MULTIMEDIA COMMUNICATIONS TECHNICAL COMMITTEE
http://www.comsoc.org/~mmc

MMTC Communications - Frontiers

Vol. 13, No. 4, July 2018

CONTENTS

Message from the MMTC Chair

Dear MMTC colleagues and friends,

It is my great honor to serve as MMTC Chair for 2018 ~ 2020 and I look forward to working with the new team of officers below:

Steering Committee chair: Sanjeev Mehrotra, Microsoft Research, US
Vice-chairs:
America: Pradeep K Atrey, University at Albany, State University of New York, USA
Asia: Wanqing Li, University of Wollongong, Australia
Europe: Lingfen Sun, niversity of Plymouth, UK
Letters & Member Communications: Jun Wu, Tongji University, China
Secretary: Shaoen Wu, Ball State University, USA
Standards Liaison: Guosen Yue, Huawei, USA

This team of officers was elected at IEEE ICC 2018 (IEEE International Conference on Communications), May 20-24, 2018, Kansas City, MO, USA. I would like take this opportunity to give great thanks to the past MMTC chair Prof. Shiwen Mao and his team for their outstanding leadership and great success during the term 2016-2018. I believe that our new team will continue the past success of the MMTC.

Further, I would like to bring to your attention of the many resources and opportunities MMTC offers to its members. You may check out the MMTC website http://mmc.committees.comsoc.org, for MMTC sponsored journals, conferences/workshops, MMTC Communications—Frontiers and Reviews.
I encourage all of you to submit papers to CSSMA symposium of IEEE ICC 2019 (May 20-24, Shanghai) and IEEE ICME (July 8-12, Shanghai), which are sponsored by our MMTC.

I would like to take this opportunity to invite all of you to attend the MMTC meeting in IEEE Globecom 2018 (December 9-13, Abu Dhabi, UAE). We will review the MMTC activities with recent updates at the meeting. I look forward to seeing you all soon in December 2018.

Have a wonderful holiday season. I wish you all the best!

Honggang Wang
Chair, Multimedia Communications Technical Committee (2018-2020)
IEEE Communications Society

SPECIAL ISSUE ON Terahertz Communication

Guest Editor: Tuncer Baykas
Istanbul Medipol Univeristy, Turkey
tbaykas@medipol.edu.tr

This special issue of Frontiers focuses on the recent progresses of an emerging technology in wireless communication, which is the Terahertz. Various research groups all around the globe are currently working on THz communication, which have recently also attracted the interest of the industry including a dedicated standard.

The first paper of the issue focuses on the channel measurements in THz band. It provides one of the first single–sweep THz measurement results within 240GHz–300GHz band and relevant statistical analysis.

In the second paper, issues related to statistical modeling of propagation channels for terahertz band are discussed. Authors concluded THz measurements require an extra depth of planning in contrast to conventional propagation channel measurements since measurement duration, measurement methodology, measurement campaign, and post-processing time are prominent parameters to be taken into account. Considering the fact that a successful communication system relies totally on analyzing the statistical characteristics of the channel and employing the optimum transceiver algorithms, each set of environment with each and every set of conditions for almost all possible transceiver interactions constitute a significant duration, effort, and processing power.

The third paper is about International Telecommunication Union's activities in THz band. The authors defined 4 use cases, close proximity mobile systems, intra-device communications, wireless links in data centers and point-to-point fronthaul and backhaul. These use cases are explained in detail.

Tuncer Baykas [SM] (tbaykas@ieee.org) works as an assistant professor at Istanbul Medipol University. He was the chair of IEEE 802.19.1 TG. He served as co-editor and secretary for 802.15 TG3c and contributed to standardization projects, including 802.22, 802.11af and 1900.7. Currently he is the vice chair of 802.19Wg and 802.11bb TG. His research interests include terahertz communication and visible light communication.

Propagation Channels in Terahertz Band[1]

Ali Rıza Ekti[1], Serhan Yarkan[2], Ali Görçin[1], Murat Uysal[3]
[1]*TÜBİTAK BİLGEM, Turkey ,*[2]*Istanbul Commerce University, Turkey,* [3]*Özyeğin University, Turkey*
aliriza.ekti@tubitak.gov.tr

1. Introduction

Systems operating at THz frequencies are attracting great interest and expected to meet the ever–increasing demand for high–capacity wireless communications as well as consumer expectations. Technological progress towards designing the electronic components operating at THz frequencies will lead to a wide range of applications especially for short–range communications such as chip–to–chip communications, kiosk downloading, device–to–device (D2D) communications, and wireless backhauling [1–3]. Although THz bands look promising to achieve data rates on the order of several tens of Gbps, realization of fully operational THz communications systems obliges to carry out a multi–disciplinary effort including statistical propagation and channel characterizations, adaptive transceiver designs (including both baseband and radio frequency (RF) front–end portions), reconfigurable platforms, advanced signal processing algorithms and techniques along with upper layer protocols equipped with various security and privacy levels. As in traditional wireless communications systems design process, realization of high–performance and reliable THz communications systems should start with obtaining detailed knowledge about the statistical properties of the propagation channel. Next, these properties are incorporated into various channel characterizations and models. Upon verification and validation of the characterizations and models under different scenarios, system design stage is initiated at the end.

Related Work

Studies that focus on modeling the channel for THz bands in the literature could be categorized in various ways. Measurement methodology; bandwidth; temporal, frequency, and spatial domain behaviors; and application–specific scenarios are some of them among others. Each and every measurement set concentrates on a specific scenario with some parameter changes [4-8]. In literature, a well-defined, systematic and comprehensive channel modeling strategy for THz bands has not been established yet. Although THz bands manifest many intrinsic propagation characteristics and mechanisms, LOS state is preeminent among others because of the following reasons: First, LOS is desired in THz bands for high–performance operation. Second, LOS presents the elementary propagation characteristics. In this regard, a detailed investigation of LOS measurements should be the first step towards acquiring a systematic and comprehensive statistical channel model.

Contributions

Contribution of this study is three–fold considering the statistical channel characterization for THz scenarios: (i) To the best knowledge of authors', this work provides one of the first single–sweep THz measurement results within 240GHz–300GHz band and relevant statistical analysis. (ii) Detailed statistical analyses of antenna–tilt measurement results under LOS conditions within large–volume anechoic chamber are provided. (iii) In addition, impact of humidity is also considered under LOS scenarios and relevant results are given.

2. Terahertz Measurement Experiment Setup

We constructed an experimental measurement setup in the Millimeter Wave and Terahertz Technologies Research Laboratories (MILTAL) at the Scientific and Technological Research Council of Turkey (TUBITAK) in Gebze, Turkey and it can be seen in Figure 1. The dimensions of the anechoic chamber used in the measurements are 7mx3mx4m (length x width x height). In this study, VNA measurements are performed with TX-RX separation up to 80cm for the frequency interval of 240GHz to 300GHz. This interval is measured using standard gain horn antennas which are attached to OML extenders. Each spectral measurement is represented with 4096 equally spaced frequency points (data points) within the interval specified by the VNA. Therefore, a spectral resolution of 14.648MHz is obtained. 240GHz to 300GHz is specially used to get better performance from the extender modules, in terms

[1] A short review for A. R. Ekti, A. Boyaci, A. Alparslan, Unal, S. Yarkan, A. Gorcin, H. Arslan, M. Uysal, "Statistical modeling of propagation channels for terahertz band" in 2017 IEEE CSCN, pp. 275-280, Sep. 2017.

of magnitude and phase stability.

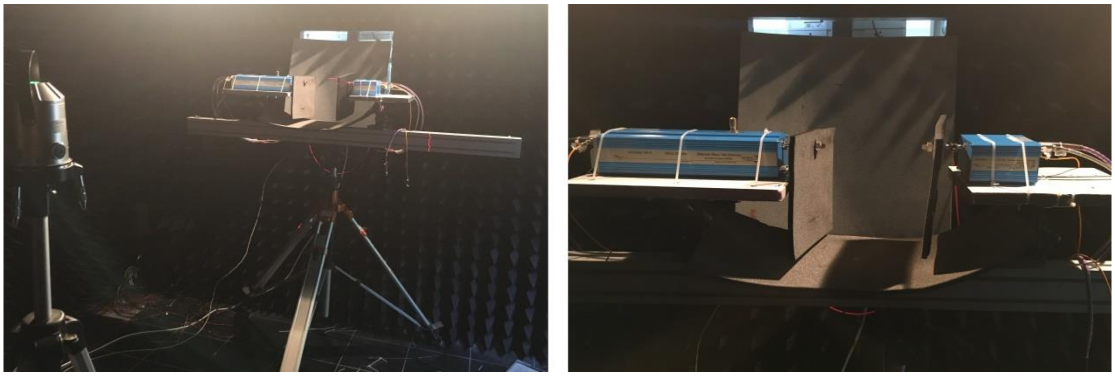


Figure 1 Measurement setup in anechoic chamber from two different angles. Note that LOS conditions are emulated by well–isolated measurement equipment in the anechoic chamber.

3. Measurement Results

Channel magnitude response for THz channels is an important qualitative tool. Especially wideband measurements reveal different mechanisms whose effects are obscured in traditional channel characterizations due to relatively narrower band measurements. Frequency–dependent loss is one of them. In this regard, channel magnitude response could be used to analyze the frequency dependency of the loss in frequency domain very easily. In Figure 2, averaged magnitude responses are given. However, first it is appropriate to evaluate the distance–dependent path loss. Overall mean path loss exponent is found to be n = 1.9704 based on 4096–point resolution with a variance of approx. 0.003 by taking into account entire 60GHz span. This result is in conformity with the LOS argument and with the measurement results reported in the literature [9-10].

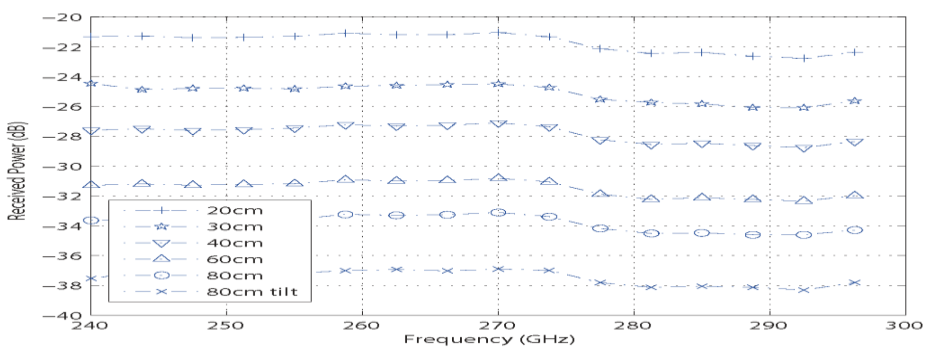


Figure 2 Averaged channel frequency responses in logarithmic scale for various LOS scenarios including the impact of antenna misalignment via antenna tilt.

In order to validate the frequency domain results, time domain analysis is carried out as well. By applying inverse fast Fourier Transform (IFFT) operation, time domain data are obtained. Raw time data are plotted in Figure 3. In Figure 3, it is seen that the delay t0 is given in terms of corresponding distance matching with the experimental setup values.

The impact of the humidity is also investigated for THz communication channels. Channel impulse response obtained under a humid environment is appended into Figure 4 as well. It is seen that peak power level of the LOS path does not exhibit a significant drop in parallel with the measurement results reported in the literature [11].

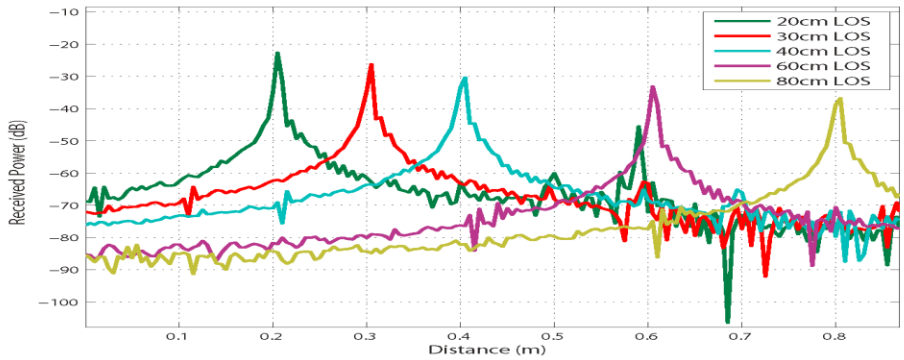


Figure 3 First arriving paths in temporal domain for different transmitter– receiver separations under LOS. The horizontal axis is given in terms of transmitter–separation to validate the experimental setup values.

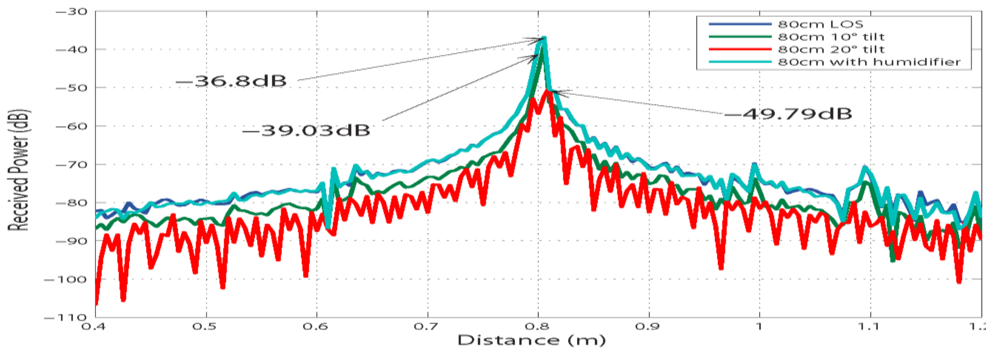


Figure 4 Measured channel impulse responses at 80cm in logarithmic scale with several antenna tilts along with a separate measurement in the presence of dense humidity.

4. Concluding Remarks and Future Directions

Deploying communication systems operating within THz bands is considered to be an alternative strategy to meet the ever–increasing data rate demands along with escalating number of devices subscribing wireless networks. Due to the technical limitations and propagation loss concerns, THz bands have not attracted a significant attention up until the last couple of years. However, with the technological advances, it is possible to migrate up to THz region. Nevertheless, a successful and reliable communication system relies heavily on well–established propagation channel models and appropriate transceiver designs.

In this study, single–sweep band measurement data for 240GHz–3000GHz band are collected in frequency domain with a very high resolution within an anechoic chamber along with a very well isolated setup to emulate the propagation. Behavior of the channel within a 60GHz span (i.e., 240GHz–300GHz interval) is captured at once. In addition, high–resolution measurement data are collected so that finer temporal details are obtained to help design reliable transceiver systems including antenna misalignment problem. Since scenario provides a theoretical borderline, collected data could be used to validate other results obtained in different measurement campaigns. Practical applications for THz communications are generally envisioned to operate in short–range such as infotainment systems. This implies that locations and spatial orientations of THz devices could be random especially in residential scenarios. In this regard, non-line–of–sight (NLOS) behavior of the channels especially for indoor applications should be examined in detail.

References

[1] I. F. Akyıldız, J. M. Jornet, and C. Hana, “Terahertz band: Next frontier for wireless communications,” Physical Communication, vol. 12, pp. 16–32, Sep. 2014.

[2] T. Ku¨rner and S. Priebe, “Towards THz Communications - Status in Research, Standardization and Regulation,” J. Infrared Milli. Terahz. Waves, vol. 35, no. 1, pp. 53–62, Jan. 2014.

[3] T. Tajima, T. Kosugi, H.-J. Song, H. Hamada, A. E. Moutaouakil, H. Sugiyama, H. Matsuzaki, M. Yaita, and O. Kagami, “Terahertz MMICs and Antenna-in-Package Technology at 300 GHz for KIOSK Download System,” J. Infrared Milli. Terahz. Waves, vol. 37, no. 2, pp. 1213–1224, Dec. 2016.

[4] S. Priebe, C. Jastrow, M. Jacob, T. Kleine-Ostmann, T. Schrader, and T. Ku¨rner, “Channel and Propagation Measurements at 300 GHz,” IEEE Trans. Antennas Propag., vol. AP-59, no. 5, pp. 1688–1698, May 2011.

[5] S. Priebe and T. Ku¨rner, “Stochastic Modeling of THz Indoor Radio Channels,” IEEE Trans. Wireless Commun., vol. WC-12, no. 9, pp. 4445–4455, Sep. 2013.

[6] N. Khalid and O. B. Akan, “Wideband THz Communication Channel Measurements for 5G Indoor Wireless Networks,” in Proc. IEEE 2016 ICC Int. Conf. Commun., Kuala Lumpur, Malaysia, Jul. 2016, pp. 1–6.

[7] ——, “Experimental Throughput Analysis of Low-THz MIMO Communication Channel in 5G Wireless Networks,” IEEE Wireless Commun. Lett., vol. WCL-5, no. 6, pp. 616–619, Dec. 2016.

[8] S. Kim and A. Zajic, “Statistical Modeling and Simulation of Short-Range Device-to-Device Communication Channels at Sub-THz Frequencies,” IEEE Trans. Wireless Commun., vol. WC-15, no. 9, pp. 6423–6433, Sep. 2016.

[9] S. Kim and A. G. Zajic, “Statistical characterization of 300-ghz propagation on a desktop,” IEEE Transactions on Vehicular Technology, vol. 64, no. 8, pp. 3330–3338, Aug 2015.

[10] H. Sawada, K. Fujii, A. Kasamatsu, H. Ogawa, K. Ishizu, and F. Kojima, “Path loss model at 300 ghz for indoor mobile service applications,” IEICE Communications Express, vol. 5, no. 11, pp. 424–428, 2016.

[11] G. A. Siles, J. M. Riera, and P. G. del Pino, “Atmospheric Attenuation in Wireless Communication Systems at Millimeter and THz Frequencies,” IEEE Antennas Propag. Mag., vol. 57, no. 1, pp. 48–61, Feb. 2015.

On Some Issues Related to Statistical Modeling of Propagation Channels for Terahertz Band

Emre Ulusoy[1], Özgür Alaca[2], Gamze Kirman[2], Ali Rıza Ekti[1], Ali Görçin[1] , Serhan Yarkan[2],
[1]*TÜBİTAK BİLGEM, Turkey*
[2]*Tapir Labs. at Istanbul Commerce University, Turkey,*
emre.ulusoy@tubitak.gov.tr, ozgur.alaca@istanbulticaret@edu.tr,
gkirman@ticaret.edu.tr, {aliriza.ekti, ali.gorcin}@tubitak.gov.tr,
syarkan@ticaret.edu.tr

1. Introduction

Underutilized spectrum constitutes a major concern in wireless communications especially in the presence of legacy systems, ever–increasing demand for high–capacity services as well as consumer expectations. From this perspective, terahertz (THz) frequencies provide a new paradigm shift in wireless communications since they have been left unexplored until recently. Such a vast frequency spectrum region extending all the way up to visible light and beyond points out significant opportunities from dramatic data rates on the order of several tens of Gbps to a variety of inherent security and privacy mechanisms and techniques that are not available in the traditional systems. On-board chip-to-chip communications, booth downloading, machine-to-machine (M2M) communications as well as wireless backhauling are just to name a few of the prominent applications of systems to operate at THz region.

Beside its apparent benefits and relevant near future anticipations, THz communications bring about significant challenges as well. Radio frequency (RF) front-end equipment including antennae, power, fidelity, and even regular environmental conditions pose serious concerns regarding high-performance communication systems operating at THz region due to its intrinsic nature. However, borrowing from system analysis strategies, THz communication systems could still be simplified as a chain of black boxes. In this chain, sophisticated propagation mechanisms along with all other concerns could be represented in channel box placed in between transmitter and receiver boxes. Such an abstraction is key to a successful communication system design for conventional technologies as well as emerging ones including THz communication systems.

Although system analysis provides almost a universal approach, there are several critical issues peculiar to THz communications systems. First and foremost, such a relatively high transmission frequency comes at the expense of dramatic attenuation levels in regular terrestrial environments (Federici, Moeller, & Su, 2013) as well as obliging the presence of optical line-of-sight (LOS) link between transmitter and receiver. Again, depending on frequency and time, certain bands become practically unusable due to ambient atmospheric absorption, pressure, temperature, and presence of various substances within such as water vapor, dust, and fine-grain particles floating around (Ma, Shrestha, Moeller, & Mittleman, 2018). From this perspective, extremely high directivity of the beams along with utmost receiver sensitivity is two interdependent but prominent mandatory design requirements (Moldovan, Ruder, Akyildiz, & Gerstacker, 2014). In conjunction with these two, theoretical assumption regarding complex baseband equivalency needs to be handled with care since generating suitable waveforms at these frequencies suffers from low-fidelity due to imperfect equipment characteristics. Furthermore, relative motion in the environment and its adversary effects on propagation should be considered as well. It is important to emphasize the fact that THz channels exhibit drastically different behaviors in terms of both spatio-temporal and spectral characteristics. Such a vast variety of behaviors and characteristics forces one to elaborate the THz channels more deeply in contrast to the contemporary measurement and characterization trend which lacks depth and detail.

2. Terahertz Propagation Channels

As electromagnetic (EM) waves propagate through a physical environment, they are affected by several mechanisms such as reflection, refraction, scattering, diffraction, and absorption. In a dynamic environment, relative mobility and/or evolving environmental conditions introduces some changes in the perceived behavior of the EM waves and leads to diverse spatio-temporal characteristics. Therefore, sophisticated propagation structures mandate advanced transceiver algorithms and designs to fully harness the received signals.

3. Fundamental Characteristics of THz Channels

Peculiar to THz channels, following parameters become vital rather than important as in conventional RF channels: transmission frequency and bandwidth; coverage; LOS; transmit power; device nonlinearities; spatio-temporal behavior of the transceivers (Piesiewicz, Jemai, Koch, & Kurner, 2005). It is clear that these vital parameters should adequately be measured, investigated, and characterized in a statistical manner.

