteristics of the LED result in some attenuation toward higher frequencies.

### PHYSICAL LAYER DESIGN ISSUES

In the discussion of the channel models in the previous section, we discussed physical layer techniques potentially considered for IEEE 802.15.7r1. As observed from the CIRs, the VLC channel is of multipath nature and exhibits frequency selectivity when high data rates are targeted. Taking into account that peak data rates up to 10 Gbit/sec are targeted in the standard, orthogonal frequency division multiplexing (OFDM) has been proposed [10, 11] as an effective mitigation technique to handle intersymbol interference (ISI) resulting from frequency-selectivity. In OFDM,

the high-rate data stream is de-multiplexed and transmitted over a number of frequency subcarriers. If the symbol duration at each sub-band is larger than the delay spread, ISI can be neglected, simplifying the receiver complexity. Different from its radio frequency counterparts, the implementation of optical OFDM requires certain modifications to ensure that the resulting signal is real and non-negative. In the literature, several variants of OFDM were introduced, including direct current optical (DCO)-OFDM, asymmetrically clipped optical (ACO)-OFDM, flip-OFDM, enhanced unipolar OFDM (eU-OFDM), and reverse polarity optical OFDM (RPO-OFDM), among others. In the ongoing standardization work [10, 11], DCO-OFDM is considered to be adopted as the mandatory waveform, while eU-OFDM and RPO-OFDM will possibly be supported as optional waveforms.

The VLC channel is of deterministic nature and does not exhibit any fading. However, the spatial characteristics might vary significantly. For example, large variations in channel gains with respect to test point locations are observed in Scenario 1 (Fig. 3). This necessitates the use of link adaptation, where transmission parameters such as modulation type/size, transmit power, etc., can be selected according to channel conditions. To demonstrate the benefits of link adaptation, we consider an adaptive modulation scheme that aims to maximize the data rate under the constraint of a targeted bit error rate (BER). We assume that binary phase shift keying (BPSK) and M-QAM (quadrature amplitude modulation) with  $M = 4, 8, 16, 32, 64, 128$ , and 256, are available as modulation schemes. We assume a transmit power of  $-30$  dBm and select the highest modulation order that satisfies BER of  $10^{-3}$  for each test point in Scenario 1. For example, at PD4, 64-QAM can be deployed since it satisfies the target BER. However, at PD7, BPSK should be deployed. The peak rates of a user walking through these points are shown in Fig. 4. A non-adaptive system with a worst case design approach provides only 14.33 Mbit/sec, while significant improvements are achieved through adaptive transmission with a peak rate reaching 85.97 Mbit/sec. Different versions of link adaptation in combination with OFDM were already proposed [10, 11] and are currently under discussion within the IEEE 802.15.7r1 Task Group.

IEEE 802.15.7r1 further considers the use of multiple-input multiple-output (MIMO) communication techniques [12]. MIMO deployment is motivated by the fact that an indoor environment typically consists of multiple luminaries. In particular, the use of spatial multiplexing is required to reach the ambitious target of 10 Gbit/sec established in the Technical Requirements Document [3]. Relay-assisted transmission (see Scenario 2) is also considered as an optional PHY mode in the standard to enhance link reliability [13]. This takes advantage of the available secondary light sources in the indoor environments and provides improvements over direct transmission.

### CONCLUSIONS

In this article, we presented an overview of reference VLC channel models adopted by the IEEE 802.15.7r1. These channel models are based on four indoor scenarios: workplace (open office

floor and cubicles); office room with secondary light; living room; and manufacturing cell. For each scenario, the CIRs were presented along with a discussion of fundamental channel characteristics such as delay spread and DC gain. These channels exhibit frequency selectivity, and their characteristics are highly location dependent. These factors motivate the deployment of advanced physical layer techniques such as optical OFDM, MIMO, and link adaptation, which were further discussed.

## REFERENCES

- [1] D. Karunatilaka et al., "LED based Indoor Visible Light Communications: State of the Art," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 3, 3rd Quarter 2015, pp. 1649–78.
- [2] H. Elgala, R. Mesleleh, and H. Haas, "Indoor Optical Wireless Communication: Potential and State-of-the-Art," *IEEE Commun. Mag.*, vol. 49, no. 9, Sept. 2011, pp. 56–62.
- [3] TG7r1, "Technical Considerations Document," doc: IEEE 802.15-15/0492r3, July 2015, available: <https://mentor.ieee.org/802.15/dcn/15/15-15-0492-03-007a-technical-considerations-document.docx>.
- [4] F. Miramirkhani and M. Uysal, "Channel Modeling and Characterization for Visible Light Communications," *IEEE Photon. J.*, vol. 7, no. 6, 2015, pp. 1–16.
- [5] M. Uysal et al., "LiFi Channel Models: Office, Home and Manufacturing Cell," doc: IEEE 802.15-15/0685r0, Sept. 2015, available: <https://mentor.ieee.org/802.15/dcn/15/15-15-0685-00-007a-lifi-reference-channel-models-office-home-manufacturing-cell.pdf>.
- [6] M. Uysal et al., "TG7r1 Channel Model Document for High-Rate PD Communications," doc: IEEE 802.15-15/0746r1, Sept. 2015, available: <https://mentor.ieee.org/802.15/dcn/15/15-15-0746-01-007a-tg7r1-channel-model-document-for-high-rate-pd-communications.pdf>.
- [7] "Lighting of Indoor Work Places," International Standard. ISO 8995:2002 CIE S 008/E-2001.
- [8] R. C. Kizilirmak, O. Narmanlioglu, and M. Uysal, "Relay-Assisted OFDM-based Visible Light Communications," *IEEE Trans. Commun.*, vol. 63, no. 10, 2015, pp. 3765–78.
- [9] L. Grobe and K. D. Langer, "Block-based PAM with Frequency Domain Equalization in Visible Light Communications," *IEEE Globecom Wksp. Optical Wireless Commun.*, 2013, pp. 1070–75.
- [10] D. Tsonev et al., "Low-Bandwidth LiFi PHY & MAC," doc: IEEE 802.15-16/0363r0, May 2016, available: <https://mentor.ieee.org/802.15/dcn/16/15-16-0363-00-007a-text-input-lifi-low-bandwidth-phy-and-mac-d0.docx>.
- [11] V. Jungnickel, "High-bandwidth PHY," doc: IEEE 802.15-16/0356r0, Mar. 2016, available: <https://mentor.ieee.org/802.15/dcn/16/15-16-0356-00-007a-text-input-for-high-bandwidth-phy.docx>.
- [12] M. Uysal et al., "Adaptive MIMO OFDM PHY Proposal for IEEE 802.15.7r1," doc: IEEE 802.15-16/0008r2, Jan. 2016, available: <https://mentor.ieee.org/802.15/dcn/16/15-16-0008-02-007a-adaptive-mimo-ofdm-phy-proposal-for-ieee802-15-7r1.pdf>.
- [13] M. Uysal et al., "PHY Proposal with Relaying Support for IEEE802.15.7r1," doc: IEEE 802.15-16/0020r2, Jan. 2016, available: <https://mentor.ieee.org/802.15/dcn/16/15-16-0020-02-007a-phy-proposal-with-relaying-support-for-ieee-802-15-7r1.pdf>.

