Cross-layer Design for Mission-Critical IoT in Mobile Edge Computing Systems

Changyang She *Member, IEEE,* Yifan Duan, Guodong Zhao *Senior Member, IEEE,*
Tony Q. S. Quek *Fellow, IEEE,* Yonghui Li *Fellow, IEEE,* and Branka Vucetic *Fellow, IEEE*

Abstract—**In this work, we propose a cross-layer framework for optimizing user association, packet offloading rates, and bandwidth allocation for Mission-Critical Internet-of-Things (MC-IoT) services with short packets in Mobile Edge Computing (MEC) systems, where enhanced Mobile BroadBand (eMBB) services with long packets are considered as background services. To reduce communication delay, the 5th generation new radio is adopted in radio access networks. To avoid long queueing delay for short packets from MC-IoT, Processor-Sharing (PS) servers are deployed at MEC systems, where the service rate of the server is equally allocated to all the packets in the buffer. We derive the distribution of latency experienced by short packets in closed-form, and minimize the overall packet loss probability subject to the end-to-end delay requirement. To solve the non-convex optimization problem, we propose an algorithm that converges to a near optimal solution when the throughput of eMBB services is much higher than MC-IoT services, and extend it into more general scenarios. Furthermore, we derive the optimal solutions in two asymptotic cases: communication or computing is the bottleneck of reliability. Simulation and numerical results validate our analysis and show that the PS server outperforms first-come-first-serve servers.**

Index Terms—**Mission-critical internet-of-things, mobile edge computing, 5G new radio, processor-sharing server, cross-layer optimization**

I. Introduction

Mission-Critical Internet-of-Things (MC-IoT) will be widely deployed in future wireless networks for remote health monitoring, haptic interaction, and factory automation [2, 3]. Achieving ultra-reliable and low-latency communications (URLLC) (e.g., 10^{-7} packet loss probability and 1 ms End-to-End (E2E) delay) for MC-IoT has been considered as one of the major goals in 5th Generation (5G) cellular networks [4]. Most existing technologies mainly focus on one of the seven layers of the open systems interconnection model, and cannot guarantee the E2E delay [5]. To satisfy the requirements of MC-IoT, we need to re-design the physical-layer resource management, the link-layer scheduling policy, and the network-layer user association from a cross-layer perspective.

One of the major differences between MC-IoT and enhanced Mobile BroadBand (eMBB) services lies in the sizes of packets. With high data rate, the packet size in eMBB services is relatively large, e.g., thousands of bytes in each packet. However, the packets generated by MC-IoT are very small, e.g., 20 or 32 bytes in each packet [6]. To achieve low latency for short packet transmissions, a short frame structure is adopted in 5G New Radio (NR) [7]. When transmitting a short packet in a short frame, the blocklength of channel codes is very limited. As a result, the decoding error probability cannot be ignored when analyzing reliability [8].

On the other hand, the computing ability at each MC-IoT device is limited. To reduce processing delay, MC-IoT devices will offload some of the packets to the Mobile Edge Computing (MEC) systems for processing [9,10]. Considering that MC-IoT services will co-exist with eMBB services, a short packet arriving at the MEC after long packets has to wait in a queue if the packets are processed with a First-Come-First-Serve (FCFS) order. To avoid long queueing delay, other scheduling orders at MEC systems should be considered.

Furthermore, the reliability and delay not only depend on the resource management and scheduling order but also depend on the traffic load. Considering that the radio resources and computing capacity at each Access Point (AP) are limited, the user association and offloading policy should be optimized to balance traffic loads. The problems for optimizing user association and offloading policy are NP-hard in general [11]. Low-complexity solutions to the NP-hard problems are in urgent need for MC-IoT since complicated searching algorithms will lead to long computation delay [12].

A. Related Works

To transmit short packets with low latency, the blocklength of channel codes is short. In the short blocklength regime, Shannon's capacity is not applicable since it cannot characterize the decoding error probability [13]. Recently, the maximal achievable rate with given decoding error probability in the short blocklength regime was obtained in multi-antenna quasi-static channel [14]. How to design transmission schemes and resource allocation in the short blocklength regime has

Manuscript received Feb. 28, 2019; revised May 1, 2019 and Jun. 22, 2019; accepted Jul. 20, 2019. This work was supported in part by the SUTD-ZJU Research Collaboration under Grant SUTD- ZJU/RES/01/2016, and in part by Australian Research Council Laureate Fellowship grant FL10100032. The work of Y. Li was supported by Australian Research Council under Grants DP150104019 and DP190101988. *(Corresponding author: Changyang She.)*

Part of this paper was presented at the International Conference on Wireless Communications and Signal Processing 2018 [1].

C. She, Y. Li and B. Vucetic are with the School of Electrical and Information Engineering, University of Sydney, Sydney, NSW 2006, Australia (email:shechangyang@gmail.com, {yonghui.li,branka.vucetic}@sydney.edu.au).

Y. Duan is with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: duanyifan@std.uestc.edu.cn).

G. Zhao is with School of Engineering, University of Glasgow, Glasgow, G12 8LT, UK. (e-mail: Guodong.Zhao@glasgow.ac.uk).

T. Q. S. Quek is with the Information Systems Technology and Design Pillar, Singapore University of Technology and Design, 8 Somapah Road, Singapore 487372 (e-mail: tonyquek@sutd.edu.sg).

been studied in existing literature [15–21]. The throughput achieved in cognitive radio channels and relay systems was studied in [15] and [16, 17], respectively. The studies in [18] optimized the scheduling of short packets to maximize energy efficiency. The authors of [19] optimized packet losses caused by decoding errors, queueing delay violation, and packet dropping over deep fading channel subject to the ultra-high reliability. Considering that the feedback of Channel State Information (CSI) leads to extra delay, the studies in [20] jointly optimized Uplink (UL) and Downlink (DL) resource configurations without CSI at the transmitters. More recently, how to optimize resource allocation among multiple users with different packet arrival processes was studied in [21].

Scheduling policies in computing systems have significant impacts on the Quality-of-Service (QoS) of MC-IoT. A near-optimal policy to minimize the average latency of short packet is the Shortest Remaining Processing Time (SRPT) first scheduler. Such a scheduler is hard to implement in practice since the remaining processing time is not available at the server, and it requires too many priority levels [22]. To reduce the latency of short packets without introducing priority levels, the Processor-Sharing (PS) server is a possible solution, where the total service rate is equally allocated to all the packets in the server [23]. Although the distribution of latency was derived in the large delay regime in the PS server [24], the latency experienced by short packets remains unclear. To derive the delay bound violation probability of URLLC services, martingales-based analysis, effective capacity, and network calculus were used in [25], [16], and [26], respectively. But all the results were obtained in the FCFS servers. Note that it is very challenging to derive the closed-form expression of the distribution of delay, how to formulate the constraints on delay and reliability of MC-IoT is still unclear.

Promising network architectures for MC-IoT were studied in [27–31]. A comprehensive overview on MC-IoT of industrial scenarios was carried out in [27], where the issues related to architecture design were discussed, such as extensibility, scalability, and modularity. To reduce the routing delay, an adaptive transmission architecture with software-defined networks and edge computing was proposed in [28]. Considering that energy consumption of IoT devices is an important issue, an energy-aware real-time routing scheme was proposed in [28] to reduce energy consumption and E2E delay in large-scale IoT networks. More recently, a fog computing architecture was proposed for 5G tactile IoT [30], where the quality-of-experience-aware model was formulated. By combining stochastic geometry and queueing theory, different QoS requirements were analyzed in ultra-dense networks [31]. The studies in [27–31] shed light upon network architecture designs for MC-IoT, but decoding errors in the physical-layer were not considered.

Computing offloading has been exhaustively studied in various MEC systems, such as wireless powered MEC [32], ultra-dense IoT networks [33, 34], and fiber-wireless networks [35, 36]. Although these works did not consider MC-IoT, they developed useful methodologies for optimizing computing offloading in MEC systems. How to improve the QoS in MEC systems by optimizing task offloading has been addressed in [37–40]. In [37], the average delay was minimized by optimizing task offloading/scheduling in MEC systems. Considering that the average delay is not suitable for URLLC services, the authors of [38] optimized task offloading and resource allocation under the constraint on a queue length violation probability. The offloading schemes for URLLC in MEC systems were optimized in [39], where a weighted sum of E2E delay and the offloading failure probability was minimized in a single-user scenario. How to analyze latency in large-scale MEC networks was studied in [40], where the average communication and computing latencies were derived.

Most of the existing studies on resource management in MEC systems only analyzed UL transmission and processing delay, and assumed DL transmission can be finished with high transmit power at the APs [37–40]. Besides, they did not take the decoding errors in the short blocklength regime into account, which is crucial for MC-IoT. Although the block error probability was considered in the optimization problem in [39], the radio resource management was not optimized and the packet losses due to the delay bound violation were not considered.

B. Our Contributions

To the best of the authors' knowledge, there is no resource allocation scheme or offloading scheme that can achieve the target E2E delay and overall packet loss probability for MC-IoT in MEC systems. Moreover, how to design scheduling policy and whether the FCFS server is suitable for MC-IoT were not addressed. In order to achieve ultra-high reliability and ultra-low E2E delay for MC-IoT in MEC systems, the following three issues will be addressed in this work: 1) *How to design scheduling and queueing policies in MEC servers and local servers to achieve ultra-high reliability and ultra-low E2E delay for short packets?* 2) *How to characterize the statistical QoS of short packets when there are both short and long packets in MEC systems?* 3) *How to improve the fundamental tradeoff between delay and reliability by optimizing user association, packets offloading, and bandwidth allocation in MEC systems?* Our major contributions are summarized as follows,

- We establish a framework for minimizing overall packet loss probability subject to E2E delay requirement in MEC systems, where processing delay and UL and DL transmission delays are taken into account. PS servers are equipped at MEC systems, where the total service rate is equally allocated to all the packets in each server, and every packet receives service at all times. As such, short packets can bypass long packets and achieve low latency.
- We derive a closed-form approximation of the Complementary Cumulative Distribution Function (CCDF) of the latency experienced by short packets in the PS server. The approximation is accurate when the number of Central Processing Unit (CPU) cycles needed to process a long packet is much larger than that needed to process a short packet.
- We optimize the user association scheme, packet offloading rates, and bandwidth allocation for short packets

in MEC systems. We propose an algorithm to solve a mixed integer problem and analyze the convergence and the complexity of it. Our analysis shows that the difference between the obtained solution and the global optimal solution only results from searching numbers of subcarriers in a continuous domain.

Furthermore, our simulation results validate the accuracy of the closed-form approximation. Numerical results indicate that when increasing the number of antennas at the AP with a fixed processing capacity, communication is the bottleneck of the reliability when the number of antennas is small, and computing is the bottleneck when the number of antennas is large. Only in a very small region (e.g., from 16 to 18 antennas), the packet loss probability in communications is comparable to the processing delay violation probability. This implies our algorithm converges to a near optimal solution in most of the cases.

The rest of the paper is organized as follows. Section II describes the system model. Section III analyzes delay and reliability. Section IV studies how to optimize the association scheme, packet offloading rates and bandwidth allocation. Section V extends the algorithm into more general scenarios. Section VI provides simulation and numerical results. Section VII concludes the paper.

II. System Model

A. The MEC System

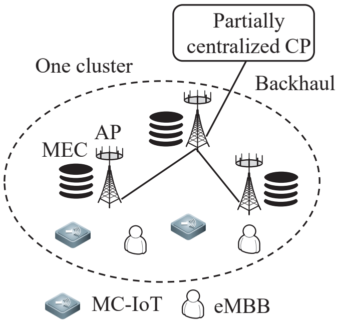


Fig. 1. System model.

As illustrated in Fig. 1, we consider a MEC system with single-antenna devices and multi-antenna APs, where the data collected by each MC-IoT device and the computation intensive tasks generated by eMBB services can be offloaded to one of the APs for processing. To provide better services in radio access networks and to avoid high backhaul overhead, the partially centralized Control Plane (CP) in [12] is considered. The whole network is decomposed into multiple clusters, each of which includes K closely located APs and one CP that optimizes user association scheme, packet offloading rates, and bandwidth allocation for M devices in the cluster. To achieve ultra-low latency and ultra-high reliability, strong co-channel interference should be avoided. To this end, orthogonal channels are allocated to different devices in each cluster, and the frequency reuse factor is less than one such that adjacent clusters use different bandwidth. In this work, we focus on one cluster of APs, and our solution is applicable for low mobility scenarios like factory automation and VR/AR applications. For high mobility scenarios, where devices travel across clusters frequently, how to reserve resources in different clusters deserves further study.

