TCCN Newsletter

Vol. 4, No. 1, May 2018

CONTENTS

Chair's Message

Dear Fellow TCCN Members,

I am very happy to write to you regarding some of our recent progresses as well as future plans of the TC.

First, the SIG. The scope of cognitive network is broad, and we have been encouraging colleagues to establish SIGs to promote emerging research directions that fit into and expand the scope of the TC. By the end of 2017, TCCN has had six SIGs. In 2018, we have established the seventh SIG on Energy-harvesting Cognitive Radio Networks. The SIG Chair is Daniel Benevides da Costa from Federal University of Ceará (UFC), Brazil. The SIG Vice-Chairs include Octavia Dobre from Memorial University, Canada, Trung Q. Duong, from Queen's University Belfast, UK, Minghua Xia from Sun Yat-sen University, China, and Phee Lep Yeoh from University of Sydney, Australia. More information regarding the information of the SIG can be found at http://cn.committees.comsoc.org/special-interest-groups-sigs/sig-on-energy-harvesting-cognitive-radio-networks/.

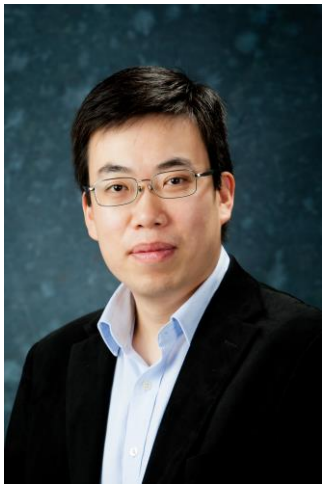
Second, the TC has been working closely with the IEEE Transactions on Cognitive Communications and Networking on several special issues. One such a special issue consists of selected papers from IEEE GLOBECOM 2017 Cognitive Radio and Networks Symposium, where three TCCN leaders (Yue Gao, Suba Subbalakshmi, and myself) have served as the Guest Editors. The special issue will be published in the June issue of the journal. There are other special issues being planned by the TCCN leaders as well.

Third, we are going to announce the call-for-nominations of the TCCN Publication and Recognition Awards for 2018 soon. These are annual awards that were reactivated last year. The call-for-nominations will come out in the summer, and we will announce the awardees in IEEE GLOBECOM 2018.

Four, the term of the current TCCN officers will come to an end by the end of 2018. We will formulate a nomination committee, and announce the call-for-nominations of the officer candidates. Following the tradition, the voting will be done electronically by all the voting members of the technical committee. The results will be announced in IEEE GLOBECOM 2018. I look forward to having more energetic and dedicated volunteers joining the leadership team.

As always, I welcome any suggestions from TCCN members regarding how to make TCCN a better community. Please feel free to contact me at jwhuang@ie.cuhk.edu.hk with your ideas and suggestions.

Thanks and best regards,



Jianwei Huang
Chair, IEEE ComSoc TCCN

Professor and IEEE Fellow
IEEE ComSoc Distinguished Lecturer
Web of Science Highly Cited Researcher
Department of Information Engineering
The Chinese University of Hong Kong
http://jianwei.ie.cuhk.edu.hk/

Director's Message

For decades, the wireless network evolution has been driven by a strive for higher data rates. Indeed, the whole premise of 4G networks had been on delivering higher rate and high network capacity. However, the advent of the Internet of Things and associated services, such as autonomous vehicles and virtual reality, has radically changed the wireless landscape. In particular, we are witnessing a major shift from data rate-centric wireless networks, to networks that require very low latency and high reliability. In this new latency-centric landscape, cognitive networking approaches will inevitably play a key role. In this regard, this TCCN Newsletter will delve into two key emerging technologies that pertain to the paradigm of highly reliable, low latency communications: a) ultra reliable low latency communication (URLLC) and b) mobile edge computing (MEC). Both URLLC and MEC will be critical components of emerging 5G networks and major contributors for the shift from rate-centric systems to latency-centric systems. Consequently, exposing their challenges and opportunities is essential. In order to do so, this first issue of the TCCN Newsletter of 2018 will bring together two feature topics on URLLC and MEC. Within each feature topic, we review the state of the art and provide an in-depth exposition of some of the recent research contributions. For URLLC, we also provide two expert interviews with Drs. Mehdi Bennis and Marios Kountouris, that provide the academic and industrial perspectives on URLLC. In the context of MEC, beyond also reviewing key papers, we provide two expert interviews with Drs. Kaibin Huang and Yang Yang.

That said, we would like to thank our two feature topic editors: Dr. Daniel Benevides da Costa from UFC - Brazil and Dr. Jie Xu, from Guangdong University of Technology, for their efforts in arranging the paper reviews and expert opinion. Moreover, we want to thank all interviewees for sharing with us their useful experience and future outlook on the discussed areas. I would finally like to acknowledge the gracious support from the TCCN chair, Dr. Jianwei Huang and all TCCN officers. As always, if you have any suggestions, feel free to contact me at: walids@vt.edu. We hope that you enjoy the material provided here!

Walid Saad (S'07, M'10, SM'15) (walids@vt.edu) received his Ph.D degree from the University of Oslo in 2010. Currently, he is an Associate Professor at the Department of Electrical and Computer Engineering at Virginia Tech, where he leads the Network Science, Wireless, and Security (NetSciWiS) laboratory, within the Wireless@VT research group. His research interests include wireless networks, machine learning, game theory, cybersecurity, unmanned aerial vehicles, and cyber-physical systems. Dr. Saad is the recipient of the NSF CAREER award in 2013, the AFOSR summer faculty fellowship in 2014, and the Young Investigator Award from the Office of Naval Research (ONR) in 2015. He was the author/co-author of six conference best paper awards at WiOpt in 2009, ICIMP in 2010, IEEE WCNC in 2012, IEEE PIMRC in 2015, IEEE SmartGridComm in 2015, and EuCNC in 2017. He is the recipient of the 2015 Fred W. Ellersick Prize from the IEEE Communications Society and of the 2017 IEEE ComSoc Best Young Professional in Academia award. From 2015-2017, Dr. Saad was named the Stephen O. Lane Junior Faculty Fellow at Virginia Tech and, in 2017, he was named College of Engineering Faculty Fellow. He currently serves as an editor for the IEEE Transactions on Wireless Communications, IEEE Transactions on Communications, IEEE Transactions on Mobile Computing, and IEEE Transactions on Information Forensics and Security.

Feature Topic: Ultra-Reliable Low-Latency Communication (URLLC)

Editor: Daniel B. da Costa

Department of Computer Engineering
Federal University of Ceará (UFC)
Sobral-CE-Brazil
danielbcosta@ieee.org

The increasing growth of data traffic witnessed by the worldwide population has spurred by the internet-of-things (IoT) applications which range from machine-type communication to mission critical communications. A successful implementation of IoT calls for a wireless communication system that is able to support a much larger number of connected devices, and that is able to fulfill much more stringent requirements on latency and reliability than what current standards can guarantee. Owing to this fact, one of the main goals of 5G communication systems is the support of ultra-reliable low-latency communications (URLLC), which refer to communication services where data packets are exchanged at moderately low throughput (i.e., around 50 Mb/s) but with stringent requirements in terms of reliability (i.e., around 99,999%) and latency (i.e., around 4ms).

Achieving URLLC introduces a myriad of challenges in terms of system design, which is further exacerbated by: (i) growing network size and increasing interactions between nodes; (ii) high level of uncertainty due to random changes in the topology; (iii) heterogeneity across applications, networks, and devices. In addition, URLCC features system design tradeoffs and these fundamental tradeoffs, such as device energy consumption vs. latency, energy expenditures vs. reliability, reliability vs. latency and rate, user density vs. dimensions (antennas, bandwidth, blocklength), deserve a study on their own.

URLLC concepts have several potential applications. For instance, virtual reality is a use case where URLLC plays an important role, due to the fact that the human eye needs to experience accurate and smooth movements with low motion-to-photon latency to avoid motion sickness. Mobile edge computing (MEC) in which MEC servers are deployed at the network edge to provide faster computation capabilities for computing mobile devices' tasks arises as another scenario of application. A third application is in multi-connectivity for ultra-dense networks, where it is interesting to investigate the fundamental problem of base station-user equipament association aiming at improving capacity in the context of ultra-dense networks via multiconnectivity.

In essence, as outlined in [1], URLCC can be broken down into three major building blocks, namely: (i) risk, (ii) tail, and (iii) scale. Risk is naturally encountered when dealing with decision making under uncertainty, when channels are time-varying, and in the presence of network dynamics. The notion of tail behavior in wireless systems is inherently related to the tail of random traffic demand, tail of latency distribution, intra/inter-cell interference, and users that are at the cell edge, power-limited, or in deep fade, that needs to be optimized. Therefore, a principled framework and mathematical tools that characterize these tails focusing on percentiles and extreme events are needed. Finally, scale is motivated by the sheer amount of devices, antennas, sensors and other nodes which pose serious challenges in terms of resource allocation and network design. In contrast to cumbersome and time-consuming Monte-Carlo simulations, mathematical tools providing a tractable formulation, analysis and crisp insights are needed. Indeed, enabling URLLC warrants a major departure from average-based performance towards a clean-slate design centered on tail, risk and scale.