4. Measurement Strategies and Relevant Concerns

One of the most important distinctions between conventional and THz propagation channel measurements is the dramatic impact of stringent operational capabilities of measurement equipment on the measurement results. In the literature, the term “THz gap” circulating around evinces this fact from the measurement perspective since THz spectrum stands between region that is reachable by electronics-based devices and photonics-based equipment. Of course, THz spectrum extends upon both sides and has overlapping parts where transition is not very well-defined. Thus, it is critical to explicitly identify the important aspects of the propagation channel mechanisms to be put forward because those aspects will determine the suitable set of measurement equipment and methodology. For instance, power measurements at THz bands could be established by electronics-based setup, whereas delay-oriented measurements require photonics-based equipment.

Wireless propagation channels exhibit diverse behaviors under different environmental conditions. THz channels are no exception. Due to their small wavelengths, miniscule changes in the propagation environment might lead to dramatic differences in the channel characterizations. Ambient temperature and dust particles present over the air are two of the prominent examples of such manifestations (Federici & Ma, 2014). It is obvious that a complete characterization of THz channels should take into account a very wide set of measurement campaigns encompassing all possible scenarios (Piesiewicz, Jansen, Mittleman, Ostmann, & M. Koch, 2007).

Mechanical stability of the measurement platforms bring about a further set of issues at THz bands. This stems from the fact any antennae misalignment, miniscule mechanical vibration and/or perturbation destroys the directivity and causes the effective power received to reduce dramatically. Even though this concern is valid almost for each and every measurement campaign, THz measurements are slightly different since such problems might not be able to be detected until measurement results are post-processed. It is evident that reconstruction and reproducibility of these type of campaigns are difficult, time consuming, and expensive.

Considering the electronics-based equipment, measurement duration is another important aspect. Because THz communications systems are considered to be of wideband type, sweep-based measurements pose two different problems in full characterization: Extremely long measurement durations and enormous measurement data to be post-processed. A very frequent approach to alleviate duration problem is to shorten measurement durations by focusing on relatively narrower but consecutive portions of the spectrum then merging together with some tapering techniques on digital domain. Another popular approach is to sweep across the entire range by hopping on certain non-adjacent bands and then interpolate the missing portions later on. Of course, chopping the entire range in consecutive blocks is more time consuming compared to hopping strategy; however, it is closer to full characterization. For the post-processing part, storage and processing power become key factors to retrieve measurement data and reduce the duration required for processing them.

5. Some Further Technical Issues

First of all, sub-THz and THz channels require some sort of RF stage right before the sounding operation. It is important to keep in mind that such high frequencies at EM spectrum are difficult to attain in a direct way; hence, several stages of upconversion operation with the aid of extender modules are almost compulsory. However, each extending stage comes at the expense of severe harmonic distortion and noise amplification. Hence, it is always beneficial to have high-precision electronics fine-tuned to operate at the desired frequency, if possible. If not, then natural choice is to employ as less number of extending stage as possible (Ekti, et al., 2017).

Even though almost-perfect calibration methods could be applied, THz measurements still suffer from several uncertainties stemming from waveguide dimensions and discontinuities and angular displacements. Furthermore, these uncertainties might lead to systematic bias at measurements, which should be taken into account.

As a final note, it should be stated that spatio-temporal measurements are affected by the shape (and parameters) of the pulses to be transmitted. Considering the bandwidth requirements, sampling capabilities and limitations, and receiver sensitivity issues, it is reasonable to employ stochastic sounding techniques.

6. Concluding Remarks

THz spectrum is considered to be a very promising solution for spectrum underutilization. On-board chip-to-chip communications, booth downloading, machine-to-machine (M2M) communications as well as wireless backhauling are just to name a few of the potential applications of THz communication system. However, there certain technical concerns regarding THz communication systems to ramp up in the near future.

First and foremost, statistical modeling of the physical propagation channel is of paramount importance for a successful transceiver design as in all of the traditional communications systems. Different from traditional system designs, THz systems bring about peculiar channel characteristics into the picture. Even though communication range is relatively short, THz systems are known to be affected severely by ambient temperature, humidity, and dust particles over the air. Therefore, a complete channel characterization should be very comprehensive such that all of the daily life scenarios under very different conditions could be verified and validated.

Second, THz measurements are dramatically equipment-dependent. Standing in the middle of electronics and photonics domain, THz propagation channel measurements should be handled with utmost care due to uncertainties, biases, and inherent systematic errors stemming from the imperfect and/or insufficient hardware driving the measurements.

Finally, THz measurements require an extra depth of planning in contrast to conventional propagation channel measurements since measurement duration, measurement methodology, measurement campaign, and post-processing time are prominent parameters to be taken into account. It is obvious that each and every item in this list corresponds to a specific set of scenarios. Considering the fact that a successful communication system relies totally on analyzing the statistical characteristics of the channel and employing the optimum transceiver algorithms, each set of environment with each and every set of conditions for almost all possible transceiver interactions constitute a significant duration, effort, and processing power.

References

[1] Ekti, A. R., Boyacı, A., Alparslan, A., Unal, I., Yarkan, S., Gorcin, A., et al. (2017). Statistical Modeling of Propagation Channels for Terahertz Band. IEEE Conference on Standards for Communications and Networking (CSCN) (s. 275-280). Helsinki: IEEE.

[2] Federici, J., & Ma, J. (2014). Comparison of terahertz versus infrared free-space communications under identical weather conditions. 39th International Conference on Infrared, Millimeter, and Terahertz waves (IRMMW-THz) (s. 1-3). Tucson, AZ: IEEE.

[3] Federici, J., Moeller, L., & Su, K. (2013). Terahertz wireless communications. Cambridge: Woodhead Publishing.

[4] Ma, J., Shrestha, R., Moeller, L., & Mittleman, a. D. (2018). Channel Characteristics for Terahertz Wireless Communications. D. Saeedkia içinde, Handbook of Terahertz Technology for Imaging, Sensing and Communications. Istanbul: Woodhead Publishing Limited.

[5] Moldovan, A., Ruder, M. A., Akyildiz, I. F., & Gerstacker, W. H. (2014). LOS and NLOS Channel Modeling for Terahertz Wireless Communication with Scattered Rays. Globecom 2014 Workshop - Mobile Communications in Higher Frequency Bands (s. 388-392). Austin, TX: IEEE.

[6] Piesiewicz, R., Jansen, C., Mittleman, D., Ostmann, T. K., & M. Koch, T. K. (2007). Scattering analysis for the modeling of THz communication systems. IEEE Transactions on Antennas and Propagation, 3002-3009.

[7] Piesiewicz, R., Jemai, J., Koch, M., & Kurner, T. (2005). THz channel characterization for future wireless gigabit indoor communication systems. INTEGRATED OPTOELECTRONIC DEVICES (s. 1-12). California: SPIE.

Developments in ITU-R on Terahertz Communications

Tuncer Baykas
Istanbul Medipol Univeristy
tbaykas@*medipol.edu.tr*

1. Introduction

World Radio Congress 19 agenda item 1.15 calls for studies to identify frequency bands for use by administrations for the land mobile and fixed services applications operating in the frequency range 275-450 GHz, in accordance with Resolution **767 (WRC-15)**. Resolution **767 (WRC-15)** *invited* ITU-R to conduct sharing and compatibility studies between land mobile service (LMS) and fixed service (FS) applications and passive services planned to operate in the frequency range 275-450 GHz and to identify candidate frequency bands for use by systems in LMS and FS applications, while maintaining protection of the passive services [1,2]. In this paper we review the THz use cases prepared by ITU-R 1A group[3].

2. Current Regulatory Information

The following frequency bands in the range 275-1 000 GHz are identified for use by administrations for passive service applications:

- radio astronomy service: 275-323 GHz, 327-371 GHz, 388-424 GHz, 426-442 GHz, 453-510 GHz, 623-711 GHz, 795-909 GHz and 926-945 GHz;
- Earth exploration-satellite service (passive) and space research service (passive): 275-286 GHz, 296-306 GHz, 313-356 GHz, 361-365 GHz, 369 392 GHz, 397-399 GHz, 409-411 GHz, 416-434 GHz, 439-467 GHz, 477-502 GHz, 523-527 GHz, 538-581 GHz, 611-630 GHz, 634-654 GHz, 657-692 GHz, 713-718 GHz, 729-733 GHz, 750-754 GHz, 771-776 GHz, 823-846 GHz, 850-854 GHz, 857-862 GHz, 866-882 GHz, 905-928 GHz, 951-956 GHz, 968-973 GHz and 985-990 GHz.

The use of the range 275-1 000 GHz by the passive services does not preclude use of this range by active services. Administrations wishing to make frequencies in the 275 1 000 GHz range available for active service applications are urged to take all practicable steps to protect these passive services from harmful interference until the date when the Table of Frequency Allocations is established in the above-mentioned 275 1 000 GHz frequency range. All frequencies in the range 1 000-3 000 GHz may be used by both active and passive services. (WRC-12)

3. Use Cases

In the report 4 use cases provided for THz communication.

- Close proximity mobile systems
- Intra-device communications
- Wireless links in data centers
- Point-to-point fronthaul and backhaul

3.1 Close proximity mobile systems

Close proximity mobile systems (CPMS) provide a means for large file sizes to be transferred in a few seconds. Some examples could be systems such as kiosk systems or ticket gate systems, which could be used for the purchase of a movie downloaded to a mobile device. These systems are typically connected to wired networks and provide the wireless data to mobile devices in public areas such as train stations, airports, etc. The distance between the user and the gate or kiosk terminal is typically less than 10 cm. The expected range of technical and operational characteristics for close proximity mobile systems planned to operate in the band 275-325 GHz and in the band 275-450 GHz are shown in Table 1**.**

TABLE 1

Expected technical and operational characteristics of a land mobile CPMS applications in the frequency range 275-450 GHz

Parameters	Values	
	CPMS application	Enhanced CPMS application
Frequency band (GHz)	275-325	275-450
Deployment density[1]	0.6 devices/km^2	0.6 devices/km^2
Tx output power density (dBm/GHz)	-3.8….6.9	-10.1…6.7
Max. e.i.r.p. density(dBm/GHz)	26.2…...36.9	19.9…36.7
Duplex Method	FDD/TDD	FDD/TDD
Modulation	OOK/BPSK/QPSK/16QAM/64QAM BPSK-OFDM/QPSK-OFDM/16QAM-OFDM/32QAM-OFDM/64QAM-OFDM	OOK/BPSK/QPSK/16QAM/64QAM/8PSK/8APSK BPSK-OFDM/QPSK-OFDM/16QAM-OFDM/32QAM-OFDM/64QAM-OFDM
Average distance between CPMS fixed and mobile devices (m)	0.1	0.1
Maximum distance between CPMS fixed and mobile devices (m)	1	1
Antenna height (m)	1…2	-
Antenna beamwidth (degree)	3…10	5…90
Antenna elevation (degree)	±90	±90
Frequency reuse	1	1
Antenna type	Horn	Horn
Antenna pattern	Gaussian	Gaussian
Antenna polarization	Linear	Linear
Indoor CPMS fixed device deployment (%)	100	90
Feeder loss (dB)	2	2
Maximum CPMS fixed/mobile device output power (dBm)	10	10
Channel bandwidth (GHz)	2.16/4.32/8.64/12.96/17.28/25.92/51.8	2.16/4.32/8.64/12.96/17.28/25.92/51.84/69.12/103.68
Maximum CPMS fixed device antenna gain (dBi)	30	30
Maximum CPMS mobile device antenna gain (dBi)	15	15
Maximum CPMS fixed device output power (e.i.r.p.) (dBm)	40	40
Maximum CPMS mobile device output power (e.i.r.p.) (dBm)	25	25
Average activity factor (%)	0.76	0.2
Average CPMS fixed device power (dBm (e.i.r.p))	20	20
Receiver noise figure typical (dB)	15	15

3.2 Intra-device communications

In intra-device communications, high speed terahertz wireless links could connect two or more PCBs or even chips on the same PCB inside a device, simplifying board design, inter module wiring harnesses, etc. Typically, these devices will be shielded, preventing ingress and egress of THz signals. The amount of shielding and the percentage of devices expected to be shielded were not available at the time of this report. Future studies should consider this information if it is available.
The expected ranges of technical and operational characteristics for wireless THz intra-device links planned to operate in the band 275-450 GHz are shown in Table 2.

TABLE 5

Expected technical and operational characteristics of wireless THz intra-device links operating in the frequency band 275-450 GHz

Parameter	Value
Frequency band (GHz)	275-450
Deployment density	0.23[1] /km^2
Maximum device output power (dBm)	10
Maximum device output power (e.i.r.p.) (dBm)	30
Maximum Tx output power density (dBm/GHz)	-10.1…6.7
Maximum e.i.r.p. density (dBm/GHz)	19.9…36.7
Indoor Deployment (%)	50
Duplex Method	TDD, FDD, SDD
Modulation	OOK/BPSK/QPSK/16QAM/64 QAM 8PSK/8APSK
Maximum distance between devices	<1 m
Antenna height (m)	1…3
Antenna beamwidth (degree)	15…180 (expected)
Frequency reuse	1
Antenna pattern	Gaussian
Antenna polarization	Linear
Channel bandwidth (GHz)	2.16/4.32/8.64/12.96/17.28//25.92/51.84/69.12/103.68
Maximum device antenna gain (dBi)	20
Typical expected device antenna gain (dBi)	6
Maximum device activity (%)	100
Receiver noise figure typical (dB)	10[2]

[1] The deployment density is estimated as an average based on assuming that every one thousandths citizen in Germany is using such a device. In highly populated cities the density could increase to e.g. 3.95/km^2 under the same assumptions.

[2] Also systems with a noise figure as low as 8 dB have been reported in publications. This value is a worst case of the published parameters.

3.3 Wireless links in data centers

The use of wireless links in data centers aims to provide flexibility by providing reconfigurable routes within a data center without extensive rewiring. The expected ranges of technical and operational characteristics for wireless links in data centers planned to operate in the band 275-450 GHz are shown in Table 3. This application is intended as a strictly indoor only application. A bandwidth of 50 GHz is necessary to achieve a data rate of at least 100 Gbit/s with a simple QPSK modulation and enable compatibility with 100 Gbit/s Ethernet links.

TABLE 3

Expected technical and operational characteristics of wireless links in data centers operating in the frequency band 275-450 GHz

Parameter	Values
Frequency band (GHz)	275-450
Deployment density	0.07 /km^2
Maximum device output power (dBm)	10
Maximum device output power (e.i.r.p.) (dBm)	40
Tx output power density (dBm/GHz)	-10.1…6.7
e.i.r.p. density (dBm/GHz)	9.9…26.7
Duplex Method	TDD, FDD, SDD
Modulation	OOK/BPSK/QPSK/16QAM/64 QAM 8PSK/8APSK
Maximum distance between devices	100 m
Antenna beamwidth (degree)	< 25 (expected)
Frequency reuse	1
Antenna pattern	Gaussian
Antenna polarization	Linear
Indoor deployment (%)	100
Channel bandwidth (GHz)	2.16/4.32/8.64/12.96/17.28/ 25.92/51.84/69.12/103.68
Maximum device antenna gain (dBi)	30
Maximum device activity (%)	100
Receiver noise figure typical (dB)	10

3.4 Wireless links in data centers

Figure 1 shows the network architecture of mobile systems, which support high-capacity transmission between a base station and a mobile terminal. The fronthaul is defined as a link connection between the base station's baseband unit (BBU) and the remote radio head (RRH), while the backhaul is a link between the base station and the higher level network elements. According to Recommendation ITU-R M.2083 and Report ITU-R M.2376, fronthaul and backhaul are critical challenges to accommodate the increase in data throughput of future mobile traffic. In order to meet the peak data rate 10–20 Gb/s of the mobile terminals in a small cell, the transmission capacity of fronthaul and backhaul may exceed tens of Gb/s substantially.

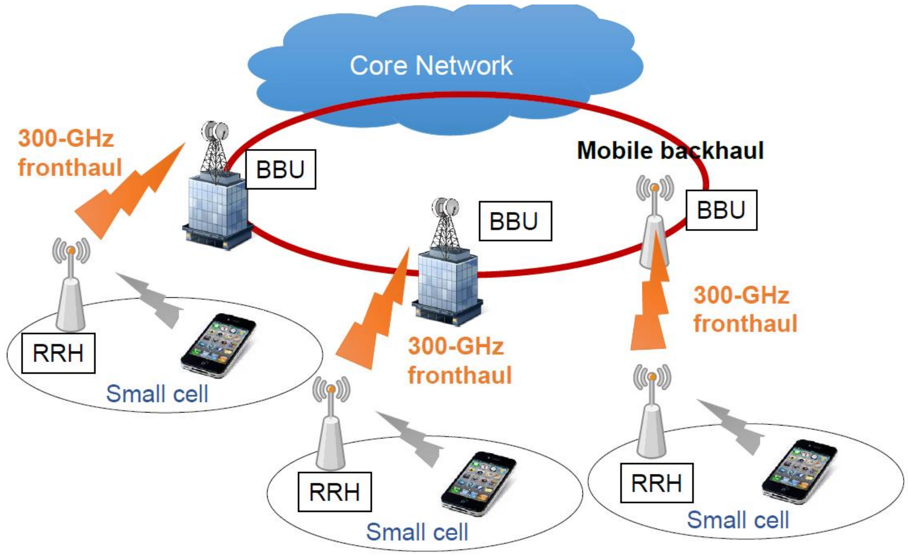


Figure 1. Fronthaul and backhaul operation to be used for mobile system network

The 275-450 GHz range provides the possibility of short range, wide bandwidth, high data rate capability for wireless systems supporting mobile terminals.
The proposed technical and operational characteristics of fixed point-to-point fronthaul and backhaul systems planned to operate in the band 275–325 GHz and 380–450 GHz are shown in Table 4.

TABLE 4

Technical and operational characteristics of the fixed service applications planned to operate

Frequency band (GHz)	**275–325**	**380–445**
Duplex Method	FDD/TDD	FDD/TDD *Editorial note: Other duplex in schemes are possible*
Modulation	BPSK/QPSK/8PSK/8APSK/16QAM/32QAM/64QAM BPSK-OFDM/QPSK-OFDM/16QAM-OFDM/32QAM-OFDM/64QAM-OFDM	BPSK/QPSK/8PSK/8APSK/16QAM/32QAM, 8PSK, 8APSK BPSK-OFDM/QPSK-OFDM/16QAM-OFDM/32QAM-OFDM
Channel bandwidth (GHz)	2…..25 (FDD) 2…..50 (TDD)	2…..32.5 (FDD) 2…..65 (TDD)
Tx output power range (dBm)	0….20	-10….10
Tx output power density range (dBm/GHz)	-17……17	-28….7
Feeder/multiplexer loss range (dB)	0 … 3	0 … 3
Antenna gain range (dBi)	24 … 50	24 … 50
e.i.r.p. range (dBm)	44…..70	37……60
e.i.r.p. density range (dBm/GHz)	30……67	19…..57

Frequency band (GHz)	275–325	380–445
Antenna pattern	Recommendation ITU-R F.699 (Single entry) Recommendation ITU-R F.1245 (Aggregate)	Recommendation ITU-R F.699 (Single entry) Recommendation ITU-R F.1245 (Aggregate)
Antenna type	Parabolic Reflector	Parabolic Reflector
Antenna height (m)	6-25	10-25
Antenna elevation (degree)	±20 (typical)	±20 (typical)
Receiver noise figure typical (dB)	15	15
Receiver noise power density typical (dBm/GHz)	-69	-69
Normalized Rx input level for 1×10-6 BER (dBm/GHz)	-61 … -54	-61 … -54
Link length (m)	100 … 300	100 … 300
I/N Protection Criteria	Recommendation ITU-R F.758	Recommendation ITU-R F.758

4. Conclusions

This letter provides a short overview of use cases provided by ITU-R. The group is planning finalize the report in 2019.

References

[1] Report ITU-R SM.2352-0- Technology trends of active services in the frequency range 275-3000 GHz; Geneva, June 2015 http://www.itu.int/dms_pub/itu-r/opb/rep/R-REP-SM.2352-2015-PDF-E.pdf

[2] Provisional Final Acts WRC-15, Geneva 2015; https://www.itu.int/dms_pub/itu-r/opb/act/R-ACT-WRC.11-2015-PDF-E.pdf

[3] A Preliminary Report ITU-R ITU-R SM.[275-450GHZ_SHARING] “Sharing and compatibility studies between land-mobile, fixed and passive services in the frequency range 275-450 GHz”

SPECIAL ISSUE ON Real-Time Multimedia Systems

Guest Editor: Dalei Wu
University of Tennessee at Chattanooga, USA
dalei-wu@utc.edu

This special issue of Frontiers focuses on the recent progresses of real-time multimedia systems which involves a variety of exciting applications, such as live streaming, video surveillance, and augmented reality. Edge computing is one of most important 5G technologies enabling next generation of real-time multimedia systems. Researchers around the world have made great effort to push the boundaries of knowledge in the related areas. Included in this special issue are some example works.

The first paper proposed an edge computing framework to enable cooperative processing on resource-abundant mobile devices for delay-sensitive multimedia IoT tasks. A deep-reinforcement-learning (DRL)-based edge caching policy was proposed by considering the freshness of IoT data and unknown request rates.

The second paper proposed a knowledge-centric cellular network architecture with device-to-device (D2D) communications. A D2D-assisted virtualized edge cellular network architecture was developed. The optimization objective functions are designed to meet the multimedia service requirements. The framework provides a new perspective in understanding the resource allocation problem in future edge cellular network.

In the third paper, the authors introduced a novel event-oriented indexable and queryable intelligent surveillance (EIQIS) system. On-site edge devices are enabled to collect the information sensed in format of frames and extracts useful features to enhance situation awareness. This intelligent surveillance system enables the operator to search for scenes or events of interest instantly.

The fourth paper designed a novel radio environment map (REM) construction architecture based on mobile crowd sensing. Ubiquitous, massive and high dimension REM-related data can be sensed by the terminals carried by mobile users. This is an important design point to deal with the large-scale distributed, high dimension and massive multimedia data sensing tasks.