B. Traffic Models

In vehicle networks and factory automation, there are two kinds of packets, i.e., periodic packets with deterministic arrivals and sporadic packets that are driven by some random events [41, 42]. Since analyzing the delay and the reliability of random packet arrival processes is more challenging than deterministic arrivals, we focus on sporadic packets in this work. The experiment in [43] indicates that the packet arrival processes of MC-IoT are very bursty, i.e., there is a high traffic state and a low traffic state. For each of the traffic states, the arrival process can be modeled as a Bernoulli process. According to 5G NR, time is discretized into slots with duration T_{s}. In each slot, a device either has a packet to transmit or stays silent. We assume the traffic state is obtained with the traffic state classification methods in [43]. When a device switches between the high and low traffic states, we only need to change the average arrival rate in our analysis. The aggregation of multiple independent Bernoulli processes at a MEC server can be accurately approximated by a Poisson process [44].

1) Short packets: The data collected by each MC-IoT device is contained in *short packets* for transmission and processing. A packet with the following three features is considered as "short",

- The number of bits in the packet is small. According to [6], the packet size in MC-IoT services is around 20 or 32 bytes. In contrast, the packets in eMBB services may include hundreds or thousands of bytes, such as video streaming.
- The blocklength of channel codes of the packet is short. For example, if quadrature phase-shift keying is used in modulation, the number of symbols required by a packet with 32 bytes (256 bits) is 128, which is the blocklength of the packet. To achieve low-latency, the blocklength of channel codes is short in MC-IoT [45].
- The number of CPU cycles required to process the packet is small. The number of CPU cycles required to process the packet depends on the number of bits in the packet and the processing algorithms. Since a packet in MC-IoT services only contains a few bits, the number of required CPU cycles is small.

Let c_{S} be the number of CPU cycles required to process a short packet. The service rate of the local server at the kth device and that of the mth MEC server are denoted as C_k and S_m (CPU cycles/slot), respectively.

2) Long packets: There are some devices requesting eMBB services in each cluster. The tasks generated by the eMBB services are packetized into *long packets*. How to optimize resource allocation and computing offloading for eMBB services has been studied in the existing literature, such as [46–48]. In our work, we focus on MC-IoT services, where eMBB services are considered as background services.

The sum of average packet arrival rates of eMBB services at the mth AP is denoted as λ_m^{L} (packets/slot). The number of CPU cycles required to process a long packet is denoted as c_{L}, which is a random variable with mean value $\bar{c}_{\mathrm{L}}$. In this work, we do not specify the distribution of c_{L}. The only assumption on c_{L} is that $c_{\mathrm{L}} \gg c_{\mathrm{S}}$, which is reasonable since the packet size of eMBB services is much larger than that of MC-IoT services and the algorithm for processing long packets (e.g., high definition pictures) is more complex than that for processing short packets (e.g., the location and velocity of a device). For example, c_{L} may follow the Pareto distribution with a heavy tail [23], i.e.,

$$\Pr\left\{\frac{c_{\mathrm{L}}}{c_{\mathrm{S}}} > x\right\} = p_{\mathrm{A}} x^{-v}, \tag{1}$$

where $1 < v < 2$, $p_{\mathrm{A}} = (c_0/c_{\mathrm{S}})^v$, and c_0 is the minimum of c_{L}. As shown in [23] and the references therein, Pareto distributions have been observed in different application scenarios, such as the service time of UNIX jobs.

C. User Association and Packet Offloading

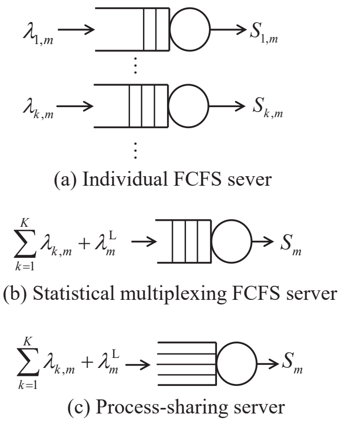
1) User association: Each device can associate with one of the APs. We leverage indicators $x_{k,m}, k = 1, ..., K, m = 1, ..., M$, to represent the user association scheme,

$$x_{k,m} = \begin{cases} 1, & \text{if device } k \text{ is associated with AP } m, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The association scheme of the kth device is denoted as $\mathbf{x}_k = [x_{k,1}, ..., x_{k,M}]^T$.

2) Packet offloading: When a short packet is generated by the kth device, the device either processes the packet with the local server or uploads the packet to an AP. The average packet rate from the kth device to the mth MEC server is denoted as $\lambda_{k,m}$ (packets/slot), where $k = 1, 2, ..., K$, and $m = 0, 1, ..., M$. $m = 0$ means that the packets are processed at the local server. If $x_{k,m'} = 1$, then $\lambda_{k,0} + \lambda_{k,m'} = \lambda_k^{\mathrm{U}}$, where λ_k^{U} is the average packet arrival rate of the kth device. If $x_{k,m} = 0$ for all $m = 1, ..., M$, then $\lambda_{k,0} = \lambda_k^{\mathrm{U}}$.

D. Queueing Models and Scheduling Policies



$\lambda_{1,m}$
$S_{1,m}$
$\lambda_{k,m}$
$S_{k,m}$
$\sum_{k=1}^{K}\lambda_{k,m}+\lambda_m^{\mathrm{L}}$
S_m
$\sum_{k=1}^{K}\lambda_{k,m}+\lambda_m^{\mathrm{L}}$
S_m

(a) Individual FCFS sever

(b) Statistical multiplexing FCFS server

(c) Process-sharing server

Fig. 2. MEC with different service orders.

As shown in Fig. 2(a), to guarantee the QoS requirements of different devices, packets from different devices are waiting in different queues at a MEC server, and are served according to the FCFS order. In the mth MEC server, the service rate allocated to the kth device is denoted as $S_{k,m}$. Once the computing resource is allocated to one device, it cannot be shared with the other devices. Such a scheduling scheme is widely used, but is not optimal in terms of minimizing the delay.

The second server in Fig. 2(b) is referred to as a statistical multiplexing FCFS server [23]. Due to statistical multiplexing gain, the average delay in the second server is much shorter than the first server when $S_m = \sum_k S_{k,m}$. Furthermore, as proved in [20], if the sizes of all the packets are identical, to achieve the same delay bound and delay bound violation probability, the required service rate in the statistical multiplexing server is less than the sum of the service rates in the individual server. However, when the distribution of the number of CPU cycles required to process the packets has a heavy-tail, some short packets arriving at the MEC server after a long packet need to wait for a long time. As a result, the delay requirement of MC-IoT services can hardly be satisfied.

The key to low latency is letting short packets bypass queued long packets. One possible solution is the PS server. As shown in Fig. 2(c), every packet in the server receives service at all times. When there are i packets in the mth MEC server, each packet is processed at rate S_m/i.

Remark 1. In practice, a server can be implemented in a time-sharing way, i.e., the service time in each slot is equally allocated to all the packets in the server. In this way, the processing delay of packets in the server is the same as that in the ideal PS server [23]. It's worth noting that there are some other possible scheduling policies if the server is aware of the diverse QoS requirements of different packets. In this work, we do not assume the computing system is aware of the types of packets in the communication systems. We will study more sophisticated scheduling policies for different types of packets in our future work.

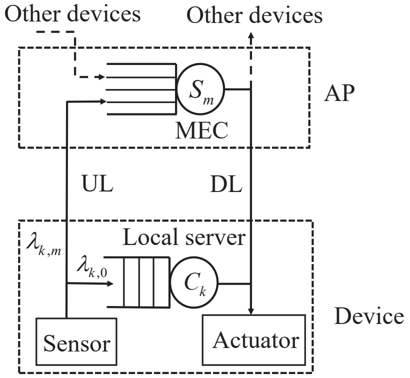


Fig. 3. Local and edge servers in our system.

We consider the scheduling policies at MEC servers and local servers in Fig. 3, where PS servers are deployed at APs and FCFS servers are deployed at devices. To avoid queueing delay, the UL and DL transmission durations of a short packet equal to one slot.[1] On the other hand, if the required CPU cycles to process different packets are identical, which is the case in the local server of each device, the FCFS server outperforms the PS server [23]. Therefore, FCFS servers are equipped at MC-IoT devices.

III. Analysis of Delay and Reliability

In this section, we study how to characterize E2E delay and overall packet loss probability. We first derive the CCDF of the processing delay of short packets in the PS server. Then, we show how to characterize the transmission delay and decoding error probability of short packets.

A. Processing Delay and Delay Violation Probability

A short packet can be processed either at the device or at the AP. The packet arrival process at the local server of the kth device is a Bernoulli process with average arrival rate $\lambda_{k,0}$. Denote the service time of a packet at the kth local server as $D_k^{\mathrm{loc}} = c_{\mathrm{S}}/C_k$. With the constant service rate at each local server, the queueing model is a Geo/D/1/FCFS model, where "Geo" means that the inter-arrival time between packets is geometric distributed, and "D" represents deterministic service processes. The CCDF of queueing delay in Geo/D/1/FCFS model has been obtained in [49]. If $i \le D_k^{\mathrm{loc}} - 1$, then

$$\Pr\{D_k^{\mathrm{q}} > i\} = 1 - (1-\lambda_{k,0})^{-i-1}\left(1 - \lambda_{k,0} D_k^{\mathrm{loc}}\right). \tag{3}$$

If $i \ge D_k^{\mathrm{loc}}$, the expression of $\Pr\{D_k^{\mathrm{q}} > i\}$ can be found in [49]. Due to the low-latency requirement, we are interested in the case $i \le D_k^{\mathrm{loc}} - 1$.

Each AP may serve multiple devices. The aggregation of multiple Bernoulli processes can be modeled as a Poisson process [40]. Thus, the MEC server can be characterized by an M/G/1/PS model, where "M" means the packet arrival process is Poisson process and "G" means that the number of CPU cycles required to process the packets can follow any distributions. To derive a closed-form CCDF of the processing delay of short packets in the M/G/1/PS model, we introduce an accurate approximation. Since the short packets are much smaller than the long packets, i.e., $c_{\mathrm{S}} \ll c_{\mathrm{L}}$, the processing delay of a short packet is much shorter than a long packet. As a result, the number of long packets in the server is nearly constant from the arrival to the departure of a short packet. When a short packet arrives at the server, the number of packets in the server is denoted as Q_m. Considering that the value of Q_m does not change significantly during the short service time of a short packet, then the service rate allocated to the short packet can be approximated by $S_m/(q+1)$ if $Q_m = q$. In this case, the processing delay of the short packet is approximated by

$$W_m^{\mathrm{S}}|_{Q_m=q} \approx \frac{c_{\mathrm{S}}(q+1)}{S_m} \quad \text{(slots)}. \tag{4}$$

According to [23], the distribution of Q_m can be expressed as follows,

$$\Pr\{Q_m = q\} = \rho_m^q(1-\rho_m), \tag{5}$$

which is the distribution of the number of packets in the M/G/1/PS model. The workload of the server is

$$\rho_m = \frac{\sum_{k=1}^{K} \lambda_{k,m} c_{\mathrm{S}} + \lambda_m^{\mathrm{L}} \bar{c}_{\mathrm{L}}}{S_m}. \tag{6}$$

From (4) and (5), we can further obtain that

$$\Pr\left\{W_m^{\mathrm{S}} = \frac{c_{\mathrm{S}}(q+1)}{S_m}\right\} \approx \Pr\{Q_m = q\} = \rho_m^q(1-\rho_m),$$

where $q = 0, 1, \ldots$. Based on the above expression, the CCDF of the processing delay of short packets in the PS server can be derived as follows,

$$\Pr\left\{W_m^{\mathrm{S}} > \frac{c_{\mathrm{S}}(q+1)}{S_m}\right\} \approx \Pr\{Q_m > q\} = \rho_m^q. \tag{7}$$

The approximation in (7) is accurate when $c_{\mathrm{S}} \ll c_{\mathrm{L}}$. We will validate the accuracy of the approximation via simulation.