[1] M. Bennis, M. Debbah, and H. V. Poor, Ultra-Reliable and Low-Latency Wireless Communication: Tail, Risk, and Scale,

https://arxiv.org/pdf/1801.01270

In the next sections, we present the review of two representative works on URLCC and two interviews with leading experts in the field, Prof. Mehdi Bennis and Dr. Marios Kountouris. I take this opportunity to thank them for their precious contributions to this feature topic.

Daniel Benevides da Costa (S'04-M'08-SM'14) was born in Fortaleza, Ceará, Brazil, in 1981. He received the B.Sc. degree in Telecommunications from the Military Institute of Engineering (IME), Rio de Janeiro, Brazil, in 2003, and the M.Sc. and Ph.D. degrees in Electrical Engineering, Area: Telecommunications, from the University of Campinas, SP, Brazil, in 2006 and 2008, respectively. His Ph.D thesis was awarded the Best Ph.D. Thesis in Electrical Engineering by the Brazilian Ministry of Education (CAPES) at the 2009 CAPES Thesis Contest. From 2008 to 2009, he was a Postdoctoral Research Fellow with INRS-EMT, University of Quebec, Montreal, QC, Canada. Since 2010, he has been with the Federal University of Ceará, where he is currently an Associate Professor.

Prof. da Costa is currently Editor of the IEEE Communications Surveys and Tutorials, IEEE Access, IEEE Transactions on Communications, IEEE Transactions on Vehicular Technology, and EURASIP Journal on Wireless Communications and Networking. He has also served as Associate Technical Editor of the IEEE Communications Magazine. From 2012 to 2017, he was Editor of the IEEE Communications Letters. He has served as Guest Editor of several Journal Special Issues. He has been involved on the Organizing Committee of several conferences. He is currently the Latin American Chapters Coordinator of the IEEE Vehicular Technology Society. Also, he acts as a Scientific Consultant of the National Council of Scientific and Technological Development (CNPq), Brazil and he is a productivity Research Fellow of CNPq. Currently, he is the Chair of the Special Interest Group on "Energy-Harvesting Cognitive Radio Networks" in IEEE Cognitive Networks Technical Committee.

Prof. da Costa is the recipient of four conference paper awards. He received the Exemplary Reviewer Certificate of the IEEE Wireless Communications Letters in 2013, the Exemplary Reviewer Certificate of the IEEE Communications Letters in 2016 and 2017, the Certificate of Appreciation of Top Associate Editor for outstanding contributions to IEEE Transactions on Vehicular Technology in 2013, 2015 and 2016, the Exemplary Editor Award of IEEE Communications Letters in 2016, and the Outstanding Editor Award of IEEE Access in 2017. He is a Distinguished Lecturer of the IEEE Vehicular Technology Society. He is a Senior Member of IEEE, Member of IEEE Communications Society and IEEE Vehicular Technology Society.

Review of: "Toward Massive, Ultrareliable, and Low-Latency Wireless Communication with Short Packets", in Proceedings of the IEEE, Sep. 2016

By G. Durisi, T. Koch, and P. Popovski

Objectives of the paper

The paper reviews recent advances in information theory which provide the theoretical principles that govern the transmission of short packets, and it shows how the developments will impact the design of future wireless communication systems. The challenges that need to be addressed to optimally design ultrareliable communication (URC) and massive machine-to-machine communication (MM2M) are highlighted by means of three examples (two-way channel, downlink broadcast channel, and uplink random access channel) which illustrate how the tradeoffs brought by short-packet transmission affect protocol design.

Relevance to the feature topic

The impressive growth of data traffic spurred by internet-of-things (IoT) applications ranging from machine-type communications (MTCs) to mission-critical communications are posing unprecedented challenges in terms of capacity, latency, reliability, and scalability. Owing to this fact, the upcoming wireless systems, notably the fifth-generation (5G) system, will address the specific needs of autonomous machines and devices by providing two wireless modes: URC and MM2M. The central challenge with these two new wireless modes is the capability to support short packet transmission since short packets are the typical form of traffic generated by sensors and exchanged in MTCs. This requires a fundamentally different design approach than the one used in currently high-data systems, such as 4G LTE and WiFi, which implies in new principles for the design of wireless protocols supporting short packets.

Major contributions

The paper first reviews the recent advances in information theory, which provide the theoretical principles that govern the transmission of short packets. Afterwards, these principles are applied to three exemplary scenarios (two-way channel, downlink broadcast channel, and uplink random access channel), thereby illustrating how the transmission of control information can be optimized when the packets are short. The insights brought by these examples suggest that new principles are needed for the design of wireless protocols supporting short packets. These principles will have a direct impact on the system design.

Novelty

The paper brings new insights for the technical literature related to massive, ultrareliable, and low-latency wireless communications with short packets. It is illustrated through some representative examples how to use the maximum coding rate performance metric to optimize the protocol design and the transmission of metadata in short-packet communications.

Key results

Two-Way Communication Protocol:

It is shown that adjusting the packet length and the coding rate has the potential to yield high reliability. However, that flexibility in the packet length necessarily implies that the receiver needs to acquire information about it. This means that the protocol needs to reserve some bits within each packet for the metadata that describes the packet length. In addition, the use of a predefined slot length yields a robust system design, since no additional error is caused by the exchange of length-related metadata. This indicates that, in designing protocols that support ultrahigh reliability, a holistic approach is required that includes all elements of the protocol/metadata that are commonly assumed to be perfectly received.

Downlink Multiuser Communication:

It is considered the scenario in which a base station (BS) transmits in the downlink to M devices. It is shown that, for short packet sizes, it may be more efficient to encode a larger number of data than the one intended to each device. Thus, instead of using TDMA, the BS may concatenate all the data packets for the individual devices.

Uplink Multiuser Communication:

It is considered the scenario in which M devices run a random access protocol in order to transmit to a common BS. Specifically, there are M users, each sending D bits to the BS. Each packet should be delivered within a time that corresponds to n channel uses. These n channel uses are

dividedinto K equally sized slots of $n_K = n/K$ channel uses. The devices apply a simple framed ALOHA protocol: each device picks randomly one of the K slots in the frame and sends its packet. If two or more users pick the same slot, then a collision occurs and none of the packets is received correctly. If only one device picks a particular slot (singleton slot), then the error probability is calculated.

The following question is addressed: given M, D, and n, how to choose the slot size n_K in order to maximize the packet transmission reliability experienced by each individual device? This problem entails a tradeoff between the probability of collision and the number of channel uses available for each packet, which affects the achievable packet error probability in a singleton slot. Indeed, if K increases, then the probability of a collision decreases, while the packet error probability for a singleton slot increases. Conversely, if K decreases, then the probability of collision increases, while the packet error probability for a singleton slot decreases.

Outlook

Motivated by the advent of novel wireless applications, such as MM2M and URC, the paper provides a review of recente advances in the theory of short-packet communications and demonstrated through three examples how this theory can help designing novel efficient communication protocols that are suitable for short-packet transmissions. The key insight is that, when short packets are transmitted, it is crucial to take into account the communication resources that are invested in the transmission of metadata. This unveils tradeoffs that are not well understood yet and that deserve further research, both on the theoretical and on the applied side.

Review of: "Fundamental Tradeoffs among Reliability, Latency, and Throughput in Cellular Networks", in IEEE GLOBECOM 2014, Workshop on Ultra-Low Latency and Ultra-High Reliability in Wireless Communications

By B. Soret, P. Mogensen, K. I. Pedersen, and M. C. Aguayo-Torres

Objectives of the paper

The key objective of the paper is to address the fundamentals tradeoffs among latency, reliability and throughput in a cellular network. The most important elements influencing the key performance indicator (KPI) of the network are identified, and the inter-relationships among them is discussed. The effective bandwidth and the effective capacity theory are used as analytical framework for calculating the maximum achievable rate for a given latency and reliability constraint. The analysis is conducted in a simplified LTE network, providing baseline – yet powerful - insight of the main tradeoffs. Guidelines to extend the theory to more complex systems are also presented, including a semi-analytical approach for cases with intractable channel and traffic models. Based on the findings, some recommendations are given for the imminent 5G technology design phase, in which latency and reliability will be two of the principal KPIs.

Relevance to the feature topic

The explosion of machine-to-machine (M2M) communications opens the possibility of implementing a myriad of applications requiring extremely low latency and ultra high reliability. LTE, the de facto standard for 4G cellular networks, is postulated as a candidate for the support of M2M. One main concern for the use of the LTE network relates to its capability of meeting the stringent reliability and latency constraints without compromising the delivery of traditional applications.

Major contributions

The paper investigates three main KPIs of LTE networks for M2M communications, namely latency, reliability and throughput. A sketch of the considered system model is shown in Figure 1. The scope is limited to the downlink and PHY/MAC procedures, i.e., higher layer procedures are not considered such as Radio Link Control and Transmission Control Protocol retransmissions in the KPIs budgets. The joint use of the effective bandwidth and the effective capacity theory as analytical or semi-analytical frameworks is proposed to investigate the tradeoffs among the KPIs. The use of system level simulations and several sources of imperfections in the system are discussed. Finally, based on the achieved results, new insights and recommendations for the design of 5G networks are provided.