## BIOGRAPHIES

MURAT UYSAL ([murat.uysal@ozyegin.edu.tr](mailto:murat.uysal@ozyegin.edu.tr)) is a full professor at Ozyegin University, Istanbul, Turkey, where he leads the Communication Theory and Technologies (CT&T) Research Group. Prior to joining Ozyegin University, he was a tenured associate professor at the University of Waterloo, Canada, where he still holds an adjunct faculty position. Dr. Uysal's research interests are in the broad areas of communication theory and signal processing, with a particular emphasis on the physical layer aspects of wireless communication systems in radio, acoustic, and optical frequency bands.

FARSHAD MIRAMIRKHANI ([fmiramirkhani@yahoo.com](mailto:fmiramirkhani@yahoo.com)) received his B.Sc. and the M.Sc. degree with high honors in electronics and communication engineering from the University of Isfahan, Isfahan,

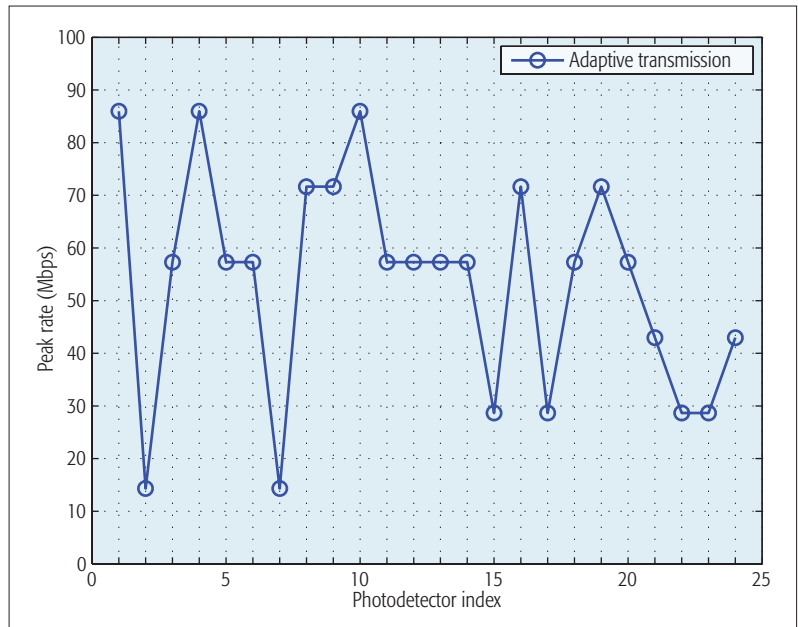


Figure 4. Peak data rates for 24 test points in Scenario 1.

an, Iran, in 2011 and 2014, respectively. He joined the Communication Theory and Technologies (CT&T) Research Group as a research assistant working toward his doctorate under the supervision of Prof. Murat Uysal at Ozyegin University, Istanbul, Turkey, in 2014. The LiFi channels developed by Prof. Murat Uysal and Mr. Miramirkhani were selected as the "LiFi Reference Channel Models" by the IEEE 802.15.7r Task Group during the IEEE's latest meeting held in Bangkok, Thailand, in September 2015. His current research interests include optical wireless communications, indoor visible light communications, underwater visible light communications, and channel modeling.

OMER NARMANLIOGLU ([omernarmanlioglu@gmail.com](mailto:omernarmanlioglu@gmail.com)) received his B.Sc. and the M.Sc. degrees from the Department of Electrical and Electronics Engineering at Bilkent University, Ankara, Turkey, in 2014, and Ozyegin University, Istanbul, Turkey, in 2016, respectively. He is currently working with P.I. Works and pursuing Ph.D. degree at Ozyegin University. His research interests are the physical layer aspects of communication systems and the software-defined networking paradigm for radio access and packet core networks.

TUNCER BAYKAS ([tbaykas@ieee.org](mailto:tbaykas@ieee.org)) works as an assistant professor and head of the Computer Engineering Department at Istanbul Medipol University. From 2007 to 2012, he worked as an expert researcher at NICT, Japan. He has served as co-editor and secretary for 802.15 TG3c, and he has contributed to many standardization projects, including 802.22, 802.11af, and 1900.7. He is the vice director of the "Centre of Excellence in Optical Wireless Communication Technologies (OKATEM)" and vice chair of 802.19 Wireless Coexistence Working Group. He contributed to the technical requirements document and the channel models of 802.15.7r1 standardization, which will enable visible light communication.

ERDAL PANAYIRCI ([eeapanay@khas.edu.tr](mailto:eeapanay@khas.edu.tr)) received the diploma engineering degree in electrical engineering from Istanbul Technical University, Istanbul, Turkey, and the Ph.D. degree in electrical engineering and system science from Michigan State University, East Lansing, MI, USA. Until 1998, he was with the Faculty of Electrical and Electronics Engineering, Istanbul Technical University, where he was a professor and head of the Telecommunications Chair. Currently, he is professor of electrical engineering and head of the Electronics Engineering Department at Kadir Has University, Istanbul. His recent research interests include communication theory, synchronization, advanced signal processing techniques and their applications to wireless electrical, and underwater and optical communications. Presently he is head of the Turkish Scientific Commission on Signals and Systems of the International Union of Radio Science (URSI).

# A Multi-Service Oriented Multiple Access Scheme for M2M Support in Future LTE

Nassar Ksairi, Stefano Tomasin, and Mérouane Debbah

The authors propose a novel multiple access technique to overcome the shortcomings of the current proposals for the future releases of LTE. They provide a unified radio access system that efficiently and flexibly integrates both traditional cellular services and M2M connections arising from IoT applications. The proposed solution, referred to as MOMA, is based on establishing separate classes of users using relevant criteria that go beyond the simple handheld-IoT device split.

## ABSTRACT

We propose a novel multiple access technique to overcome the shortcomings of the current proposals for the future releases of LTE. We provide a unified radio access system that efficiently and flexibly integrates both traditional cellular services and M2M connections arising from IoT applications. The proposed solution, referred to as MOMA, is based on establishing separate classes of users using relevant criteria that go beyond the simple handheld-IoT device split; service-dependent hierarchical spreading of the data signals; and a mix of multiuser and single-user detection schemes at the receiver. Signal spreading in MOMA allows densely connected devices with different QoS profiles to be handled, and at the same time its flexible receiver structure allows the receiver computational resources to be allocated to the connections that need them most. This yields scalable and efficient use of the available radio resources and better service integration. While providing significant advantages for key future communications scenarios, MOMA can be incorporated into LTE with a limited impact on the protocol structure and the signaling overhead.