In the fifth paper, the authors studied the edge computing scheme for 5G heterogeneous networks to improve the computation efficiency through offloading computing task. The resource allocation involving the computing mode selection and power allocation is considered to improve the computing efficiency through mixed integer programing problem formulation and Lagrangian dual method based solution.

Dalei Wu (dalei-wu@utc.edu) is an assistant professor with the Department of Computer Science and Engineering at the University of Tennessee at Chattanooga. He is currently serving IEEE MMTC as the co-director of the board of Frontiers. His research interests include multimedia computing, intelligent systems, and cyber-physical systems.

Edge Computing and Caching for Video Processing in Multimedia IoT systems

Hao Zhu[1], Yang Cao[1], Tao Jiang[1]
[1] School of Electronic Information and Communications,
Huazhong University of Science and Technology, Wuhan, China
zhuhao@hust.edu.cn; ycao@hust.edu.cn; Tao.Jiang@ieee.org

1. Introduction

Recently, the Internet of things (IoT) is developing rapidly in areas including intelligent transportation, smart grid, industrial and home automation, e-Health, and so on. IoT enables the interconnection of physical objects and human by delivering, processing and analyzing the tremendous data which are collected by the ubiquitous connected devices (e.g., tags, sensors, embedded devices, and hand-held devices) [1], [2]. Multimedia IoT system is an emerging type of IoT, which integrates image processing, computer vision and network capabilities. It has been widely used in surveillance (e.g., human/ vehicle detection), event recognition, and automatic behavior analysis [3], [4]. There are two traditional paradigms for multimedia IoT system. In the first paradigm, captured video chunks are preprocessed (e.g., features are extracted from videos) at the camera node. In the second paradigm, video chunks are transmitted to the remote IoT server, and then processed straightly. However, significant delay would be caused by these two traditional paradigms, which has been demonstrated by the measurements in [5]. This is due to the fact that the limited computational resources of camera nodes may cause computation delay when video is processed locally, while the delivering of original video chunks to the remote server may cause congestions and delays because of the limited network bandwidth. Thus, these two paradigms cannot satisfy the requirement of delay-sensitive video processing and analyzing tasks.

Edge/fog computing has been introduced as an emerging technique of enabling distributed computing for the preprocessing of video chunks, which reduces the transmission delay [6], [7]. By leveraging the edge computing technique, redundant computation and communication capabilities of multiple mobile devices in the proximal can be utilized to handle delay-sensitive video processing and analyzing tasks through short-range wireless communications. By sending back only a few video features to the remote servers, the bandwidth starvation of delivering original video chunks can be avoided. In addition, video data in multimedia IoT systems are collected at specific locations and are popular among many location-based services which deliver local information to end users. If popular IoT data (e.g., collected video data or processed video data) can be temporarily cached at edge nodes, requests for these data do not have to be answered all the way by IoT data sources, and in-network traffic is thus reduced [8]. Moreover, local data retrieval rather than server retrieval could allow faster response to requests [9], [10].

Our research aims at leveraging the idle computing and caching resources of edge nodes to reduce the network traffic and the latency of services in multimedia IoT systems. On the one hand, we have proposed an edge computing framework to enable cooperative processing on resource-abundant mobile devices for delay-sensitive multimedia IoT tasks. This work is different from most existing works which mainly adopt the "partition and allocation" strategy without cooperative processing and do not consider the group formation as well as the video–group matching. On the other hand, we have proposed a deep-reinforcement-learning (DRL)-based edge caching policy which is aware of the freshness of IoT data and has the ability to deal with unknown request rates. The proposed policy is able to intelligently perceive the environment and then automatically learns caching policy from history and current raw observations of the environment. This is different from most existing works which have explicit assumptions about the operating environment.

2. Edge Computing Framework for Cooperative Video Processing in Multimedia IoT Systems

In [11], we have studied a cooperative video processing scheme in an edge computing framework. The architecture of the proposed framework is shown in Fig. 1, which consists of three main components, namely camera node, edge node and server.

1) Camera node: This can be a static camera device fixed at the top of a street lamp, which invokes video tasks, divides them into smaller sub-tasks (video chunks), compresses video chunks and finally transmits them to edge nodes within the scope of a certain distance via device-to-device (D2D) communications.

2) Edge node: This is a mobile device with sufficient computational ability and storage capacity, helping process video

sub-tasks, e.g., image feature detection and extraction. Edge nodes form cooperative groups based on the proposed group formation algorithm and receive the compressed video chunks according to the video-group matching algorithm.

3) Server: This is a static device that collects the processing results from edge nodes and performs further video analysis, which has powerful computational abilities.

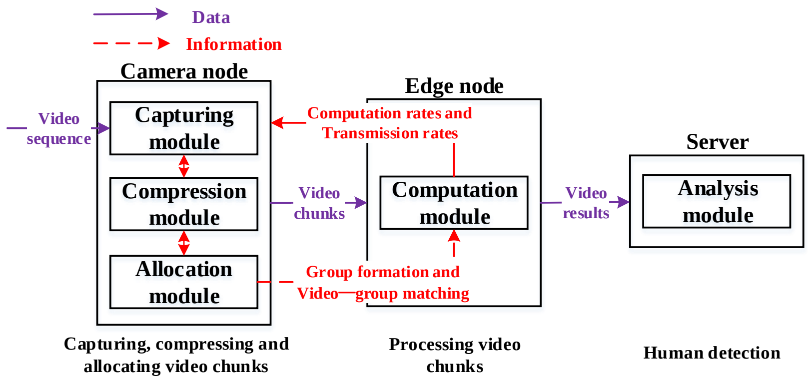


Figure 1. Architecture of the edge computing framework.

The camera node captures video sequences periodically, and then does some operations on them. Take a video task for example, the camera node divides it into a fixed number of video chunks with the same size, compresses video chunks at different video coding ratios (vary from 0 to 1, the larger the value, the less the compression loss) and allocates the compressed video chunks among all the edge nodes according to our proposed scheme. The camera node transmits video chunks to edge nodes in terms of two transmission modes, namely the multicast mode and the unicast mode. In the multicast mode, a video chunk is simultaneously transmitted to multiple edge nodes in a group, where these edge nodes cooperatively processes the different parts of a video chunk. In the unicast model, a video chunk is transmitted to only one edge node. For the sake of simplicity, the concrete operations about how to coordinate the cooperation among edge nodes in a group are not considered. Meanwhile, we assume that edge nodes in a group process a non-overlapping partition with the same size. After processing the assigned video task, edge nodes transmit the results to the server via cellular links. Finally, the server performs video analysis (e.g., human detection), and the video analysis performance is determined by the average video coding rate of all the video chunks.

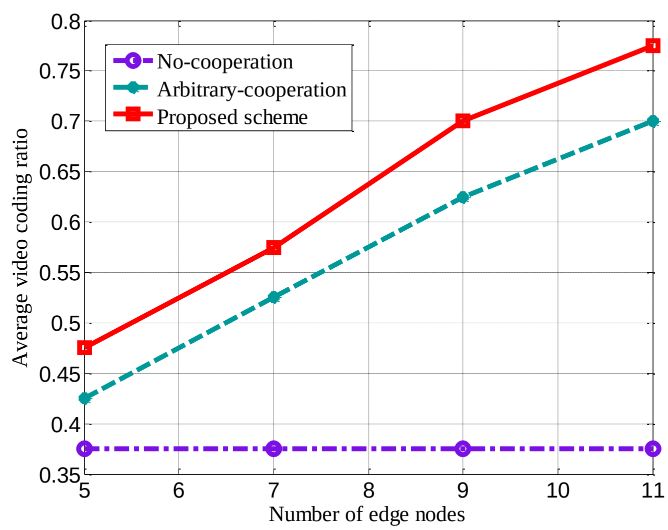


Figure 2. Average video coding ratio versus the number of edge nodes for no-cooperation scheme, arbitrary-cooperation scheme and the proposed scheme.

To maximize the average video coding rate of all the video chunks within the deadline, we formulate the problem of jointly partitioning tasks, compressing and allocating sub-tasks as an integer non-linear programming problem. This problem is decomposed into two subproblems, namely the group formation problem and the video–group matching problem. We transform the group formation problem into a Winner Determination Problem (WDP) and solve it with an effective algorithm, which is 2-approximation and can significantly reduce the complexity. Further, based on the derived optimal matching theorem, a low-complex algorithm is proposed for the video–group matching problem. The proposed scheme has been evaluated under diverse parameter settings, and compared with two baseline schemes. One of the baseline schemes is the no-cooperation scheme, in which there are no cooperative groups, and the L video chunks are randomly transmitted to the former L edge nodes sorted in descending order of the transmission rates. Another baseline scheme is the arbitrary-cooperation scheme, in which edge nodes randomly form cooperative groups and video chunks are arbitrarily matched with the formed groups. As shown in Fig. 2, extensive simulation confirms the superiority of the proposed scheme over other two baseline schemes.

3. Intelligent Edge Caching Policy for Transient Multimedia IoT Data

In multimedia IoT systems, data are transient. In other words, an IoT data file has a lifetime, during which it is useful. When the IoT data file is expired, it becomes useless and must be discarded. Thus, the requirements of IoT applications on data freshness needs to be taken into consideration when designing caching policies for IoT. We define a cost function, which makes a tradeoff between data freshness and communication cost when fetching IoT data. The cost function is denoted as

$$C(d) = \alpha \cdot c(d) + \left(1-\alpha\right) \cdot l(d),$$

where $c(d)$ is the communication cost of fetching data item d from the data source or the edge node, $l(d)$ is defined as the freshness loss of data item d, and $\alpha \in (0,1)$ is a coefficient weighting the relative importance of the communication cost. A larger value of α means a larger weight of communication cost and indicates that an IoT application does not prefer frequent data retrieval from the data producer. As shown in Fig. 3, the freshness loss is defined as

$$l(d) = \frac{t_{\text{age}}(d)}{T_{\text{life}}(d)},$$

where the age of data item d is denoted by $t_{\text{age}}(d) = t - t_{\text{gen}}(d)$, the time of generating d at the data producer is denoted by $t_{\text{gen}}(d)$, and the lifetime of d is denoted by $T_{\text{life}}(d)$.

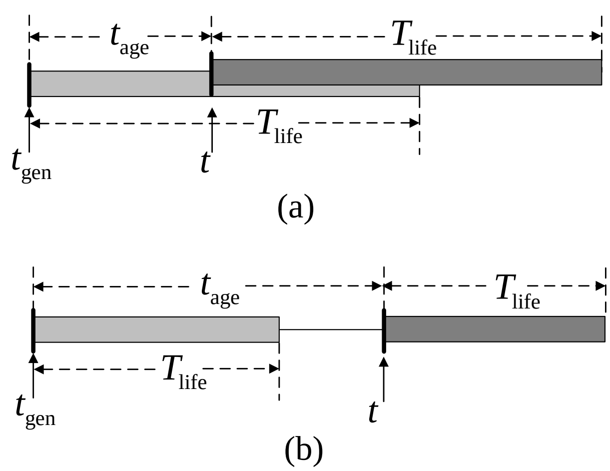


Figure 3. Data item freshness. (a) Fresh. (b) Non-fresh.

To minimize the long-term cost of fetching IoT data, we formulate the cache replacement problem as a Markov Decision Process (MDP) problem. The MDP model can be defined by the tuple $\{\mathcal{S}, \mathcal{A}, \mathcal{M}(s_{n+1} \mid s_n, a_n), \mathcal{R}(s_n, a_n)\}$.

- $\mathcal{S}$ is the set of states for the edge caching based IoT system. We define s_n as the state at time step *n*.

- $\mathcal{A}$ is the set of caching actions. The action selected by the edge node at time step *n* is denoted by a_n.
- $\mathcal{M}(s_{n+1} | s_n, a_n)$ is the state transition probability that maps a state-action pair at time step *n* onto a distribution of states at time step $n+1$.
- $\mathcal{R}(s_n, a_n)$ is the immediate/instantaneous reward function that determines the reward fed back to the edge node when performing action a_n at state s_n.

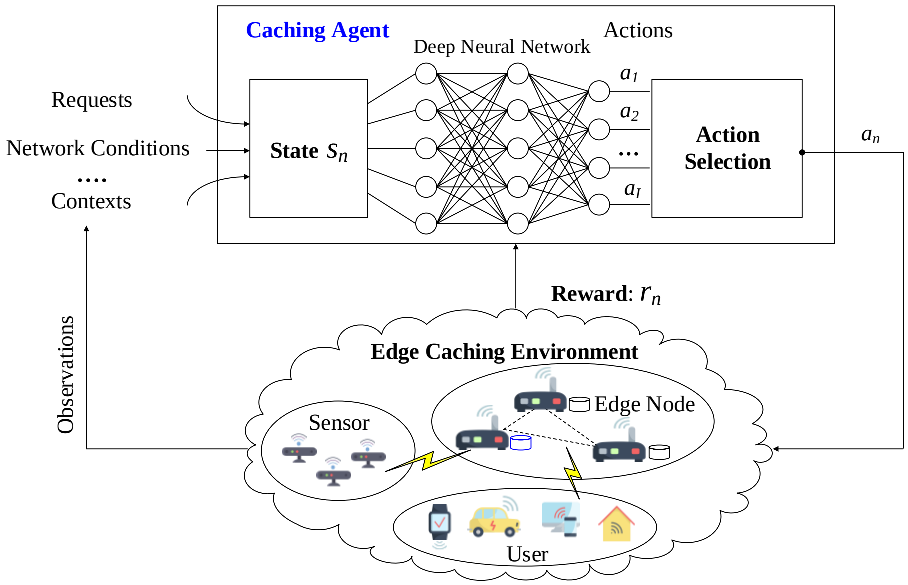


Figure 4. Applying DRL to caching IoT data.

Since there are no explicit models for $\mathcal{R}$ and $\mathcal{M}$, we use reinforcement learning (RL) to learn the caching policy from experience. Moreover, since the operating environment of the edge caching based IoT system is complex and dynamic, and it is difficult to manually extract all useful features of the environment as low-dimensional state spaces. As shown in Fig. 4, DRL is adopted to directly train caching agents on raw high-dimensional observations, rather than handcraft useful features or low-dimensional state spaces. We adopt the A3C algorithm [12]. The caching agent takes state input $s_n = (\vec{x}_n^0, \vec{x}_n^1, \cdots, \vec{x}_n^I, y_n^0, y_n^1, \cdots, y_n^I, z_n^0, z_n^1, \cdots, z_n^I)$ to its neural networks. We extract features from the currently requested data item, whose index is 0, and cached data items, whose indexes range from 1 to the cache size *I*. $\vec{x}_n^i = (x_n^i[1], x_n^i[2], \cdots, x_n^i[J])$ is a vector which represents the number of requests for content f_n^i within past *J* groups of requests, where each group consists of *G* requests. Upon obtaining s_n, the caching agent selects an action a_n based on policy $\pi(a_n | s_n)$, which is the probability of selecting action a_n in state s_n. The action space $\mathcal{A} = \{a^0, a^1, \cdots, a^I\}$, where a^0 means that the cached data items keep unchanged and a^i $(1 \le i \le I)$ means that the new data item is cached in the edge node by replacing the position of f_n^i. As shown in Fig.5, simulation results show that the proposed DRL-based caching policy outperforms other baselines (i.e., the LRU policy and the LFF policy). In the LRU policy, the new data item is cached by replacing the content which is requested least times in the cache. In the LFF policy, the new data item is cached by removing a cached data item which is the least fresh.

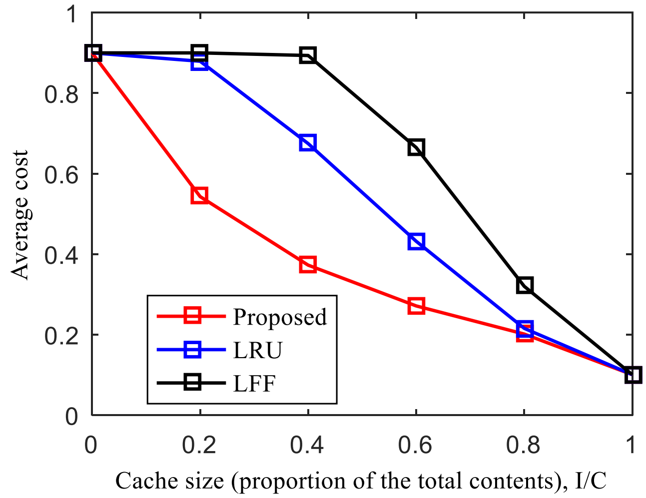


Figure 5. Average cost versus the cache size for the proposed policy, LRU policy and LFF policy.

5. Conclusion

In this letter, we advocated the use of edge computing and edge caching to reduce the network traffic and the service latency in multimedia IoT systems, followed by our research on this topic. We introduced the proposed edge computing framework for cooperative video processing in multimedia IoT systems and the proposed intelligent edge caching policy for transient multimedia IoT data.

Acknowledgement

This work was supported in part by the National Science Foundation of China with Grant number 61729101, 61601193, 61720106001, and Major Program of National Natural Science Foundation of Hubei in China with Grant 2016CFA009, and the Fundamental Research Funds for the Central Universities with Grant number 2015ZDTD012.

References

[1] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey," Comput. Netw., vol. 54, no. 15, pp. 2787–2805, Oct. 2010.
[2] Y. Cao, T. Jiang, and Z. Han, "A survey of emerging M2M systems: Context, task, and objective," IEEE Internet Things J., vol. 3, no. 6, pp. 1246–1258, Dec. 2016.
[3] X. Zhang et al., "Enhancing video event recognition using automatically constructed semantic-visual knowledge base," IEEE Trans. Multimedia, vol. 17, no. 9, pp. 1562–1575, Sep. 2015.
[4] A. Floris and L. Atzori, "Quality of experience in the multimedia Internet of Things: Definition and practical use-cases," in Proc. IEEE Int. Conf. Commun. Workshop, 2015, pp. 1747–1752.
[5] A. Redondi et al., "A visual sensor network for object recognition: Testbed realization," in Proc. Int. Conf. Digital Signal Process., 2013, pp. 1–6.
[6] P. Lopez et al., "Edge-centric computing: Vision and challenges," ACM SIGCOMM Comput. Commun. Rev., vol. 45, no. 5, pp. 37–42, Sep. 2015.
[7] X. Sun and N. Ansari, "EdgeIoT: Mobile edge computing for the Internet of Things," IEEE Commun. Mag., vol. 54, no. 12, pp. 22-29, Dec. 2016.
[8] M. I. Naas et al., "iFogStor: an IoT data placement strategy for fog infrastructure," IEEE Commun. Mag., vol. 54, no. 12, pp. 22-29, Dec. 2016.
[9] H. Zhu et al., "QoE-aware resource allocation for adaptive device-to-device video streaming," IEEE Network, vol. 29, no. 6, pp. 6-12, Nov.-Dec. 2015.
[10] H. Zhu et al., "Multi-bitrate video caching for D2D-enabled cellular networks," appear to IEEE Multimedia, 2018.
[11] C. Long et al., "Edge computing framework for cooperative video processing in multimedia IoT systems," IEEE Trans. Multimedia, vol. 20, no. 5, pp. 1126-1139, May 2018.
[12] V. Mnih et al., "Asynchronous methods for deep reinforcement learning," Proc. ICML, June 2016.

Hao Zhu received the B.S. degree in information and communication engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2014, where he is currently pursuing the Ph.D. degree with the School of Electronic Information and Communications. His research interests focus on mobile edge networks and device-to-device communication.

Yang Cao (S'09–M'14) received the B.S. and Ph.D. degrees in information and communication engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2009 and 2014, respectively. From 2011 to 2013, he was a Visiting Scholar with the School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, AZ, USA. He is currently an Associate Professor with the School of Electronic Information and Communications, Huazhong University of Science and Technology. His research interests include 5G cellular networks, Internet of Things, and future networks. He has coauthored 40 papers on refereed IEEE journals and conferences. He received the CHINACOM Best Paper Award in 2010 and the Microsoft Research Fellowship in 2011.

Tao Jiang (M'06–SM'10) received the B.S. and M.S. degrees in applied geophysics from the China University of Geosciences, Wuhan, China, in 1997 and 2000, respectively, and the Ph.D. degree in information and communication engineering from the Huazhong University of Science and Technology, Wuhan, in 2004. From 2004 to 2007, he was with some universities, such as Brunel University and the University of Michigan at Dearborn, respectively. He is currently a Distinguished Professor with the School of Electronics Information and Communications, Huazhong University of Science and Technology. He has authored or co-authored over 300 technical papers in major journals and conferences and nine books/chapters in the areas of communications and networks. He received the NSFC for 1103 Distinguished Young Scholars Award in 2013, the Young and Middle-Aged Leading Scientists, Engineers and Innovators by the Ministry of Science and Technology of China in 2014, and the Cheung Kong Scholar Chair Professor by the Ministry of Education of China in 2016. He received the Most Cited Chinese Researchers in Computer Science announced by Elsevier in 2014, 2015, and 2016, respectively. He served or is serving as symposium technical program committee membership of some major IEEE conferences, including the INFOCOM, GLOBECOM, and ICC. He is invited to serve as the TPC Symposium Chair for the IEEE GLOBECOM 2013, IEEE WCNC 2013, and ICCC 2013. He is serving as an Associate Editor-in-Chief for China Communications, served or serving as an Associate Editor for some technical journals in communications, including the IEEE TRANSACTIONSON SIGNAL PROCESSING, the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and the IEEE INTERNET OF THINGS JOURNAL.