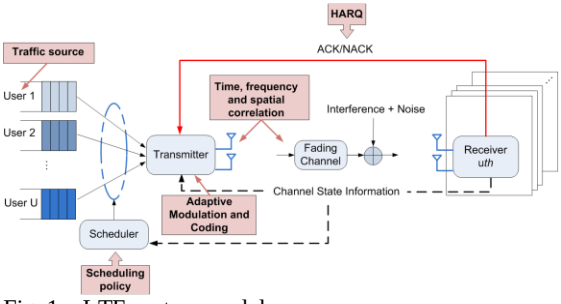


Fig. 1 – LTE system model

Novelty

The paper provides good insights related to the main mechanisms impacting the KPIs of interest. Guidelines to extend the theory to cases in which the channel and/or the traffic model are not tractable from a mathematical point of view are also provided, as well as the advisability of system-level simulations in such a complex system with several random elements and related sources of non-idealities. Finally, the paper identifies different candidate techniques to improve the tradeoff among the KPIs in future 5G systems.

Key results

One of the key results of the paper is that in a multiuser environment the three considered KPIs can be guaranteed only for a fraction of the load in the system, and at the expenses of larger latencies for the rest of best effort users in the network. Moreover, the main procedures relevant for the study can be incorporated into the system model to get a good approximation to the final values. However, simplifications and idealizations limit the scope of purely analytical studies when dealing with very complex systems, as it is the case of LTE. In this sense, system-level simulations can model not only all the relevant elements but also different sources of imperfection. Besides, it is shown that several sources of imperfection at the PHY/MAC level

are also important for the latency and reliability performance. For example, Adaptive Modulation and Coding (AMC) makes use of the received channel state information (CSI) feedback, which is subject to various imperfections such as measurement imperfections, quantization, reporting delay, and reception errors. All of them can be represented by a random process, which essentially means that there is certain probability that the AMC selected by the base station deviates from the ideally desired selection. In addition, the hybrid automatic repeat request (HARQ) performance is influenced by the randomness associated with occasionally rare mis-detection of ACK/NACK at the base station from the terminals. The composite variability of all of the sources of imperfection can indeed lead to long tails of the transmission delay, although at low probabilities

Outlook

The next generation of mobile radio access Technologies (5G), expected to become available for commercial launch around 2020, is right now in its early exploration phase with several unknowns regarding the requirements and the Technologies to be used. It is indicated in the paper the various sources of variability with impact on the link performance in 4G LTE and how to further improve them for 5G. Hence, pointing to the candidate techniques that could help in enabling the support of ultra low latency and high reliable communications in the future.

In the frequency and spatial domain, the increased diversity given by larger bandwidths and higher order MIMO (massive MIMO) will help in improving the tradeoff between latency and reliability. As for the interference, the goal is to achieve a more stable interference footprint. The use of advanced interference management comprising both network-based coordination and receiver-based techniques is expected to be an integral part of 5G, providing both general capacity benefits and improved reliability by reducing the interference vulnerability. A promising enhancement in the network side is the use of multi-cell baseband pooling, which offers opportunities for centralized multi-cell scheduling, reducing some of the uncertainties that would otherwise be present if conducting independent scheduling and resource allocation per cell. Further enhancements aiming at reducing the latency include the use of a shorter subframe duration for a reduced HARQ transmission delay.

Interview with Prof. Mehdi Bennis

Understanding URLLC for 5G

How do you define future ultra-reliable low-latency communications (URLLC) networks, and what will be the business cases for such 5G networks? What are the sought after reliability and latency requirements?

MB: 5G is broken down into eMBB, machine type communication (MTC) and URLLC. URLLC warrants a major departure from average-based performance (average delay, average throughput, etc.) towards a design in which guarantees in terms of latency and reliability are crucial.

URLLC will pervade every 5G vertical where low-latency and high-reliability requirements are important. These include smart factories, vehicular communication, virtual reality, drones, remote surgery and many other unforeseen applications. Owing to these heterogeneous requirements, lots remains to be done across academia, industry, and regulators.

Requirements are very much use case dependent. For instance, V2V traffic safety requires very low latency and high-reliability with low rates. However, exchanging high-definition maps requires also high-bandwidth. Likewise, in VR high-bandwidth low-latency and high-reliability are needed.

Could you please explain the most pertinent URLLC techniques/scenarios for 5G?

MB: As opposed to the eMBB design, URLLC warrants a major departure from average-based performance (average delay, average throughput, etc.) towards a clean-slate design centered on tail, risk and scale. The notion of risk is encountered when dealing with decision making under uncertainty; scale is motivated by the sheer amount of devices, antennas, sensors and other nodes which pose unprecedented challenges in resource optimization and network design. URLLC scenarios are limitless, among the pertinent ones virtual reality, self-driving vehicles, drones, remote control for surgery, multi-sensory AI, and so forth.

There are several mathematical methodologies tailored to the unique requirements of URLLC. For instance, for tail-centric design in which providing latency/reliability guarantees is key, extreme value theory helps characterizing statistics of extreme events which are then incorporated into the system design (e.g. power minimization subject to URLLC constraints). For decision making under risk, game theory and reinforcement learning are useful tools. Finally, for addressing the scaling part, tools from mean field theory, statistical physics are of interest.

Could you please elaborate on the main role of URLLC in key vertical industries? How does the telecoms sector engage with them during the standardization phase? How does it actually build these state-of-the-art networks? Is 2020 a realistic deadline?

MB: URLLC requirements are vertical-specific. In this respect, 3GPP has defined early requirements, namely that the minimum requirement for reliability is $1-10-5$ success probability of transmitting a layer 2 protocol data unit of 32 bytes within 1 ms. Worth noting is that URLLC service requirements are end-to-end, whereas the 3GPP and ITU requirements focus on the one-way radio latency over the 5G radio network. To address these heterogeneous requirements within URLLC 3GPP started discussing with different verticals such as V2X and VR to understand joint end to end requirements.

5G standardization is going full steam with the recent release of the first 5G NR milestone (for non standalone 5G). While the evolutionary part of 5G (eMBB) has gained significant momentum with substantial research targeting the use of high frequency bands (6GHz-300GHz), the promised revolutionary path of 5G has not lived up to the expectations. Said otherwise URLLC is not well understood and hence more work is needed beyond 2020, with the need to address even lower requirements.

Could you please briefly introduce the most recent research project that you have done in URLLC?

MB: Within my group and our collaborators, we have been working on the fundamentals of URLLC from a networking perspective. As opposed to the cumbersome Monte-Carlo simulations lacking insights our aim is to deconstruct URLLC in terms of building blocks, namely risk, tail and scale. We have also been working on several use cases, including VR, V2X, UAVs, and smart factories. Currently, we are exploring the use of distributed wireless AI/ML to solve some of the learning URLLC problems.

Beyond your own work, are there any resources that you would like to recommend, specially to those who are new in the field and want to learn more about URLLC?

MB: For relevant works in URLLC, short-packet communication works are very important. These go back to the information theoretic works of Polyansky, Verdu and Poor, then recent works of Durisi and Popovski. Several recent works can be found in URLLC from a system level simulations perspective (3GPP oriented), whose focus is basically on introducing new frame structure and numerology (shorter TTI). From a networking perspective, there are very few works but this is increasing gradually as evidenced by the latest IEEE Network special issue on URLLC and the upcoming JSAC on URLLC. In this respect our recent works aim at filling the void in terms of techniques geared towards the URLLC requirements at the network level.

What are the most important open problems and future research directions towards URLLC?

MB: End-to-end latency and reliability are still not well understood. This is due to the fact that current designs are not jointly optimized. For instance, if one takes virtual reality 3GPP focuses on the wireless part of the equation and overlooks the specific requirements of VR in terms of sensing, perception and tracking. Likewise for UAVs, V2X. The right approach is to jointly consider perception, sensing, computing and wireless as a control loop subject to motion-to-photon constraints.

Few open problems:
- Short packet communication provides a closed-form expression of the maximum coding gain for several fading channels and interference-free settings. The more practical case with burst interference remains an open problem.
- Current channel models do not reflect the unique requirements of URLLC. There is a need to better characterize the tails of these fading distributions.
- Joint design across all 5G verticals is still an open problem. Thus far, control/communication requirements have been designed sequentially.
- Data driven/optimized URLLC where depending on latency/reliability budget, how to optimize the modulation coding scheme as a function of interference, network density and application requirements.

Dr. Mehdi Bennis is an Associate Professor at the Centre for Wireless Communications, University of Oulu, Finland and an Academy of Finland Research Fellow. His main research interests are in radio resource management, heterogeneous networks, game theory and machine learning in 5G networks and beyond. He has co-authored one book and published more than 200 research papers in international conferences, journals and book chapters. Dr. Bennis has been the recipient of several awards including the 2015 Fred W. Ellersick Prize from the IEEE Communications Society, the 2016 Best Tutorial Prize from the IEEE Communications Society and the 2017 EURASIP Best paper Award for the Journal of Wireless Communications and Networks.