## INTRODUCTION

The provisioning of Internet of Things (IoT) services is now widely seen in the telecommunications sector as one of the major features in the evolution of cellular systems. Indeed, having a unified cellular system capable of handling both IoT machines and handheld mobile devices would be greatly advantageous for both operators and users. Toward this goal, the design of an integrated radio access solution is a challenging problem. Major issues are the large number of IoT devices requiring to be served simultaneously, and the difference between their traffic patterns/quality of service (QoS) requirements and those of mobile broadband services [1]. Another issue is that IoT applications do not all have the same QoS and traffic characteristics [2], with the implication that future M2M-related system optimization should have inherent flexible support for several types of IoT services.

To address some of these challenges, the Third Generation Partnership Project (3GPP) has started to add machine-to-machine (M2M)-type communications into the radio access subsystem of Long Term Evolution (LTE) starting from Release

12 [3] by introducing a new user equipment (UE) category, Category 0 (Cat. 0). Cat. 0 devices are characterized by their low cost due to their by-design lack of support for high peak rates and multiple antennas. For Release 13 of the LTE standard, 3GPP is working on providing further cost reductions for M2M communications [4]. The proposals that emerged within this work are referred to as *clean-slate narrowband cellular IoT* (NB CIoT) [5], and most of them advocate orthogonal physical layer transmission schemes that are a mixture of frequency-division multiple access (FDMA) and time-division multiple access (TDMA). This is the case, for example, of LTE for Machine-Type Communications (LTE-M) and Narrowband LTE-M (NB LTE-M). Each of these two proposals introduces a new UE category, the so-called Cat. 1.4 MHz for LTE-M and Cat. 200 kHz for NB LTE-M [6]. As their respective names indicate, these new UE categories restrict the device transceiver bandwidth to 1.4 MHz and 200 kHz, respectively. Other proposals focused on upgrading the LTE random access procedure for better support of massive M2M transmissions [7] by overcoming the so-called physical random access channel (PRACH) overloading issue.

## LIMITATIONS OF M2M PROPOSALS FOR THE NEXT LTE RELEASES

None of the existing M2M-related proposals for LTE is able to meet the following crucial requirements all at once.

### MULTI-CLASS USERS/SERVICES

While most of the existing proposals treat IoT devices as a single class of users, not enough attention has been paid to the different QoS and traffic profiles within this class. Indeed, IoT services *will not be limited to data collection from simple sensors and will not only emit small data packets* [2, 4]. One example is mobile video surveillance, which is expected to use medium to high-end devices that do not have battery life constraints [2]. Moreover, there should be a distinction within the services running on handheld devices between mobile broadband applications and the applications that have traffic characteristics and data rate requirements resembling those typical of IoT services. In the latter group we have, for example, the messages generated by social networking and chat applications.

Significant gains in resource utilization efficiency are expected from a multiple access scheme that treats the services with similar QoS profiles, whether running on handhelds or IoT machines, as belonging to the same user class.

### DENSE IOT DEPLOYMENT

The existing M2M proposals for LTE allow its IoT capabilities to be enhanced, but their reliance on FDMA limits the number of simultaneous M2M connections to the (typically moderate) number of frequency sub-channels they reserve for IoT communications. There is a need to support a much larger number of simultaneously connected IoT devices (and possibly mobile services with low QoS requirements). This goal should be met without sacrificing the QoS of mobile broadband services.

### FLEXIBILITY IN RESOURCE ASSIGNMENT

In existing M2M proposals, there is no way to dynamically adjust the respective proportions of resources assigned to broadband services and IoT devices. This could lead to wasting resources or to denials of service depending on the current traffic demands. Moreover, these proposals have limited flexibility in resource allocation within the M2M frequency band, which only comes from varying the number of sub-channels occupied by each device from within a limited number of possible values (six in LTE-M). Any new multiple access scheme should be more flexible in assigning resources to different classes and to users within each class.

### EFFICIENCY IN RESOURCE UTILIZATION

Assigning orthogonal frequency sub-channels to devices with low QoS requirements is not the most efficient way to use the available radio resources. First, this will bound the maximum number of simultaneously connected devices by the number of available sub-channels. Second, it is known that the boundary of the multiple access channel (MAC) capacity region is not achieved with orthogonal transmission schemes. Third, achieving robustness against timing and carrier frequency offsets in time- and/or frequency-division orthogonal schemes requires the use of guard intervals and/or bands around each sub-channel, further reducing their resource utilization efficiency.

In the sequel, we show that multi-service oriented multiple access (MOMA) can overcome these limitations while being compatible with low-cost devices having simple transceivers and long battery life requirements. Indeed, both the (narrow) bandwidth values that were originally proposed for LTE-M (1.4 MHz) and NB LTE-M (200 kHz) as a way to reduce transceiver complexity are supported in MOMA.

### MULTI-SERVICE ORIENTED MULTIPLE ACCESS

MOMA is a multiple access scheme conceived for scenarios where users are grouped into different classes. It reveals all its potentials when the BS is equipped with a large number  $M$  of antennas. In this article we assume that classes are defined based on users' QoS requirement profiles. For example, we define  $L \geq 2$  classes of users as follows.

**One Maximum Data Rate (HD) Class of Users:** Here, HD stands for *high data rate*. For the HD class the objective is to obtain a *data rate as high as possible for  $K^{HD}$  simultaneous transmissions*. Typically, these users are associated with data-hungry applications on handheld devices such as videoconferencing and media streaming.

**$L - 1$  Constant Low-to-Moderate Data Rate (LMD) Classes of Users:** Here LMD stands for *low-to-moderate data rate*. The  $l$ th class, with  $l \in \{1, \dots, L - 1\}$ , includes users requesting services with a relatively low or moderate data rate  $r_l^{LMD}$ . In the sequel we assume that LMD classes are ordered from  $l = 1$  to  $l = L - 1$  with increasing target data rates. Services with low target rates could originate from applications running on either handheld devices (e.g., social messaging) or machines (e.g., M2M light-duty data collection from smart meters and remote sensors). The same applies to services with higher target data rates such as moderate-quality live streaming from handheld devices and M2M heavy-duty data collection from mobile video surveillance machines. Class  $l \in \{1, \dots, L - 1\}$  aims at accommodating the maximum number of simultaneous transmissions  $K_l^{LMD}$  at the granted data rate  $r_l^{LMD}$ .

Since what matters for HD users is maximizing their respective throughput, proper scheduling techniques will typically limit the number of simultaneous HD transmissions, exactly as in current wireless communications standards. It is thus reasonable to assume that  $K^{HD}$  is small and that multiuser multiple-input multiple-output (MU-MIMO) techniques implemented on top of orthogonal FDMA (OFDMA) in the downlink and single-carrier FDMA (SC-FDMA) in the uplink can be used for HD/HD signal separation. We also propose the use of these frequency domain transmission schemes for the separation between the HD and other user classes. This choice allows full compatibility with the LTE standard to be maintained. As for the  $L - 1$  LMD classes, due to both their specific data rate requirements and the objective of massive M2M deployment, we propose to overload radio resources. Note that this overloading can be achieved, for instance, by MU-MIMO techniques also operating in the code domain. The way we propose to access the code domain is dubbed *service-dependent hierarchical spreading*, which can be thought of as a *layered* or *hierarchical* spreading with a class-dependent overloading factor. The MOMA uplink transmission scheme is illustrated on the left of Fig. 1.