The processing delay of short packets in the mth MEC server can be characterized by a delay bound and a delay bound violation probability, D_m^{mec} and $\varepsilon_m^{\mathrm{mec}}$. From the CCDF in (7), the relationship between $\varepsilon_m^{\mathrm{mec}}$ and $D_{\mathrm{m}}^{\mathrm{mec}}$ can be expressed as follows,

$$\varepsilon_m^{\mathrm{mec}} = \rho_m^{\left(\frac{S_m D_m^{\mathrm{mec}}}{c_{\mathrm{S}}} - 1\right)}. \tag{8}$$

B. Transmission Delay and Decoding Error Probability

If a packet is processed at the MEC server, the device first uploads the packet to the AP. After the MEC server finishes the processing, the result is sent back to the device. We introduce a superscript of parameters X^ξ, where $\xi \in \{\mathrm{u}, \mathrm{d}\}$. If $\xi = \mathrm{u}$, X is a parameter in UL transmissions. Otherwise, it is a parameter in DL transmissions. We consider Orthogonal Frequency Division Multiple Access (OFDMA) systems, which will be used to support MC-IoT services in 5G NR [7]. The total bandwidth is equally allocated to $N_{\max}$ subcarriers, each with a bandwidth of W_0. Denote the number of subcarriers allocated to the kth device for UL and DL transmissions as N_k^ξ, $\xi \in \{\mathrm{u}, \mathrm{d}\}$, respectively. Since the packet size is small, it is reasonable to assume that $N_k^\xi W_0$ is smaller than the coherence bandwidth. As mentioned in the previous section, to avoid queueing delay before UL and DL transmissions, the transmission duration of each packet is one slot, which is smaller than channel coherence time. Thus, each packet is transmitted over a flat fading quasi-static channel. Considering that feedback from receivers to transmitters may cause large overhead and extra delay, CSI is not available at the transmitters. According to [14], the achievable rate in the short blocklength regime over quasi-static flat fading channel can be accurately approximated by

$$R_{k,m}^\xi \approx \frac{N_k^\xi W_0}{\ln 2}\left[\ln\left(1 + \frac{\alpha_{k,m} g_{k,m}^\xi P_{\mathrm{s}}^\xi}{N_0 W_0}\right) - \sqrt{\frac{V_{k,m}^\xi}{T_{\mathrm{s}} N_k^\xi W_0}}\, f_{\mathrm{Q}}^{-1}\left(e_{k,m}^\xi\right)\right] \text{ bits/s}, \tag{9}$$

where $\alpha_{k,m}$ is the large-scale channel gain from the kth device to the mth AP, $g_{k,m}^\xi$ is the UL or DL small-scale channel

[1] Since the packet arrival rate of each device is less than one packet per slot, there is no queue before UL and DL transmissions.

fading between the kth device and the mth AP, $P_{\rm s}^{\xi}$ is the UL or DL transmit power of one antenna on each subcarrier, N_0 is the single-side noise spectral density, $f_{\rm Q}^{-1}(.)$ is the inverse of Q-function, e_k^{ξ} is the decoding error probability, and $V_{k,m}^{\xi} = 1 - 1\Big/\left(1 + \frac{\alpha_{k,m} g_{k,m}^{\xi} P_{\rm s}^{\xi}}{N_0 W_0}\right)^2$.

Remark 2. Due to the following two reasons, we only consider the flat fading channel, and do not consider frequency-selective channels. First, the maximal achievable rate over a frequency-selective channel in the short blocklength regime has not been derived in existing studies. Although the upper and lower bounds were obtained in [50], there is no closed-form expression. Second, as shown in [20], when the number of antennas is large (e.g., 16 antennas), frequency diversity is not necessary for URLLC. Therefore, we focus on the multi-antenna flat fading channel.

Let b_k^{ξ} be the number of bits in a short packet of the kth device. When sending a packet of b_k^{ξ} bits within one slot, the decoding error probability can be obtained from (9) by setting $T_{\rm s} R_{k,m}^{\xi} = b_k^{\xi}$. According to the law of total probability, the packet loss probability due to decoding errors can be expressed as follows [14],

$$\varepsilon_{k,m}^{\xi} \approx \mathbb{E}_{g_{k,m}^{\xi}}\{e_{k,m}^{\xi}\} = \mathbb{E}_{g_{k,m}^{\xi}}\left\{ f_{\rm Q}\left(\sqrt{\frac{T_{\rm s} N_k^{\xi} W_0}{V_k^{\xi}}} \times \left[\ln\left(1 + \frac{\alpha_{k,m} g_{k,m}^{\xi} P_{\rm s}^{\xi}}{N_0 W_0}\right) - \frac{b_k^{\xi} \ln 2}{T_{\rm s} N_k^{\xi} W_0}\right]\right)\right\}, \tag{10}$$

where the distribution of small-scale channel gain depends on the number of antennas at each AP, which is denoted as $N_{\rm t}$. To compute (10), we need to calculate the integral for a given distribution of $g_{k,m}^{\xi}$. For Rayleigh fading, we can apply the closed-form result in [51].

C. E2E Delay and Overall Packet Loss Probability

1) Delay and Reliability at Local Servers: If a packet is processed at the device, then the E2E delay is equal to the sum of the service time and the queueing delay at the local server of the device. Given the E2E delay requirement $D_{\max}$, the delay violation probability at local servers, $\varepsilon_k^{\rm loc}$, can be obtained by substituting $i = D_{\max} - D_k^{\rm loc}$ into (3). When $D_{\max} - D_k^{\rm loc} \le D_k^{\rm loc} - 1$,

$$\varepsilon_k^{\rm loc} = 1 - (1 - \lambda_{k,0})^{-(D_{\max} - D_k^{\rm loc}) - 1}\left(1 - \lambda_{k,0} D_k^{\rm loc}\right). \tag{11}$$

2) Delay and Reliability at the MEC server: If a packet is processed at a MEC server, then the UL and DL transmission delays and the processing delay in the server should be considered. The E2E delay can be satisfied under the following constraint,

$$2 + D_m^{\rm mec} \le D_{\max}, \tag{12}$$

where two slots are occupied by the UL and DL transmissions.

Due to decoding errors and processing delay violations, the overall packet loss probability can be expressed as follows,

$$\begin{aligned}\varepsilon_{k,m}^{\rm A} &= 1 - (1 - \varepsilon_{k,m}^{\rm u})(1 - \varepsilon_{k,m}^{\rm d})(1 - \varepsilon_m^{\rm mec}) \\ &\approx \varepsilon_{k,m}^{\rm u} + \varepsilon_{k,m}^{\rm d} + \varepsilon_m^{\rm mec},\end{aligned} \tag{13}$$

where the approximation is very accurate since $\varepsilon_{k,m}^{\rm u}$, $\varepsilon_{k,m}^{\rm d}$, and $\varepsilon_m^{\rm mec}$ are extremely small in MC-IoT. Upon substituting (6) and (12) into (8), we can get the expression of $\varepsilon_m^{\rm mec}$, i.e.,

$$\varepsilon_m^{\rm mec} = \left(\frac{\sum_{k=1}^{K} \lambda_{k,m} c_{\rm S} + \lambda_m^{\rm L} \bar{c}_{\rm L}}{S_m}\right)^{\frac{S_m (D_{\max} - 2)}{c_{\rm S}} - 1}. \tag{14}$$

Remark 3. The factors that lead to packet losses or errors depend on network architectures [52]. For the considered MEC system, reliability only includes queueing delay violations in computing systems and packet losses in radio access networks.

IV. Cross-Layer Optimization

In this section, we study how to optimize the association scheme, packet offloading rates, and UL and DL bandwidth allocation to minimize the overall packet loss probability subject to the E2E delay requirement.

A. Problem Formulation

Note that the packet loss probabilities at the local server and the MEC can be different, the reliability is determined by the worse one. Thus, the packet loss probability of the kth device is characterized by

$$\begin{aligned} & f_k(\mathbf{x}_k, \lambda_{k,m}, N_k^{\rm u}, N_k^{\rm d}) \\ &= \max\left[\varepsilon_k^{\rm loc}, x_{k,m}(\varepsilon_{k,m}^{\rm u} + \varepsilon_{k,m}^{\rm d} + \varepsilon_m^{\rm mec}), \forall m\right]. \end{aligned} \tag{15}$$

The problem that minimizes the maximal packet loss probability experienced by the K devices can be formulated as follows,

$$\min_{\substack{\mathbf{x}_k, \lambda_{k,m}, N_k^{\rm u}, N_k^{\rm d} \\ k=1,...,K}} \; \max_{k=1,...,K} f_k(\mathbf{x}_k, \lambda_{k,m}, N_k^{\rm u}, N_k^{\rm d}) \tag{16}$$

$$\text{s.t.} \; \sum_{m=1}^{M} x_{k,m} \le 1, x_{k,m} \in \{0, 1\}, \tag{16a}$$

$$0 \le \lambda_{k,m} \le x_{k,m}, \tag{16b}$$

$$\sum_{m=1}^{M} \lambda_{k,m} + \lambda_{k,0} = \lambda_k^{\rm U}, \tag{16c}$$

$$\sum_{k=1}^{K} N_k^{\rm u} + \sum_{k=1}^{K} N_k^{\rm d} \le N_{\max}, N_k^{\rm u}, N_k^{\rm d} \in \{1, 2, ..., N_{\rm c}\}, \tag{16d}$$

$$\max\{\varepsilon_k^{\rm loc}, x_{k,m}(\varepsilon_{k,m}^{\rm u} + \varepsilon_{k,m}^{\rm d} + \varepsilon_m^{\rm mec}), \forall m\} \le 1, \tag{16e}$$

where $k = 1, ..., K$, $m = 1, ..., M$, $N_{\rm c}$ is the maximum number of subcarriers that can be allocated to a device without exceeding the coherence bandwidth, and the expressions of $\varepsilon_{k,m}^{\xi}$, $\varepsilon_k^{\rm loc}$, and $\varepsilon_m^{\rm mec}$ can be found in (10), (11), and (14), respectively. Constraint (16a) guarantees that a device can only associate with one AP. If $\sum_{m=1}^{M} x_{k,m} = 0$, then all the packets are processed at the local server.

With constraint (16a), each device cannot be served by two or more APs. Constraint (16b) ensures that the packet offloading rate $\lambda_{k,m}$ is zero if the kth device is not served by the mth AP. Constraint (16c) guarantees that the sum of the packet offloading rates at the APs and the packet arrival rate at the local server is equal to the total packet arrival rate of a device. The constraint on the maximal number of subcarriers of the system is given in (16d), where the UL and DL bandwidth allocated to each device does not exceed the coherence bandwidth. When constraint (16e) is satisfied, $\varepsilon_k^{\rm loc}$ and $\varepsilon_m^{\rm mec}$ are smaller than 1, and hence the local and MEC servers are stable. By minimizing the objective function, we can check whether constraint (16e) can be satisfied or not. If it cannot be satisfied, the problem is infeasible.

Since CSI is not available at the transmitters, the transmit power on each subcarrier is fixed. In UL transmission, the maximal transmit power of a device, $P_{\max}^{\rm U}$, is equally allocated to $N_{\rm c}$ subcarriers, $P_{\rm s}^{\rm u} = \frac{P_{\max}^{\rm U}}{N_{\rm c}}$. In DL transmission, the maximal transmit power of an AP, $P_{\max}^{\rm A}$, is equally allocated to $N_{\rm t}$ antennas. Considering that the number of subcarriers for DL transmission can be up to $N_{\max}$, to satisfy maximal transmit power constraint, the transmit power on each subcarrier is fixed as $P_{\max}^{\rm A}/N_{\max}$. Thus, we have $P_{\rm s}^{\rm d} = \frac{P_{\max}^{\rm A}}{N_{\max} N_{\rm t}}$.

Problem (16) is a mixed integer optimization problem, which is non-convex. In typical scenarios, the throughput of eMBB services is much higher than the throughput of MC-IoT services, and hence the number of CPU cycles required to process long packets are much larger than that required to process short packets. In the rest part of this section, we first consider the scenario that $(\sum_{k=1}^{K} \lambda_k^{\rm U} c_{\rm S})/(\lambda_m^{\rm L} \bar{c}_{\rm L}) \to 0$, and then extend our algorithm into more general scenarios.