Interview with Dr. Marios Kountouris

Understanding URLLC for 5G

How do you define future ultra-reliable low-latency communications (URLLC) networks, and what will be the business cases for such 5G networks? What are the sought after reliability and latency requirements?

MK: Fifth generation (5G), the next generation mobile communication system, is currently being standardized and the specifications of 5G New Radio (NR) in standalone system (Release 15) will be completed in June 2018. One major difference of 5G compared to previous mobile generations is its target to support a broader spectrum of applications and use cases. Mobile networks have been driven so far by human-centric communications, delay-tolerant content, and non-critical services. The major objectives have been to boosting data rate and increasing coverage; the latency requirements of different applications have mostly been an after-thought. 5G envisions to provide wireless connectivity for massive machine-type communications and to support ultra-reliable, low latency communication (URLLC) for mission-critical services. As the term states, URLLC introduces a new performance metric related to reliability guarantees and latency bounds.

Reliability is defined as the success probability of transmitting a given amount of traffic within a certain latency at a certain channel quality. Latency in 5G URLLC refers to user plane latency, which is defined as the one-way time it takes to successfully deliver an application layer packet from the radio protocol layer 2/3 ingress point to the radio protocol layer 2/3 egress point of the radio interface in unloaded conditions.

The URLLC requirement defined by ITU and 3GPP is to transmit a layer 2 protocol data unit of 32 bytes over the 5G radio interface with a user plane latency of 1 ms and with a 1-10-5 success probability in channel quality of coverage edge for an urban macrocell environment. 3GPP further imposes 5G to achieve a user plane latency of 0.5 ms for uplink/downlink. The minimum requirement for control plane latency is 10 ms.

It is important to emphasize that the URLLC requirements specified in ITU and 3GPP consider only the one-way latency, which represents a certain fraction of end-to-end (E2E) latency. E2E latency accounts for delays at various layers in the protocol stack and includes over-the-air transmission and retransmission delays, queuing delay, core network delay, as well as processing delay. Furthermore, E2E latency has both deterministic and random components, which may scale or not with the number of devices, the topology, and the communication range. Another relevant URLLC metric – often neglected or incorporated in the reliability – is network availability. It refers to the probability (and period) that predefined service requirements are satisfied over the coverage area. For mission-critical services, the required network availability is 99.999%. Lastly, security and safety are key to several URLLC applications but there are no formal definitions or tangible requirements in ITU and 3GPP.

There is a plethora of socially useful services and business domains that would benefit greatly from URLLC. The major and most challenging scenarios are factory automation and industrial control, intelligent and autonomous transportation, and remote healthcare. URLLC is also expected to revolutionize processes in the areas of smart cities, smart farming, smart grid, remote manufacturing, and algorithmic trading. The stringent URLLC constraints and the nature of real-time mission-critical applications imply the predominance of short packets and low-rate transmissions. Nevertheless, future releases of URLLC services may also consider rate requirements. The emergence of immersive services, such as augmented and virtual reality (AR/VR), high-definition entertainment and gaming, tactile Internet, and consumer robotics, calls for real-time, high-fidelity, broadband communications operating at latencies of few milliseconds.

Could you please explain the most pertinent URLLC techniques/scenarios for 5G?

MK: 5G NR introduces several novel techniques and changes compared to LTE in order to support URLLC services.

For low-latency transmissions, 5G NR defines flexible/scalable numerology with subcarrier spacing of 15, 30, and 60 kHz below 6 GHz, and 60 and 120 kHz above 6 GHz. Higher subcarrier spacing decreases the OFDM symbol duration, which in turn reduces the slot duration. Other promising techniques are transmission time interval (TTI) shortening to 0.125 ms (at 120 kHz SCS) and the use of mini-slot, which allows

URLLC transmissions to occur at any time. A mini-slot can start at any OFDM symbol and can have lengths of 2, 4, or 7 symbols. The HARQ roundtrip time is reduced to 3-4 TTIs, enabling more retransmissions within the latency constraint. In time division duplex (TDD) mode, NR will also benefit from self-contained slot design, which enables flexible TDD switching and faster turn-around. To alleviate the delay from scheduling requests and grants in the uplink, 5G NR advocates for fast uplink access using grant-free/contention-based access and/or semi-persistent scheduling. To complete the picture, 5G NR propounds efficient multiplexing of long and short transmissions and preemption for mission-critical traffic.

The role of diversity for ultra-reliable transmissions is cardinal. Although diversity can be exploited in time, frequency and space domains, spatial diversity using multiple antennas (MIMO) stands out as the most promising solution. The difficulty of exploiting time diversity under tight latency constraints and frequency diversity under scarce spectrum resources makes MIMO an essential component for achieving high reliability. Defining additional modulation and coding schemes to support operations at very low code rates and with low constellation orders will play a complementary role in reducing the block error rate. Finally, multi-connectivity either via multiple sites (macrodiversity) or across multiple carrier frequencies (inter-frequency) is primordial to ensure steady network availability and persistent connectivity.

In addition to the new air interface, 5G introduces several novel architectural elements and capabilities to reduce processing time, bypass several protocol layers, and support heterogeneous service requirements. Radio access virtualization and mobile edge computing are two key enabling technologies for URLLC. Together with network slicing, which benefits from recent developments in software defined network and network function virtualization, they will provide service level agreement (SLA) guarantees for verticals.

Could you please elaborate on the main role of URLLC in key vertical industries? How does the telecoms sector engage with them during the standardization phase? How does it actually build these state-of-the-art networks? Is 2020 a realistic deadline?

MK: The vision of 5G is to provide a vertical industry optimized platform that will support the very diverse – even the most stringent and business critical – service requirements of each sector. As mentioned above, URLLC is central to vertical markets with time-critical applications involving real-time closed-loop control and remote monitoring.

Without having a complete view, it seems that there is no direct involvement of vertical industries into the standardization discussions. Engagement with the public safety, broadcasting and automotive sectors is judged satisfactory. The picture is a bit darker when it comes to engaging the factory automation, agriculture and mining industries. ETSI and 3GPP, being fully aware of the importance of direct engagement from verticals, have intensified their efforts. 5G-PPP is endeavoring to create a platform where verticals can use 5G applications in a more integrated way. 5G validation trials across multiple vertical industries are targeted for 2019 (https://5g-ppp.eu/wp-content/uploads/2017/05/5GInfraPPP-Trials-Roadmap-Strategy_Short_28-February-2017.pdf).

It seems that 2020 is a realistic deadline for 5G commercialization, despite the diverging opinions during the past 2-3 years. The telecom sector has realized that meeting the 2020 deadline is crucial to avoid a fragmentation of standards and to support the tangible needs of upcoming Olympics in Japan and Korea. This is also the main, not to say the sole, reason why standardization bodies have accelerated the completion of the first phase of 5G (Release 15). However, all indications corroborate that the first 5G deployments will focus on mobile broadband. Verticals are not expected to adopt 5G at its early stage and 2022+ sounds more realistic for real URLLC deployments. Verticals may wait for the 5G benefits to be established in order to decide whether 5G can meet their needs and can replace services for which significant investment has already been made.

There are various non-technical reasons why URLLC is not included in the initial deployment plans. First, bringing demand and supply together in the market is always challenging. Second, there are still several critical questions to be answered. The biggest question is who will build and pay for URLLC networks. Let's take the case of industrial automation, where operators may not be allowed to have full access in the factory production line due to security and reliability reasons. Does this mean that private 5G networks are necessary for

industry 4.0 factories, and if so, how the business models and opportunities will evolve in this new ecosystem? Will private companies solely lease capacity or slices from an operator? Will mobile operators seize the opportunities and take the risk of providing URLLC services in environments they do not fully control? It is also primordial to clarify who takes responsibility for costly production stops or for catastrophic events and complications during a remote surgery due to URLLC requirement violations. Moreover, providing URLLC services is a complete mindset shift for operators, which are used to reason in terms of capacity, average delay, and probabilistic guarantees rather than reliability and worst-case latency.

Could you please briefly introduce the most recent research project that you have done in URLLC?

MK: My group has been heavily involved in the development of fundamental theory, new transceiver techniques and resource allocation algorithms for URLLC.

A deep understanding of the delay performance in wireless networks is essential for efficient URLLC systems. Identifying the lack of powerful theoretical approaches for URLLC system optimization, we developed a mathematical framework for resource planning. Capitalizing on stochastic network calculus and effective bandwidth theory, we proposed new methodologies for modeling and evaluating the network layer performance of 5G networks. Our approach allows to characterize the delay violation probability (worst-case latency) and the backlog for various technologies, including MIMO transmissions, multicast beamforming, and small cells networks. In another theoretical work, we studied the fundamental energy-latency tradeoff in the finite blocklength regime for URLLC systems employing incremental redundancy hybrid automatic repeat request (IR-HARQ). Establishing the non-convexity of the average energy minimization problem under latency constraints and feedback delay, we used dynamic programming for energy efficient IR-HARQ optimization in terms of number of retransmissions, blocklength and transmit power. Our results showed that it is beneficial to split the packet into sub-codewords and use IR-HARQ rather than do one-shot long transmission.