### MOMA FEATURES

- MOMA is based on service-dependent hierarchical spreading. This new transmission scheme has the advantages of efficiently using available resources, being scalable with the number of connected devices, and allowing flexible resource allocation among the different user classes.

- MOMA can easily be integrated into LTE systems as it can be implemented on a sub-band of the LTE resource grid without affecting the legacy connections occupying the rest of the bandwidth. Furthermore, MOMA signals can be transmitted using modulation and coding schemes (MCSs) and transport block (TB) sizes that are taken from the LTE standard. Finally, MOMA can make use of some advanced features of LTE such as transmission time interval (TTI) bundling.

MOMA is based on service dependent hierarchical spreading. This new transmission scheme has the advantages of efficiently using available resources, being scalable with the number of connected devices and allowing flexible resource allocation among the different user classes.

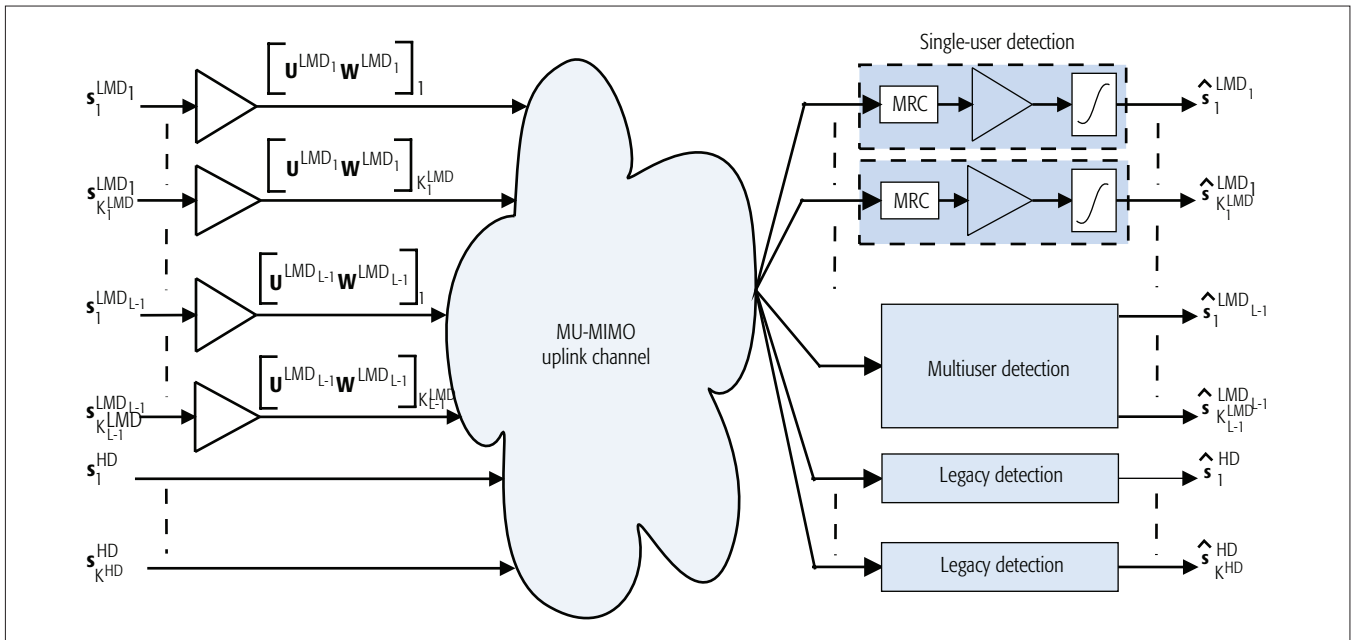


Figure 1. MOMA transceivers for  $L$  classes of users.  $[\mathbf{M}]_j$  designates the  $j$ th column of matrix  $\mathbf{M}$ .

- A narrowband implementation of MOMA is possible, thus making it compatible with low-cost battery constrained devices.

- MOMA exploits massive MIMO to both increase user multiplexing capabilities and simplify the receiver structure. Thanks to these properties, low-complexity detection is used for most of the users in MOMA, while the more complex detection schemes are only applied for the highest data rate classes.

- Using MOMA entails a gain in coverage, making it advantageous for connecting devices deployed in remote or bad coverage areas.

- Several random access mechanisms with different degrees of signaling overhead are compatible with MOMA.

With these features, which are discussed in more detail later, MOMA can overcome the shortcomings of the current proposals for the evolution of LTE in supporting IoT. For instance, features 1, 4, and 6 are essential for enabling real massive M2M deployments.

### SERVICE DEPENDENT HIERARCHICAL SPREADING

For the sake of clarity, we only consider the case  $L = 3$  from now on. The first LMD class ( $l = 1$ ) will be referred to simply as the LD class, where LD stands for *low data rates*, with target data rate  $r^{LD}$ . Similarly, the second LMD class ( $l = 2$ ) will be referred to as the MD class, where MD stands for *moderate data rates*, with target data rate  $r^{MD}$ .

To get a more precise description of MOMA, let  $\mathbf{U}$  be an  $N \times N$  code matrix (e.g., a Walsh-Hadamard matrix or a discrete Fourier transform matrix). In MOMA, the set of columns of matrix  $\mathbf{U}$  is divided into two disjoint subsets, matrices  $\mathbf{U}^{MD}$  and  $\mathbf{U}^{LD}$ , with dimensions  $N \times N^{MD}$  and  $N \times N^{LD}$ , respectively. Now assume that a maximum number  $K^{MD}$  ( $K^{LD}$ ) of simultaneously connected MD (LD) users are to be served within the current sub-frame. Since we want to overload the MD and LD radio resources, we typically have  $K^{MD} > N^{MD}$  and  $K^{LD} > N^{LD}$ . Finally, since

we want to guarantee higher data rates for the MD class compared to the LD class, we impose  $K^{LD}/N^{LD} > K^{MD}/N^{MD}$ .