B. Solution in the Typical Scenario

1) Simplified Optimization Problem: According to (16c), $\lambda_{k,m} \le \lambda_k^{\rm U}$. Thus, we have

$$\rho_m = \frac{\sum_{k=1}^{K} \lambda_{k,m} c_{\rm S} + \lambda_m^{\rm L} \bar{c}_{\rm L}}{S_m} \le \frac{\sum_{k=1}^{K} \lambda_k^{\rm U} c_{\rm S} + \lambda_m^{\rm L} \bar{c}_{\rm L}}{S_m} \triangleq \rho_m^{\rm ub}. \tag{17}$$

When $(\sum_{k=1}^{K} \lambda_k^{\rm U} c_{\rm S})/(\lambda_m^{\rm L} \bar{c}_{\rm L}) \to 0$, the equality in (17) holds, and $\varepsilon_m^{\rm mec}$ in (14) is a constant that does not depend on packet offloading rates. Moreover, the packet loss probabilities due to decoding errors in UL and DL transmissions, $\varepsilon_{k,n}^{\rm u}$ and $\varepsilon_{k,n}^{\rm d}$, do not change with packet offloading rates. Thus, the second term in $\max \left[\varepsilon_k^{\rm loc}, x_{k,m}(\varepsilon_{k,n}^{\rm u} + \varepsilon_{k,n}^{\rm d} + \varepsilon_m^{\rm mec}), \forall m\right]$ does not depend on packet offloading rates. By setting $\lambda_{k,m} = \lambda_k^{\rm U}$, all the packets are offloaded to the MEC servers. Then, $\varepsilon_k^{\rm loc} = 0$ and

$$f_k(\mathbf{x}_k, \lambda_{k,m}, N_k^{\rm u}, N_k^{\rm d}) = \max_{m=1,...,M} x_{k,m}(\varepsilon_{k,m}^{\rm u} + \varepsilon_{k,m}^{\rm d} + \varepsilon_m^{\rm mec}). \tag{18}$$

With (18), problem (16) can be simplified as follows,

$$\min_{\substack{\mathbf{x}_k, N_k^{\rm u}, N_k^{\rm d} \\ k=1,...,K}} \; \max_{\substack{k=1,...,K \\ m=1,...,M}} x_{k,m}(\varepsilon_{k,m}^{\rm u} + \varepsilon_{k,m}^{\rm d} + \varepsilon_m^{\rm mec}) \tag{19}$$

$$\text{s.t.} \;\; \text{(16a)}, \text{(16d)}, \text{ and (16e)}.$$

2) Packet Loss Balance Algorithm: Denote the optimal solution and the minimal packet loss probability of problem (19) as $(\tilde{\mathbf{x}}_k, \tilde{N}_k^{\rm u}, \tilde{N}_k^{\rm d})$ and $\tilde{\varepsilon}^{\rm A}$, respectively. In the following, we propose a binary search algorithm to find the optimal solution. The basic idea of the algorithm is to keep $\varepsilon_{k,m}^{\rm u} + \varepsilon_{k,m}^{\rm d} + \varepsilon_m^{\rm mec}, k = 1, ..., K, m = 1, ..., M$, below a threshold $\varepsilon_{\rm th}$, and search the minimal $\varepsilon_{\rm th}$ in the regime $(0, \varepsilon_{\rm in}]$, where $\varepsilon_{\rm in} \le 1$ is an initial upper bound of the overall packet loss probability. We refer to the algorithm as the Packet Loss Balance (PLB) Algorithm.

For a given threshold of overall packet loss probability $\varepsilon_{\rm th}$, we search for the optimal association scheme and subcarrier allocation that minimize the total number of subcarriers. If the minimum number of subcarriers exceeds $N_{\max}$, then $\tilde{\varepsilon}^{\rm A} > \varepsilon_{\rm th}$. Otherwise, $\tilde{\varepsilon}^{\rm A} \le \varepsilon_{\rm th}$.

The problem that minimizes the total number of subcarriers can be expressed as follows,

$$\min_{\substack{\mathbf{x}_k, N_k^{\rm u}, N_k^{\rm d} \\ k=1,...,K}} \sum_{k=1}^{K} N_k^{\rm u} + \sum_{k=1}^{K} N_k^{\rm d} \tag{20}$$

$$\text{s.t.} \;\; x_{k,m}(\varepsilon_{k,m}^{\rm u} + \varepsilon_{k,m}^{\rm d} + \varepsilon_m^{\rm mec}) \le \varepsilon_{\rm th}, \tag{20a}$$

$$N_k^{\rm u}, N_k^{\rm d} \in \{1, 2, ..., N_{\rm c}\}, \text{ and (16a)},$$

where constraint (16e) is removed since $\varepsilon_{\rm th} < 1$. The above problem can be decoupled into K problems since the association schemes and bandwidth allocation of different devices are independent.

Given that the kth device is served by the m'th MEC server, $x_{k,m'} = 1$, the required number of subcarriers can be found from the following problem,[2]

$$\min_{N_k^{\rm u}, N_k^{\rm d}} N_k^{\rm u} + N_k^{\rm d} \tag{21}$$

$$\text{s.t.} \;\; \varepsilon_{k,m'}^{\rm u} + \varepsilon_{k,m'}^{\rm d} + \varepsilon_{m'}^{\rm mec} \le \varepsilon_{\rm th}, \tag{21a}$$

$$N_k^{\rm u}, N_k^{\rm d} \in \{1, 2, ..., N_{\rm c}\}.$$

To solve the inter programming problem, we first relax $N_k^{\rm u}$ and $N_k^{\rm d}$ as continuous variables, and find the optimal subcarrier allocation. Then, we discretize the number of subcarriers used in UL and DL transmissions. Note that only the discretization step will cause performance loss, which is minor as shown in [53].

To solve problem (21), we need the following property of (10).

Property 1. The packet loss probabilities $\varepsilon_{k,m'}^{\rm u}$ and $\varepsilon_{k,m'}^{\rm d}$ in (10) are convex in N_k^{ξ}.

Proof. See proof in Appendix A. □

Further considering that $\varepsilon_{m'}^{\rm mec}$ in (14) does not change with N_k^{ξ}, constraint (21a) is convex. Therefore, problem (21) is a convex problem, and can be solved by techniques such as the interior-point method [54]. The algorithm for solving problem (20) is provided in Table I, where $\lceil x \rceil$ is the minimum integer that is equal to or higher than x.

Note that problem (20) could be infeasible if $\varepsilon_{\rm th}$ is too small. In this case, the minimal overall packet loss probability is higher than $\varepsilon_{\rm th}$. Based on the algorithm in Table I, the PLB algorithm for solving problem (19) is shown in Table II.

[2] By solving problem (21) with different $m' = 1, ..., M$, we can obtain the optimal user association scheme and related bandwidth allocation that minimize $N_k^{\rm u} + N_k^{\rm d}$.

TABLE I
ACCESS SCHEME AND BANDWIDTH ALLOCATION

```
Input: Threshold of overall packet loss probability $\varepsilon_{\rm th}(i)$ (in the $i$th
    step of the binary search).
Output: Access scheme, $\mathbf{x}_k(i)$, and bandwidth allocation, $N_k^{\rm u}(i)$
    and $N_k^{\rm d}(i)$ (optimal solution of problem (20) in the $i$th step of
    the binary search).
1:  Set $k := 1$ and $m := 1$.
2:  while $k \leq K$ do
3:      while $m \leq M$ do
4:          Set $x_{k,m}(i) := 1$.
5:          Relaxing $N_k^{\rm u}(i)$ and $N_k^{\rm d}(i)$ as continuous variables $\hat{N}_k^{\rm u}(m)$
            and $\hat{N}_k^{\rm d}(m)$, respectively.
6:          Solve convex optimization problem (21), and obtain
            $\hat{N}_k^{\rm u}(m)$ and $\hat{N}_k^{\rm d}(m)$.
7:          Discretize the numbers of subcarriers, $\hat{N}_k^{\rm u}(m) :=$
            $\left\lceil \hat{N}_k^{\rm u}(m) \right\rceil$ and $\hat{N}_k^{\rm d}(m) := \left\lceil \hat{N}_k^{\rm d}(m) \right\rceil$.
8:          Set $\hat{N}_k^{\rm tot}(m) := \hat{N}_k^{\rm u}(m) + \hat{N}_k^{\rm d}(m)$.
9:      end while
10:     Set $m' := \arg\min_{m} \hat{N}_k^{\rm tot}(m)$.
11:     Set $x_{k,m'}(i) := 1$ and $x_{k,m}(i) := 0$, $\forall m \neq m'$.
12:     Set $N_k^{\rm u}(i) := \hat{N}_k^{\rm u}(m')$ and $N_k^{\rm d}(i) := \hat{N}_k^{\rm d}(m')$
13: end while
14: return $\mathbf{x}_k(i)$, $N_k^{\rm u}(i)$ and $N_k^{\rm d}(i)$, $k = 1, ..., K$.
```

TABLE II
PACKET LOSS BALANCE ALGORITHM

```
Input: Total number of subcarriers, $N_{\max}$, the bandwidth of each
    subcarrier, $W_0$, coherence bandwidth, $W_0 N_{\rm c}$, UL and DL trans-
    mit power on each subcarrier, $P_{\rm s}^{\rm u}$ and $P_{\rm s}^{\rm d}$, large-scale channel
    gains of devices, $\alpha_k$, the initial search area, $(0, \varepsilon_{\rm in})$, required
    accuracy of packet loss probability, $\Delta_\varepsilon$.
Output: Access scheme, $\tilde{\mathbf{x}}_k$, bandwidth allocation, $\tilde{N}_k^{\rm u}$ and $\tilde{N}_k^{\rm d}$, and
    packet loss probability $\tilde{\varepsilon}^{\rm A}$.
1:  Set $i := 1$, $\varepsilon^{\rm LB}(i) := 0$, $\varepsilon^{\rm UB}(i) := \varepsilon_{\rm in}$, and $\varepsilon_{\rm th}(i) := (\varepsilon^{\rm LB}(i) +$
    $\varepsilon^{\rm UB}(i))/2$.
2:  while $\varepsilon^{\rm UB}(i) - \varepsilon^{\rm LB}(i) > \Delta_\varepsilon$ do
3:      Solve problem (20) with the algorithm in Table I, and obtain
        $\mathbf{x}_k(i)$, $N_k^{\rm u}(i)$ and $N_k^{\rm d}(i)$.
4:      if $\sum_{k=1}^{K} \left[ N_k^{\rm u}(i) + N_k^{\rm d}(i) \right] > N_{\max}$ or problem (21) is
        infeasible then
5:          Set $\varepsilon^{\rm LB}(i+1) := \varepsilon_{\rm th}(i)$ and $\varepsilon^{\rm UB}(i+1) := \varepsilon^{\rm UB}(i)$.
6:      else
7:          Set $\varepsilon^{\rm UB}(i+1) := \varepsilon_{\rm th}(i)$ and $\varepsilon^{\rm LB}(i+1) := \varepsilon^{\rm LB}(i)$.
8:      end if
9:      Set $\varepsilon_{\rm th}(i+1) := (\varepsilon^{\rm LB}(i+1) + \varepsilon^{\rm UB}(i+1))/2$.
10:     $i := i + 1$.
11: end while
12: Set $\tilde{\varepsilon}^{\rm A} := \varepsilon_{\rm th}(i-1)$, $\tilde{\mathbf{x}}_k := p_k(i-1)$, $\tilde{N}_k^{\rm u} := N_k^{\rm u}(i-1)$, and
    $\tilde{N}_k^{\rm d} := N_k^{\rm d}(i-1)$, $k = 1, ..., K$.
13: return $\tilde{\varepsilon}^{\rm A}$, $\tilde{\mathbf{x}}_k$, $\tilde{N}_k^{\rm u}$, and $\tilde{N}_k^{\rm d}$, $k = 1, ..., K$.
```

3) Convergence of the PLB Algorithm: To prove that the PLB algorithm converges to the minimal packet loss probability of problem (19), we first prove the following proposition,

Proposition 1. The minimal packet loss probability $\varepsilon^{\rm A*}$ lies in the region $(\varepsilon^{\rm LB}(i), \varepsilon^{\rm UB}(i)]$, $\forall i \in \{1, 2, 3, ...\}$.