From an algorithmic perspective, we designed low-complexity MIMO scheduling and precoder computation schemes for scenarios with both continuous long transmissions and bursty short-packet traffic. The key idea is that ongoing downlink transmissions with channel conditions similar to the URLLC receiver larger than a predefined threshold are preempted to meet the latency constraint. Additionally, one can perform persistent MIMO scheduling, i.e., the MIMO precoder of the preempted transmissions can be kept and used for URLLC transmissions, resulting in remarkable computational complexity reduction. For supporting the 5G standardization activities on ultra-reliable transmissions, we developed an adaptive weighted MIMO transmission scheme, which combines MIMO precoding and space-time/frequency coding depending on the reliability required and the quality of channel knowledge. The scheme is compatible with type-I and type-II codebook feedback and no change to downlink DMRS-RS and CSI-RS pilots is needed.

Finally, we investigated multi-user scheduling in URLLC systems, which turned out to be an NP–hard problem. For that, we formulated and solved the associated problem of URLLC SLA satisfaction, which is an infinite horizon constrained Markov Decision Process. To alleviate the curse of dimensionality, we proposed a class of knapsack-inspired computationally efficient - but not necessarily optimal – scheduling policies. Interestingly, every policy in that class becomes optimal in a fluid regime and all policies are shown to perform very well even in small practical instances of the URLLC scheduling problem.

Beyond your own work, are there any resources that you would like to recommend, especially to those who are new in the field and want to learn more about URLLC?

MK: First, I would recommend to read the overview/tutorial papers on URLLC that have appeared in the past 2-3 years - together with earlier ones on ultra-reliable communications and finite-blocklength information theory. For technical details, there are several academic groups researching and publishing on URLLC, such as the Berkeley Wireless Research Center, the MassM2M group, the CWC Finland, and the 5G Lab Germany in Dresden, to name a few.

Second, the results disseminated by related projects funded by 5G-PPP could be a valuable source (e.g. NORMA, FANTASTIC, SLICENET, MONARCH), together with the outcome of the ERC-funded WILLOW project.

Third, the upcoming IEEE JSAC special issue on URLLC (https://www.comsoc.org/jsac/cfp/ultra-reliable-low-latency-communications-wireless-networks) is expected to collect the latest and most valuable results and developments on URLLC from both academia and industry. Lastly, two interesting venues are the NSF Workshops on Low-Latency Wireless and the URLLC 2017 industrial conference (http://urllc2017.executiveindustryevents.com).

What are the most important open problems and future research directions towards URLLC?

MK: Delay characterization is a long standing challenge in wireless networking research. Despite various analytical approaches that have been proposed, we are still lacking of a solid mathematical theory, which will enable us to quantify and understand the fundamental limits of low-latency communications. A neat and insightful theory, of similar beauty and utility to Shannon theory, is instrumental for efficient and optimized URLLC operation.

Emerging 5G use cases may drive some of the innovation and open problems. For instance, VR is considered to be the "killer app" of 5G networks due to the high rate and low-latency requirements. We should seek out novel advanced technologies that will provide at the same time high throughput and unprecedented levels of reliability with reasonably latency. New uses of multi-antenna technology ("extreme MIMO") and extremely high frequency communication (even up to terahertz) are two promising directions, which in turn will bring a variety of technical problems.

With the exception of real-time control applications, several URLLC services could tolerate larger latency. For instance, the latency requirement for AR/VR is around 13ms, telesurgery with haptics requires E2E latency of 10 ms, tactile Internet needs sub 10ms and video-enabled autonomous driving is feasible with 50 ms E2E latency. Designing wireless networks for extremely low latency may lead to ineffective or conservative system operation. Future URLLC research directions may relax the challenging 1 ms latency constraint opening up the path for high-rate URLLC.

Finally, E2E network orchestration and optimization is required for coping with the heterogeneous service requirements. However, the majority of E2E optimization problems are highly complex (often NP-hard), for which efficient, fast, and implementable algorithms have to be developed. Machine learning and algorithms and artificial intelligence could be of great assistance with this task.

Marios Kountouris (S'04-M'08-SM'15) received the Diploma in Electrical & Computer Engineering from the National Technical University of Athens, Greece in 2002 and the M.S. and Ph.D. degrees in Electrical Engineering from the Ecole Nationale Supérieure des Télécommunications (Télécom ParisTech), France in 2004 and 2008, respectively. From February 2008 to May 2009, he has been with the Department of ECE at The University of Texas at Austin as a research associate, working on wireless ad hoc networks under DARPA's IT-MANET program. From June 2009 to July 2016 he has been an Assistant and Associate Professor at SUPELEC (now CentraleSupélec), France. From March 2014 to February 2015, he has been an Adjunct Professor in the School of EEE at Yonsei University, S. Korea. Since January 2015, he has been a Principal Researcher at the Mathematical and Algorithmic Sciences Lab, Paris Research Center, Huawei Technologies, France. He currently serves as Associate Editor for the IEEE Transactions on Wireless Communications, the IEEE Transactions on Signal Processing, and the IEEE Wireless Communication Letters. He received the 2016 IEEE ComSoc Communication Theory Technical Committee Early Achievement Award, the 2013 IEEE ComSoc Outstanding Young Researcher Award for the EMEA Region, the 2014 Best Paper Award for EURASIP Journal on Advances in Signal Processing (JASP), the 2012 IEEE SPS Signal Processing Magazine Award, the IEEE SPAWC 2013 Best Student Paper Award, and the Best Paper Award in Communication Theory Symposium at IEEE Globecom 2009. He is a Professional Engineer of the Technical Chamber of Greece.

Feature Topic: Mobile Edge Computing (MEC)

Editor: Jie Xu

School of Information Engineering, Guangdong University of Technology
Email: jiexu@gdut.edu.cn

1. Introduction

Recent technological advancements have enabled various emerging applications such as Internet-of-things (IoT), augmented reality (AR), virtual reality (VR), autonomous driving, unmanned aerial vehicles (UAVs), surveillance, and health monitoring. These applications critically rely on the ubiquitous sensing, communication, computation, and control among massive wireless devices including sensors, actuators, etc. In many applications, such wireless devices need to handle latency-critical and computation-intensive tasks such as the implementation of sophisticated artificial intelligence (AI) algorithms; however, wireless devices are each with small size and only have limited power. Therefore, how to provide them with enhanced communication and computation capabilities is a challenging task faced for making these emerging applications a reality. Towards this end, mobile or multi-access edge computing (MEC) or fog computing has been recognized as a promising solution by pushing computation, storage, and network control to the network edge such as WiFi access points (APs) and cellular base stations (BSs). With MEC, wireless devices can offload computation tasks to close-proximity edge servers for remote execution, thus helping enhance their computation capability and improve the energy efficiency. As compared to the conventional mobile cloud computing (MCC) technique with computation resources located at centralized cloud that is far apart, in MEC the computation resources are distributed at the network edge that is in close proximity to wireless devices. Therefore, the MEC can achieve significantly lower end-to-end computation latency than the MCC. Due to such benefits, MEC has attracted explosively growing interests in both academia and industry.

2. Benefits of MEC

MEC is anticipated to bring various benefits for future wireless networks. First, MEC provides wireless devices with remote computation and storage resources at the network edge. This can not only improve the computation capacity and reduce the computation latency for enabling task-intensive applications at low-power and small-size devices, but also help reduce devices' energy consumption for task execution and extend their lifetime. Next, in MEC, the computation task offloading is performed only at the network edge. This can avoid data transmission at the backhaul networks (for task offloading) and considerably reduce the traffic therein, as compared to the conventional MCC. Furthermore, by exploiting the proximity of edge servers to wireless devices, the MEC can infer users' real-time information about their behaviors and locations, thus enabling advanced location-aware wireless services. Last but not the least, in MEC the computation tasks are processed at nearby edge servers that can be privately owned. Therefore, as compared to the MCC in which the computation tasks are offloaded to clouds that are generally publicly accessible, the MEC can also help improve the security and privacy for computation.

3. Challenges of MEC

Despite the benefits mentioned above, MEC faces various technical challenges in the network design. First, to gain the benefit offered by the remote computation execution at edge servers, wireless devices need to consume additional communication resources for task offloading. Therefore, there generally exists an interesting trade-off between computation and communication for task offloading. How to reveal such a trade-off for characterizing the theoretical limits of computation task offloading is a fundamental but open problem, and solving this problem calls for a *joint design of communication and computation*. Next, both task arrivals and wireless channels fluctuate significantly over time, and different tasks can be dependent with each other in general. These issues make the task offloading/scheduling (over time) a non-trivial problem. Moreover, the task arrival information and channel state information (CSI) are causal in practice, i.e., at each time instant the MEC system is only aware of the current information but does not know that in the future. Therefore, how to balance the (*a-priori* known) utility at the current time instant versus the (*a-priori* unknown) utilities in the future via *online* task

offloading/scheduling is an even more difficult problem to be tackled. In addition, besides dedicatedly deployed servers at APs and BSs, nowadays wireless devices such as smartphones are also attached with rich computation resources. Crowdsourcing these wireless devices to help in the remote task execution is a viable way to further improve the computation performance. This, however, requires carefully designed incentive mechanism for motivating self-interest users to participate in, and also needs careful resource allocation design by potentially considering the new device-to-device (D2D) task offloading. In addition, some location-based applications (e.g., AR) may generate same computation tasks among different wireless devices at different time. How to *cache* repeatedly executed tasks *a-priori* to further reduce the computation latency is also an interesting but challenging problem.