Instead of assigning the orthogonal spreading codes to individual users, the  $N^{MD}$  ( $N^{LD}$ ) columns of  $\mathbf{U}^{MD}$  ( $\mathbf{U}^{LD}$ ) are simultaneously used in MOMA by the  $K^{MD}$  ( $K^{LD}$ ) users. Indeed, each MOMA transmitter applies as spreading code a linear combination of the columns of the code matrix corresponding to its class. The coefficients of this linear combination serve as a signature sequence to separate the signals of the users belonging to the same class. More precisely, the data symbols of each MD (LD) user are spread using one column of the product matrix  $\mathbf{U}^{MD}\mathbf{W}^{MD}$  ( $\mathbf{U}^{LD}\mathbf{W}^{LD}$ ) where  $\mathbf{W}^{MD}$  ( $\mathbf{W}^{LD}$ ) is an *overloading* matrix of dimensions  $N^{LD} \times K^{LD}$  ( $N^{LD} \times K^{LD}$ ) the columns of which are referred to in the sequel as the *overloading sequences*. In principle,  $\mathbf{W}^{MD}$  ( $\mathbf{W}^{LD}$ ) can be constructed by selecting  $K^{MD}$  ( $K^{LD}$ ) points from the surface of the  $N^{MD}$ -dimensional ( $N^{LD}$ -dimensional) complex sphere with radius 1. The resulting spread symbols of each user are then mapped to the elements of the OFDMA/SC-FDMA time-frequency grid elements that fall within the frequency band assigned to the MD and LD classes before being transmitted on the radio channel. Finally, since the BS is equipped with a number  $M > 1$  of antennas, we know from the literature [8] that the effective spreading gain of MD transmissions (LD transmissions) is, roughly speaking, proportional to  $M N^{MD}$  ( $M N^{LD}$ ). This intuition was confirmed by the analysis done in [9].

The main advantage of this multiple access scheme is an *efficient, scalable, and flexible* use of the available radio resources. MOMA *efficiency* is shown by its overloading of the available radio resources in order to connect a large number of IoT machines and of handhelds requiring low to moderate data rates. The *scalability* of MOMA with respect to increasing device densities is simply a matter of applying a larger value

for the MD (LD) overloading factor defined as  $K^{MD}/N^{MD}$  ( $K^{LD}/N^{LD}$ ) and/or of employing a larger  $N$ . MOMA flexibility is manifested by the ease with which the network can dynamically adjust the proportion of resources assigned to each class of devices/services and the degree to which these resources are overloaded with in each class by means of simply updating the values of parameters  $N^{MD}$ ,  $K^{MD}$ ,  $N^{LD}$ , and  $K^{LD}$ . Finally, by properly mapping MOMA signals to the LTE time-frequency resource grid, MOMA combines the benefits of both OFDM (e.g., robustness against timing errors) and frequency-domain spreading (e.g., the ability to harvest the frequency diversity of the channel and the robustness against carrier frequency shifts).

## MOMA INTEGRATION INTO FUTURE LTE

Let  $B$  be the total system bandwidth and denote by  $B^{HD}$  ( $B^{LMD}$ ) the bandwidth assigned to the HD (MD and LD) class such that  $B^{HD} + B^{LMD} = B$ . In order to apply MOMA in the future evolution of LTE, we need to set  $B^{HD}$  and  $B^{LMD}$ , the spreading factor  $N$ , and the map of the spread data symbols to the OFDMA/SC-FDMA time-frequency grid. We also need to determine which new signaling messages could be needed and what effect MOMA could have on the LTE system protocols. First, let us recall how the available time-frequency resources are structured into sub-frames in LTE. The smallest item in the time-frequency grid in LTE is the resource element (RE) defined as one subcarrier within one OFDM symbol of a duration equal to 66.7  $\mu$ s. However, the basic unit for scheduling and resource allocation is the physical resource block (PRB), which is composed of 12 REs in 14 consecutive OFDM symbols covering 180 kHz over 1 ms. The duration of the basic period of data scheduling in LTE, called a sub-frame, is also equal to 1 ms. Finally, the duration of one sub-frame is also referred to as the data transmission time interval (TTI).

### MOMA ON 1.4 MHz BANDWIDTH

One possible implementation of MOMA provides that  $B^{LMD}$  coincides with the frequency band occupied by the six PRBs that are destined for M2M communications in LTE-M. This implementation is illustrated in Fig. 2. In this implementation, at the beginning of each TTI each active MD or LD transmitter extracts from its transmission queue enough data bits that, after coding and mapping, results in a number of data symbols equal to the size of one PRB. The motivation behind this choice is to retain full compatibility with LTE MCSs and TB sizes. Next, each data symbol is spread using service-dependent hierarchical spreading with  $N = 6$ . The resulting spread symbol is finally mapped to six consecutive resource elements (REs) from one PRB, as shown in Fig. 2.

### MOMA ON 200 kHz BANDWIDTH

Another possible MOMA implementation is obtained by letting  $B^{LMD}$  coincide with the one PRB reserved for M2M communications in NB LTE-M. Except for the difference in bandwidth and resource allocation granularity, this implementation is not different from the 1.4 MHz implementation.

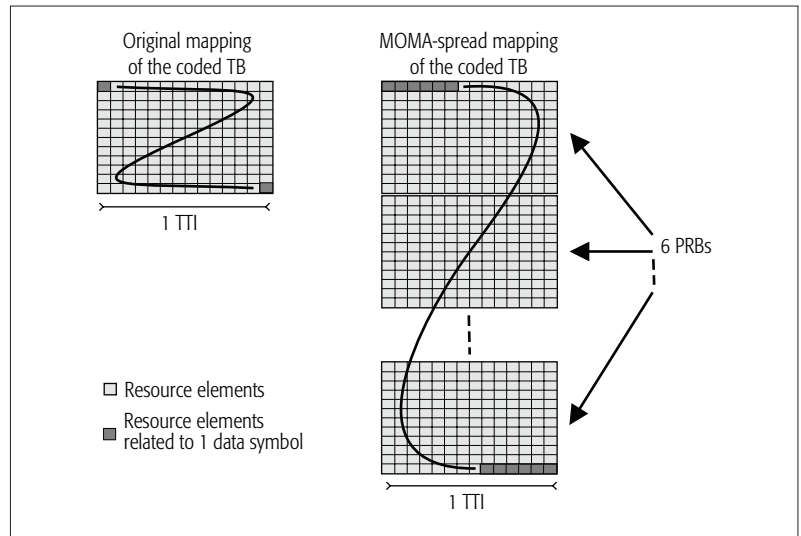


Figure 2. Implementing MOMA on six PRBs.

## MOMA WITH TTI BUNDLING

TTI bundling is a transmission technique that was originally proposed for coverage enhancement in delay-limited applications such as voice over LTE (VoLTE) [10]. It consists of allowing users to transmit the four redundancy versions (RVs) of their current codewords in one shot using four consecutive TTIs, instead of waiting for the BS acknowledgment (ACK) or negative ACK (NACK) message after each RV transmission. When applied along with MOMA, the already existing control messages and protocol structure that were introduced to support TTI bundling can be re-purposed in order to increase the multiplexing capabilities for IoT devices while getting a coverage enhancement gain. In a 1.4 MHz implementation of MOMA with TTI bundling, the MD and LD spreading code length is  $4N$ , where  $N$  is the original spreading factor without TTI bundling. Each sequence resulting from spreading an MD or LD symbol is mapped to  $4N$  consecutive REs from one PRB, as shown in Fig. 3.