Proof. See proof in Appendix B. □

According to Proposition 1, the minimal packet loss probability lies in the region $(\varepsilon^{\rm LB}(i), \varepsilon^{\rm UB}(i)]$. After i steps of searching, the gap between $\varepsilon^{\rm A*}$ and the output of the PLB algorithm, $\tilde{\varepsilon}^{\rm A}$, is smaller than $0.5[\varepsilon^{\rm UB}(i) - \varepsilon^{\rm LB}(i)]$. In addition, with the binary search algorithm (i.e. from Line 1 to Line 11 in Table II), the range of $(\varepsilon^{\rm LB}(i), \varepsilon^{\rm UB}(i)]$ decreases according to the following expression, $\varepsilon^{\rm UB}(i) - \varepsilon^{\rm LB}(i) = \varepsilon_{\rm in}/2^i$. When i is large enough, $\tilde{\varepsilon}^{\rm A}$ approaches to $\varepsilon^{\rm A*}$.

The above proof holds when the performance loss caused by the discretization step in Line 7 of Table I is negligible. Since the discretization step inevitably causes some performance loss, the related association scheme and bandwidth allocation are near optimal.

4) Complexity of the PLB Algorithm: With the PLB algorithm, we need to solve problem (20) around $\log_2(\varepsilon_{\rm in}/\Delta_\varepsilon)$ times. Problem (20) is decoupled into K single-device problem in (21). With the algorithm in Table I, the convex optimization problem in (21) is solved KM times for K devices with M possible APs. The complexity of solving the convex optimization problem is denoted as Ω_0, which is not high. Therefore, the complexity of the PLB algorithm is $\mathcal{O}\left(\log_2(\varepsilon_{\rm in}/\Delta_\varepsilon) KM\Omega_0\right)$. Considering that a device will not associate with an AP that is very far from it, M will not be very large. For example, a device can only be connected to one of the three or four APs with the highest large-scale channel gains. As a result, the complexity of the PLB algorithm increases linearly with the number of devices.

V. SOLUTION IN GENERAL SCENARIOS

To solve problem (16), we extend the PLB algorithm into general scenarios without the assumption $(\sum_{k=1}^{K} \lambda_k^{\rm U} c_{\rm S})/(\lambda_m^{\rm L} \bar{c}_{\rm L}) \to 0$.

A. Extended PLB Algorithm

Although problem (16) cannot be simplified as problem (19), we can still use the algorithm in Table II. The difference between the general scenarios and the scenario with the assumption $(\sum_{k=1}^{K} \lambda_k^{\rm U} c_{\rm S})/(\lambda_m^{\rm L} \bar{c}_{\rm L}) \to 0$ lies in Line 3 of the algorithm, where problem (20) is obtained from (19). In general scenarios, given the threshold of overall packet loss probability, $\varepsilon_{\rm th}$, the optimization problem that minimizes the total number of subcarriers can be expressed as follows,

$$\min_{\substack{\mathbf{x}_k, \lambda_{k,m}, N_k^{\rm u}, N_k^{\rm d} \\ k=1,...,K}} \sum_{k=1}^{K} N_k^{\rm u} + \sum_{k=1}^{K} N_k^{\rm d} \tag{22}$$

$$\text{s.t.} \quad \max[\varepsilon_k^{\rm loc}, x_{k,m}(\varepsilon_{k,m}^{\rm u} + \varepsilon_{k,m}^{\rm d} + \varepsilon_m^{\rm mec}), \forall m] \leq \varepsilon_{\rm th}, \tag{22a}$$

$$N_k^{\rm u}, N_k^{\rm d} \in \{1, 2, ..., N_{\rm c}\}, \text{(16a)}, \text{(16b)}, \text{ and (16c)}.$$

From the expression of $\varepsilon_m^{\rm mec}$ in (14), we can see that $\varepsilon_m^{\rm mec}$ increases with the packet offloading rate $\lambda_{k,m}$. Besides, the expressions of $\varepsilon_{k,m}^{\rm u}$ and $\varepsilon_{k,m}^{\rm d}$ in (10) show that the required number of subcarriers increases as $\varepsilon_{k,m}^{\xi}$ decreases. Therefore, to satisfy constraint (22a), the number of subcarriers increases with the packet offloading rate $\lambda_{k,m}$. To minimize the total number of subcarriers, the first step is minimizing packet offloading rates.

Step 1: Optimize offloading rates. To minimize packet offloading rates, we find the maximal packets arrival rate at the local server, denoted as $\tilde{\lambda}_{k,0}(i)$. Note that the delay violation probability at each local server should satisfy $\varepsilon_k^{\rm loc} \leq \varepsilon_{\rm th}$, by substituting the expression of $\varepsilon_k^{\rm loc}$ in (11) into the constraint, $\tilde{\lambda}_{k,0}(i)$ can be obtained via binary search. If $\tilde{\lambda}_{k,0}(i) > \lambda_k^{\rm U}$, all the packets are processed at the local server, $\lambda_{k,m}(i) = 0$, $x_{k,m}(i) = 0$, $m = 1, ..., M$ and $N_k^{\rm u} = N_k^{\rm d} = 0$. Otherwise, the packet offloading rate of the kth device is $\sum_{m=1}^{M} \lambda_{k,m}(i) = \lambda_k^{\rm U} - \tilde{\lambda}_{k,0}(i)$. As such, the constraint on the packet offloading rate in (16c) can be expressed as follows,

$$\sum_{m=1}^{M} \lambda_{k,m} = \max[0, \lambda_k^{\rm U} - \tilde{\lambda}_{k,0}(i)], k = 1, ..., K. \tag{23}$$

According to (16a) and (16b), each device only offload it's packets to one AP. Thus, the value of $\lambda_{k,m}$ is determined by $x_{k,m}$. If $x_{k,m} = 1$, $\lambda_{k,m} = \max[0, \lambda_k^{\rm U} - \tilde{\lambda}_{k,0}(i)]$. Otherwise, $\lambda_{k,m} = 0$.

With the minimal packet offloading rates, problem (22) can be simplified as follows,

$$\min_{\substack{\mathbf{x}_k, N_k^{\rm u}, N_k^{\rm d} \\ k=1,...,K}} \sum_{k=1}^{K} N_k^{\rm u} + \sum_{k=1}^{K} N_k^{\rm d} \tag{24}$$

$$\text{s.t.} \;\; x_{k,m}(\varepsilon_{k,m}^{\rm u} + \varepsilon_{k,m}^{\rm d} + \varepsilon_m^{\rm mec}) \leq \varepsilon_{\rm th}, \tag{24a}$$

$$\lambda_{k,m} = x_{k,m} \max[0, \lambda_k^{\rm U} - \tilde{\lambda}_{k,0}(i)] \qquad (24a)$$

$$N_k^{\rm u}, N_k^{\rm d} \in \{1, 2, ..., N_{\rm c}\}, \text{ and (16a).}$$

where $k = 1, ..., K$, and $m = 1, ..., M$. Different from problem (20), problem (24) cannot be decoupled into K subproblems. This is because the workloads of the APs depend on association schemes of all the devices. As a result, $\varepsilon_m^{\rm mec}$ changes with $x_{k,m}$. Changing the association scheme of one device will lead to different overall packet loss probabilities of all the other devices.

Based on this fact that the throughput of eMBB services is higher than MC-IoT services in most of the scenarios, we optimize the association scheme given the optimal bandwidth allocation obtained from problem (20), and then update bandwidth allocation according to the association scheme and related workloads at MEC servers.

Step 2: Optimize the association scheme $\mathbf{x}_k(i)$. We set $N_k^{\rm u}$ and $N_k^{\rm d}$ as the values that are obtained under the assumption $(\sum_{k=1}^{K} \lambda_k^{\rm U} c_{\rm S})/(\lambda_m^{\rm L} \bar{c}_{\rm L}) \to 0$, and compute $\varepsilon_{k,m}^{\rm u} + \varepsilon_{k,m}^{\rm d}$, $k = 1, ..., K$, $m = 1, ..., M$. The initial workloads of the MEC servers are $\hat{\rho}_m = \lambda_m^{\rm L} \bar{c}_{\rm L} / S_m$, $m = 1, ..., M$. From (8) we can obtain the initial delay bound violation probability $\hat{\varepsilon}_m^{\rm mec}$. Then, from the 1st device to the Kth device, each device selects one AP that can minimize $\varepsilon_{k,m}^{\rm u} + \varepsilon_{k,m}^{\rm d} + \hat{\varepsilon}_m^{\rm mec}$. Denote $\hat{m}_k$ as the AP that minimizes $\varepsilon_{k,m}^{\rm u} + \varepsilon_{k,m}^{\rm d} + \hat{\varepsilon}_m^{\rm mec}$. The $\hat{m}_k$th element of $\mathbf{x}_k(i)$ equals to one and $\lambda_{k,\hat{m}_k} = \max[0, \lambda_k^{\rm U} - \tilde{\lambda}_{k,0}(i)]$, $\lambda_{k,m} = 0$, $\forall m \neq \hat{m}_k$. After the kth device is associated with the $\hat{m}_k$th AP, the workload is updated according to $\hat{\rho}_m = (\lambda_{1,m} c_{\rm S} + \lambda_{2,m} c_{\rm S} + ... + \lambda_{k,m} c_{\rm S} + \lambda_m^{\rm L} \bar{c}_{\rm L}) / S_m$.

Step 3: Update bandwidth allocation $N_k^{\rm u}$ and $N_k^{\rm d}$. Given $\mathbf{x}_k(i)$, we solve problem (21) for each device, and obtain the bandwidth allocation.

Remark 4. The packet offloading rates obtained in Step 1 and the bandwidth allocation obtained in Step 3 are optimal with given $\mathbf{x}_k(i)$. If we can obtain the optimal association scheme in Step 2, then we can obtain the optimal solution of problem (22). However, the final workloads of the APs are not exactly the same as the initial values. Thus, $\mathbf{x}_k(i)$ obtained in Step 2 is not optimal. However, $\mathbf{x}_k(i)$ is near optimal if the association scheme of MC-IoT services has little impacts on the workloads of the APs. To provide more insights, we will prove that $\mathbf{x}_k(i)$ is optimal in two asymptotic cases in the sequel.

B. Optimal Access Schemes in Two Asymptotic Cases

In this subsection, we derive the optimal association scheme of problem (16) in the two asymptotic cases: communication or computing is the bottleneck of the overall packet loss probability. Whether communication or computing is the bottleneck depends on the number of antennas at each AP and the processing ability of the MEC server. In the next section, we will show the assumption that either communication or computing is the bottleneck is reasonable for various system parameters.

1) Communication is the Bottleneck: When the MEC servers have enough computing resources, the processing delay bound violation probability is much smaller than packet loss due to decoding errors, i.e., $\varepsilon_m^{\rm mec} \ll \varepsilon_{k,m}^{\xi}$. In this case, communication is the bottleneck of reliability. Denote $\mathbf{x}_k^{\rm comm}, k = 1, ..., K$, as the association scheme that the kth device is served by the m_k^*th AP, which has the highest large-scale channel gain among all the APs, i.e., $\alpha_{k,m_k^*} > \alpha_{k,m'_k}, \forall m'_k \neq m_k^*$. The following proposition indicates that $\mathbf{x}_k^{\rm comm}$ is the optimal association scheme if $\sum_{m=1}^{M} x_{k,m} = 1$.

Proposition 2. If $\sum_{m=1}^{M} x_{k,m} = 1$, then for any solution of problem (16), $(x_{k,m'_k} = 1, \lambda_{k,m}, N_k^{\rm u}, N_k^{\rm d})$, we can always find another solution, $(x_{k,m_k^*} = 1, \lambda_{k,m}, N_k^{\rm u}, N_k^{\rm d})$, that can achieve smaller packet loss probability.