4. Organization of this Feature Topic

This feature topic brings together a number of contributions that aim to address the above challenges faced in MEC. Specifically, in the following we first review three representative works on MEC, which deal with the stochastic joint communication and computation resource allocation, the D2D computation resource sharing, and the computation task caching, respectively. Following those works, we present two interviews with Dr. Kaibin Huang from The University of Hong Kong and Dr. Yang Yang from Shanghai Research Center for Wireless Communications (WiCO), who are leading experts in the field of MEC. We hope that you will find this feature topic useful in spurring research on MEC. Finally, I would like to take the opportunity to thank all the contributors to this feature topic.

Jie Xu (S'12–M'13) received the B.E. and Ph.D. degrees from the University of Science and Technology of China in 2007 and 2012, respectively. From 2012 to 2014, he was a Research Fellow with the Department of Electrical and Computer Engineering, National University of Singapore. From 2015 to 2016, he was a Post-Doctoral Research Fellow with the Engineering Systems and Design Pillar, Singapore University of Technology and Design. He is currently a Professor with the School of Information Engineering, Guangdong University of Technology, China. His research interests include energy efficiency and energy harvesting in wireless communications, wireless information and power transfer, wireless securities, UAV communications, and mobile edge computing. He was a recipient of the IEEE Signal Processing Society Young Author Best Paper Award in 2017. He is currently an Editor of the IEEE Wireless Communications Letters, an Associate Editor of the IEEE Access, and a Guest Editor of the IEEE Wireless Communications.

Review of "Stochastic Joint Radio and Computational Resource Management for Multi-User Mobile-Edge Computing Systems", in IEEE TWC, Sep. 2017

By Y. Map, J. Zhang, S. Song, and K. B. Letaief

Objectives of the Paper

The objective of this paper is to minimize the long-term average weighted sum power consumption of a multi-user MEC system subject to a task buffer stability constraint, where the MEC server (attached to an access point) only has limited computing capability. The radio and computation resources are jointly managed to optimize the long-term average MEC performance. Based on Lyapunov optimization, a low complexity online algorithm is developed. Furthermore, for computation offloading, the optimal transmit power and bandwidth allocations are determined with the Gauss-Seidel method.

Relevance to Feature Topic

Mobile-edge computing (MEC) has been recognized as a promising technology to provide cloud-like computing at the edge of radio access networks closer to the user devises, which can significantly decrease the computing latency for the users offloading the computation tasks to the MEC server.

For multiuser MEC system designs, the optimal operations are temporally and spatially correlated due to the stochastic computation task arrivals for each user and the competition among multiple users for the limited resources.

This paper considers the long-term MEC performance, and stochastic control algorithms are applied for online resource allocation towards energy-efficient MEC designs.

Major Contributions

Based on Lyapunov optimization, the challenging long-term average weighted sum power consumption minimization problem is solved by a low-complexity online algorithm, with a delay-improved approach designed to reduce the execution delay. Performance analysis shows the asymptotic optimality of the proposed online algorithm and numerical results further corroborate the analysis.

Novelty

The Lyapunov optimization based online algorithm for resource allocation in multi-user MEC systems is the main novelty of this paper. The MEC server is assumed to have limited computation resources, and the users perform computation offloading based on frequency-division multiple access (FDMA). Unlike some work focused on the sole computational resource scheduling, this paper considers the joint management of both radio and computational resources to improve the performance of the multi-user MEC systems.

Key Results

Performance analysis shows the performance of the proposed algorithm in the realization of asymptotic optimality and describes the tradeoff between the weighted sum power consumption and the computation delay. Simulation results demonstrate that the proposed framework can deal with the tradeoff between the weighted sum power consumption and computation delay performance. In addition, the impacts of various parameters (e.g., delay, time, and queen length) are revealed, which indicate the importance of a joint radio and computation resource optimization of multi-user MEC systems

Outlook

This paper investigates a stochastic joint radio and computation resource management for multi-user MEC systems and proposes a low-complexity online algorithm based on Lyapunov optimization to minimize the weighted sum power consumption. For the future work, it will be an interesting direction to further take into account the fairness issue among multiple devices and apply more advanced multiple access protocols such as the non-orthogonal multiple access (NOMA). Furthermore, some other research problems, such as mobility-aware resource management, dynamic access control, and user-server associations, should also be investigated to improve the performance of multi-user systems. In addition, the energy-efficient MEC designs should also consider the users' channel uncertainty, as well as the heterogeneity of users' computation/ communication resources and their tasks.

Review of "D2D Fogging: An Energy-Efficient and Incentive-Aware Task Offloading Framework via Network-assisted D2D Collaboration", in IEEE JSAC, Dec. 2016

By L. Pu, X. Chen, J. Xu, and X. Fu

Objectives of the Paper

This paper proposes a D2D Fogging framework via the assistance of the network operators, to enable joint computation and network resources sharing among mobile users. This paper also presents an online low complexity algorithm for task offloading based on Lyapunov optimization. The main objective of the paper is to minimize the time-average energy consumption for task execution for all users, while satisfying the incentive of long-term collaboration to prevent the over-exploiting and free-riding behaviors. Several task scheduling policies are proposed for the designed online task offloading.

Relevance to Feature Topic

Fog computing has been recognized as an emerging paradigm that enables user devices which are closer to the edge of radio access networks to execute massive communication and computation tasks. Differently, in network-assisted D2D collaboration, users can dynamically and beneficially share the computation and communication resources with each other to support collaborative task execution for a variety of services.

Major Contributions

The main contribution of this paper is to propose a novel D2D Fogging architecture, enabling the sharing of both computation and network resources among users. Besides, this paper studies an interesting optimization problem that minimizes the time-average energy consumptions for all users' task execution, under their incentive constraints. An online task offloading algorithm is presented based on the Lyapunov optimization method, utilizing the network information in the current time frame only. In the corresponding time frame, efficient task scheduling policies are devised. Numerical results show that the algorithm can achieve significant performance gain in the energy consumption minimization.

Novelty

The main novelty of this paper is to establish a novel D2D Fogging framework and develop an online computationally-efficient algorithm for mobile task offloading to minimize the time-average energy consumption. Further, the proposed D2D Fogging framework is motivated by both local cloud and user cooperation based architectures, which is the first work to jointly consider the user incentive and cooperation and task scheduling policies.

Key Results

This paper developed a Lyapunov based online task offloading algorithm, which asymptotically achieves the offline optimum. In order to evaluate the performance of the proposed online task offloading algorithm, some numerical results are presented, which demonstrate that the algorithm can achieve superior performance in achieving the minimization of energy cost. Moreover, additional numerical experiments are further provided to validate that the proposed algorithm adapts to different situations, e.g., with different task types.

Outlook

This paper discusses a D2D Fogging framework in multi-user systems, which enables the resource sharing among these users. This D2D Fogging framework can achieve a win-win situation for both network operators and mobile users due to the incentive mechanism for the users and the service coverage extension for the operators. For future investigations, it would be more interesting to take the user-server association into consideration to better exploit potentialities of both computation and communication in the network. Some other interesting application scenarios, such as the heterogeneous networks (HetNets), the server-server association, should also be taken into consideration. Motivated by the novel computation cooperation framework, a possible research trend is how to design and implement user computation cooperation by removing the participation/control of network operators, e.g., in peer-to-peer networks and D2D communication networks.

Review of "Proactive Edge Computing in Latency-Constrained Fog Networks", in Proc. EuCNC, 2017

By M. S. Elbamby, M Bennis, and W. Saad

Objectives of the Paper

The main objective of the paper is to study the distribution and proactive caching of computing tasks in fog networks under latency and reliability constraints. Under such a scenario, users can execute their computation tasks either locally or via offloading to nearby cloudlets at the network edge, and these cloudlets are proposed to proactively cache popular tasks' computed results for further reducing the computing latency. Specially, the total computation latencies minimization under the tasks' reliability constraints is formulated as a nonconvex optimization problem and is approximately solved by decoupling into two separate and tractable subproblems: users' task distribution and popular task results caching.

Relevance to Feature Topic

Fog computing has been investigated to bring computing resources closer to the user devices in order to minimize the computing latency with joint communication and computing resource allocation schemes. In addition to abundant computation and storage resources, the advanced MEC/Fog systems would require proactive caching capacity for the cacheable computation results for multiple users, in order to further reduce computing latency. The caching for computation results requires the MEC/Fog servers to prefetch users' tasks based on learning and predicting users' behavior/environment. Therefore, the challenge here is how to model such complicated relationship between learning users' environment and prefetching tasks, as well as designing joint prefetching, caching, and offloading schemes. With the distributed characteristics of the fog computing, users can attain personalized services of computing. In addition, the fog network can exploit the benefit of proactive network, which has been extensively studied in the wireless content caching, to follow the popularity patterns of user tasks and properly cache the computing results of popular tasks.