Intelligent Edge Computing for Real-time Multimedia Communication

Junjie Yan[1], Dapeng Wu[1], Ruyan Wang[1]
[1] *Dep. of Information and Communication Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China*
cqupt2013yjj@sina.com; wudp@cqupt.edu.cn; wangry@cqupt.edu.cn

1. Introduction

According to the visual networking index (VNI) released by Cisco in 2017, mobile data traffic by 2021 will be about 7 times of the traffic in 2016, where mobile video traffic accounted for 60 percent of total mobile data traffic in 2016. The explosive demand for multimedia services in a mobile network increases network traffic between data centers and base stations inevitably. To meet the increasing multimedia service demand, the computational and service capability provided by mobile network must be significantly improved to offer a better quality of service (QoS). One possible solution to enhance network capacity is through the mobile edge computing (MEC) technology which is coined by the European telecommunications standards institute (ETSI). Mobile-edge cloud computing is a new paradigm to provide cloud computing capabilities at the edge of pervasive radio access networks in close proximity to mobile users [2]. In order to integrate and coordinate wireless network resources and provide unified multimedia services to users, wireless network virtualization (WNV) and network slicing are introduced in edge network [3]. It can be considered as the process of dynamically slicing network resources into different virtual slices, based on different service requirements, to achieve optimized resource utilization [4]. However, there are significant research challenges in applying both of technologies in edge networks.

There are several existing works exploiting software defined network (SDN) technique to integrate WNV and slicing network for MEC [5]-[7]. The work in [5] adopted a hybrid control model with two hierarchical control levels. The proposed hybrid control model simplified network management and configuration, and improved the network scalability. The authors in [6] proposed a scalable SDN-enabled architecture that integrated a heterogeneous vehicular network with reliable communication services, based on precise application requirements. The work in [7] applied both WNV and SDN in MEC. WNV allows physical wireless network to be virtualized and sliced into multiple virtual resources. Meanwhile, the SDN paradigm was introduced to integrate various protocols and standards, including MEC, WNV and M2M communications. Among the existing technologies, wireless network resources can be reasonably used. However, the resources of the user are not fully exploited and developed in the previous work. In particular, the D2D-assisted virtualized cellular network has not been investigated under the MEC context. In a D2D-assistant virtualized cellular network, user equipment (UE) resources including storage capacity, energy, services provided by UEs, can be virtualized and requested by other users. D2D-assisted virtualized cellular network can not only support seamless integration of different technologies, but also improve network performance, e.g., reduced transmission latency [8]-[9]. In addition, to meet the requirements of different users for video services, as well as the dynamically changing network environment, Scalable Video Coding (SVC) can be introduced in user side. The basic idea is that the video is encoded into one base layer (BL) and one or more enhancement layers (ELs), and each layer containing the information relative to a quality level. The basic quality of the video stream is guaranteed by providing the user with the base layer. The quality of video can be improved as the number of received enhancement layers increases. An enhancement layer can be decoded if and only if all enhancement layers below it are received. Therefore, SVC-based sharing mechanism provides a degree of freedom by matching video stream characteristics to user device capability and the social characteristics of different users.

When D2D-assisted virtualized cellular network and SVC are investigated under the MEC context, intelligent edge systems need to be established, because of the limited processing capacity, energy, and potential service of user and the social relations among mobile users [10]-[11]. In particular, due to above reasons, users may present a non-cooperative behavior in communications, not all users can be virtualized to provide trustworthy network resource access for others. Apparently, non-cooperative behaviors can severely degrade the network's QoS performance. Given these challenges and facts, the traditional edge control and management approaches will not be sufficient and flexible enough. To support trustworthy network access and satisfy various multimedia service requirements, we propose a new architecture, called knowledge centric edge (KCE), to combine virtualized D2D communication systems and learning-based D2D communication systems such as social and smart transportation D2D systems to support MEC. The key of the architecture is that insightful knowledge such as user relationship and trust information are obtained and computed from the data collected via D2D communications so that network structure can accurately detected by the knowledge and network resources are dynamically allocated in an optimized manner.

2. System Architecture

In a D2D-assisted cellular network, UEs' resources, including storage, energy and services provided, are virtualized to support flexible network access for users. In other words, UE resources are dynamically allocated to requested users to meet multimedia service requirements. To effectively manage network structure and allocate network recourse, we proposed an innovative edge computing architecture, called knowledge-centric edge (KCE). As shown in Fig.1, three layers are proposed in the KCE architecture, namely physical layer, knowledge edge layer and virtual management layer. Physical layer is mainly responsible for collecting and sharing data among users. Knowledge layer is used to establish and control the connections between users based on learning-based D2D communication systems such as social and smart transportation D2D networking systems, which requires the integration of multidisciplinary theories and methods, e.g., sociology, management and psychology. The virtual management layer aims at creating optimal path(s) that connects users.

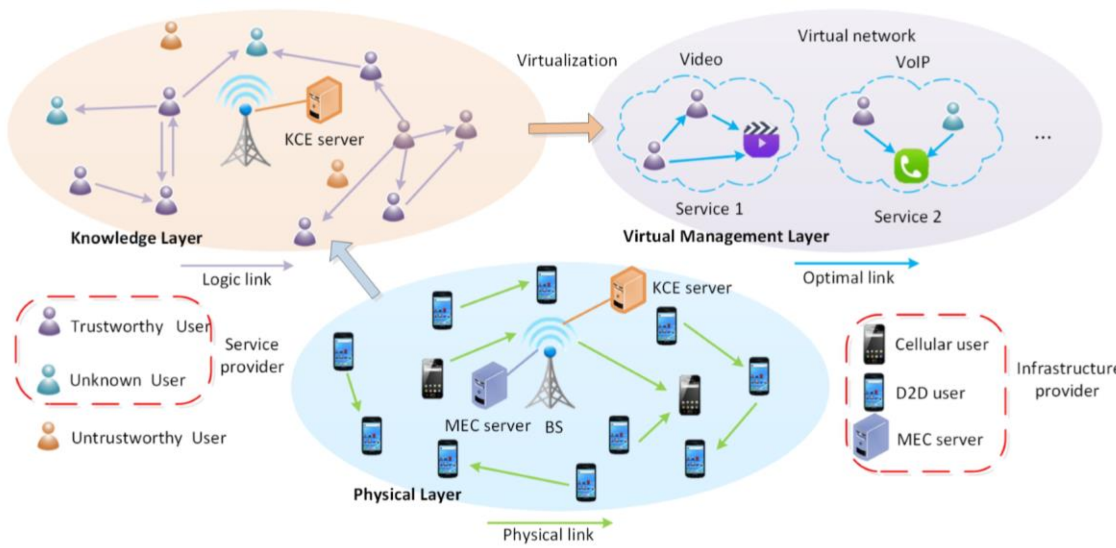


Figure 1. Architecture for KCE

Physical Layer: The physical layer contains various types of infrastructure providers (InPs), e.g., typical cellular users, D2D users, MEC servers and KCE servers. It is responsible for offering physical resources to users, at the same time, collecting data from users and reporting users' features to KCE servers.

Knowledge Layer: To dynamically manage the network connections among users, knowledge layer is introduced in our architecture. Based on users' data collected from the physical layer, KCE servers can derive new knowledge such as the social connection among users, and use the new knowledge to control the users' connection on the physical layer. Such as, several InPs can collect the data for the KCE servers and be selected as the service providers (SPs) by the KCE server.

Virtual Management Layer: In order to to optimize resource allocation, network resources from both base stations and UEs will be virtualized in the D2D-assisted network. Through WNV, physical resources on UEs are virtualized. In particular, knowledge obtained by the KCE server, e.g., tie strength among users, connection frequency between UEs, and trustworthiness between users, will be exploited to establish and manage the virtual network. After the resource virtualization of SPs, the KCE server will assist requested users to establish links with the corresponding SPs satisfying the multimedia service requirements.

3 Physical Layer

While considering resource management and allocation on UEs, UEs' limited capacity and their ability of providing service must be considered. To promote data sharing among users, UE resource allocation is performed by the KCE server in the physical layer. A few research questions need to be answered here. What contents will be cached on which devices so that the overall network performance is optimized? How to effectively support the multimedia service requirements of requested users?

3.1 Caching Strategy

SPs will be identified by the KCE server, based on users' social connections, similarity of attributes, and interacting history. What contents will be cached on which SPs, however, is still a challenging problem, due to the dynamic nature of users accessing contents in the network. We propose an innovative caching strategy to cache contents on users based on their potential in providing cached data for other users.

To efficiently cache contents, we propose a caching strategy on mobile users, considering their social connections. According to the social connections between users, users are divided into different social groups (communities). Users who are strongly associated with a social group are regarded as members of the corresponding community. These users can cache contents with a high local popularity because it might provide cached data to others in the same community. On the contrary, users belonged to several communities will cache contents with high global popularity. Because these users connect to several different groups, they can better assist global data dissemination and the globally popular contents can be quickly accessed by many users. Caching contents with different (local or global) popularities can effectively improve the cache hit rate on individual users.

3.2 Serving Capacity

Another important factor affecting network performance is whether SPs have the capacity to provide requested services. The serving capacity of selected SPs directly influences whether a task will be completed in the proposed architecture. For a MEC server, its serving capacity, e.g., the resource needed to process input data, should be consider in the physical layer. On the other hand, a UE's serving ability is not only related to physical capacity (e.g., storage, energy and communication resources) but also its social attributes, e.g., social connections in the network. In order to compute the serving ability of UEs, models are needed to understand the willingness of a UE, regarding to providing services for other users. In addition, both vertical and horizontal relationship strength between users needs to be fully exploited. Horizontal and vertical relationships connect users with similar behaviors and users with different characteristics, respectively. Overall, the servicing capacity of different SPs should be accurately evaluated so that a requested task can be accomplished at the earliest time.

4 Knowledge Layer

As mentioned above, based on users' data collected from the physical layer, KCE servers can derive new knowledge through learning. Take social computing as an example, after users are selected as SPs by the physical layers, they may be non-cooperative in terms of providing services for others. User's non-cooperative behaviors are mainly caused by either objective reasons (e.g., users run out of storage) or subjective reasons (e.g., users are selfish and not willing to provide service for others). As shown in Fig.2, an effective incentive mechanism is needed to encourage users to participate in the data forwarding and transmission. On the other hand, trust management should be deployed to choose trustworthy users to be the SPs.

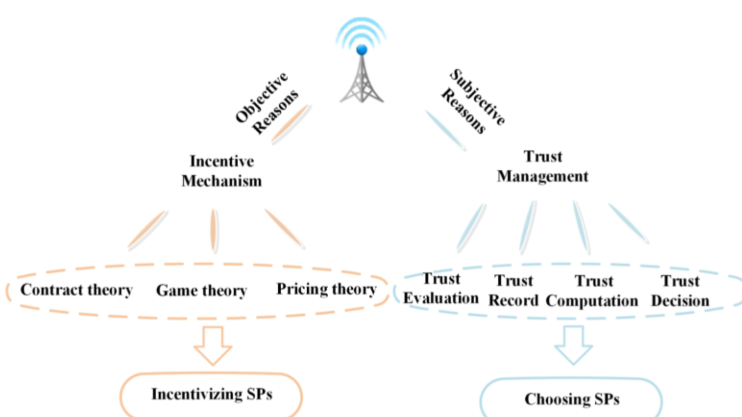


Figure 2. User Incentive Mechanism and Trust Management in Knowledge-Centric Edge Computing.

The goal of introducing a knowledge layer in the architecture is twofold. In this layer, users' social connections (existing and potential) are analyzed so that users can be divided into multiple groups. Users in the same groups are believed to have the potential to share data in the future. On the other hand, to promote regular users to provide service for others, an incentive scheme and a trust management are needed. In this section, we will elaborate on how these two goals are realized in the knowledge layer.

4.1 Social Group Identification

By collecting and sharing data via D2D communication, KCE servers can derive the knowledge about users social group information. In particular, by exploiting users' social attributes, such as interests in common, background and similarity, users can be divided into several overlapping groups. Within a certain group, users tend to encounter with each other more frequently than users in other groups. In addition, the encountering duration between users in the same group is relatively long. The encountering probability and encountering duration among users can be computed by the KCE server, based on the social features of users.

Users may have multi-dimensional social attributes, so a user may belong to multiple groups/communities. By analyzing network structure of users with different attributes, overlapping among various groups can be detected. Therefore, detecting the overlapping communities is crucial in the knowledge layer, and users within the overlapping groups can better assist data dissemination in the whole network.

4.2 Incentive Mechanism

In a mobile edge network, the resources of users are allocated dynamically, according to the services requested by other users. The achievable benefit is significantly influenced by whether users behave in a cooperative manner. From the economic perspective, the profit of all stakeholders in the system should be balanced. Therefore, an incentive mechanism is needed on the KCE server to facilitate the content sharing among users.

Targeting at the aforementioned problems, game-based, pricing-based, and contract-based incentive mechanisms can be applied on the KCE server. For example, the Stackelberg game theory can be used to encourage users to cooperatively participate in data transmission. A Stackelberg game consists of two game players: the leader and the follower. The leader (or the KCE server) first provides a strategy in the game, and the follower (or the chosen SPs) gives its optimal strategy. The leader observes the optimal strategy given by the followers and constantly adjusts its optimal strategy. Through negotiations and adjustments, the Nash equilibrium of system can be finally achieved. Besides, the benefits on SPs are directly proportional to the amount of forwarded data, which further encourages users to participate in cooperative data transmission.

4.3 Trust Management

Due to selfishness, users in the network may present a non-cooperative behavior [13]-[14]. To select the trustworthy SPs from InPs, trust management is critical in the proposed architecture. In general, the trust management module mainly includes four components: trust evaluation, trust recording, trust computation and trust decision. In the following, we will describe the functionality of each component in detail.

Trust evaluation is the first step that considers the trust relationships between SPs and the requested users. Psychological researches have shown that the trustworthiness between users can be divided into three aspects: cognition, emotion and behavior. Therefore, multi-dimensional trust relationships between SPs and requested users need to be considered in trust evaluation, including cognitive trust, emotional trust and behavioral trust. Cognitive trust is defined as a user's belief of the ability and reliability of cooperation of other users. For example, a trustworthy SP may be acquainted with the requested user as they interacted with each other on social networks. Emotional trust expresses the degree of interest of a user, regarding to data relay. Emotional trust is usually established among the like-minded people. Behavioral trust is established by mutual interaction between users, e.g., helping each other on relaying and forwarding data. Mutual interactions can be divided into two categories: direct interaction and indirect interaction. When mutual interactions exist in users, there is certain behavioral trust between them. By recording users' interaction history, the KCE server can evaluate the trustworthiness between users by various trust models, e.g., the three-valued subjective logic model [15]. Considering the above-mentioned three types of trust, the KCE server can evaluate the trustworthiness between potential SPs and the request user. It can make trust decisions based on decision theories, e.g., decision-theoretic rough set based on the naive bayesian. As a result, users can be divided into trustworthy users, uncertain users and untrustworthy users, respectively.

5 Virtual Management Layer

In virtual management layer, UE resources are virtualized and managed by the KCE server. To ensure the optimized physical resources allocation, the KCE server will decide how D2D-assisted virtual UE resources are allocated for different communications. In particular, we design a delay-aware resource allocation schemes that mainly matches UE resources to satisfy the multimedia service requirements, by dynamically slicing UE resources into virtual groups. As for multimedia services, a low latency is always expected. As is shown in Fig. 3, to meet multimedia service requirements, we firstly divide UE resources into different groups, and then share the physical resources to obtain the minimum. In particular, users who are trustworthy and have higher serving potentials are chosen for faster transmission in virtual management layer.

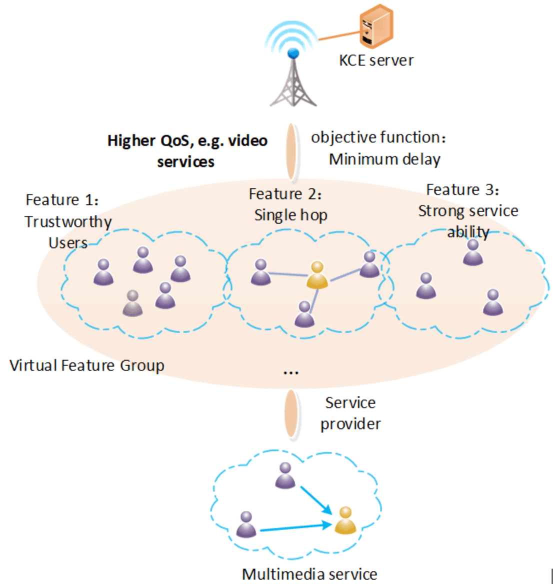


Figure 3. Structure for Delay-aware Resource Allocation

In addition, as is shown in fig.4, the SVC sharing mechanism based on the social group is proposed in virtual management layer, which can transmit the video data between users at various bitrates. The user can decide what to share in the local cache according to the similarity of proximity ones, which not only provides differentiated mutimedia services, but also effectively reduces the transmission energy and transmission delay.

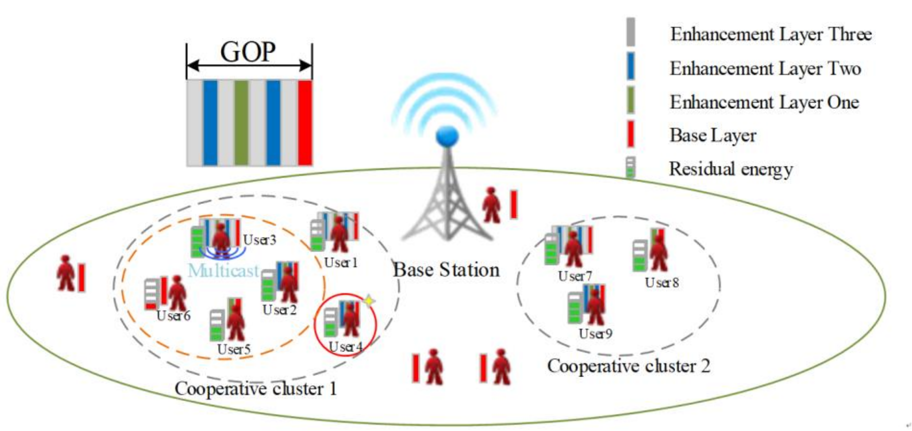


Figure 4. SVC-based edge cooperation for video transmission

6 Evaluation and Result analysis

In this section, we evaluate the performance of the proposed KCE system combining the social and virtualized D2D communication (SVDC).

6.1 Simulation Settings

We consider that users are randomly deployed in a 100 × 100 m^2 area. There is a total of 1000 multimedia files available for transmission in the system. The MEC server is able to store 1000 multimedia files, and each UE can cache 10-50% of these multimedia files. We assume that the maximum transmission range of D2D communication is R = 50m. Other simulation parameters are given in Table I.

TABLE I: Simulation Paremeters

Paremeter	Value
Carrier frequency	2.5GHz
System bandwidth	5MHz
eNodeB Tx power	46dBm
Maximum UE Tx power	29dBm (D2D mode: 20dBm)
RB size	12 sub-carriers 0.5ms
Path loss (cell link)	128.1+37.6log(d),d[km]
Path loss (D2D link)	40log(d)+30log(f)+49, d[km],f[Hz]
Thermal Noise	-174dBm/Hz

6.2 Result Analysis

The effectiveness of caching strategy in edge network directly affects the network performance of MEC, such as task offloading ration. The delay of requested users can be greatly improved by caching contents on mobile devices, if the caching strategy is carefully designed. To evaluate the performance of proposed architecture, we evaluate the system's performance with two caching strategies. In the randomness based caching strategy, the probability of a file being cached on a user is proportional to probability that this particular file is requested. In the popularity based caching strategy, each device only stores the most popular contents.

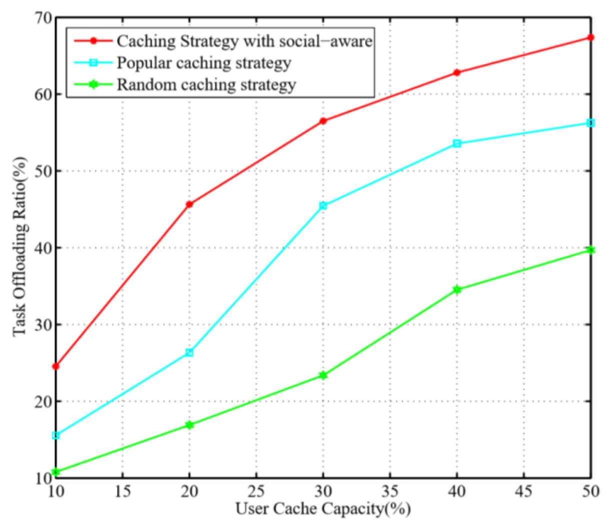


Figure 5. The relationship between task offloading radio and UE's cache capability

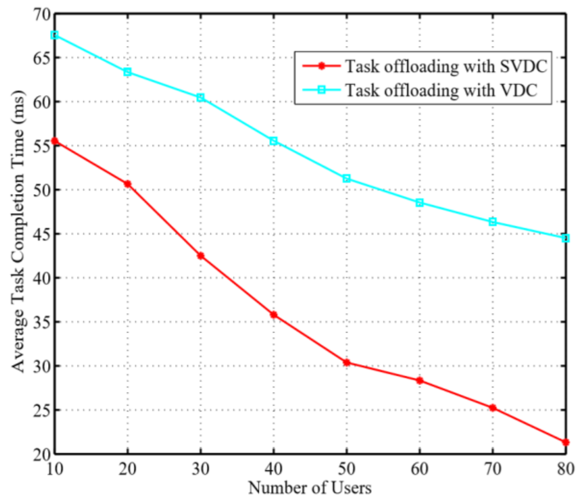


Figure 6. The relationship between number of users and average task completion time

Fig. 5 reflects the relationship between the task offloading radio and the UE's cache capability. As the cache capability increases, the task offloading radio of proposed system is 27.5% and 56.2% higher than those obtained from random and popularity based caching strategies. This is because unpopular contents are more likely to be shared in local regions, where users are closer to each other. These contents are simply ignored in random and popularity based caching strategies. For those globally popular contents, as they are cached on users in the overlapping communities, they can be quickly disseminated in the networks.