Proof. Since $\alpha_{k,m_k^*} \geq \alpha_{k,m'_k}$ and the decoding error probability in (10) decreases with the large-scale channel gain, we have $\varepsilon_{k,m_k^*}^{\xi} \leq \varepsilon_{k,m'_k}^{\xi}$. Therefore,

$$\max_{k=1,...,K} \max\left[\varepsilon_k^{\rm loc}, x_{k,m'_k}(\varepsilon_{k,m'_k}^{\rm u} + \varepsilon_{k,m'_k}^{\rm d}), \forall m\right]$$
$$\leq \max_{k=1,...,K} \max\left[\varepsilon_k^{\rm loc}, x_{k,m_k^*}(\varepsilon_{k,m_k^*}^{\rm u} + \varepsilon_{k,m_k^*}^{\rm d})\right].$$

This completes the proof. □

When all the packets of the kth device are processed at the local server, $\lambda_{k,0} = \lambda_k^{\rm U}$, if $\varepsilon_k^{\rm loc} \leq \varepsilon_{k,m^*}^{\rm u} + \varepsilon_{k,m^*}^{\rm d}$, then $x_{k,m} = 0, \forall m$, and $N_k^{\rm u} = N_k^{\rm d} = 0$. Otherwise, the device uploads some packets to the MEC servers to achieve better reliability, and hence $\sum_{m=1}^{M} x_{k,m} = 1$.

2) Computing is the Bottleneck: When there are very limited computing resources at the APs and no processing unit at devices, the packet loss probability due to decoding errors is much smaller than the processing delay violation probability, i.e., $\varepsilon_{k,m}^{\xi} \ll \varepsilon_m^{\rm mec}$. In this case, all the packets are processed at the MEC servers, and the objective function of problem (16) can be simplified as follows,

$$\max_{k=1,...,K} x_{k,m}\varepsilon_m^{\mathrm{mec}}. \tag{25}$$

Let $\mathbf{x}_k^{\mathrm{comp}}, k = 1, ..., K$, be the optimal association scheme when computing is the bottleneck, respectively. To find the optimal solution, we denote the average packet arrival rate at the mth MEC server as λ_m^{A}. Then, $\lambda_m^{\mathrm{A}} = \sum_{k=1}^{K} \lambda_{k,m}$ and

$$\sum_{m=1}^{M} \lambda_m^{\mathrm{A}} = \sum_{m=1}^{M}\sum_{k=1}^{K} \lambda_{k,m}. \tag{26}$$

Given the average arrival rate of each MEC server, the processing delay bound violation probability can be expressed as $\varepsilon_m^{\mathrm{mec}} = \rho_m^{\frac{S_m(D_{\max}-2)}{c_{\mathrm{S}}}-1}$, where $\rho_m = \frac{\lambda_m^{\mathrm{A}} c_{\mathrm{S}} + \lambda_m^{\mathrm{L}} \bar{c}_{\mathrm{L}}}{S_m}$.

Let $\varepsilon^{\mathrm{A}*}$ and ρ^* be the minimal value of (25) and the workload achieved by the optimal association scheme, respectively. If $\frac{\lambda_m^{\mathrm{L}} \bar{c}_{\mathrm{L}}}{S_m} \geq \rho^*$, then $x_{k,m} = 0, \forall k$, and $\lambda_m^{\mathrm{A}^*} = 0$. Let $\mathcal{M}$ be the set of MEC servers with $\frac{\lambda_m^{\mathrm{L}} \bar{c}_{\mathrm{L}}}{S_m} < \rho^*$. From $\rho^* = \frac{\lambda^{\mathrm{A}^*} c_{\mathrm{S}} + \lambda_m^{\mathrm{L}} \bar{c}_{\mathrm{L}}}{S_m}$, we can derive

$$\lambda_m^{\mathrm{A}^*} = \frac{\rho^* S_m - \lambda_m^{\mathrm{L}} \bar{c}_{\mathrm{L}}}{c_{\mathrm{S}}}, \forall m \in \mathcal{M}. \tag{27}$$

Substituting $\lambda_m^{\mathrm{A}^*}$ into (26), we can derive that

$$\rho^* = \frac{\sum\limits_{m=1}^{M}\sum\limits_{k=1}^{K} \lambda_{k,m} c_{\mathrm{S}} + \sum\limits_{m\in\mathcal{M}} \lambda_m^{\mathrm{L}} \bar{c}_{\mathrm{L}}}{\sum\limits_{m\in\mathcal{M}} S_m}. \tag{28}$$

To obtain ρ^* and the related $\lambda_m^{\mathrm{A}^*}$, we need to obtain $\mathcal{M}$. Without loss of generality, we assume $\frac{\lambda_1^{\mathrm{L}} \bar{c}_{\mathrm{L}}}{S_1} \leq \frac{\lambda_2^{\mathrm{L}} \bar{c}_{\mathrm{L}}}{S_2} \leq ... \leq \frac{\lambda_M^{\mathrm{L}} \bar{c}_{\mathrm{L}}}{S_M}$. Then, $\mathcal{M}$ can be expressed as $\mathcal{M} = \{m = 1, ..., M_{\mathrm{th}}\}$. Then, we can use the binary search algorithm to find the maximal M_{th} that satisfies $\lambda_m^{\mathrm{A}^*} > 0, \forall m \leq M_{\mathrm{th}}$. As illustrated in Fig. 4, the basic idea of the optimal solution is offloading packets to the MEC servers with lower workloads.

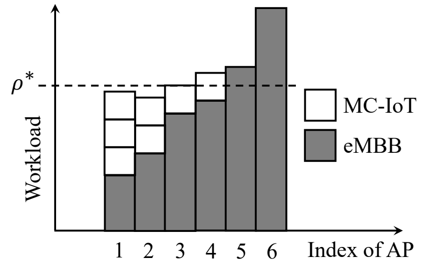


Fig. 4. Optimal association scheme when computing is the bottleneck.

On the one hand, since the association scheme is discretized, perhaps there is no association scheme that can keep the workloads of the M_{th} MEC servers exactly the same. On the other hand, if it is possible to satisfy $\rho_m = \rho^*$ in all the M_{th} servers, x_k^{comp} may not be unique. For example, if $\lambda_1^{\mathrm{U}} = \lambda_2^{\mathrm{U}}$, exchanging $\mathbf{x}_1^{\mathrm{comp}}$ and $\mathbf{x}_2^{\mathrm{comp}}$ does not change the workloads of the servers. Any association schemes that satisfy $\rho_m = \rho^*, \forall m \leq M_{\mathrm{th}}$, are optimal.

C. Convergence of the Extended PLB Algorithm

When communication is the bottleneck, $\varepsilon_{k,m}^{\mathrm{u}} + \varepsilon_{k,m}^{\mathrm{d}} + \hat{\varepsilon}_m^{\mathrm{mec}} \approx \varepsilon_{k,m}^{\mathrm{u}} + \varepsilon_{k,m}^{\mathrm{d}}$. Since the decoding error probability in (10) decreases with the large-scale channel gain, each device is associated with the MEC server with the largest large-scale channel gain. Thus, the association scheme obtained in Step 2 of the extended PLB algorithm is the same as the optimal association scheme when communication is the bottleneck, $\mathbf{x}_k^{\mathrm{comm}}$.

When computing is the bottleneck of reliability, $\varepsilon_{k,m}^{\mathrm{u}} + \varepsilon_{k,m}^{\mathrm{d}} + \varepsilon_m^{\mathrm{mec}} \approx \varepsilon_m^{\mathrm{mec}}$. In the second step of the extended PLB algorithm, each device connected to the MEC server with the lowest workload. Then, the association scheme is the same as the optimal association scheme in Fig. 4, $\mathbf{x}_k^{\mathrm{comp}}$.

D. Complexity of the Extended PLB Algorithm

In the first step, we can use the binary search algorithm to find each $\lambda_{k,0}(i)$ with low complexity Ω_1. In the second step, we find $\hat{m}_k$ from M MEC servers for each of the K devices. Denote the complexity for computing $\varepsilon_{k,m}^{\mathrm{u}} + \varepsilon_{k,m}^{\mathrm{d}} + \hat{\varepsilon}_m^{\mathrm{mec}}$ as Ω_2. Then, the complexity of the second step is around $MK\Omega_2$. In the third step, we need to solve problem (21) K times. Thus, the complexity is $K\Omega_0$. Therefore, the complexity of the extended PLB algorithm is $\mathcal{O}\left((K\Omega_0 + K\Omega_1 + MK\Omega_2)\log_2(\varepsilon_{\mathrm{in}}/\Delta_\varepsilon)\right)$, which is linear with the number of devices.

VI. Simulation and Numerical Results

In this section, we validate the approximation of the CCDF of the processing delay of short packets in the PS server. To show the performance gain of our proposed framework, we compare the distributions of delay with PS servers and that with FCFS servers.[3] Besides, we illustrate the near optimal association scheme obtained with the extended PLB algorithm, and compared it with the optimal solution in the asymptotic cases. Finally, the reliability achieved by the extended PLB algorithm is illustrated in the scenarios with different communication and computing resources.

A. Simulation Setup

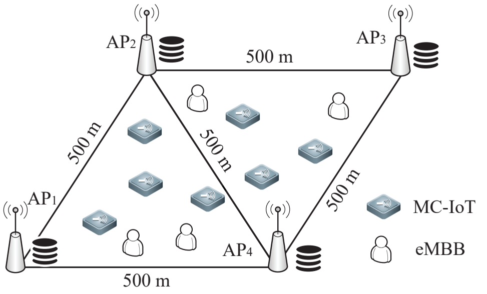


Fig. 5. Simulation Scenario

[3] Although there are some related works, none of them took both uplink and downlink transmissions of short packets into account, and few of them optimized offloading policy from multiple devices to multiple APs with random packets arrival processes.

TABLE III
System Parameters [6, 7]

Transmit power of each device $P_{\text{tot}}^{\text{M}}$	23 dBm
Transmit power of each AP $P_{\text{tot}}^{\text{A}}$	46 dBm
Duration of each slot T_{s}	0.125 ms
E2E delay requirement $D_{\max}$	1 ms
Bandwidth of each subcarrier W_0	120 kHz
Number of subcarriers in each cluster $N_{\max}$	256
Coherence bandwidth $N_{\text{c}}W_0$	1.2 MHz
Packet size b_k^{ξ}	32 bytes
The required minimal CPU cycles of long packets c_0/c_{S}	10
The average arrival rate of long packets at each AP	0.1 packets/slot
Single-sided noise spectral density N_0	-174 dBm/Hz

The simulation scenario is shown in Fig. 5, where 4 APs serve multiple MC-IoT and eMBB devices. The distance between two APs is $d_{\text{ap}} = 500$ m. The path loss model is $35.3 + 37.6\log_{10}\{d\ (\text{m})\}$, where d is the distance between an AP and a device served by it [55]. The shadowing is lognormal distributed with 8 dB standard deviation. Half of the devices need $D_k^{\text{s}} = 5$ slots to process a packet at their local servers, and the other half of the devices need $D_k^{\text{s}} = 6$ slots. The packet arrival rate of each device, λ_k^{U}, is uniformly distributed between 0.05 and 0.1 packets/slot [43].

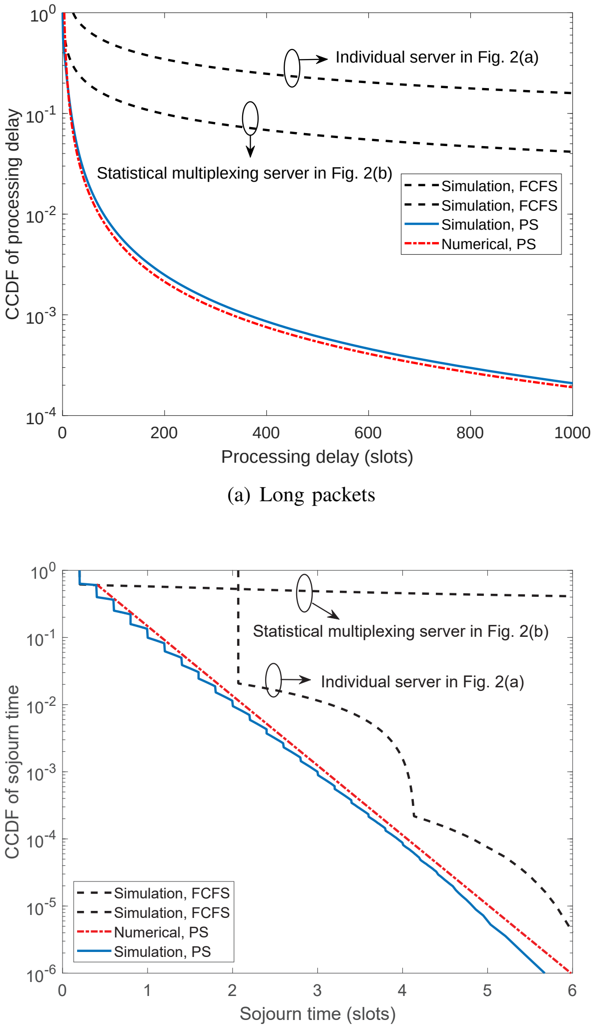
The required CPU cycles for processing a long packet, c_{L}, follows the distribution in (1). Denote the processing delay of long packets at the mth MEC server as W_m^{L}. According to the result in [24], we have

$$\Pr\{W_m^{\text{L}} > x\} \sim p_{\text{A}}(S_m/c_{\text{S}})^{-v}(1-\rho_m)^{-v}x^{-v}, \qquad (29)$$

where $f(x) \sim h(x)$ means that $\lim_{x\to\infty}\frac{f(x)}{h(x)} = 1$. In our simulation, $v = 1.5$. Other parameters are listed in Table III, unless otherwise specified.