Major Contributions

The main contribution of this paper is to consider the unified design of the edge computing and proactive caching in fog networks, by jointly exploiting the computation offloading and proactive caching of popular and cacheable tasks. For a tractable treatment of the formulated optimization problem, the Markov's inequality is applied to approximate the probabilistic constraints. Further, the considered problem is further decoupled into a task distribution subproblem and a popular task caching subproblem. Efficient algorithms are developed to solve the two subproblems. Numerical results are provided to show the effectiveness of the proposed scheme.

Novelty

The system model for cache-enable Fog networks is interesting, where the edge computing and the proactive caching are jointly considered. Depending on the task computation results are cacheable or not, a popular computation results caching scheme is proposed.

Key Results

The paper aims to minimize the computation latency by joint optimizing users' task distribution, offloading, and proactively caching computation results. A low-complexity algorithm is developed to solve the challenging problem, where a clustering scheme is introduced to group users into disjoint sets and then the task distribution is solved by a one-to-one matching game. Simulation results show substantial performance gains of the proposed scheme in terms of the average computing delay, as compared to the baseline scheme without considering the caching capability and that without considering the latency/reliability constraints.

Outlook

For this new type of cache-enabled MEC/Fog systems, a joint design of computing and caching is crucial. To this end, the proposed joint offloading and proactive computing method opens a new path to decrease the latency and meet the stringent requirements on computation and communication, by jointly exploiting the computation and storage resources of the fog network. In the future work, it would be interesting to further explore such a joint design for the cache-enabled MEC/fog systems by considering the user mobility and task partitioning, as well as the coordination of cloud and fog. This new design is anticipated to offer

remarkable opportunities for a wide range of ultra-reliable and ultra-low-latency applications such as autonomous driving, unmanned aerial vehicles (UAVs), and virtual reality (VR).

Interview with Dr. Kaibin Huang

Department of Electrical & Electronic Engineering, The University of Hong Kong
Email: huangkb@eee.hku.hk

How do you define future mobile edge computing networks, and what will be the key business cases and application scenarios for such networks? (JX)

KH: Mobile edge computing (MEC) networks are an emerging type of mobile networks that aim at enabling ubiquitous cloud-computing at the edge of cellular networks, thereby enhancing the computation capacities of mobile devices and reducing their energy consumption. By pushing computing towards the edge, MEC can scavenge the available computation resources at the edge (e.g., idling base stations, access points and personal computers). At the same time, it also resolves the problem of core-network congestion faced by traditional cloud computing. Furthermore, due to its proximity to users, MEC features low-latency, high bandwidth and location/context awareness. The above advantages of MEC networks make them a cost-effective and enabling platform for implementing a wide range of 5G applications including VR/AR, auto-pilot, smart home and cities, industrial automation, eBanking, video stream analysis, and multimedia content delivery. Due to its impact on next-generation wireless applications, MEC platforms are being standardized by European Telecommunications Standards Institute (ETSI) and will likely be included by the 3rd Generation Partnership Project (3GPP) as part of the 5G standard.

What are the most essential challenges faced in mobile edge computing? (JX)

KH: The computer-science aspect of MEC is not entirely new. Many topics, such as computation offloading, parallel computing and live migration, are in fact classic topics in the area of mobile computing. However, in computer science, wireless links are typically abstracted as "bit pipes". Due to this rather coarse abstraction, existing research on mobile computing has largely overlooked most of the advanced techniques in wireless communication such as MIMO, OFDMA, coding, and adaptive transmission. Existing computing technology thus designed is far from being able to tackle some key challenges faced by 5G, namely ultra-low latency, massive access, and gigabit data rates. In my view, the essential challenges faced in mobile edge computing share the theme of how to meet the 5G requirements. To be specific, the challenges of designing MEC techniques lie in minimizing mobile energy consumption and provide reliable-and-fast edge-computing even under hostile conditions of high mobility and unreliable edge servers/helpers. Tackling the challenges requires the seamless integration of wireless communication and mobile-and-cloud computing as elaborated in the sequel. From the perspective of network design, the challenges are to provide an efficient platform for large-scale deployment of MEC services. Specific research issues include MEC network architecture and deployment, network function virtualization, software defined networks, network slicing, and geographical load balancing.

Could you please explain the most pertinent techniques for mobile edge computing? (JX)

KH: MEC is a cross-disciplinary area lying at the boundary between two classic areas: one is wireless communication and the other computer science. There exist an extremely rich literature on wireless communication techniques that can be grouped into physical layer, radios resource management, multi-access control and network designs. On the other hand, existing mobile-computing techniques have addressed diversified issues including network function virtualization, computation offloading, cloud computing, program partitioning, parallel computing, and live migration. A typical approach for designing MEC would involve the joint design of several techniques from the two areas which otherwise may appear to be disconnected. This has led to the emergence of many new techniques targeting MEC. Among them, perhaps one of the most pertinent techniques for MEC is computing offloading, referring to the offloading of computation-intensive tasks from mobiles to edge severs. A focus of relevant research is the joint allocation of radio-and-computing resources for multiple offloading users. Optimizing the resource allocation is rather complex as it involves the interplay of adaptive transmission, scheduling, parallel computing, and program partitioning. There exist many other MEC techniques covering issues such as handover with

live migration, green MEC (energy harvesting, wireless power, and dynamic right-sizing), peer-to-peer offloading, data prefetching, just to name a few.

Could you please briefly introduce the most recent research project that you have done in mobile edge computing? (JX)

KH: My group has completed several projects covering the topics of multi-user computation offloading, wirelessly powered MEC, peer-to-peer cooperative MEC, predictive data prefetching, and large-scale MEC networks. The most recent MEC project we have done focuses on developing "MIMO over-the-air computation" techniques for sensor networks. Instead of decoupling data transmission and computing in the traditional approach, our techniques exploit the superposition nature of multi-access channels to compute multiple functions of distributed sensor data "in the air". This leads to dramatic reduction of multi-access latency in a dense IoT network.

Beyond your own work, are there any resources that you would like to recommend, especially to those who are new in the field and want to learn more about the mobile edge computing? (JX)

KH: Recently, many tutorial and overview papers have been published on MEC addressing difference aspects of the broad area. Several collaborators and I have written a survey paper entitled "A survey on mobile edge computing: The communication perspective" summarizing recent advancements in wireless communication techniques for MEC. Another paper entitled "On multi-access edge computing: A survey of the emerging 5G network edge cloud architecture and orchestration" coauthored by Taleb Tarik et al. focuses on 5G network architectures for implementing MEC. Moreover, numerous FAQs regarding fog computing (similar to MEC) are answered in a precisely written magazine article "Clarifying fog computing and networking: 10 questions and answers" by Mung Chiang et al. Due to limitation of space, I cannot list all the tutorial papers on MEC but most of them can be found by a keyword search in IEEExplore. These materials provide a good starting point for newcomers who are interested in knowing more about the area or even pursue some relevant research.

What, in your opinion, are the most important open problems and future research directions in mobile edge computing? (JX)

KH: Designing MEC techniques is highly complex as it involves both wireless communication and mobile computing. For the sake of tractability, simplified mathematical models are commonly adopted in MEC research. For example, computation tasks are assumed to have either a sequential or a parallel topologies and the computation load is measured by a fixed number of CPU cycles per bit. However, practical applications are complex and their tasks may not have such simple structures. In my personal view, the most important open problems and most promising research directions are those concerning how to transform the MEC theory and techniques into practice by considering specific applications such as AR and online gaming. Among others, one direction particularly interesting to me is how to design MEC techniques for enabling machine learning at the edge, or called "an intelligent edge", which is being actively pursued by leading Internet companies such as Microsoft and Google. Another important research direction is security and privacy. The recent Facebook data crisis has raised public concerns about privacy and prompted governments all over the world to start defining their data policies. This will inevitably make security and privacy for mobile data offloading in MEC an increasingly important research topic.

Kaibin Huang (M'08–SM'13) received the B.Eng. degree (Hons.) and the M.Eng. degree from the National University of Singapore, and the Ph.D. degree from The University of Texas at

Austin (UT Austin), Austin, TX, USA, all in electrical engineering. Since 2014, he has been an Assistant Professor with the Department of Electrical and Electronic Engineering (EEE), The University of Hong Kong. He used to be a Faculty Member with the Department of EEE, Yonsei University, South Korea, where he is currently an Adjunct Professor. His research interests focus on the analysis and design of wireless networks using stochastic geometry and multi-antenna techniques. Dr. Huang is an elected member of the SPCOM Technical Committee of the IEEE Signal Processing Society. He serves on the technical program committees of the major IEEE conferences in wireless communications. He has been the Technical Chair/Co-Chair for the IEEE CTW 2013, the Wireless Communications Symposium of IEEE GLOBECOM 2017, the Communication Theory Symposium of IEEE GLOBECOM 2014, and the Advanced Topics in Wireless Communications Symposium of IEEE/CIC ICCC 2014 and has been the Track Chair/Co-Chair for the IEEE PIMRC 2015, IEE VTC Spring 2013, Asilomar 2011, and the IEEE WCNC 2011. He was a recipient of the 2015 IEEE ComSoc Asia Pacific Outstanding Paper Award, Outstanding Teaching Award from Yonsei, Motorola Partnerships in Research Grant, the University Continuing Fellowship from UT Austin, and a Best Paper Award from IEEE GLOBECOM 2006. He is currently an Editor for the IEEE Journal on Selected Areas in Communications Series on Green Communications and Networking, the IEEE Transactions on Wireless Communications, and the IEEE Wireless Communications Letters. He was also a Guest Editor for the IEEE Journal on Selected Areas in Communications Special Issues on Communications Powered by Energy Harvesting and an Editor for the IEEE/KICS Journal of Communication and Networks from 2009 to 2015.