Besides, in order to demonstrate effectiveness of our proposed resource allocation schemes for computation offloading in MEC, we evaluate the average task completion time of individual users. Fig. 6 shows the relationship between the number of users and the average task completion time. When the number of users increases, the average task completion time of the proposed system is 36.6% lower than that obtained from pure virtualized D2D communication (VDC) systems. This is because that not all users can be virtualized as the SPs to enable trustworthy network access. Some non-cooperative users have a severe impact on the average task completion time.

7. Conclusion

In this paper, we propose a knowledge-centric cellular network architecture with D2D communications. First, we design a D2D-assisted virtualized edge cellular network architecture, consisting of physical layer, knowledge layer, and virtual management layer. We then describe the challenges and potential solutions in the proposed system, including knowledge extraction, social awareness, incentive schemes, trust management mechanism, and optimized resource allocation on UEs. As such, the optimization objective functions are designed to meet the multimedia service requirements. Simulation results show that the proposed architecture can greatly improve the network performance, in terms of the less backhaul consumption and reduced transmission latency. Future work includes applying some intelligent technologies such as machine learning and search technologies in edge network for smart transmission and IoT applications. For example, mobile edge computing will play a vital role in future connected vehicular system in which vehicles exchange information among each other via the help of mobile edge nodes available on the road. The information obtained from the knowledge layer, e.g., the trustworthiness between devices, could be used to facilitate the communications among IoT devices and to realize a trustworthy and reliable data transmission. Last but not the least, the proposed framework provides a new perspective in understanding the resource allocation problem in future cellular network, which is unavailable in previous literature.

Acknowledgement
This work is supported in part by the Natural Science Foundation of China (61771082) and Program for Innovation Team Building at Institutions of Higher Education in Chongqing (CXTDX201601020).

References

[1] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016-2021, Cisco Syst., San Jose, USA, 2017.

[2] K. Zhang, Y. Mao, S. Leng, Q. Zhao, "Energy-Efficient Offloading for Mobile Edge Computing in 5G Heterogeneous Networks," IEEE Access, vol. 4, pp. 5896-5907, 2016.
[3] M. Richart, J. Baliosian, J. Serrat, J. Gorricho, "Resource Slicing in Virtual Wireless Networks: A Survey," IEEE Transactions on Network and Service Management, vol. 13, no. 3, pp. 462-472, 2016.
[4] C. Liang, F. Richard Yu, "Wireless virtualization for next generation mobile cellular networks," IEEE Wireless Communications, vol. 22, no. 1, pp. 61-69, 2015.
[5] J. Liu, J. Wan, B. Zeng, Q. Wang, H. Song, M. Qiu, "A Scalable and Quick-Response Software Defined Vehicular Network Assisted by Mobile Edge Computing," IEEE Communications Magazine, vol. 55, no. 7, pp. 94-100, 2017.
[6] K. Liang, L. Zhao, X. Chu, H. Chen, "An Integrated Architecture for Software Defned and Virtualized Radio Access Networks with Fog Computing," IEEE Network, vol. 31, no. 1, pp. 80-87, 2017.
[7] M. Li, F. Richard Yu, P. Si, H. Yao, "Energy-efficient M2M Communications with Mobile Edge Computing in Virtualized Cellular Networks," in Proceedings of 2017 IEEE International Conference on Communications (ICC), Paris, France, pp. 1-6, May. 2017.
[8] D. Wu, Q. Liu, H. Wang, D. Wu, and R. Wang, "Socially Aware Energy Efficient Mobile Edge Collaboration for Video Distribution," IEEE Transactions on Multimedia, vol. 19, no. 10, pp. 2197-2209, 2017.
[9] D. Wu, J. Yan, H. Wang, D. Wu, and R. Wang, "Social Attribute Aware Incentive Mechanism for Device-to-Device Video Distribution," IEEE Transactions on Multimedia, vol. 19, no. 8, pp. 1908-1920, 2017.
[10] D. Wu, L. Zhou, Y. Cai, "Social-Aware Rate Based Content Sharing Mode Selection for D2D Content Sharing Scenarios," IEEE Transactions on Multimedia, vol. 19, no. 11, pp. 2571-2582, 2017.
[11] Z. Zhang, L. Luo, L. Wang, "D2D multicast retransmission algorithm in mobile cloud based on SINR constraint," China Communications, vol. 13, no. 8, pp. 41-52, 2016.
[12] Z. Wang, L. Sun, M. Zhang, H. Pang, E. Tian, W. Zhu, "Propagation- and Mobility-Aware D2D Social Content Replication," IEEE Transactions on Mobile Computing, vol. 16, no. 4, pp. 1107-1120, 2017.
[13] M. J. Wang, Z. Yan, V. Niemi, "UAKA-D2D: Universal Authentication and Key Agreement Protocol in D2D Communications," Mobile Networks and Applications, Springer, Vol. 22, no. 3, pp. 510-525, 2017.
[14] M. J. Wang, Z. Yan, "A Survey on Security in D2D Communications," Mobile Networks and Applications, Springer, Vol. 22, no. 2, pp. 195-208, 2017.
[15] G. Liu, Q. Yang, H. Wang, X. Lin, M. P. Wittie, "Assessment of Multi-hop Interpersonal Trust in Social Networks by Three-Valued Subjective Logic," Proceedings IEEE INFOCOM, Toronto, ON, pp. 1698-1706, 2014.

Toward an Event-Oriented Indexable and Queryable Intelligent Surveillance System

Seyed Yahya Nikouei[1], Yu Chen[1], Alexander Aved [2], Erik Blasch[2],
[1]Dep. of Electrical and Computer Engineering, Binghamton University, Binghamton, NY, USA
[2]The U.S. Air Force Research Laboratory, Rome, NY, USA
{snikoue1, ychen}@Binghamton.edu; {alexander.aved, erik.blasch.1}@us.af.mil

1. Introduction

Advances in networking, intelligence, and media available in urban areas attracts people towards a more comfortable lifestyle. Urbanization at an unprecedented scale and speed incurs significant challenges to city administrators, urban planners and policy makers. In order to efficiently manage the cities functions and be responsive to dynamic transitions, surveillance systems are essential for situational awareness (SAW). Nowadays, a prohibitively large amount of surveillance data is being generated every second by ubiquitously distributed video sensors. For example, North America alone has more than 62 million cameras in the year 2016. These cameras are connected to powerful data centers through communication networks and the delivery of surveillance video streams creates a heavy burden on the network. Researchers have shown that video streaming accounts for 74% of the total online traffic in 2017 [1].

Since the first generation video surveillance systems known as Close Circuit TV (CCTV) were introduced in 1960s, urban surveillance mechanisms adapted to the changing technology. Compared with today's edge computing paradigm, CCTV-like surveillance systems are limited because:

- The network is "best effort" based which means not only transmission of the video data suffers delays and jitters, the data may get lost or dropped because of network congestion.
- The raw-data transmission is "dedicated" which wastes resources in the communication network and at the data center, because not all data is globally significant or worthy to be stored for long time.
- An agent needs to pay "full attention" to the video to capture any emergency in real-time. Obviously this naïve approach is not scalable, and there are several architectures introduced based on computer vision techniques and make decisions based on machine learning algorithms. However, to date there is not a system that is able to meet the performance requirements like real-time, good scalability, and robustness [2].
- An agent employs "working memory" as computing capabilities afforded only searching for a specific target of interest or focusing on a special feature. Meanwhile, today's multimedia forensics desires real-time or near real-time searching by scanning through the large surveillance video record base.

It is very challenging to immediately analyze the objects of interest or zoom in on suspicious actions from thousands of video frames. Making the big data indexable is critical to tackle the object analytics problem. It is ideal to generate pattern indexes in a real-time, on-site manner on the video streaming instead of depending on the batch processing at the cloud centers. The modern edge-fog-cloud computing paradigm allows implementation of time sensitive tasks at the network edge. In this paper, a novel event-oriented indexable and queryable intelligent surveillance (EIQIS) system is introduced leveraging the on-site edge devices to collect the information sensed in format of frames and extracts useful features to enhance situation awareness.

The rest of this paper is organized as follows. Section 2, briefly discusses background knowledge and relative work. Section 3 highlights the main challenges in the real-time surveillance. Section 4 introduces the rationale of the proposed indexable and queryable surveillance system. A preliminary study is presented in Section 5, which validates the concept and shows the feasibility of the system architecture. Finally, Section 6 concludes the paper with future research directions.

2. Background Knowledge and Related Work

Today, most available surveillance systems archive streaming video footage to be used off-line for forensics analysis [3]. Communication delays and uncertainties associated with the data transfer from image sensors to a remote computing facility limit implementation of the online surveillance tasks. However, delay sensitive applications require on-line processing. Thanks to the recent development of lightweight machine learning (ML) algorithms that require less computing power and storage space, more processing can be migrated to the edge of the network [4], where no more delay is incurred for data transmission. For tasks like anomalous behavior detection that is not affordable at the edge, instead of directly outsourcing the job to the remote cloud, near-site fog nodes are powerful enough for complex data analytics tasks. For instance, in a smart transportation application following a hierarchical system architecture, data is accessed by the sensors implemented on buses and transferred to a fog node where contextualization and decision making happens [5]. For video surveillance systems, the remote cloud is mainly used for profile building, pattern analysis, and long term historical record analysis.

In general, a smart surveillance system includes three layers as shown in Fig. 1. In the first layer, image analysis, the input camera frame is given to an edge device and the low-level features are extracted [6], [7]. The edge devices are able to conduct object detection and object tracking tasks [8], [9]. The intermediate-level, considered as the fog stratum, is in charge of mode recognition for action recognition, behavior understanding, and abnormal event detection. Finally, the high-level, cloud center, is focused on systems analysis including historical profile building, global statistical analysis, and narrative reporting. Connections among the edge, fog and cloud nodes present challenges in terms of overall platform, connections, quality of service (QoS) requirements, and preserving privacy and security.

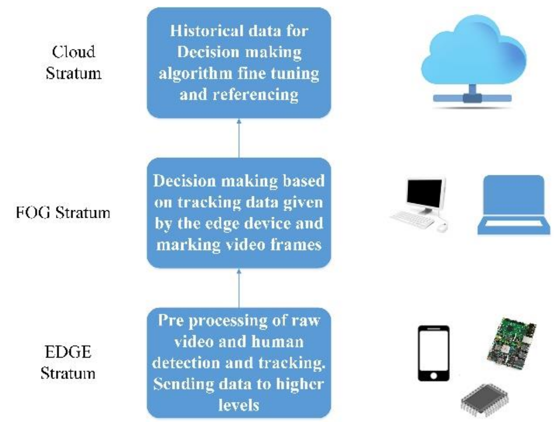


Figure 1. Hierarchy architecture of a smart security camera implemented in fog networking architecture.

The first step of a video surveillance system is to simultaneously track and identify (ID) (STID) the objects of interest in the video [18], [19]. STID continues to be a challenging task performed on the edge of the network [10]. Nowadays, once an event incurred, the operators need to spend considerable amount of time to go through the footage and look at videos from different cameras in order to find a specific target. Even in the next generation surveillance systems that are combined with image processing techniques for better decision making, performing a search in real-time or near real-time is very challenging [2], [20].

Ideally, the surveillance system is expected to be able to quickly and automatically identify the clips of interest based on a given query. Earlier researchers have proposed to adopt video parsing techniques that automatically extract index data from video and store the index data in relational tables [11], [21], [22]. The index is used through SQL queries to retrieve events of interest quickly. However, this approach cannot meet the performance requirements of online, real-time, operator-in-loop interactions. Future smart surveillance video streams have to be indexable and queryable such that the operator is able to obtain the information of interest instantly.

3. Real-time Queryable Surveillance: Architecture and Challenges

This section introduces an edge-fog-cloud computing based system architecture to achieve event-based indexable and queryable intelligent surveillance (EIQIS). It is non-trivial to extract features in real-time and use them as indexes to conduct online query on surveillance video streams [23]. Advances in machine learning, multi-modal data fusion, and physics-based and human-derived information fusion (PHIF) show promise for EIQIS. Current systems are designed to enhance user responsibilities to include security, surveillance, and forensics. Typically, the user provides a standing query that the image processing is to provide event triggers [14], [15]. The user would like the system to do the functions autonomously, however, the ultimate design would include a combination of humans in, on, or out-of the loop (HIL, HON, HOON).

In order to have a smart surveillance system raise an alarm when something abnormal is detected, each captured frame that is processed requires knowledge of the proceeding frames. A three layer edge-fog-cloud hierarchical architecture reduces the delays that are incurred when the frame is transferred to a remote cloud center. The more processing that is migrated to the network edge, the faster the features are obtained and indexes are constructed because of the close proximity of the edge node to the geo-location of the camera. Meanwhile, due to the constraints on computation and storage capacity at the edge devices, more computing or data intensive tasks are outsourced to more powerful cloud.

The first layers is the edge camera, it should be mentioned that most reliable detection and tracking algorithms are dedicated for specific surveillance applications. Running them in a resource constrained environment that requires the algorithm to be a light weight version of the original does not help the accuracy. Thus, finding better methods is a contemporary research topic [24].

Once a frame is captured by the image sensor, it will be either transferred to the edge device that is connected via a local area network (LAN) connection or processed on-site if it is a smart camera (edge device) with sufficient computing power. The edge node has limited computing power and so all computing intensive event detections cannot be executed at this level. The edge device conducts pre-processing using a convolutional neural network (CNN), which will identify the objects of interest and give their positions in the image frame. Even with small architectures with few layers that reduce the overall computation complexity, CNNs are heavy for the edge device [10]. The edge device cannot afford to execute the CNNs more than couple of times per second. Therefore, in order to reach a higher resolution of the detection, the bounding box around the object of interest is given to a tracker algorithm that uses an online learning algorithm to follow the object in each frame until it moves out of the frame. Each time the CNN runs, the newly found bounding boxes are sent to a fast tracker such as the Kernelized Correlation Filter (KCF), improving the speed. It should be noted that although newer and powerful edge nodes are made every day, with more features to be extracted, a longer processing time is needed. Consequently, the key for the real-time application is a trade-off between the speed and the amount of features to be extracted in each frame.

After each object is detected and tracked, features can be extracted. These features might include, but are not limited to the current position and speed the object is walking, the direction of the walk and some other physical features such as the angles the other parts of the upper body parts create and so the pose of the pedestrian [26]. For each detected pedestrian, there is a table that is updated with each frame and includes a key and value for features extracted from the video. The actual video may not be needed to be transferred to the fog level device where the decision making code is executed.

The edge device is designed to conduct immediate techniques such as feature extraction, while the advanced analytics is outsourced to a more powerful, near-site node. Several edge devices from several camera feeds can be connected to a fog node, which conducts feature contextualization, indexing, and storage. One of the challenges in a surveillance system is the security of the connection between the edge and fog. Although there are new promising technologies to address privacy/security, like blockchain technology [13], more development is needed to make them light weight and robust for the smaller networks with low power. The features transmitted to the fog node can be contextualized to support decision making [25]. Valuable data in the contextualization include: The location of the camera, time of the footage, terrain information, semantic ontologies of descriptors, etc. For example, while it is normal for people to walk and stand in a campus building, it can be considered as abnormal if it is late at night when the building should be close. Also, connecting several cameras in the same area to the same fog node will give the fog the ability to look at the monitored area from different perspectives, illuminations, and contexts.

Another challenge that the surveillance community faces is the decision support algorithm, which includes supervised, unsupervised, and semi-supervised methods. The, the lack of labelled data for unknown situations, requires methods in semi-supervised training to better characterize abnormal situations. The answer may include the location and several other factors and sequence of events lead to abnormal behavior detection. Also, the security camera and the functionality of the place surveillance may differ from one to the other which makes it very difficult to differentiate between normal and abnormal activity.

The historical analysis, profile building, and situation analysis are conducted by the most powerful node in the edge-fog-cloud architecture hierarchy, the cloud. The decisions making and the detection of false alarm and the features that raised the alarm are sent for future fine tuning of the algorithms and also some analytical studies. Figure 1 shows the interconnections of the nodes in the network described in this section.

4. Making the Video Streams Indexable

The usability of any exploited video is based on what is stored or indexed or fast retrieval, such as content-based image retrieval. The surveillance video streamed to the edge device enables features extracting for decision making. Decision making is based on the real-time search query. The real-time video search will make the job of the operator/user easier by giving instances of the video that are asked for in a query to the system. Search string is the query that is given to the fog node. The fog node is the ideal level to handle search requests where contextualized information from close by cameras is stored. The following describes how a query is handled at the fog layer:

1- The fog node receives the query and will check the eligibility of the machine asking for the information. The access level of the nodes in such a network is defined in a smart contract in a blockchain enabled security platform.
2- The fog node searches for the query in the index table to find the corresponding camera, timestamp and other information based on the real-time features provided and select them if any.

3- The fog node answers the search requester based on the information found.
4- Then the operator selects the cameras with the query and has the live feed or recorded clips (it is assumed that the operator has access to the edge device in charge of the camera of interest if he/she has access to the higher-level fog).

The operator thus can search the video streams in real-time.

Indexing requires the association of complementary information (hashed, correlated, and linked) with the video frame for storage. Using the mapping table affords fast information retrieval. Considering the indexing table the same as the features simplifies the search operation. While there are many features extracted from the video, there might be several different indexes that are required by the system administrator. Features are generated in order to make a decision for the actions of the object in the video. However, indexes that are based on features might include more options. There are two scenarios that are plausible. First, the fog node uses the same features and adds context to make the data useable as the index table. Second, the fog node uses several edge devices (perform as microservices) to extract features required and creating a table to be used as indexes based on the resulting features.

4.1 Indexing

In order to facilitate faster search results, one known method used today in search engines and operating systems, is to create an index table which is used later for finding search queries. Indexing means to have a key and value table of features that are of interest and once the keys are searched for (in query format), the corresponding values are the results of the search that gives certain files that contain the query. This way the search is faster and there is no need to scan all files for the key values that are searched for. The same principle applied to the video file captured by the surveillance cameras results in efficient and real time operations. Based on the index table points to the corresponding edge device, the camera live or recorded footage clips are identified and sent to the query sender.

Once the camera captures each frame, an edge device extracts features in real-time or near real-time from the video and the features are transferred to a fog node. After the contextualization of the features, they can be used as the indexes for querying when the operator needs to find something instantly. For example, if the operator is looking for moments that there are congestion of people on the campus in the late night hours. The search can be directed to the exact hours and locations, then look for features that report more than ten people or more at the same frame. Using the query-based parameters inherent in the index table will lead to the corresponding video clips faster and the operator can look for incidents that have the exact search keys. The EIQIS method is obviously more efficient than having to check all the camera footage security systems to find what imagery is of interest.

4.2 Features VS. Indexes

Creating the indexes for the extract features that are useful for video search supports historical analytics. However, the features that are of interest in the abnormal behavior detection may not support an operator search, be enough, or exactly the same as the indexes (key values) that are applicable in usual search. Figure 2 shows a scenario in which more feature extraction from the video is needed. The job can be divided into more than one edge devices and each feature can be handled as a microservice [12]. Microservices is defined as a separate piece of program that provides a service to a bigger piece of program. In this case the feature extraction can be considered as the microservice that is used in the video indexing platform. More features can be extracted as a result of this architecture. If any indexes need to be added, simply adding the service to the platform can expand the scope of the indexes that are used.

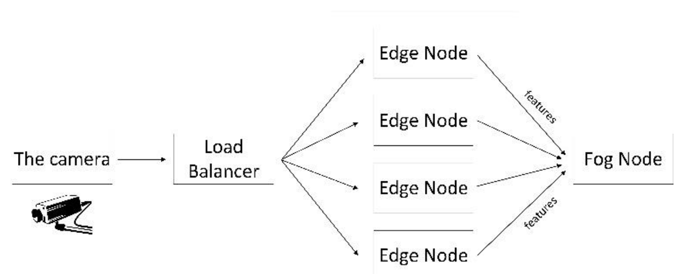


Figure 2. Edge feature extraction as microservices for indexing purposes.

5. A Preliminary Case Study

A preliminary proof-of-concept prototype has been built to validate the feasibility of EIQIS [13]. It shows that the edge devices are capable of extracting and sending features in real-time to the fog layer. The features are written into a text file and sent to fog through a secure channel. The features are synchronized with every node of the network for added security. Figure 3 is an example of features stored in the fog in a key value manner and Fig. 4 is graphical output of the edge device, where the device adds a bounding box around the object (e.g., person, vehicle, other) of interest and the box follows the object. Figure 4 presents several moments that are challenging to be detected. It is a proof showing an acceptable performance of the edge device. The real environment validates the feasibility of the proposed system. The prototype model run on two Asus Tinker Boards with the configuration as follows: 1.8 GHz 32-bit quad-core ARM Cortex-A17 CPU, the memory is 2GB of LPDDR3 dual-channel memory and the operating system is the TinkerOS based on the Linux kernel. The fog layer functions are implemented on a laptop, in which the configuration is as follows: the processor is 2.3 GHz Intel Core i7 (8 cores), the RAM memory is 16 GB and the operating system is the Ubuntu 16.04. A private blockchain network is implemented to secure the feature data transferring from edge to fog. Our private Ethereum network includes four miners, which are distributed to four desktops that are empowered with the Ubuntu 16.04 OS, 3 GHz Intel Core TM (2 cores) processor and 4 GB memory. Each miner uses two CPU cores for mining task to maintain the private blockchain network and the resulting blocks are synchronized through the whole network so every node has a copy of the latest block. The data transfer between the fog node and the miner is carried through an encrypted channel. Before the fog node can secure the features, there should be no adversaries who can temper with the surveillance data. Python based socket programming language is used for both ends of the channel. More details of the prototype are reported in [13].

```
14.46.33;10;326;1;(360, 203);West;1.64012194669
14.46.33;10;327;1;(352, 203);South-West;3.71618083521
14.46.33;10;328;1;(345, 204);West;3.30151480384
14.46.34;10;329;1;(336, 210);West;3.86393581727
14.46.34;10;330;1;(325, 206);West;4.31045241245
14.46.34;10;331;1;(338, 206);West;2.22036033112
14.46.34;10;332;1;(325, 222);South-West;3.30151480384
14.46.34;10;333;1;(315, 236);South-West;4.38634243989
14.46.34;10;334;1;(305, 249);South-West;4.98196748283
14.46.34;10;335;1;(306, 250);South-West;4.79270278653
14.46.34;10;336;1;(307, 251);South-West;5.46443043693
14.46.34;10;337;1;(307, 250);South-West;3.32866339542
14.46.34;10;338;1;(291, 263);South-West;3.61247837364
14.46.35;10;339;1;(292, 264);South-West;1.98494332413
14.46.35;10;340;1;(292, 263);South-West;1.91049731745
14.46.35;10;341;1;(292, 263);South-West;1.92093727123
```

Figure 3. Feature table for each camera (time, camera id, frame, pedestrain_num, possition, direction, reletive speed).