B. CCDFs of Processing Delay

The CCDFs of processing delay in the FCFS servers and the PS server at the mth AP are illustrated in Fig. 6. In the simulation, 20 devices are served by the AP, where the first 10 of them send short packets and the other 10 send long packets. The average packet rate from each device is $\lambda_{k,m} = 0.01$ packets/slot. The individual and statistical multiplexing servers are illustrated in Fig. 2(a) and Fig. 2(b), respectively. In the individual server, the service rate of the MEC server is allocated to the devices according to their packet arrival rate, i.e., $\sum_{k=1}^{20} S_{k,m} = S_m$, $\frac{\lambda_{k,m}c_{\text{S}}}{S_{k,m}} = \rho_m$, $k = 1,...,10$, and $\frac{\lambda_{k,m}\bar{c}_{\text{L}}}{S_{k,m}} = \rho_m$, $k = 11,...,20$. The simulation results are obtained by generating 10^8 packets and computing their processing delay. The numerical results in Fig. 6(a) and Fig. 6(b) are obtained from (29) and (7), respectively.



(a) Long packets

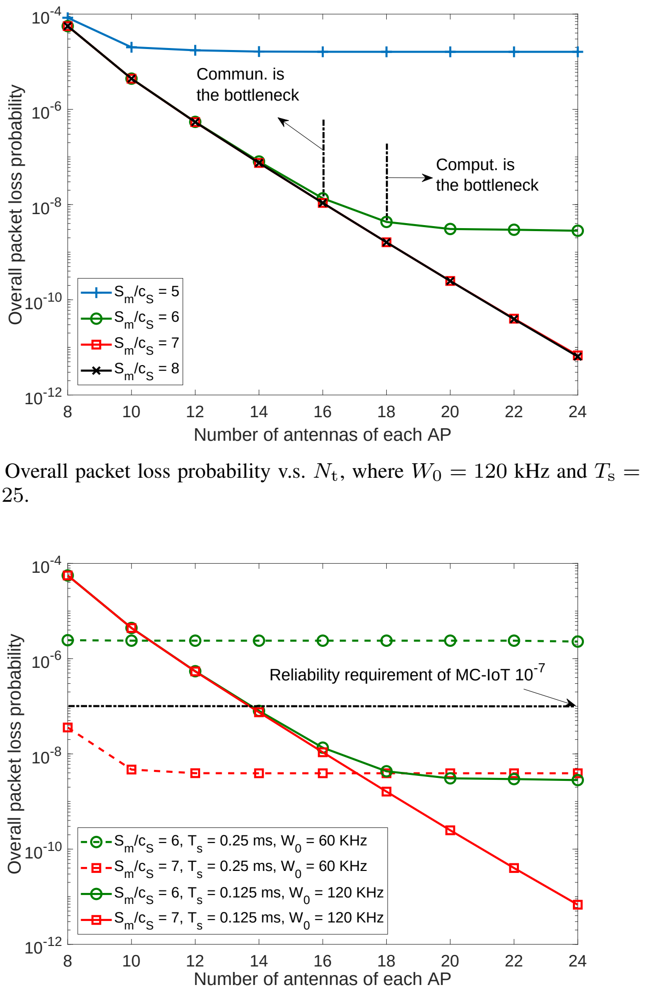
(b) Short packets

Fig. 6. CCDFs of processing delay in the AP, where the service rate of the MEC is $S_m/c_{\text{S}} = 5$ packets/slot.

The results in Fig. 6(a) indicate that to achieve the same delay bound for long packets, the delay bound violation probability of PS server is much smaller than that of the FCFS servers. The results in Fig. 6(b) indicate that the approximation in (7) is very accurate for short packets, and the PS server outperforms the FCFS servers in the short delay regime. Besides, we can see that compared with the statistical multiplexing FCFS server, the individual FCFS server achieves better QoS for short packets by sacrificing the QoS of long packets. However, the PS server can achieve better QoS than the FCFS servers for both long and short packets when the distribution of the number of CPU cycles required to process the packets has a heavy tail.

C. Overall Packet Loss Probabilities

The overall packet loss probabilities achieved by the optimized association scheme, packet offloading rates, and bandwidth allocation are illustrated in Fig. 7, where the solution and the minimal overall packet loss probability are obtained with the extended PLB algorithm. As shown in Fig. 7(a), better reliability can be achieved with more antennas or with higher service rate at the MEC servers. To show when communication or computing is the bottleneck of reliability, we consider the case $S_m/c_{\text{S}} = 6$. When $N_{\text{t}} \leq 16$, increasing the service rate of the MEC servers does not help improving reliability, and hence communication is the bottleneck. When $N_{\text{t}} \geq 18$,



(a) Overall packet loss probability v.s. $N_{\rm t}$, where $W_0 = 120$ kHz and $T_{\rm s} = 0.125$.

(b) Overall packet loss probability v.s. $N_{\rm t}$ with different W_0 and $T_{\rm s}$.

Fig. 7. Overall packet loss probabilities achieved by the optimized association scheme, packet offloading rates, and bandwidth allocation, where $K = 20$.

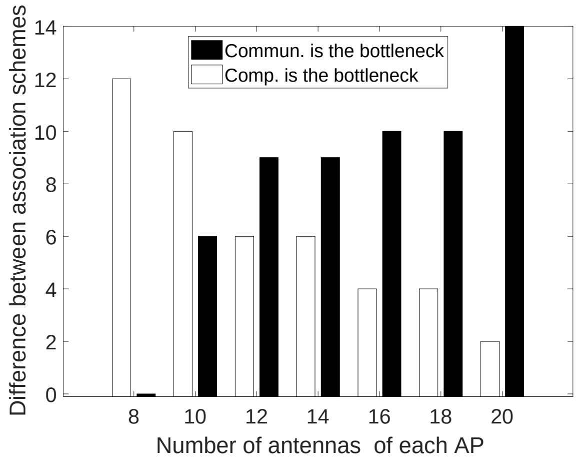


Fig. 8. The difference between the association schemes, where $S_m/c_{\rm S} = 6$ packets/slot and $K = 20$.

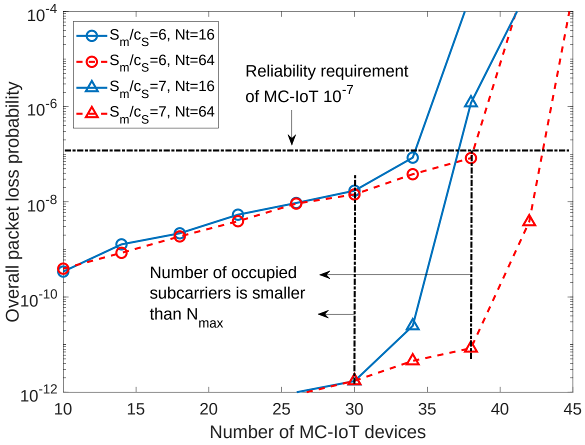


Fig. 9. Overall packet loss probability v.s. the number of MC-IoT devices in one cluster.

overall packet loss probability does not decrease with $N_{\rm t}$, and hence computing is the bottleneck. As a result, only when $N_{\rm t} \in [16, 18]$, the packet loss in UL and DL transmissions are comparable to the processing delay violation. In the cases that $S_m/c_{\rm S} > 6$, communication is always the bottleneck, because the processing delay violation probability is much smaller than the packet loss due to decoding errors. These results imply that the extended PLB algorithm converges to the optimal solutions in most of the scenarios.

According to 5G NR in [7], the bandwidth of each subcarrier, W_0, and the duration of each slot, $T_{\rm s}$, can be adjusted according to the requirements of services. In Fig. 7(b), we illustrated the impact of W_0 and $T_{\rm s}$ on the reliability, where $T_{\rm s}W_0$ is fixed as a constant such that there are 14 symbols transmitted in each slot with one subcarrier. The results in Fig. 7(b) indicate that when $N_{\rm t}$ is small (i.e., communication is the bottleneck), increasing $T_{\rm s}$ is helpful for increasing reliability. The total bandwidth allocated to each device does not exceed coherence bandwidth 1.2 MHz, which does not change with W_0. By increasing $T_{\rm s}$, the maximal blocklength of each packet increases. As a result, the packet loss probability due to decoding errors decreases with $T_{\rm s}$. However, when computing is the bottleneck, increasing $T_{\rm s}$ leads to higher overall packet loss probability.

The differences between the association scheme obtained with the extended PLB algorithm and the optimal association schemes (i.e., $||\mathbf{x}_k^{\rm PLB} - \mathbf{x}_k^{\rm comm}||$ with the legend “Commun. is the bottleneck” and $||\mathbf{x}_k^{\rm PLB} - \mathbf{x}_k^{\rm comp}||$ with the legend “Comp. is the bottleneck”) are shown in Fig. 8. When $N_{\rm t} = 8$, communication is the bottleneck and $\mathbf{x}_k^{\rm PLB}$ and $\mathbf{x}_k^{\rm comm}$ are the same. When $N_{\rm t}$ is large, $\mathbf{x}_k^{\rm PLB}$ approaches to $\mathbf{x}_k^{\rm comp}$. The results in Fig. 8 are consistent with our analysis in Section V.C that the extended PLB algorithm converges to the optimal solutions in the two asymptotic cases.

The relation between the overall packet loss probability and the density of devices is illustrated in Fig. 9. The curves are not smooth, because problem (16) is a mixed integer optimization problem. The results in Fig. 9 indicate that there is a tradeoff between the overall packet loss probability and the density of devices. When the number of occupied subcarriers is less than the maximal number of subcarriers, i.e., $\sum_{k=1}^{K}(N_k^{\rm u} + N_k^{\rm d}) < N_{\max}$, the overall packet loss probability increases slowly as K increases. When $\sum_{k=1}^{K}(N_k^{\rm u} + N_k^{\rm d}) = N_{\max}$, the overall packet loss probability increases extremely fast as K increases. By increasing $N_{\rm t}$ and S_m, the density of devices can be improved (i.e., the number of devices with overall packet loss probability less than 10^{-7}), but the performance gain in Fig. 9 is only around 25%.

VII. Conclusion

In this work, we analyzed the processing delay of short packets in the M/G/1/PS server. By introducing an accurate approximation, the CCDF of the processing delay of short packets was derived in closed-form. We then formulated an optimization problem that minimizes the overall packet loss probability under the constraints on ultra-low E2E delay in the MEC system, where the association scheme, packet offloading rates, and bandwidth allocation for MC-IoT services are optimized. The problem is a mixed integer problem and is non-convex. To solve the problem, we proposed a PLB algorithm in the scenario that the data rates of eMBB services are much higher than that of MC-IoT services. We further extended the algorithm into more general scenarios, where we can obtain a near optimal solution with low complexity, i.e., the complexity increases linearly with the number of devices. To analyze the performance of the extended PLB algorithm, we derived the optimal solutions of the problem in two asymptotic cases: communication or computing is the bottleneck of reliability, and proved that the extended PLB algorithm converges to the optimal solution in these two asymptotic cases. Simulation and numerical results validated our analysis and showed that the PS server outperforms FCFS servers.