Interview with Dr. Yang Yang

Shanghai Research Center for Wireless Communications (WiCO)
CAS Key Lab of Wireless Sensor Network and Communication
Shanghai Institute of Microsystem and Information Technology
Chinese Academy of Sciences, China
Shanghai Institute of Fog Computing Technology (SHIFT)
Email: yang.yang@wico.sh

How do you define future mobile edge computing networks, and what will be the key business cases and application scenarios for such networks? (JX)

YY: The future mobile edge computing network will be an intelligent computing paradigm which autonomously harvests the vast amount of the idle computation power and storage space distributed at the network edges to yield sufficient capacities for performing computation-intensive and latency-critical tasks at mobile devices, with significantly improved the energy efficiency and user experience. Key business cases and application scenarios will include large-scale IoT networks, drone networks and UAV networks.

What are the most essential challenges faced in mobile edge computing? (JX)

YY: First of all, the key challenges include edge resource discovery, pooling and cooperation that facilitate efficient 3C (computing, communication, caching) resource sharing among multiple users at the network edge. This is challenging due to the huge scale of edge devices, limited resource of edge devices, as well as the joint bottlenecks on both computation and communication. Also, the service handover among multi-edge environment is challenging. Considering the ubiquitous mobility or user devices, the handover between multiple edges which follows the user mobility and thus to maintain low-latency is crucial. Last but not the least, in the era of AI and big data, how to support real-time and energy-efficient computation-intensive intelligent applications at the resource-constrained edge servers can also be a challenging but vital problem.

Could you please explain the most pertinent techniques for mobile edge computing? (JX)

YY: The first one should be task offloading/migration, which offloads the data- or computing intensive tasks from the resource and energy constrained devices to the nearby but more powerful edge nodes/devices for distributed and collaborative task execution. The second one can be lightweight 3C resource virtualization, in order to enable efficient edge resource sharing and pooling across a multitude of collaborative edge nodes and devices.

Could you please briefly introduce the most recent research project that you have done in mobile edge computing? (JX)

YY: My research group is mainly focused on the task scheduling problem in fog computing networks, with mobile edge computing as a special application case. Our recent publications include:

- Y. Yang, S. Zhao, W. Zhang, Y. Chen, X. Luo, and J. Wang, "DEBTS: Delay Energy Balanced Task Scheduling in Homogeneous Fog Networks," IEEE Internet of Things Journal, in print, Mar. 2018.
- Y. Yang, K. L. Wang, G. W. Zhang, X. Chen, X. Luo, and M. T. Zhou, "MEETS: Maximal Energy Efficient Task Scheduling in Homogeneous Fog Networks," IEEE Internet of Things Journal, in print, May 2018.
- N. Chen, Y. Yang, T. Zhang, M. Zhou, X. Luo, and J. Zao, "FA2ST: Fog As A Service Technology," in submission.
- S. Zhao, Y. Yang, Z. Shao, X. Yang, H. Qian, and C. X. Wang, "FEMOS: Fog-Enabled Multi-tier Operations Scheduling in Dynamic Wireless Networks," IEEE Internet of Things Journal, vol. 5, no. 2, pp. 1169-1183, Apr. 2018.

Beyond your own work, are there any resources that you would like to recommend, especially to those who are new in the field and want to learn more about the mobile edge computing? (JX)

YY: Prof. Xu Chen's research group has recently conducted a set of works on mobile edge

computing. Specifically, (1) on edge resource pooling and sharing, they have developed some resource scheduling frameworks to boost resource cooperation among edge devices, from various perspectives such as social ties, crowdsourcing, incentive mechanism, cross-layer collaboration, etc. (2) On service orchestration among multiple edges, they have developed mobility-aware and online service placement framework which dynamically migrate the service profiles of users to follow the mobility of users, thereby improving the user QoE. (3) On edge intelligence, they have built a prototype named Edgent to facilitate real-time ML-based applications on mobile device, the key idea of Edgent is to collaboratively execute the deep neural network (DNN) with adaptively optimized model size, among the edge server and the mobile devices.

- X. Chen, Z. Zhou, W. Wu, D. Wu and J. Zhang, "Socially-Motivated Cooperative Mobile Edge Computing," accepted by IEEE Networks, Jan. 2018.
- L. Pu, X. Chen, J. Xu, X. Fu, "Crowd Foraging: A QoS-oriented Self-organized Mobile Crowdsourcing Framework over Opportunistic Networks," accepted by IEEE Journal on Selected Areas in Communications, Jan. 2017.
- L. Pu, X. Chen, J. Xu, X. Fu, "D2D Fogging: An Energy-efficient and Incentive-aware Task Offloading Framework via Network-assisted D2D Collaboration," IEEE Journal on Selected Areas in Communications, Vol. 34, No. 12, pp. 3887 – 3901, Dec. 2016.
- X. Chen, L. Pu, L. Gao, W. Wu, and D. Wu, "Exploiting Massive D2D Collaboration for Energy-Efficient Mobile Edge Computing," IEEE Wireless Communications, Vol. 24, No. 4, pp. 64 - 71, Aug. 2017.
- T. Ouyang, Z. Zhou, and X. Chen, "Follow Me at the Edge: Mobility-Aware Dynamic Service Placement for Mobile Edge Computing", IEEE/ACM IWQoS 2018.
- E. Li, Z. Zhou, and X. Chen, "Edge Intelligence: On-Demand Deep Learning Model Co-Inference with Device-Edge Synergy", ACM MECOMM 2018 (in conjunction with SIGCOMM 2018).

Here are some good tutorials, surveys and whitepapers on edge and fog computing.

- Mung Chiang, Sangtae Ha, Chih-Lin I, Fulvio Risso, and Tao Zhang: Clarifying Fog Computing and Networking: 10 Questions and Answers. IEEE Communications Magazine 55(4): 18-20 (2017).
- Weisong Shi, Jie Cao, Quan Zhang, Youhuizi Li and Lanyu Xu, Edge Computing: Vision and Challenges, IEEE Internet of Things Journal, Vol. 3, No. 5, October 2016, pp. 637-646.
- "Fog Network and Internet of Things (IoT) in Wireless 5G Environments", IEEE GLOBECOM 2015 Tutorials
- "Fog Computing and Networking: A New Paradigm for 5G and IoT Applications", IEEE ICC 2017 Tutorials
- "Fog as a Service Technology (FA2ST): a New Approach for the Development of 5G Applications", IEEE GLOBECOM 2017 Tutorials
- "Fog Services and Enabling Technologies", IEEE ICC 2018 Tutorials
- OpenFog Architecture Overview, OpenFog whitepaper.
- Mobile-Edge Computing, ETSI whitepaper.

What, in your opinion, are the most important open problems and future research directions in mobile edge computing? (JX)

YY: In my opinion, one of the critical issue is edge intelligence. That is, how to support the ubiquitous machine learning applications which are typically resource and energy-consuming on mobile devices which are typically resource and energy-constrained. Beyond edge intelligence, efficient system management and operation is also an open yet critical problem. In particular, since the limited edge resource is shared by a multitude amount of edge devices/users, how to virtualize the resource (e.g., via virtual machines, containers, or others?) and how to monitor/diagnose/eliminate performance interference among devices/users are worthy of study.

Dr. Yang Yang is now a professor with Shanghai Institute of Microsystem and Information Technology (SIMIT), Chinese Academy of Sciences, serving as the Director of CAS Key Laboratory of Wireless Sensor Network and Communication, and the Director of Shanghai Research Center for Wireless Communications (WiCO). He is also a Distinguished Adjunct Professor with the School of Information Science and Technology, ShanghaiTech University, serving as a Co-Director of Shanghai Institute of Fog Computing Technology (SHIFT). Prior to these, he has held various faculty positions at the Chinese University of Hong Kong (CUHK), Brunel University (UK), and University College London (UCL, UK). He received the BEng and MEng degrees from Southeast University, China, in 1996 and 1999, respectively; and the PhD degree from the Chinese University of Hong Kong in 2002.

Yang is a member of the Chief Technical Committee of the National Science and Technology Major Project "New Generation Mobile Wireless Broadband Communication Networks" (2008-2020), which is funded by the Ministry of Industry and Information Technology (MIIT) of China. In addition, he is on the Chief Technical Committee for the National 863 Hi-Tech R&D Program "5G System R&D Major Projects", which is funded by the Ministry of Science and Technology (MOST) of China. Since January 2017, he has been elected as the Director for Greater China Region of the OpenFog Consortium.

Yang's current research interests include wireless sensor networks, Internet of Things, Fog computing, Open 5G, and advanced wireless testbeds. He has published more than 150 papers and filed over 80 technical patents in wireless communications.

TCCN Newsletter Editorial Board

TCCN Officers