Figure 4. Viualized features in real-time.

6. Conclusions

Many surveillance systems available today cannot meet the performance requirements raised from real-time, human-in-loop interactive operations. The event-oriented indexable and queryable intelligent surveillance (EIQIS) edge-fog-cloud hierarchical architecture is promising for real-time or near real-time applications, which allows instant querying on the online surveillance video streams to give more time to first responders. In this paper, the architecture toward an event-oriented, indexable, queryable smart surveillance system is introduced. The proposed system enables query of video in real-time based on an index table, which is created on top of the features that are extracted on-site by edge computing nodes. This intelligent surveillance system enables the operator to search for scenes or events of interest instantly. A preliminary study has validated the feasibility of the proposed architecture.

References

[1] N. Chen, Y. Chen, E. Blasch, H. Ling, Y. You, and X. Ye, “Enabling smart urban surveillance at the edge,” In *Smart Cloud (SmartCloud), 2017 IEEE International Conference on*, pp. 109-119. IEEE, 2017.

[2] V Tsakanikas, T Dagiuklas, “Video surveillance systems-current status and future trends”, *Computers & Electrical Engineering*, Nov 2017.

[3] N. Chen and Y. Chen, “Smart city surveillance at the network edge in the era of iot: Opportunities and challenges,” in *Smart Cities*. Springer, 2018, pp. 153–176 .

[4] A. Ouaddah, A. Abou Elkalam, and A. Ait Ouahman, “Fairaccess: a new blockchain-based access control framework for the internet of things,” *Security and Communication Networks*, vol. 9, no. 18, pp. 5943–5964, 2016.

[5] F. F. Chamasemani and L. S. Affendey, “Systematic review and classification on video surveillance systems,” *International Journal of Information Technology and Computer Science* (IJITCS), vol. 5, no. 7, p. 87, 2013.

[6] S. Y. Nikouei, Y. Chen, S. Song, R. Xu, B.-Y. Choi, and T. Faughnan, “Real-time human detection as an edge service enabled by a lightweight cnn,” in *the IEEE International Conference on Edge Computing*. IEEE, 2018.

[7] S. Penmetsa, F. Minhuj, A. Singh, and S. Omkar, “Autonomous uav for suspicious action detection using pictorial human pose estimation and classification,” *ELCVIA: electronic letters on computer vision and image analysis*, vol. 13, no. 1, pp. 18–32, 2014. .

[8] N. Khanezaei and Z. M. Hanapi, “A framework based on RSA and AES encryption algorithms for cloud computing services,” in *Systems, Process and Control (ICSPC), 2014 IEEE Conference on*. IEEE, 2014, pp. 58–62.

[9] D. Yu, Y. Jin, Y. Zhang, and X. Zheng, “A survey on security issues in services communication of microservices-enabled fog applications,” *Concurrency and Computation: Practice and Experience*, p. e4436, 2018.

[10] S. Y. Nikouei, Y. Chen, S. Song, R. Xu, B.-Y. Choi, and T. R. Faughnan, "Intelligent surveillance as an edge network service: from harr-cascade, svm to a lightweight cnn," *arXiv preprint arXiv*:1805.00331, 2018.

[11] A. Hampapur, L. Brown, R. Feris, A. Senior, C.-F. Shu, Y. Tian, Y. Zhai, and M. Lu, "Searching surveillance video," in *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on. IEEE,* 2007, pp. 75–80.

[12] D. Nagothu, R. Xu, S. Y. Nikouei, and Y. Chen, "A Microservice-enabled Architecture for Smart Surveillance using Blockchain Technology," *arXiv preprint arXiv*:1807.07487 (2018).

[13] S. Y. Nikouei, R. Xu, D. Nagothu, Y. Chen, A. Aved, and E. Blasch, "Real-Time Index Authentication for Event-Oriented Surveillance Video Query using Blockchain," *arXiv preprint arXiv*:1807.06179 (2018).

[14] A. J, Aved, E. Blasch, "Multi-INT Query Language for DDDAS Designs," *Procedia Computer Science*, Vol. 51, pp. 2518-2523, 2015.

[15] E. Blasch, A. Aved, "URREF for Veracity Assessment in Query-Based Information Fusion Systems," *International Conf. on Information Fusion*, 2015.

[16] P. Liang, E. Blasch, H. Ling, "Encoding Color Information for Visual Tracking: Algorithms and Benchmark," *IEEE Trans. On Image Processing,* Vol. 24, No. 12, Dec. 5630-5644, 2015.

[17] Y. Zheng. E. Blasch, Z. Liu, Multispectral Image Fusion and Colorization, *SPIE*, 2018.

[18] E. Blasch and B. Kahler, "Multiresolution EO/IR Tracking and Identification" *International Conference on Information Fusion*, 2005.

[19] E. Blasch and L. Hong, "Data Association through Fusion of Target track and Identification Sets," *International Conference on Information Fusion*, 2000.

[20] E. Blasch, J. Nagy, A. Aved, W. M. Pottenger, M. Schneider, R. Hammoud, E. K. Jones, A. Basharat, A. Hoogs, G. Chen, D. Shen, H. Ling, "Context aided Video-to-Text Information Fusion," *Int'l.. Conf. on Information Fusion*, 2014.

[21] R. I. Hammoud, C. S. Sahin, E. P. Blasch, B. J. Rhodes, and T. Wang, "Automatic Association of Chats and Video Tracks for Activity Learning and Recognition in Aerial Video Surveillance," *Sensors*, 14, 19843-19860, 2014.

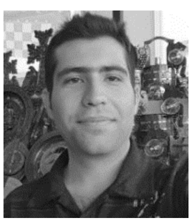
[22] E. P. Blasch, S. K. Rogers, H. Holloway, J. Tierno, E. K. Jones, R. I. Hammoud, "QuEST for Information Fusion in Multimedia Reports," *International Journal of Monitoring and Surveillance Technologies Research* (IJMSTR), Vol. 2 (3), pp. 1-30, Sept. 2014.

[23] K Palaniappan, F Bunyak, P Kumar, I Ersoy, "Efficient feature extraction and likelihood fusion for vehicle tracking in low frame rate airborne video," *International Conference on Information Fusion*, 2010.

[24] H. Li, K. Sudusinghe, Y. Liu, J. Yoon, et al., "Dynamic, Data-Driven Processing of Multispectral Video Streams," *IEEE Aerospace and Electronics Systems Magazine*, Vol. 32, No. 7, pp. 50-57, July 2017.

[25] L. Snidaro, J. Garcia Herrero, et al Context-Enhanced Information Fusion: Boosting Real-World Performance with Domain Knowledge, *Springer*, 2016.

[26] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Transactions on Circuits and Systems for Video technology.* Vol. 18, No. 11 (2008): 1473.



Mr. Sayed Yahya Nikouei earned his Bachelor's Degree from the Electrical & Electronics Department at Shahrekord University, Iran in 2014. He received his Master's degree in 2016 from University of Isfahan, Iran also in electrical & electronics specializing in random hierarchy algorithms to optimize Inverter performance. Currently he is pursuing his doctoral degree at Binghamton University and his research focuses on Machine Learning and pattern recognition methods and Fog/Edge Computing and the applications in Smart Cities such as the smart surveillance system.

Dr. Yu Chen is an Associate Professor of Electrical and Computer Engineering at the Binghamton University - SUNY. He received the Ph.D. in Electrical Engineering from the University of Southern California (USC) in 2006. His research interest lies in Trust, Security and Privacy in Computer Networks, focusing on Edge-Fog-Cloud Computing, Internet of Things (IoTs), and their applications in smart and connected environments. His publications include over 100 papers in scholarly journals, conference proceedings, and books. He has served as reviewer for NSF panels and for journals, and on the TPC of prestigious conferences.

Dr. Alex J. Aved received the BA degree in Computer Science and Mathematics in 1999 from Anderson University in Anderson, Indiana, an MS in Computer Science from Ball State University and PhD in Computer Science in 2013 (focus area: real-time multimedia databases) from the University of Central Florida. He is currently a technical advisor at the Air Force Research Laboratory Information Directorate in Rome, NY. Alex's research interests include multimedia databases, stream processing and dynamically executing models with feedback loops incorporating measurement and error data to improve the accuracy of the model.

Dr. Erik Blasch is a member of the Air Force Research Laboratory (AFRL). He received his B.S. from MIT and PhD from Wright State University (1999) in addition to seven Master's Degrees. He has been with the Air Force Research Laboratory since 1996 compiling over 750 papers, 25 patents, and 5 books. He is a Fellow of IEEE, SPIE, and Associate Fellow of AIAA. His areas of research include target tracking, image fusion, information fusion performance evaluation, and human-machine integration.

A Novel Architecture for Radio Environment Map Construction Based on Mobile Crowd Sensing

Yuanni Liu [1],Jianli Pan[2], Guofeng Zhao[1,3]
[1]*Future Network Research Center of Chongqing University of Posts and Telecommunications, Chongqing ,China*
[2]*Department of Mathematics and Computer Science, University of Missouri-Sanit Louis*
[3]*Optical Communication and Network Key Laboratory of Chongqing, Chongqing,China*
liuyn@cqupt.edu.cn; pan@umsl.edu; zhaogf@cqupt.edu.cn

1. Introduction

It is with great significance to improve the utilization of radio resources. In order to characterize the situation of radio resources timely and accurately, it is necessary to understand the radio information and share it with some applications. Radio Environment Map (REM) is such a feasible system, which covers a large scale of radio environment information, such as available spectrum, geographic information, strategy, geographical features, available services, spectrum regulations, locations and activities of radios, relevant policies and past experiences [1]. Based on these information, further details of the radio environment can be measured, modeled, and then applied to a variety of upper-layer applications.

Currently, most of the REMs aim at small-scale applications. The universal methods to build a REM are deploying sensors in a certain environment to collect the sensing data. The REM is applied to different kinds of networks and applications, which requires the networks and applications to collect different types of data separately. Moreover, the same data can hardly be shared and reused among different applications, resulting in a duplication of data collection and a waste of resources. So there is a great significance to construct a large scale and universal REM, which can integrate data sources of radio environment and avoid the cost of the re-constructing databases. Concerned with the problems stated above, we propose to leverage Mobile Crowd Sensing (MCS) for REM data collection. MCS is a novel sensing paradigm that empowers everybody to contribute sensed or generated data from their mobile devices, aggregates and combines the data in the cloud for crowd intelligence extraction and people-centric service delivery [2]. Compared with the traditional data collecting technologies, MCS collects the environment information by built-in sensing modules in the mobile terminals, thus has the properties of mobility, node ubiquity, powerful storing and computing abilities. Presently, MCS has been widely used in many applications including measuring pollution [2], analyzing social behaviors [3-4] and detecting traffic condition [5]. Such application cases have proved that MCS is an excellent solution for large scale, high dimension data collection. With the advantages of low cost of network deployment, better system scalability and the mobility of the terminals, MCS could be an excellent scheme for REM data collection, which is the motivation of our work.

This paper is about the details of designing the REM construction architecture based on MCS. Our design has several novel contributions to deal with large scale and high dimension data collection. Our key contributions are as follows:

- We design a novel REM construction architecture based on MCS, where the ubiquitous, massive and high dimension REM-related data can be sensed by the terminals carried by mobile users.
- We discuss some design issues related to the life cycle of the MCS for REM data collection including REM task creation, REM task assignment, individual task execution and REM data integration. In these issues, we also describe our mobile sensing Android based applications. We also define the collected data types and some REM data parameters.

The rest of the paper is organized as follows: Section Ⅱ outlines the architecture of the system and the system functions. Section Ⅲ describes our method for several design issues. Section Ⅳ shows some results of our system and section Ⅴ is the conclusion.

2. Related works

In the related studies, REM has been widely applied to interference management and users' coexistence in various wireless networks (WLAN (802.11), WiMAX (802.16), WRAN (802.22)) [6], radio resource management in 3G [7], high-speed trains LTE management [8], dynamic spectrum access and sensing [9-10], network integration and collaborative communications [11], and localization [12].

The primary problem to build a REM is how to collect a large amount of data. Firstly, the radio signal is ubiquitous, so it is very difficult for the sensors to cover the targeted radio environment [13]. The higher accuracy of the REM, the more sensor nodes are needed to be deployed. Moreover, more than a dozen of data types are required to build a universal REM and each type has more than one attribute. Therefore, a general sensor node can hardly complete this complex data collecting task.

The current data collection methods for REM can mainly be categorized into three types: (1) integrating or accessing the related information directly from existing databases; (2) estimating radio propagation characteristics by software tools; (3) leveraging cognitive radios devices or networks to sense data. We will discuss these methods in details.

First, gathering data from the existing database is a relatively convenient way, while the data updating time depends on the updating period of the underlying database. Moreover, the historical information is not stored in the underlying database. Riihijärvi uses external datasets to build REM, but the update cycle of the external datasets is very long, which makes datasets unable to meet the real-time requirement of REM [14]. Constructing REM in this way is difficult to satisfy the upper-layer applications with the requirement for real-time and historical information.

Second, the way to characterize and estimate the properties of radio transmission based on software is to calculate the signal attenuation by modeling so that we can better plan the radio environment [15-16]. The model in [17] clearly gives a solution to the signal diffraction problem caused by the occlusion, but this requires an accurate vector model of all three-dimensional structures, with limited data and resolution in most experimental environments, it cannot be applied to applications that require high accuracy. The above-mentioned estimation method usually provides limited data, bad accuracy of the data.

Third, the method based on wireless device or external network mainly uses the information sensing ability of heterogeneous spectrum sensor network to collect data [18-19]. According to network structures, the wireless sensor network can be divided into the direct-connect wireless sensor network, multi-hop wireless sensor network, cluster-based wireless sensor network and wireless sensor network based on mobile sensors [20]. The direct-connect wireless sensor network has a simple structure and a small coverage area, which is suitable for small-scale applications. The multi-hop wireless sensor network and Ad-Hoc networks can support large network scale than the direct-connect network, but it is still not suitable for large scale network because of the bandwidth bottlenecks and "hot spots" around the sink nodes and congestion problems [16,21]. In the wireless sensor networks based on mobile nodes, mobile nodes move in the network according to certain rules, and the nodes in the mobile process will simultaneously transmit the data. The key of wireless sensor networks based on mobile nodes is how to achieve a specific optimization goal by controlling the movement of the mobile nodes.

3. System architecture

3.1 The REM Based on MCS Architecture

Figure1 shows the overview of our system based on MCS. From bottom to upper layer, the system includes data sensing layer, data collection layer, data processing layer, data analysis layer and visualization layer. In the data sensing layer, a large number of mobile terminals constitute the mobile crowd sensing network, and they play the role of data sensing by running our data collecting APP named wireless detect. The mobile terminals upload the sensing data to our cloud servers via Wi-Fi/3G/4G networks. The data collection layer is mainly responsible for receiving data, sensing node selection, task allocation, making incentive mechanism to recruit enough interested nodes to participate into the sensing tasks. The data preprocessing like arranging the data format, data fusion, etc. The data analysis layer is responsible for the statistical analysis and calculation of the radio environment relevant parameters. At last, the visualization layer shows the REM relating results in the forms of the field strength map, heat map, and some other maps.

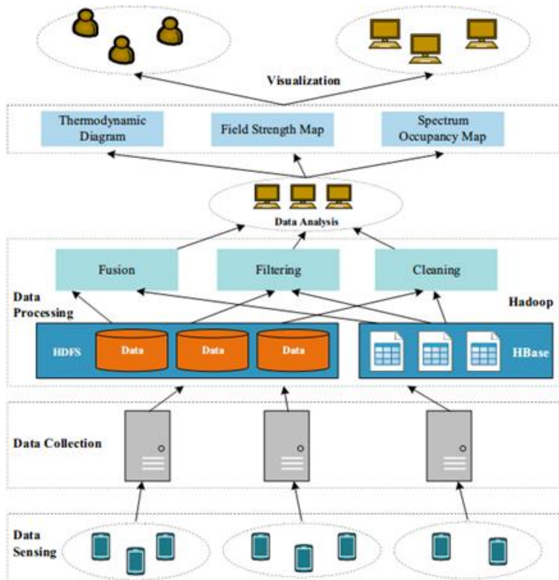


Figure 1. Systems network architecture

3.2 System Function

Our design involves various functional blocks, communicating via well-specified interfaces. To establish a complete radio environment map, the fundamental problem is the collection of a large number of data with complex types and data processing and visualization. Our system consists of five different function modules, data sensing, data collection, data processing, data analysis and visualization, each of them has its own function. In this section, we provide the main components of our system architecture, their functionality and interactions.

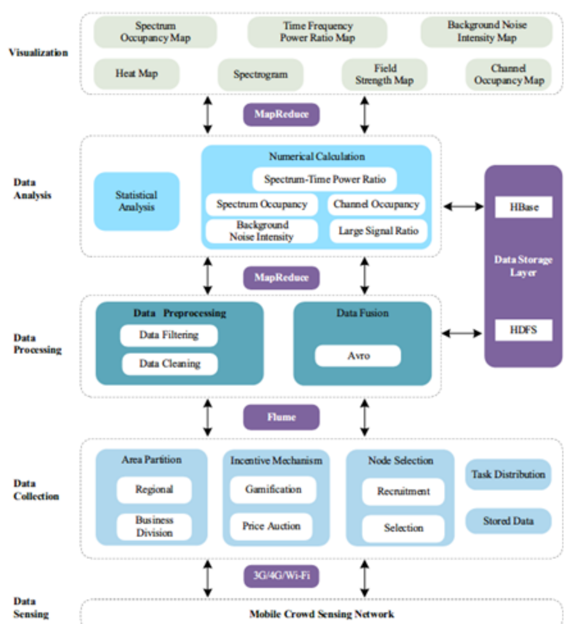


Figure 2. System functional architecture

3.2.1 Data Sensing Module

The data sensing function is operated by the MCS network, which is organized by mobile terminals carried by mobile users. When a mobile user receives a data sensing task, it will determine whether or not to involve in the task. If so, it will collect the required data by the sensing module embedded in the terminal. Moreover, it will also upload the data to the web server by different types of network accessing technologies like Wi-Fi/3G/4G. In our system, the perception of user-uploaded data and call the mobile phone Baidu API real-time construction of heat map and signal strength map. Users can use wireless detect real-time view of the environment in which the use of radio spectrum resources.

3.2.2 Data Collection Module

This module mainly includes area partition, incentive mechanism, nodes selection, task distribution, data storage, data distribution, etc. The area partition is designed to identify whether a sensing task refers to a geographical location or is based on some social relationships. In our system, we divided it into regional division and business division. The incentive mechanism is used to reduce the cost of the platform as well as attracting enough sensing users. Furthermore, node selection mechanism needs to select some appropriate nodes for the data sensing, and also needs to assign the sensing nodes to the corresponding sensing tasks if there is more than one task.

3.2.3 Data Processing Module

It mainly includes two functions: data preprocessing (filtering and cleaning) and data fusion, which is implemented by the MapReduce workflow. The data processing flow is as follows. Firstly, the Avro in the data fusion module compresses various types of formats of the data and merges massive small files into large files to improve the efficiency of MapReduce. Secondly, as the raw data is varying in data types, the data cleaning and filtering can play an important role to remove the noise and interference such as error data. Thirdly, these data are processed by our Hadoop cluster, and the processing results are stored in HDFS.

3.2.4 Data Analysis Module

It is responsible for the statistical analysis and calculation after the data pre-proceeding. In order to exhibit the radio environment on the map, it needs to analyze and calculate the data to get related parameters such as the channel occupation, frequency band occupancy, background noise intensity, large signal ratio, etc. In this module, the systems MapReduce program call for the pre-processed data from HDFS and send the computing results to HBase for the ultimate visualization.

3.2.5 Data Visualization Module

The Visualization module is responsible for the REM-related data parameters exhibition. We designed the visualization for the REM properties. The system can show the Wi-Fi signal coverage, cellular signal coverage heat map, Wi-Fi channel occupation ratio map. The visual REM makes it easy to identify the radio environment of the target area. The REM can provide different parameter maps as shown in part 5.

4. Design issues

In [17], the author proposed 4W1H model in mobile sensing and divided the MCS life cycle into four phases: task creation, task assignment, individual task execution, and crowd data integration. Based on this, we will discuss the following key design issues, REM task creation, REM task assignment, participants recruiting and participants' selection.

4.1 REM Task Creation

The task creation is to specify the sensing time and coverage area for the REM. In our system, the web server releases the sensing tasks in our website. REM is to support long spatio-temporal information for the upper layer applications, so the sensing time is continuous.

4.2 REM Task Assignment

In this stage, our system is responsible for recruiting and selecting participants for the MCS Task. Correspondingly, this stage includes participants recruiting, participants' selection and incentive mechanism.

4.2.1 The Participants Recruiting

The main goal of participants recruiting is to encourage enough people joining the sensing task and get more radio environment data. The system will publish a sensing task notice in the website and then push the notice to the APP installed in the users' mobile phones. These participants are seeds. In order to inform as many as volunteers, these initial participants will send the sensing task to other people, like nearby users or friends based on social relationships.