### MOMA RECEIVER COMPLEXITY AND PERFORMANCE

We propose a receiver structure for MOMA that takes advantage of the large number,  $M \gg 1$ , of BS antennas. This massive MIMO scenario, which is expected to be prevalent in next-generation cellular networks, proves to be advantageous for MOMA from both the performance and receiver complexity perspectives. The proposed receiver is illustrated on the right of Fig. 1 and consists in performing the following two steps.

**Spatial Combining:** When the BS has multiple antennas, linear receive combining in the uplink is typically utilized. When the number of BS antennas  $M$  is large enough, maximum ratio combining (MRC), which is a low-complexity scheme, has been shown [11] to achieve a spectral efficiency not far from that achieved with more involved linear combining methods thanks to the asymptotic (with respect to the number of BS antennas) orthogonality of users' channel vectors in massive MIMO. We thus propose to apply MRC for the detection of both LD and MD spread signals on the chip level.

**Code Spreading:** Since the target data rate  $r^{\text{MD}}$  for the MD class is relatively high as compared to  $r^{\text{LD}}$ , the number of devices that can be simultaneously served in that class is expected to be smaller than its LD counterpart. The BS can thus typically afford for this class the use of multiuser detection techniques such as successive interference cancellation (SIC). On the other hand, we consider only single-user detection for the LD class. This choice is motivated by the need to maintain a reasonable detection complexity while serving a large number of LD users with low target data rates.

Interestingly, this simple receiver structure, which does not involve any inter-class multiuser detection, was shown [9] to achieve asymptotic MD/LD orthogonality even on fast-varying frequency-selective channels and even when the number  $K^{\text{LD}}$  of LD users grows to infinity. This is due to a favorable property of massive MIMO channels. Indeed, as an effect of combining a large number  $M$  of signals in the case of a rich scattering propagation environ-

ment, the small-scale fading averages out over the array in the sense that the variance of the resulting scalar channel decreases with  $M$ . This effect is known as channel hardening and is a consequence of the law of large numbers [11]. Most importantly, in our case, the frequency response of the effective channel is asymptotically flat and asymptotically constant over several consecutive OFDM symbols, as illustrated in Fig. 4 when  $M = 100$ . The channel realizations used in this figure were generated using the Extended Type Urban (ETU) channel model [12].

In Figs. 5 and 6, we plotted the number of MD and LD connections that can be simultaneously served with and without TTI bundling, respectively, as a function of their respective target data rates ( $r^{\text{MD}}$  and  $r^{\text{LD}}$ ) for both NOMA and LTE-M. The values of  $r^{\text{MD}}$  are taken from the range [30, 60] kb/s, while  $r^{\text{LD}} \in [10, 25]$  kb/s. The higher value in these two intervals is dictated by the maximum per-link data rate achievable with orthogonal (thus underloaded) access schemes. From the figures we can notice the significant advantage of using NOMA as opposed to orthogonal access NB-IoT solutions in terms of the capability of serving densely deployed IoT devices. For instance, four times more simultaneous MD and LD connections can be served while both  $r^{\text{MD}}$  and  $r^{\text{LD}}$  are approximately at half their respective upper bound values. Note that this performance has been obtained on doubly dispersive channels generated using the Extended Vehicular A (EVA) model [12], which is characterized with a relatively long delay spread and short coherence time. Also note that the four times larger spreading gain resulting from TTI bundling allows four times more MD and LD simultaneous connections to be served while meeting their respective target data rates, in the same range as in the absence of TTI bundling. The figures were obtained using 100 realizations of users' distances to the BS randomly chosen in the interval [25, 100] m assuming that the BS is equipped with  $M = 64$  antennas and that users' transmit power is equal to 23 dBm.

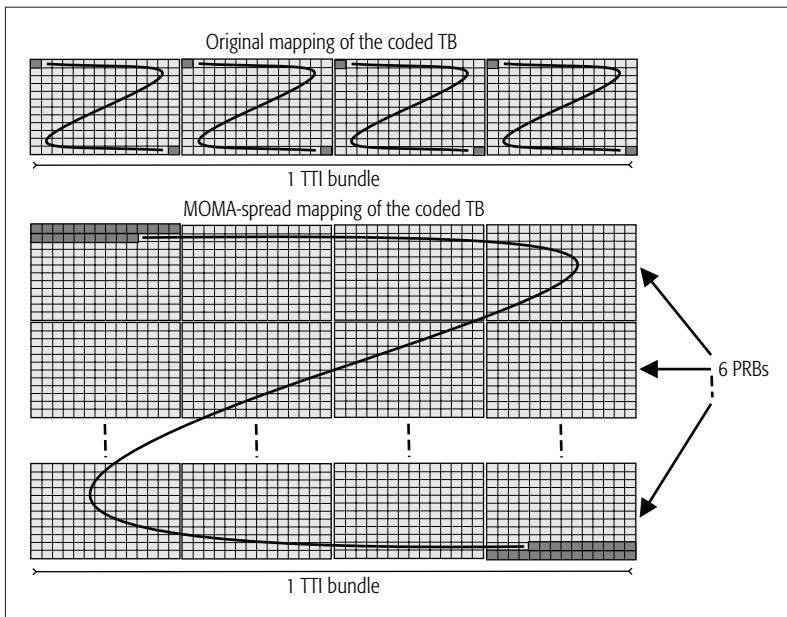


Figure 3. Implementing NOMA on six PRBs with TTI bundling.

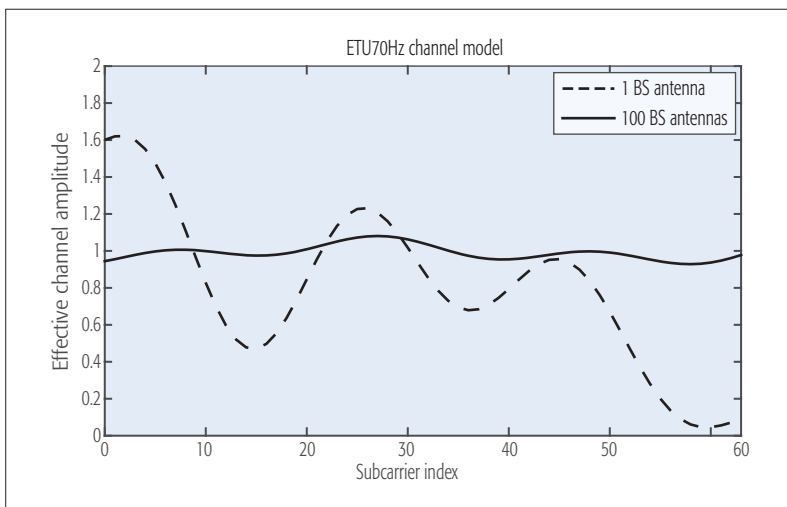


Figure 4. Channel hardening effect in MRC combining.