We have designed a short-range task spreading mechanism to recruit more participants based on Wi-Fi Direct communication technology. As is shown in Figure 3, the MCS network constructed by participants who carried mobile phones is divided into clusters based on the properties of a sensing task. Each cluster has a cluster header and other nodes. The nodes of the clusters need to complete the sensing task according to the rules of a sensing task and upload the sensing data to the data server of the MCS platform by TCP/IP connection. The head of the cluster is responsible for receiving the sensing tasks from the MCS platform, sending to nearby users and recruiting more participants joining the sensing task. When the cluster heads need to recruit more users, they will scan nearby users and send the sensing tasks to the users by Wi-Fi Direct. If a user is interested in the sensing tasks, they will negotiate with cluster headers and join the sensing task. When the new participants join in a cluster, the cluster header will send more information about the sensing task, the way to download the sensing APP and upload the sensing data.

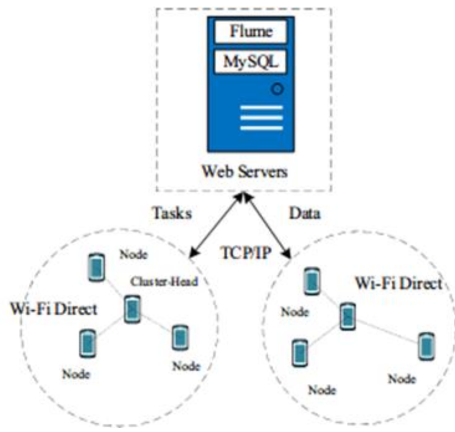


Figure 3. Short-range task spreading mechanism based on Wi-Fi-Direct

4.2.2 The Participants Selection

The participants' selection step needs to select the appropriate users to join in the sensing process. It contains how to assign users to the correspondent sensing tasks subjected to some constraints: reducing the whole cost, minimum sensing delay, etc. In our system, we select the nodes based on the trade-off between cost and the balancing of node participation.

Sometimes, the sensing capacity of a user's terminal is relatively weak to finish the high dimension data sensing. They can only operate part types of the data that needed by the REM. Therefore, we propose to divide the high-dimensional data sensing task into several sub-tasks. Then, massive nodes with different sensing capacities may complete the whole task cooperatively. As shown in Figure 4, in order to finish the sensing task with higher data dimension (m dimensions), we have to divide the m dimensions sensing task into k sub-tasks. Consequently, with each node sensing parts of the data, the whole m dimensions' data collection task could be finished. Here, the participant's selection problems can be abstracted to divide multiple nodes into corresponding sub-tasks. In order to assign each node to an appropriate subtask, such as reducing the whole sensing cost or making nodes participating sub-tasks evenly, we present to make a trade-off between the whole sensing cost of the platform and the equality of node participation.

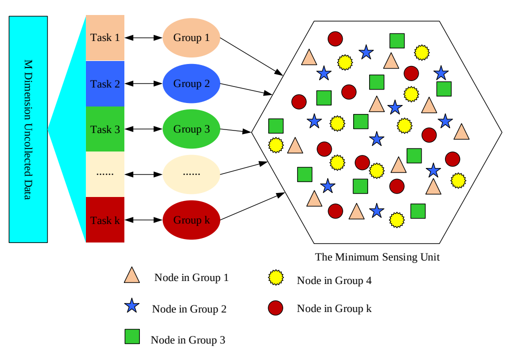


Figure 4. k groups cooperatively complete the m dimensions' data sensing

4.2.3 Incentive Mechanism

In order to encourage enough users to join in the sensing period, the platform needs to leverage some incentive mechanisms. In our system, volunteers send their confirmations back to the platform with their price for the task. The platform receives the confirmation, and in node selection phase, the platform pays rewards in forms of money or other ways to the volunteers for the data sensing. The incentive mechanisms used in our project are monetary reward incentive. Based on the node selection stage, the platform and users negotiate the sensing price. In order to reduce the cost of the whole sensing cost, the platform is prone to choose the lowest quotations and make users distributed to different sub-sensing tasks evenly.

4.3 Individual Task Execution

In this stage, participants conduct sensing tasks and upload the sensed data to the MCS platform. The participants receive the sensing tasks and then collect the radio environment data. There are enough participants distributed in the target places collecting data. After radio environment data collection, the participants upload the data to the MCS platform server. The design issues in this stage are MCS module design, data acquisition frequency, data collect types and data upload.

4.3.1 Mobile Crowd Sensing Module Design

The sensing module is mainly composed of functions like data collection, data communication , control, data storage and interface structure. As illustrated in Figure 5, we develop the data sensing APP by Android SDK (Software Development Kit). for data collection, the API of the android system is called to make the sensing module to collect data. In data communication module, we develop a method of File_upload (), uploading the data files to servers by HTTP protocol. We also using the HTTP protocol to receive data from the server to build the heat map and signal strength map of the radio environment information. During the process of file upload, the raw data would be encapsulated, error handled and parsed. In the control module, a control class is used to manage all other modules. In data store module, the method of write_sensor_info() in service writes data into SD card by Java data stream. In Interface module, we create an interface based on UI (User Interference) and set all buttons in the layout. The heat map and signal strength map is passed though the interface to show the environment where the user perceives the radio spectrum resource usage.

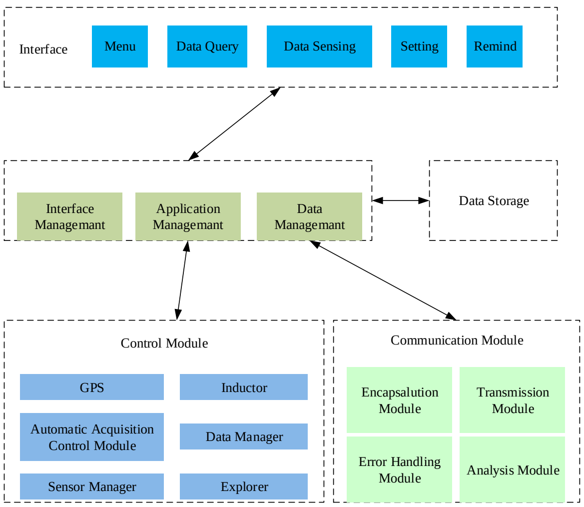


Figure 5. The sensing architecture in a mobile terminal

Concerning the ANN, the main input features are depths and RD values of neighboring CUs as well as a MergeFlag binary feature. The results reported list a complexity reduction of 47.5% for a BD-Rate increase of 1.17%. However, these results do not consider the extra computation involved in the retraining of the ANN by the secondary threads, which in all likelihood contribute a non-negligible amount of complexity to the overall encoding procedure.

4.3.2 Data Acquisition Frequency

The user consumes the power of the mobile terminal and network traffic (even the time cost of the user) during the radio environmental information collection process. If the data acquisition frequency is too tight, it will cause data redundancy. And if too sparse, it will cause data integrity problems. Therefore, in the system, to infer the user's current environment through the users' history data, and the data acquisition frequency is adjusted automatically according to the user's current environment information. In order to optimize a user's mobile terminal resources and improve the quality of data, the system divides users' current environment into three kinds of scenes: indoor, outdoor, and in car. The scene is judged as follows.

Table 1. Sense frequency

Types	Sense Frequency	Judgments based
Indoor	Low	SSID: Repeat with the SSID in the history data; Wi-Fi hotpots: The total number of Wi-Fi hotpots scanned is almost constant; Speed: Less than 1.5m/s.
Outdoor	Medium	SSID: Repeat with the SSID in the history data; Wi-Fi hotpots: The total number of Wi-Fi hotpots changes; Speed: 1.5~5m/s
Driving	High	SSID: The SSID repetition rate in historical perceptual data is low. Wi-Fi hotpots: The total number of Wi-Fi hotpots changes; Speed: higher than 5m/s

4.3.3 Data Collection Types

These types include the basic data types such as, Wi-Fi network data and cellular network data, etc. The REM-related parameters like Wi-Fi and cellular network data are shown in Table Ⅱ and Table Ⅲ respectively. As shown in Table Ⅱ, Wi-Fi network data is divided into Wi-Fi connection data and Wi-Fi scanning data. The Wi-Fi connection data refers to the information of the Access Point (AP) connected to a sensing node, and Wi-Fi scanning data refers to the information of all the APs scanned by the sensing node. Table Ⅲ shows the cellular network data, such as information of GSM, CDMA, LTE, etc.

Table 2. Wi-Fi network data

Data Types	Parameters	Parameter Specification
Wi-Fi Connection	Network_ID	Network identification
	Link_SSID	Service Set Identifier of Link
	Link_BSSID	Basic Service Set Identifier
	Supplicant State	Connection status of the AP (connecting/completing)
	RSSI	Received Signal Strength Indication.
	Link Speed	Link Speed
	IP	Internet Protocol of Mobile phone.
Wi-Fi Scanning	Frequency	The primary frequency of the channel.
	Level	Signal strength
	Capabilities	EncryMode
	Scan_SSID	Service Set Identifier of Scanning.
	Scan_BSSID.	Basic Service Set Identifier of Scanning

Table 3. Cellular network data

Data Types	Parameters	Parameter Specification
GSM	Type	Network Type
	CID	Cell Identity
	LAC	Location area code
	RSSI	Received Signal Strength Indication
	BER	Bit Error Rate
	Valid Cellular	The number of the adjacent base stations.
	PSC	The primary scramble code of adjacent base stations.
CDMA	CDMA_RSSI	CDMA Signal Strength
LTE	LTE_RSSI	LTE Signal Strength

4.3.4 Data Uploading

As shown in Figure 2, the terminals connect to the network through 3G/4G/Wi-Fi and other ways, and they establish TCP/IP connections to the Web server so that the data is transmitted to the Web server by exploiting HTTP protocol. The Flume component installed in the Web server is used to monitor file update. If there is a file updating event, the Web server will automatically upload the new data to the data center built by the Hadoop cluster.

As shown in Figure 3, the participants will upload the sensed data to the MCS platform server. First the users will connect the server by establishing TCP/IP connections. Then the participants will transfer the radio environment data to the server and put the data in database for further processing

4.4 Crowd Data Integration

In general, the main issue of the MCS platform that is to process, analyze the raw data received from mobile terminals, and visualize the required results eventually. Correspondingly, the crowd data integration mainly includes data processing and data analysis.

4.4.1 Data Processing

We leverage the Flume software to automatically upload the new sensing data from Web server to big data processing platform which is composed of the Hadoop cluster. The small files are sequenced and then constituted into large files by the Avro plugin, which can improve the efficiency of the subsequent steps such as data filtering, data cleaning, data processing, and analysis. Here, the Avro compression and small-files-tuning methods are used to improve the efficiency of MapReduce. Finally, the MapReduce programs are developed to remove the redundant sensing data.

4.4.2 Data Analysis

Data analysis module is responsible for the statistical analysis and the calculation of radio environment. As shown in Table Ⅳ, the relevant parameters, according to the requirements of the upper arm need to be computed. Later, all of the analysis results are stored in HBase for data visualization.

Table 4. Numerical index

Numerical Index	Formula	Parameter Definition
Channel Occupancy	$\rho = \frac{T_1}{T} \times 100\% = \frac{c}{n} \times 100\%$	T_1, duration of the signal exceeds the receiver threshold level. T, measurement time. n, scanning times per channel. c, occupancy times.
Band Occupancy	$\theta = \frac{m_1}{m} \times 100\%$	m_1, the number of channels whose level is bigger than the threshold level in the frequency band. m, denotes the total number of channels measured.
Background Noise Strength	$B_n = \frac{\sum(rb_n - rb_0)/(40 - rb_0)}{FS_n}$	rb_n, value of noise strength. rb_0, sensitivity test system.FS_n, total number of channels.
Large-signal Ratio	$S_s = \left(\frac{E_{max}}{90}\right)^3 \times \sqrt[3]{\frac{SS_n}{FS_n}}$	E_{max}, maximum signal field strength, the strength which more than 50% of the average strength.FS_n total number of channels.
Time Band Power Ratio	$P_t = 1 - \frac{15 \times t \times FS_n}{\iint E_f}$	t, channel measured time.E_f, measured field strength value

4.4.3 Visualization

Our REM visualization is based on Baidu Map, by calling the Baidu Map API, radio environment map is built based on the analysis results stored in MySQL. Where the MySQL is in the Web server, and the data in MySQL is automatically imported via Sqoop from HBase.

5. Results

We designed and implemented prototype REM system based on MCS, then collected the radio environment data of Chongqing University of Posts and Telecommunications (CQUPT). The collected data was uploaded to our system and analyzed, several heat-maps and figures are as follows.

Figure 6 shows a prototype heat map which provides the overall of Wi-Fi signal strength in CQUPT sensed by mobile terminals. Good signal strength occupied zones appeared as islands of red or yellow in otherwise green or blue settings. The initial visualization made it easy to identity good signal strength areas, as is shown Block A and Block B. As we can see a high occupancy of Wi-Fi is shown in the map, while only a few areas of the school have good signal strength as Block A.

Figure 6. The heat map of Wi-Fi signal strength

Figure 7. The heat map of user density

Figure 7 shows a prototype heat map for user density, which is constructed by the user geographic information. The relatively high user density areas are in red or orange as Block A/B/C, while the relatively low user density areas are in blue or green. As is shown in the figure, the distribution of the user density in the school area is not uniform. However, the results are as expected. The high user density areas are teaching area and living area, so most of the students and staffs are studying or working in these areas which are occupied by red or orange.

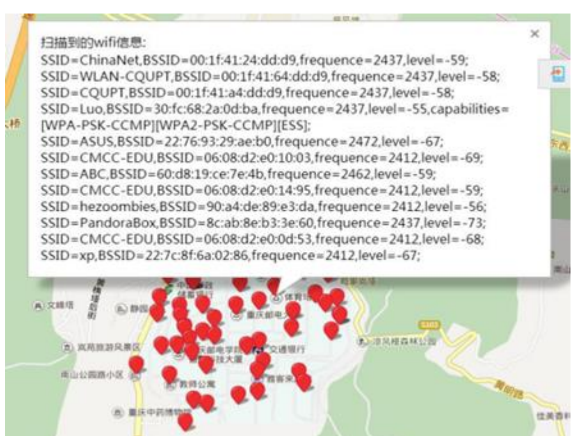


Figure 8. The information of Wi-Fi connected and scanned

As is shown in Figure 8, lots of Wi-Fi properties can be seen on the banner. The properties collected by participants are as SSID, BSSID, frequency and the Wi-Fi signal strength level. This information belongs to the Wi-Fi signal sources sensed by the nearby participants. The density of the red nodes represents the density of the Wi-Fi signal sources. As we can see in Figure8, the density of the Wi-Fi signal source is not uniform. This result is expected. The large density areas have more classroom and offices, which requires more Wi-Fi sources are. The small density areas are outdoor playgrounds, that's why less W-Fi are needed.

Figure 9. The distribution of LTE signal strength

Figure 9 is a heat map of LTE shows the coverage of LTE and signal strength in the school. The color gradient of the shades can be seen in the map. The pink shades indicate the bad LTE signal which is less than -90dbm, while the dark red shows better LTE signal which is more than -40dbm and the red gradients denotes the signal strength between -90dbm and -40dbm. As we can see in the picture, there is still lots places can't be covered by LTE signals. This result can be used to help the ISPs to build more LTE base stations until the school is all covered by the LTE network.

Figure 10. The distribution of LTE signal strength around the subscriber

Figure 10 shows the LTE signal strength map within 5 km where the user is located, which provides a new solution for radio network optimization. The different colors represent different signal strength. As we can see in Figure 10, most of the signal nodes are orange, which represents that the LTE signal strength of these places are between -60-70dBm. Some signal nodes are pink, which represents that the LTE signal strength of these places are between -90-100dBm. These places need a better solution to optimize the LTE network.

6. Conclusion

In this paper, we introduced MCS to collect data for REM construction. We believe this is an important design point to deal with the large-scale distributed, high dimension and massive data sensing tasks. We designed five layers reference architecture for REM. We also considered some design issues based on MCS life cycle: REM task creation, REM task assignment, individual task execution and crowd data integration. Based on our analysis, MCS has several advantages over other traditional data collecting methods in constructing the REM. In the future, we plan to design more effective incentive mechanisms for participants recruiting, MCS networking, and collaborative data collecting, etc.

Acknowledgement

This work is partially supported by NSFC China (No.61501075).

References

[1] Y. Zhao, S. Mao, J. O. Neel, and J. H. Reed, "Performance Evaluation of Cognitive Radios: Metrics, Utility Functions, and Methodology," Proceedings of the IEEE, vol. 97, no. 4, pp. 642-659, April. 2009.

[2] Y. Zhao, B. Le, and J. H. Reed. "Network Support: The Radio Environment Map," Cognitive Radio Technology, pp. 337-363, May. 2009.

[3] T. Ikuma, and M. Naraghi-Pour, "A Comparison of Three Classes of Spectrum Sensing Techniques," in Proc. of 2008 IEEE Global Telecommunications Conference, New Orleans, LO, 2008, pp. 1-5.

[4] J. Caffery, "Wireless Location in CDMA Cellular Radio Systems" [M]. Kluwer Academic Publishers, 2001.

[5] P. C. Chen, "Location estimation in CDMA systems: enhanced measurement on pilot channels," in Proc. of 1999 IEEE International Conference on Communications, Vancouver, BC, 1999, pp. 1784-1788.

[6] B. Guo, Z. Yu, X. Zhou, and D. Zhang, "From participatory sensing to Mobile Crowd Sensing," in Proc. of the 2014 IEEE International Conf erence on Pervasive Computing and Communications Workshops, Budapest, 2014, pp. 593-598.

[7] R. K. Rana, C. T. Chou, S. S. Kanhere, N. Bulusu, and W. Hu, "Ear-phone: an end-to-end participatory urban noise mapping system," in Pro c. of the 2010 ACM/IEEE International Conference on Information Processing, New York, NY, 2010, pp. 105-116.

[8] Z. Hou, Y. Zhou, L. Tian, J. Shi, Y. H. Li, and B. Vucetic, "Radio environment map-aided doppler shift estimation in LTE railway," IEEE Transactions on Vehicular Technology, vol. 66, no. 5, pp. 4462-4467, August. 2017.

[9] F. Paisana, Z. Khan, J. Lehtomäki, L. A. DaSilva, and R. Vuohtoniemi, "Exploring radio environment map architectures for spectrum sharing in the radar bands," in Proc. of the 2016 23rd International Conference on Telecommunications (ICT), Greece, 2016, pp. 1-6.

[10] A. Chakraborty, S. R. Das, and M. Buddhikot, "Radio environment mapping with mobile devices in the TV white space," in Proceedings of the 19th annual international conference on Mobile computing & networking, New York, NY, 2013, pp. 247-250.

[11] D. Denkovski, V. Rakovic, M. Angjelicinoski, and V. Atanasovski, "Small-cells radio resource management based on Radio Environmental Maps," in Proc. of the 2014 IEEE International Conference on Computer Communications Workshops, Toronto, ON, 2014, pp. 155-156.

[12] D. Y. Seol, H. J. Lim, and G. H. Im, "Optimal threshold adaptation with radio environment map for cognitive radio networks," in Proc. of the 2009 IEEE International Symposium on Information Theory, Seoul, 2009, pp. 2527- 2531.

[13] C. Chen, J. Yan, N. Lu, Y. Y. Wang, X. Yang, and X. P. Guan, "Ubiquitous monitoring for industrial cyber-physical systems over relay-assisted wireless sensor networks," IEEE Transactions on Emerging Topics in Computing, vol. 3, no. 3, pp. 352-362, January. 2015.

[14] J. Riihijärvi, J. Nasreddine, and P. Mähönen, "Demonstrating radio environment map construction from massive data sets," in Proc. of the 2012 IEEE International Symposium on Dynamic Spectrum Access Networks, Bellevue, WA, 2012, pp. 266-267.

[15] S. Ulaganathan, D. Deschrijver, M. Pakparvar, I. Couckuyt, W. Liu, D. Plets, W. Joseph, T. Dhaene, L. Martens, and I. Moerman, "Building accurate radio environment maps from multi-fidelity spectrum sensing data," Wireless Networks, vol. 22, no. 8, pp. 2551-2562, November, 2015.

[16] S. Üreten, A. Yongaçoğlu, and E. Petriu, "A comparison of interference cartography generation techniques in cognitive radio networks," in Proc. of the 2012 IEEE International Conference on Communications (ICC), Ottawa, ON, 2012, pp. 1879-1883.

[17] J. B. Liang, M. Liu, and X. Y. Kui, "A survey of coverage problems in wireless sensor networks," Sensors & Transducers, vol. 163, no. 1, pp. 240-246, January. 2014.

[18] Z. Q. Wei, Q. X. Zhang, Z. Y. Feng, W. Li, and T. A. Gulliver, "On the construction of radio environment maps for cognitive radio network s," in Proc. of the 2013 IEEE Wireless Communications and Networking Conference (WCNC), Shanghai, 2013, pp. 4504-4509.

[19] V. Atanasovski, J. V. D. Beek, A. Dejonghe, D. Denkovski, L. Gavrilovska, S. Grimoud, and P. Mähönen, "Constructing radio environment maps with heterogeneous spectrum sensors," in Proc. of the 2011 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN), Aachen, 2011, pp. 660-661.

[20] A. H. Dehwah, S. B. Taieb, J. S. Shamma, and C. G. Claudel, "Decentralized energy and power estimation in solar-powered wireless sensor networks," in Proc. of the 2015 International Conference on Distributed Computing in Sensor Systems, Fortaleza, 2015, pp. 199-200.

[21] D. Ciullo, G. D. Celik, and E. Modiano, "Minimizing transmission energy in sensor networks via trajectory control," in Proc. of the 8th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks, Avignon, 2010, pp. 132-141.

Yuanni Liu received her M.E. degree from Zhengzhou Univ. in 2008 and Ph.D. degree from Beijing Univ. of Posts and Telecommunications in 2011. Now she serves as a Associate Professor in School of Communication and Information Engineering, Chong Qing University of Posts and Telecommunications. Her research interest focus on Mobile Crowd Sensing, data analysis, IP routing technology and complex networks.