### COVERAGE ENHANCEMENT WITH NOMA

The coverage measure adopted for LTE channels is the *maximum coupling loss* (MCL), defined as the difference in logarithmic scale between the maximum transmission power and the receiver sensitivity [10]. A higher MCL value indicates that the transmitter-receiver distance can be made larger while still meeting the target received signal-to-noise ratio (SNR) and hence the target block error rate (BLER). This translates into better cellular service coverage.

As a consequence of the  $N$ -long spreading used in NOMA, IoT users benefit from a coverage enhancement gain. This gain can be further increased if TTI bundling is applied due to the higher processing gain (equal to  $4N$ ) resulting in this case from spreading over a larger number of resource blocks. Indeed, the increase in MCL due to the use of NOMA is on the order of  $10\log_{10}N$  without TTI bundling and of  $10\log_{10}(4N)$  with TTI bundling, corresponding to a gain of 7.78 dB and 13.80 dB, respectively.



## SIGNALING AND RANDOM ACCESS FOR MOMA

To implement MOMA, matrix  $\mathbf{U}$  should be known in advance to all MD and LD users, for instance, in the form of a lookup table (LUT). Moreover, the resource allocation parameters  $N^{\text{MD}}$  and  $N^{\text{LD}}$  are dynamic and need to be broadcast by the BS. This can be done by making use of the LTE broadcast control channel (BCH). Note that the values of  $N^{\text{MD}}$  and  $N^{\text{LD}}$  can typically be kept fixed for a relatively long time as they only need to be changed when the traffic characteristics change. As for the overloading sequences, their allocation is intimately related to the kind of random access (RA) scheme we choose.

**Contention-Free RA:** In a contention-free access scheme, the BS is in full control of the assignment of the available radio resources to the active users in the cell area [7]. In MOMA, this translates into the BS choosing which MD (LD) user is assigned which column of the overloading matrix  $\mathbf{W}^{\text{MD}}$  ( $\mathbf{W}^{\text{LD}}$ ). Note that these two matrices could be generated in advance (i.e., offline) for different combinations of the values of  $N^{\text{MD}}$ ,  $N^{\text{LD}}$ ,  $K^{\text{MD}}$ , and  $K^{\text{LD}}$  and made available in the form of a LUT to the relevant users. Contention-free RA has the advantage of eliminating collisions among concurrent connections, and hence eliminating the need for collision detection and resolution. However, this comes at the price of a relatively large protocol overhead, especially in the case of a highly overloaded system. Contention-free RA is thus more suitable for systems with low to moderate user densities.

**Contention-Based RA:** In a contention-based access scheme, we let MD and LD users compete within their respective classes for the overloading sequences. On one side, this helps cut the protocol overhead, thus making contention-based RA relevant for systems with high user densities. On the other side, it comes at the price of eventual collisions between concurrent uplink transmissions and ensuing retransmissions. A conflict/collision occurs when at least two MD or LD users choose the same overloading sequence, making it difficult for the BS to correctly detect their transmitted data symbols. In contention-based MOMA, there is no need to store LUTs corresponding to the overloading matrices as the overloading sequences could be locally generated when needed. In this case, collision resolution is left entirely to the higher network layers. Otherwise, the probability of collisions could be reduced by using preamble transmission as in LTE PRACH and/or by using contention transmission unit (CTU) messages [13] that have the overloading sequence as one of their fields.

**Hybrid RA Scheme:** Interestingly, LD collisions that take place in contention-based MOMA cannot affect MD connections, and neither MD nor LD collisions can affect HD (legacy) connections thanks to inter-class quasi-orthogonality in MOMA. This observation can be used to motivate the use of a contention-free RA scheme for HD and MD connections and a contention-based RA scheme for LD connections. Such a hybrid scheme has the advantage of reducing protocol overhead in systems characterized by a relatively high LD and a lower MD device density.

Note that in all three cases uplink synchroniza-

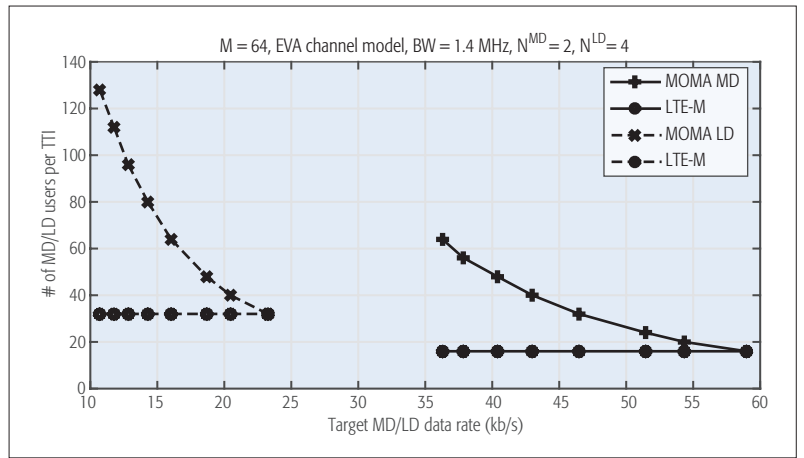


Figure 5. Number of served MD/LD users vs. target data rate without TTI bundling.

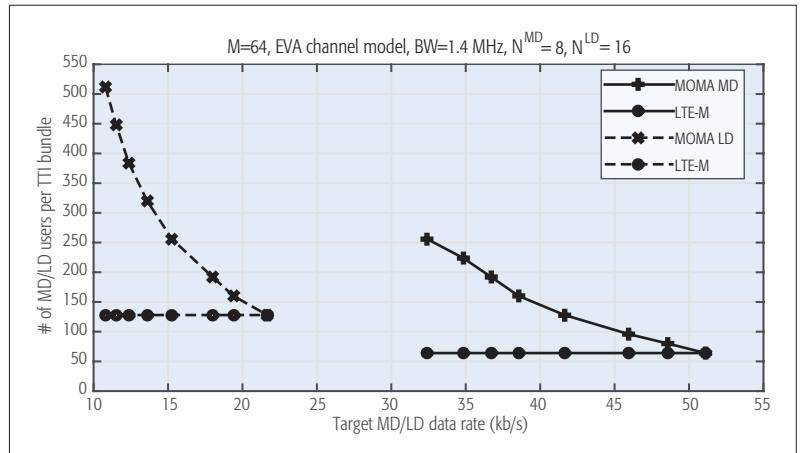


Figure 6. Number of served MD/LD users vs. target data rate with TTI bundling.

tion is needed before initiating the RA procedure and the actual data transmission. As in LTE, this can be maintained with a timing advance procedure [14].

## CONCLUSIONS AND PERSPECTIVES

MOMA is a novel multiple access scheme compatible with massive MIMO, which can be integrated into the evolution of LTE to enhance its support for a wide range of services including M2M communications. MOMA is based on assigning, in a flexible and dynamic manner, different code resources and different degrees of resource overloading to different classes of users, each representing a different data rate requirement, a different service type, and/or a different traffic pattern. Code assignment in MOMA is conceived in such a way that overloading the resources of the lower data rate classes would only slightly affect the higher data rate classes, dropping the need for wasteful guard bands and steep transmit filters for uplink transmission. Moreover, the different QoS requirements in the system can be satisfied in a flexible and efficient fashion by reserving higher-complexity detection schemes at the BS only for classes that need them. Finally, we show that MOMA outperforms the other M2M-related proposals for LTE in most

One research direction for MOMA consists in analyzing its performance using different MIMO channel models reflecting the diverse propagation environments and practical BS array configurations. Another research direction is conducting a higher-level assessment of MOMA using advanced data traffic models.

of the M2M-relevant performance measures while requiring comparable signaling and protocol overhead.