Jianli Pan is currently an Assistant Professor in the Department of Mathematics and Computer Science at the University of Missouri, St. Louis. He obtained his Ph.D. degree from the Department of Computer Science and Engineering of Washington University in St. Louis. He also holds a M.S. degree in Computer Engineering from Washington University in Saint Louis and a M.S. degree in Information Engineering from Beijing University of Posts and Telecommunications (BUPT), China. He is currently an associate editor for both IEEE Communication Magazine and IEEE Access. His current research interests include edge clouds, Internet of Things (IoT), Cybersecurity, Network Function Virtualization (NFV), and smart energy.

Guofeng Zhao received his M.E. degree from Northwest Polytechnic Univ. in 1996 and Ph.D. degree from the Univ. of Chongqing in 2003. Now he serves as a professor in the School of Communication Engineering, and the director of the Research center of Future Internet (RFI) at Chongqing Univ. of Posts and Telecommunications (CQUPT). He has undertaken 20+ projects or programs, including the National Basic Research Program of China and Natural Science Foundation of China. He has published more than 100 papers and held 6 patents. His research interests are in Future Internet, Mobile Internet, Network management, Network security, etc.

Optimized Resource Allocation for Edge Computing Offloading in 5G Networks

*Qiang Wang, and Tiejun Chen**
School of Electronic and Communication Engineering, Yulin Normal University, Yulin, China
chentiejun2000@163.com

1. Introduction

With the explosive increase of the number of mobile terminal, massive data that needed to be processed is generated. It will cause serious delay if these mobile devices depend on themselves to process these data. To address this problem, the concept of mobile cloud computing is proposed in which data can be computed at third party server. Since there is stronger computing power in cloud, cloud computing can bring great benefits to wireless networks' data processing [1].

Cloud computing is a centralize computation method. There are some shortcomings for cloud computing applied in 5G networks. For example, If all mobile data is transmitted to cloud to process, it will decrease the computation efficiency due to the massive data. In addition, serious delay and energy consumption are caused in cloud computing since the cloud is far away from the data source. Based on this, recently, the edge computing is proposed to improve the computing efficiency [2-7]. In edge computing, the computation process is executed at the networks edge. Due to the distribution, the edge computing has higher computing efficiency than traditional cloud computing. Meanwhile, edge computing could effectively decrease the energy consumption and network delay since it is more closer to mobile terminals. In [8], the authors study the application of edge computing for internet of things (IoT), and the edge computing is adopted for emerging IoT applications that leverage sensor streams to augment interactive applications. The authors in [9] propose an computing task offloading for edge computing in 5G heterogeneous networks. In [10], the authors investigate several key issues on cooperative sharing of large volume of vehicular data in edge computing assisted 5G enabled vehicular ad hoc network. In [11], the authors analyzes the mobile edge computing reference architecture and main deployment scenarios.

In edge computing, mobile data is firstly transmitted to edge computing server, and then edge computing server forwarded the computation output to mobile terminals. Therefore, the data transmission will consume some certain power. Due to different computing and communication abilities, for some terminals, computing the task locally may consume lower power than the task is performed at edge server. Thus, the computing mode should be carefully chosen to improve the energy efficiency. In addition, the transmission power should also be optimized due to the co-channel interference. In this paper, we consider the computing mode selection and power allocation simultaneously for edge computing in 5G networks to maximize the energy efficiency. Due to the combinatorial characteristic of computing mode selection, we formulate the joint optimization problem as a mixed-integer programming problem which is difficult to optimize. To address it, we propose a improved evolutionary algorithm based on Lagrangian dual method. It is well known that the evolutionary algorithm is a powerful tool to tackle the combinatorial optimization problem. Therefore, we adopt the evolutionary algorithm to optimize the discrete variables (computing mode selection), and the Lagrangian dual method is used to solve the continuous variables (power allocation).

The rest of the paper is organized as follows. Section II introduces the system model. In section III, we describe the proposed evolutionary algorithm based on Lagrangian dual domain method. In section IV, simulation results are provided to evaluate the performance of the proposed algorithm. Finally, we conclude this paper in section V.

2. System Model and Problem Formulation

In this paper, we consider a 5G heterogeneous networks with one edge computing server, one macro cell, M small cells and K mobile terminals. Mobile terminals can offload their computation task to the edge computing server either through macro base station (MBS) or small base station (SBS). In the considered communication system, the MBS can share the same frequency band with SBS, and the spectrum is divided into N subcarriers. If terminal i transmit data to edge computing server through MBS over subcarrier n, the transmission rate can be given as:

$$R_{i,n}^{M}=\log\left(1+\frac{p_{i,n}^{M}g_{i,n}^{M}}{I_{n}^{S}+N_{0}}\right) \tag{1}$$

where $p_{i,n}^{M}$ is the transmission power from terminal *i* to MBS on channel *n*. $g_{i,n}^{M}$ is the channel gain from terminal *i* to MBS. $I_{n}^{S}=\sum_{i}p_{i,n}^{S}$ means the interference from SBS to MBS on channel *n*, and $p_{i,n}^{S}$ is the transmission power from terminal *i* to SBS on channel *n*. N_0 is the Gaussian white noise.

In this paper, the computing mode selection is considered. That is, terminals can compute the task locally, or offload computing task to edge computing server through MSB or SBS. We introduce the task offloading variables $s_{i,j,n}$, and $s_{i,j,n}\in\{0,1\}$, where $j\in\{1,2,3\}$ denotes the chosen mode. $j=1$ means the task is computed locally. $j=2$ means the task is computed at edge computing server through MBS. Otherwise, the task is computed at edge computing server through SBS. Then, the total transmission rate from terminal *i* to MBS is is given as:

$$R_{i}^{M}=\sum_{n=1}^{N}s_{i,2,n}R_{i,n}^{M}=\sum_{n=1}^{N}s_{i,2,n}\log\left(1+\frac{p_{i,n}^{M}g_{i,n}^{M}}{I_{n}^{S}+N_{0}}\right) \tag{2}$$

Similarly, the transmission rate from terminal *i* to SBS on channel *n* can be given as:

$$R_{i,n}^{S}=\log\left(1+\frac{p_{i,n}^{S}g_{i,n}^{S}}{I_{n}^{M}+N_{0}}\right) \tag{3}$$

where $p_{i,n}^{S}$ is the transmission power from terminal *i* to SBS on channel *n*. $g_{i,n}^{S}$ is the channel gain from terminal *i* to SBS. $I_{n}^{M}=\sum_{i}p_{i,n}^{M}$ means the interference from MBS to SBS on channel *n*. The total transmission rate from terminal *i* to SBS is:

$$R_{i}^{S}=\sum_{n=1}^{N}s_{i,3,n}R_{i,n}^{S}=\sum_{n=1}^{N}s_{i,3,n}\log\left(1+\frac{p_{i,n}^{S}g_{i,n}^{S}}{I_{n}^{M}+N_{0}}\right) \tag{4}$$

In the case that terminal *i* transmit data to edge computing server through MBS, the total energy consumption can be given as:

$$E_{i}^{M}=\sum_{n=1}^{N}s_{i,2,n}p_{i,n}^{M}d_{i}/R_{i}^{M}+c_{i}\delta \tag{5}$$

where d_i is the size of input data needed to be process by terminal *i*. c_i is the computing ability required for accomplishing the input data of terminal *i*. δ denotes the energy consumption for one CPU cycle of device *i*.

Similarly, in the case that terminal *i* transmit data to edge computing server through SBS, the total energy consumption can be shown as:

$$E_{i}^{S}=\sum_{n=1}^{N}s_{i,3,n}p_{i,n}^{S}d_{i}/R_{i}^{S}+c_{i}\delta \tag{6}$$

When the task is computed locally, the energy consumption can be given as:

$$e_{i}^{L}=c_{i}\delta \tag{7}$$

In this paper, our objective is to minimize the total energy consumption of the networks. The optimization problem can be formulated as:

$$\min \sum \left(s_{i,1,n} e_i^L + E_i^M + E_i^S \right) \qquad (8)$$

s.t.

$$\sum_{i=1}^{K} s_{i,1,k} \cdot \sum_{i=1}^{K} s_{i,2,k} = 0 \qquad (9)$$

$$\sum_{i=1}^{K} s_{i,2,k} \cdot \sum_{i=1}^{K} s_{i,3,k} = 0 \qquad (10)$$

$$\sum_{i=1}^{K} s_{i,1,k} \cdot \sum_{i=1}^{K} s_{i,2,k} = 0 \qquad (11)$$

$$\sum_{n=1}^{N} s_{i,2,n} p_{i,n}^M + \sum_{n=1}^{N} s_{i,3,n} p_{i,n}^S \le P_{\max}, \ \forall i \qquad (12)$$

where constraints (9)-(11) means that each terminal can only transmit one computing mode. Constraint (12) means the transmission power of each terminal cannot exceed the maximum transmission power.

3 Energy Consumption Mininization

As aforementioned, the formulated optimization problem is a mixed integer programming problem. In this section, we propose an evolutionary algorithm based on Lagrangian dual method to solve this problem. The evolutionary algorithm is adopted to optimize the computing mode selection, and Lagrangian method is applied to address the power allocation.

When the computing mode selection is fixed, there is only power allocation variables in the original problem. Meanwhile, the energy consumption minimization problem can be converted as sum rate maximization problem in this case. Thus, the original problem can be rewritten as:

$$\min \sum \left(s^*_{i,1,n} e_i^L + E_i^M + E_i^S \right) \qquad (13)$$

$$\sum_{n=1}^{N} s^*_{i,2,n} p_{i,n}^M + \sum_{n=1}^{N} s^*_{i,3,n} p_{i,n}^S \le P_{\max} \qquad (14)$$

As shown above, each terminal can only choose one computing mode. If the task is computed locally, there is no power transmitted to MBS OR SBS. Therefore, for power allocation, it is only necessary to consider the edge computing through MBS or SBS, respectively.

In the case that the computing task is offloaded to edge computing server through MBS, the power allocation problem can be given as:

$$\max \sum_{i=k}^{K} \sum_{n=1}^{N} s_{i,2,n} \log\left(1 + \frac{p_{i,n}^M g_{i,n}^M}{I_n^S + N_0} \right) \qquad (15)$$

s.t.

$$\sum_{n=1}^{N} s^*_{i,2,n} p_{i,n}^M \le P_{\max} \qquad (16)$$

It is worth nothing that problem (15) is not a convex optimization problem. To solve it , we first fix the transmit power of SBS. Therefore, problem (15) can be rewritten as follows:

$$\max \sum_{i=k}^{K} \sum_{n=1}^{N} s_{i,2,n} \log\left(1+\frac{p_{i,n}^{M} g_{i,n}^{M}}{I_n^{S*}+N_0}\right) \qquad (17)$$

s.t.

$$\sum_{n=1}^{N} s^{*}_{i,2,n} p_{i,n}^{M} \le P_{\max} \qquad (18)$$

The above problem is a convex optimization problem. Lagrangian dual method can be used to solve it.

The Lagrangian function for (16) is:

$$L(p_{i,n}^{M}) = \sum_{n=1}^{N} s^{*}_{i,2,n} \log\left(1+\frac{p_{i,n}^{M} g_{i,n}^{M}}{I_n^{S*}+N_0}\right) + \lambda_i^{M}\left(P_{\max} - \sum_{n=1}^{N} s^{*}_{i,2,n} p_{i,n}^{M}\right) \qquad (19)$$

where λ_i^M is the Lagrangian multiplier.

Then, the dual function can be obtained:

$$g(\lambda_i) = \max L\left(p_{i,n}^{M}\right) \qquad (20)$$

Differentiating (19) with $p_{i,n}^{M}$, the optimal power allocation for MBS can be obtained:

$$p_{i,n}^{M*} = \left[\frac{1}{\alpha\lambda_i^{M}} - \frac{N_0 + I_n^{M}}{g_{i,n}^{M}}\right]^{+} \qquad (21)$$

where $\alpha = 2\ln 2$, $[a]^{+} = \max\{0, a\}$.

Then, the subgradient algorithm is used to optimize the dual function as follows:

$$\Delta\lambda_i^{M} = P_{\max} - \sum_{n=1}^{N} p_{i,n}^{M*} \qquad (22)$$

In the same way, the optimal power allocation for SBS can be obtained:

$$p_{i,n}^{S*} = \left[\frac{1}{\alpha\lambda_i^{S}} - \frac{N_0 + I_n^{S}}{g_{i,n}^{S}}\right]^{+} \qquad (23)$$

Accordingly, the Lagrangian multiplier λ_i^S can be updated as follows:

$$\Delta\lambda_i^{S} = P_{\max} - \sum_{n=1}^{N} p_{i,n}^{S*} \qquad (24)$$

4 Simulation Results

In this section, we intend to evaluate the performance of the proposed edge computing scheme for 5G networks.In the simulation, one MBS and one SBS are considered. 60 terminals uniformly distribute in the macro cell and small cell. The cost-231 is adopted as the large scale path loss model, and the small scale fading is modeled as Raleigh fading process. The computation capability and unit energy consumption of edge computing server are set as 4GHz/s and δ = 1W/GHZ [12], respectively.

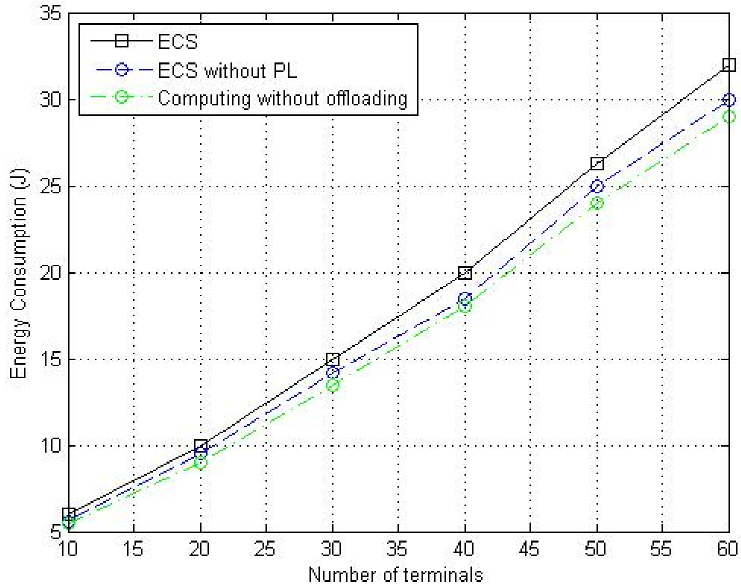


Fig.1 Energy Consumption vs different number of users.

Fig.1 depicts the energy consumption with different user number with the proposed edge computing offloading scheme and the computing without offloading scheme, respectively. From Fig.1, it can be observed that the total energy consumption increases with the increase of the number of terminals. It can also be seen that the energy consumption performance of the proposed edge computing offloading scheme outperforms computing without offloading scheme. This result illustrates that the edge computing can effectively improve the computation efficiency.

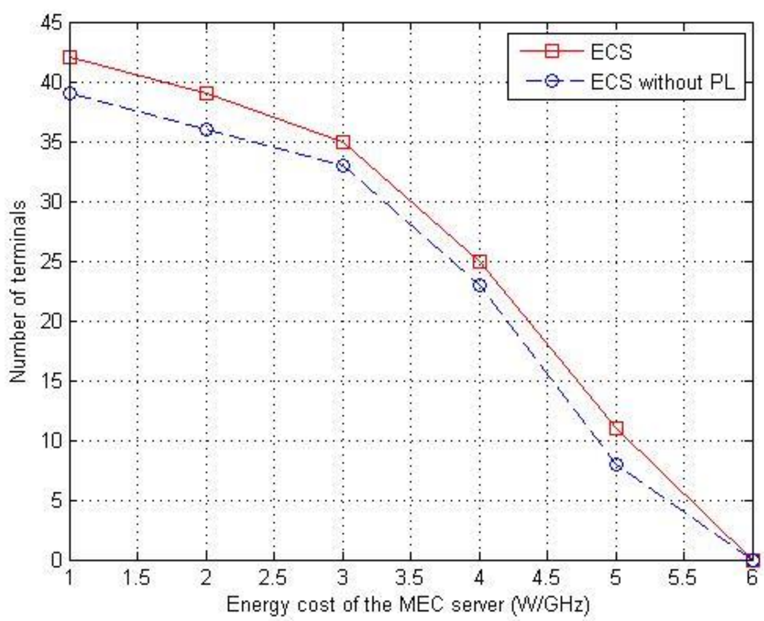


Fig. 2. The Number of Offloading Terminals with different Energy Cost of the MEC server

Fig.2 shows the impacts of the mobile edge computing server energy cost on the number of the offloading devices with different computing offloading scheme. It can be observed that the number of offloading terminals decrease as the edge computing server energy cost grows. This illustrates that when the edge computing server energy cost is high, the terminals will process the computing task locally. It can also be seen that the proposed edge computing offloading scheme has better performance than the computing scheme without power allocation. This result shows that the power allocation can bring benefits for edge computing.

5. Conclusion

In this paper, we study the edge computing scheme for 5G heterogeneous networks to improve the computation efficiency through offloading computing task. The resource allocation involving the computing mode selection and power allocation is considered to improve the computing efficiency. We formulate the edge computing offloading problem as a mixed integer programming problem which is difficult to address in general. To solve it, we propose an evolutionary algorithm based on Lagrangian dual method. In the proposed algorithm, the evolutionary algorithm is adopted to optimize the computing mode selection since evolutionary algorithm is one of the most powerful tool to tackle the discrete problem. When the computing mode selection is fixed, there is only power allocation which is continuous. The Lagrangian dual method is adopted to optimize the power allocation. Simulation results show the effectiveness of the proposed edge computing offloading scheme.

Acknowledgement

This work is part by the Projects of Education Department of Guangxi under Grant 2018KY0472.

References

[1] Bowen Zhou, Amir Vahid Dastjerdi, Rodrigo N. Calheiros, Satish Narayana Srirama, and Rajkumar Buyya, "mCloud: A Context-Aware Offloading Framework for Heterogeneous Mobile Cloud", IEEE Transactions on Services Computing, vol. 10, no. 5, pp. 797-810, 2017.

[2] Peter Corcoran, and Soumya Kanti Datta, "Mobile-Edge Computing and the Internet of Things for Consumers: Extending cloud computing and services to the edge of the network," IEEE Consumer Electronics Magazine, vol. 5, no. 4, pp. 73-74, 2016.

[3] Yuyi Mao, Changsheng You, Jun Zhang, Kaibin Huang, and Khaled B. Letaief, "A Survey on Mobile Edge Computing: The Communication Perspective," IEEE Communications Surveys & Tutorials, vol. 19, no. 4, pp. 2322-2358, Dec. 2017.

[4] Syed Noorulhassan Shirazi, Antonios Gouglidis, Arsham Farshad, and David Hutchison, "The Extended Cloud: Review and Analysis ofMobile Edge Computing and Fog From a Security and Resilience Perspective,"Journal on Selection Areas in Communications, vol. 35, no. 11, pp. 2586-2598, 2017.

[5] Shuo Wang, Xing Zhang, Yan Zhang, Lin Wang, Juwo Yang and Wenbo Wang, "A Survey on Mobile Edge Networks: Convergence of Computing, Caching and Communications,"IEEE Access, vol. 5, pp. 6757 - 6779, 2017.

[6] Weisong Shi, Schahram Dustdar, "The Promise of Edge Computing," Computer, vol. 45, no. 9, pp. 78-81, 2016.

[7] Syed Noorulhassan Shirazi, Antonios Gouglidis, Arsham Farshad, and David Hutchison, "The Extended Cloud: Review and Analysis of Mobile Edge Computing and Fog From a Security and Resilience Perspective," Journal on Selection Areas in Communications, vol. 35, no. 11, pp. 2586-2598, 2017.

[8] Gopika Premsankar, Mario Di Francesco, and Tarik Taleb, "Edge Computing for the Internet of Things: A Case Study," IEEE Internet of Things Journal, Vol. 5, no. 2, 1275-1284, 2018.

[9] Ke Zhang, Yuming Mao, Supeng Leng, Quanxin Zhao, Longjiang Li, Xin Peng,Li Pan, Sabita Maharjan, and Yan Zhang, "Energy-efficient Offloading for Mobile Edge Computing in 5G Heterogeneous Networks," IEEE Access, vol. 4, pp. 5896-5907, 2017.

[10] Quan Yuan, Haibo Zhou, Nan Cheng, Zhihan Liu, Fangchun Yang, and Xuemin (Sherman) Shen, "Cooperative Vehicular Content Distribution in Edge Computing Assisted 5G-VANET," China Communications, vol. 15, no. 7, pp. 1-17, 2018.

[11] Tarik Taleb, Konstantinos Samdanis, Badr Mada, Hannu Flinck, Sunny Dutta, and Dario Sabella, "On Multi-Access Edge Computing: A Survey of the Emerging 5G Network Edge Cloud Architecture & Orchestration," IEEE Communications Surveys & Tutorials, vol. 19, no. 3, pp. 1657-1681, 2017.

[12] M. Wittmann, G. Hager, T. Zeiser, J. Treibig and G. Wellein, "Chiplevel and multi-node analysis of energy-optimized lattice Boltzmann CFD simulations," Concurrency and Computation: Practice and Experience, vol. 28, no. 1, pp. 2295-2315, 2016.

Qiang Wang received the B.S. degree from Qufu Normal University, China in 2010, and received the M.Sc. and Ph.D. degree from Guangdong University of Technology, China. Now he is working in Guangxi Colleges and Universities Key Laboratory of Complex System Optimization and Big Data Processing, Yulin Normal University, Yulin, China. His research interests include wireless communications, signal processing, and evolutionary algorithm.

Tiejun Chen received the B.S. degree from University of Electronic Science and technology of China, China in 1988, and received the M.Sc. degree from Guilin University of Electronic Technology, China in 2007. He is currently professor of School of Electronic and Communication Engineering, Yulin Normal University, Yulin, China, and he is the senior member of China Electronics Society. His research interests include wireless communications, signal processing, and Embedded system.