One research direction for MOMA consists of analyzing its performance using different MIMO channel models reflecting the diverse propagation environments and practical BS array configurations. Another research direction is conducting a higher-level assessment of MOMA using advanced models, such as those introduced in [15], for the data traffic generated by the different classes of users. Finally, the integration into MOMA of other coverage enhancement techniques such as relaying has yet to be investigated.

#### REFERENCES

- [1] L. Vangelista, A. Zanella, and M. Zorzi, "Long-Range IoT Technologies: The Dawn of LoRaTM," *FABULOUS*, Ohrid, Sept. 2015, pp. 51–58.
- [2] A. Meader and P. Rost, "The Challenge of M2M Communications for the Cellular Radio Access Network," *EuroView*, Aug., 2011, pp. 1–2.
- [3] 3GPP Tech. Rep. 36.888, "Study on Provision of Low-Cost Machine-Type Communications (MTC) User Equipments (UEs) Based on LTE, v. 12.0.0," June 2013.
- [4] 3GPP Tech. Rep. 22.368, "Service Requirements for Machine-Type Communications (MTC); Stage 1, v. 13.0.0," Dec. 2014.
- [5] W. Guibene, K. E. Nolan, and Mark Y. Kelly, "Survey on Clean Slate Cellular-IoT Standard Proposals," CIT/IUCC/DASC/PICOM, Liverpool, U.K., Oct. 2015, pp. 1596–99.
- [6] R. Ratasuk et al., "Narrowband LTE-M System for M2M Communication," VTC-Fall, Vancouver, Sept. 2014, pp. 1–5.
- [7] M. Hasan, E. Hossain, and D. Niyato, "Random Access for Machine-to-Machine Communications in LTE-Advanced Networks: Issues and Approaches," *IEEE Commun. Mag.*, vol. 51, no. 6, June 2013, pp. 86–93.
- [8] S. V. Hanly and D. Tse, "Resource Pooling and Effective Bandwidths in CDMA Networks with Multiuser Receivers and Spatial Diversity," *IEEE Trans. Info. Theory*, vol. 47, no. 4, May 2001, pp. 1328–51.
- [9] N. Ksairi, S. Tomasin, and M. Debbah, "A Multi-Service Oriented Multiple-Access Scheme for Next-Generation Mobile Networks," *EUCNC*, Athens, Greece, June 2016.
- [10] G. Naddafzadeh-Shirazi et al., "Coverage Enhancement Techniques for Machine-to-Machine Communications over LTE," *IEEE Commun. Mag.*, vol. 53, no. 7, July 2015, pp. 192–200.
- [11] E. Björnson, E. G. Larsson, and T. L. Marzetta, "Massive MIMO: Ten Myths and One Critical Question," *IEEE Commun. Mag.*, vol. 54, no. 2, Feb. 2016, pp. 114–23.
- [12] 3GPP Tech. Rep. 36.104, "Evolved Universal Terrestrial Radio Access (E-UTRA); Base Station (BS) Radio Transmission and Reception, v. 13.3.0," Mar. 2016.
- [13] K. Au et al., "Uplink Contention Based SCMA for 5G Radio Access," *Proc. IEEE GLOBECOM*, July 2014, pp. 900–05.
- [14] F. Khan, *LTE for 4G Mobile Broadband: Air Interface Technologies and Performance*, Cambridge Univ. Press, 2009.
- [15] D. Niyato, P. Wang, and D. I. Kim, "Performance Modeling and Analysis of Heterogeneous Machine Type Communications," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, May 2014, pp. 2836–49.

#### BIOGRAPHIES

NASSAR KSAIRI received his M.Sc. degree from CentraleSupélec, France, in 2006 and his Ph.D. degree from the University of Paris-Sud XI, France, in 2010. From 2010 to 2012, he was an assistant professor at the Higher Institute for Applied Sciences and Technology, Damascus, Syria. From 2012 to 2014, he was a postdoctoral researcher at Télécom ParisTech, France. Since December 2014, he has been a researcher at Huawei's Mathematical and Algorithmic Sciences Lab, France.

STEFANO TOMASIN received his Ph.D. in 2003 from the University of Padova, Italy, which he then joined in 2002, and is currently an associate professor. He has spent leaves and sabbaticals at Qualcomm in California, Polytechnic University in New York, and the Mathematical and Algorithmic Sciences Laboratory of Huawei Technologies, France. Since 2011 he is an Editor of both *IEEE Transactions of Vehicular Technologies* and *EURASIP Journal of Wireless Communications and Networking*.

MÉROUANE DEBBAH entered the Ecole Normale Supérieure de Cachan, France, in 1996, where he received his M.Sc and Ph.D. degrees. He worked for Motorola Labs, Saclay, France, from 1999 to 2002 and the Vienna Research Center for Telecommunications, Austria, until 2003. From 2003 to 2007, he joined the Mobile Communications Department of the Institut Eurecom, Sophia Antipolis, France, as an assistant professor. Since 2007, he is a full professor at CentraleSupélec. From 2007 to 2014, he was the director of the Alcatel-Lucent Chair on Flexible Radio. Since 2014, he is vice-president of the Huawei France R&D Center and director of the Mathematical and Algorithmic Sciences Lab.

Now...

# 2 Ways to Access the IEEE Member Digital Library

With two great options designed to meet the needs—and budget—of every member, the IEEE Member Digital Library provides full-text access to any IEEE journal article or conference paper in the IEEE *Xplore*<sup>®</sup> digital library.

Simply choose the subscription that's right for you:

## IEEE Member Digital Library

Designed for the power researcher who needs a more robust plan. Access all the IEEE content you need to explore ideas and develop better technology.

- 25 article downloads every month

## IEEE Member Digital Library Basic

Created for members who want to stay up-to-date with current research. Access IEEE content and rollover unused downloads for 12 months.

- 3 new article downloads every month

Get the latest technology research.

**Try the IEEE Member Digital Library—FREE!**

[www.ieee.org/go/trymdl](http://www.ieee.org/go/trymdl)



IEEE Member Digital Library is an exclusive subscription available only to active IEEE members.

---

# 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 their productivity. Using the NI software defined radio platform and LabVIEW Communications, leading researchers are innovating faster and building 5G proof-of-concept systems to demonstrate new technologies first.

---

Accelerate your innovation at [ni.com/5g](https://ni.com/5g)



LabVIEW Communications System Design Software, USRP-2943R SDR Hardware