Appendix A
Proof of Property 1

Proof. Denote $f_N(N_k^\xi) = \sqrt{\frac{T_\mathrm{s} N_k^\xi W_0}{V_k^\xi}}\left[\ln\left(1+\frac{\alpha_{k,m} g_{k,m}^\xi P_\mathrm{s}^\xi}{N_0 W_0}\right) - \frac{b_k^\xi \ln 2}{T_\mathrm{s} N_k^\xi W_0}\right]$. The second order derivative of $f_N(N_k^\xi)$ can be derived as follows,

$$\begin{aligned} f_N''(N_k^\xi) = & -\frac{1}{4}\left(N_k^\xi\right)^{-\frac{3}{2}}\sqrt{\frac{T_\mathrm{s} W_0}{V_k^\xi}}\ln\left(1+\frac{\alpha_{k,m} g_{k,m}^\xi P_\mathrm{s}^\xi}{N_0 W_0}\right) \\ & -\frac{3}{4}\left(N_k^\xi\right)^{-\frac{5}{2}}\frac{b_k^\xi \ln 2}{\sqrt{T_\mathrm{s} W_0 V_k^\xi}} < 0. \end{aligned} \tag{A.1}$$

Thus, $f_N(N_k^\xi)$ is concave in N_k^ξ. Moreover, Q function $f_\mathrm{Q}(x)$ is a convex and decreasing function when $f_\mathrm{Q}(x) < 0.5$, which is the case in MC-IoT. According to [54], the composite function $f_\mathrm{Q}\left(f_N(N_k^\xi)\right)$ is convex in N_k^ξ. This completes the proof. □

Appendix B
Proof of Proposition 1

Proof. We apply the mathematical induction to prove Proposition 1. When $i = 1$, $\varepsilon^{\mathrm{LB}}(1) = 0$ and $\varepsilon^{\mathrm{UB}}(1) = \varepsilon_{\mathrm{in}}$, we have $\tilde{\varepsilon}^{\mathrm{A}} \in (\varepsilon^{\mathrm{LB}}(1), \varepsilon^{\mathrm{UB}}(1)]$. We assume that when $i = j$, $\tilde{\varepsilon}^{\mathrm{A}} \in (\varepsilon^{\mathrm{LB}}(j), \varepsilon^{\mathrm{UB}}(j)]$, and prove $\tilde{\varepsilon}^{\mathrm{A}} \in (\varepsilon^{\mathrm{LB}}(j+1), \varepsilon^{\mathrm{UB}}(j+1)]$.

In the case that $\sum_{k=1}^{K}\left[N_k^{\mathrm{u}}(i) + N_k^{\mathrm{d}}(i)\right] \leq N_{\max}$, $\varepsilon_{\mathrm{th}}(i)$ can be achieved by a solution that lies in the feasible region of problem (19). Thus, $\tilde{\varepsilon}^{\mathrm{A}} \leq \varepsilon_{\mathrm{th}}(i)$. According to the algorithm in Table II, we have $\varepsilon^{\mathrm{UB}}(j+1) = \varepsilon_{\mathrm{th}}(j)$ and $\varepsilon^{\mathrm{LB}}(j+1) = \varepsilon^{\mathrm{LB}}(j)$. Further considering that $\tilde{\varepsilon}^{\mathrm{A}} > \varepsilon^{\mathrm{LB}}(j)$ with the assumption in the case $i = j$, we have $\tilde{\varepsilon}^{\mathrm{A}} \in (\varepsilon^{\mathrm{LB}}(j+1), \varepsilon^{\mathrm{UB}}(j+1)]$.

In the case that $\sum_{k=1}^{K}\left[N_k^{\mathrm{u}}(i) + N_k^{\mathrm{d}}(i)\right] > N_{\max}$, $\varepsilon_{\mathrm{th}}(i)$ cannot be achieved with $N_{\max}$ subcarriers. Thus, $\tilde{\varepsilon}^{\mathrm{A}} > \varepsilon_{\mathrm{th}}(i)$. According to the algorithm in Table II, we have $\varepsilon^{\mathrm{LB}}(j+1) = \varepsilon_{\mathrm{th}}(j)$ and $\varepsilon^{\mathrm{UB}}(j+1) = \varepsilon^{\mathrm{UB}}(j)$. Further considering that $\tilde{\varepsilon}^{\mathrm{A}} \leq \varepsilon^{\mathrm{UB}}(j)$ with the assumption in the case $i = j$, we have $\tilde{\varepsilon}^{\mathrm{A}} \in (\varepsilon^{\mathrm{LB}}(j+1), \varepsilon^{\mathrm{UB}}(j+1)]$. The proof ends here. □

References

[1] Y. Duan, C. She, G. Zhao, and T. Q. S. Quek, "Delay analysis and computing offloading of URLLC in mobile edge computing systems," in *Proc. WCSP*, 2018.

[2] E. K. Markakis, I. Politis, A. Lykourgiotis, Y. Rebahi, G. Mastorakis, C. X. Mavromoustakis, and E. Pallis, "Efficient next generation e-mergency communications over multi-access edge computing," *IEEE Commun. Mag.*, vol. 55, no. 11, pp. 92–97, 2017.

[3] O. N. C. Yilmaz, Y.-P. E. Wang, N. A. Johansson, N. Brahmi, S. A. Ashraf, and J. Sachs, "Analysis of ultra-reliable and low-latency 5G communication for a factory automation use case," in *IEEE ICC Workshop*, 2015.

[4] P. Schulz, M. Matthe, H. Klessig, M. Simsek, G. Fettweis, J. Ansari, S. A. Ashraf, B. Almeroth, J. Voigt, I. Riedel, A. Puschmann, A. Mitschele-Thiel, M. Muller, T. Elste, and M. Windisch, "Latency critical IoT applications in 5G: Perspective on the design of radio interface and network architecture," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 70–78, 2017.

[5] X. Jiang, et al., "Low-latency networking: Where latency lurks and how to tame it," *Proc. IEEE*, vol. 107, no. 2, pp. 280–306, Feb. 2019.

[6] 3GPP TR 38.913, "Study on scenarios and requirements for next generation access technologies." Tech. rep. Release 14. 3GPP, Jun. 2017.

[7] 3GPP, "Study on new radio (NR) access technology; physical layer aspects (release 14)." TR 38.802 V2.0.0, Apr. 2017.

[8] C. She, C. Yang, and T. Q. S. Quek, "Radio resource management for ultra-reliable and low-latency communications," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 72–78, Jun. 2017.

[9] Y.-Y. Shih, W.-H. Chung, A.-C. Pang, T.-C. Chiu, and H.-Y. Wei, "Enabling low-latency applications in fog-radio access networks," *IEEE Netw.*, vol. 31, no. 1, pp. 52–58, Jan./Feb. 2017.

[10] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2322–2358, 2017.

[11] T. Q. Dinh, J. Tang, Q. D. La, and T. Q. Quek, "Offloading in mobile edge computing: Task allocation and computational frequency scaling," *IEEE Trans. on Commun.*, vol. 65, no. 8, pp. 3571–3584, Aug. 2017.

[12] 5GPPP Architecture Working Group, "View on 5G architecture," in *5G Architecture White Paper*, Dec. 2017.

[13] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[14] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4232–4264, Jul. 2014.

[15] G. Ozcan and M. C. Gursoy, "Throughput of cognitive radio systems with finite blocklength codes," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 11, pp. 2541–2554, Nov. 2013.

[16] Y. Hu, A. Schmeink, and J. Gross, "Blocklength-limited performance of relaying under quasi-static Rayleigh channels," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4548–4558, Jul. 2016.

[17] Y. Hu, M. C. Gursoy, and A. Schmeink, "Relaying-enabled ultra-reliable low-latency communications in 5G," *IEEE Network*, vol. 32, no. 2, pp. 62–68, Mar.-Apr. 2018.

[18] S. Xu, T.-H. Chang, S.-C. Lin, C. Shen, and G. Zhu, "Energy-efficient packet scheduling with finite blocklength codes: Convexity analysis and efficient algorithms," *IEEE Trans. Wireless Commun.*, vol. 15, no. 8, pp. 5527–5540, Aug. 2016.

[19] C. She, C. Yang, and T. Q. S. Quek, "Cross-layer optimization for ultra-reliable and low-latency radio access networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 127–141, Jan. 2018.

[20] ——, “Joint uplink and downlink resource configuration for ultra-reliable and low-latency communications,” *IEEE Trans. Commun.*, vol. 66, no. 5, pp. 2266–2280, May 2018.

[21] Y. Hu, M. Ozmen, M. C. Gursoy, and A. Schmeink, “Optimal power allocation for QoS-constrained downlink multi-user networks in the finite blocklength regime,” *IEEE Transactions on Wireless Communications*, vol. 17, no. 9, pp. 5827–5840, Sep. 2018.

[22] B. Montazeri, Y. Li, M. Alizadeh, and J. Ousterhout, “Homa: A receiver-driven low-latency transport protocol using network priorities,” in *Proc. ACM SIGCOMM*, 2018. [Online]. Available: https://arxiv.org/pdf/1803.09615.pdf

[23] M. Harchol-Balter, *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*. Cambridge University Press, 2013.

[24] A. P. Zwart and O. J. Boxma, “Sojourn time asymptotics in the M/G/1 processor sharing queue,” *Queueing Systems*, vol. 35, pp. 141–166, 2000.

[25] L. Zhao, X. Chi, and Y. Zhu, “Martingales-based energy-efficient D-ALOHA algorithms for MTC networks with delay-insensitive/URLLC terminals co-existence,” *IEEE Internet of Things J.*, vol. 5, no. 2, pp. 1285–1298, Apr. 2018.

[26] S. Schiessl, J. Gross, and H. Al-Zubaidy, “Delay analysis for wireless fading channels with finite blocklength channel coding,” in *Proc. ACM MSWiM*, 2015.

[27] E. Sisinni, A. Saifullah, S. Han *et al.*, “Industrial internet of things: Challenges, opportunities, and directions,” *IEEE Trans Ind. Informat.*, vol. 14, no. 11, pp. 4724–4734, Nov. 2018.

[28] X. Li, D. Li, J. Wan *et al.*, “Adaptive transmission optimization in SDN-based industrial internet of things with edge computing,” *IEEE Internet of Things J.*, vol. 5, no. 3, pp. 1351–1360, Jun. 2018.

[29] N. B. Long, H. Tran-Dang, and D.-S. Kim, “Energy-aware real-time routing for large-scale industrial internet of things,” *IEEE Internet of Things J.*, vol. 5, no. 3, pp. 2190–2199, Jun. 2018.

[30] M. Aazam, K. A. Harras, and S. Zeadally, “Fog computing for 5G tactile industrial internet of things: QoE-aware resource allocation model,” *IEEE Trans Ind. Informat., early access*, 2019.

[31] Y. Zhong, X. Ge, H. H. Yang, T. Han, and Q. Li, “Traffic matching in 5G ultra-dense networks,” *IEEE Commun. Mag.*, vol. 56, no. 8, pp. 100–105, Aug. 2018.

[32] S. Bi and Y. J. Zhang, “Computation rate maximization for wireless powered mobile-edge computing with binary computation offloading,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 4177–4190, Jun. 2018.

[33] H. Guo, J. Liu, J. Zhang *et al.*, “Mobile-edge computation offloading for ultradense IoT networks,” *IEEE Internet of Things J.*, vol. 5, no. 6, pp. 4977–4988, Dec. 2018.

[34] H. Guo, J. Zhang, J. Liu, and H. Zhang, “Energy-aware computation offloading and transmit power allocation in ultra-dense IoT networks,” *IEEE Internet of Things J., early access*, 2019.

[35] B. P. Rimal, D. P. Van, and M. Maier, “Cloudlet enhanced fiber-wireless access networks for mobile-edge computing,” *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 3601–3618, Jun. 2017.

[36] H. Guo and J. Liu, “Collaborative computation offloading for multiaccess edge computing over fibercwireless networks,” *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4514–4526, May 2018.

[37] J. Liu, Y. Mao, J. Zhang, and K. B. Letaief, “Delay-optimal computation task scheduling for mobile-edge computing systems,” in *Proc. IEEE ISIT*, 2016.

[38] C.-F. Liu, M. Bennis, and H. V. Poor, “Latency and reliability-aware task offloading and resource allocation for mobile edge computing,” in *IEEE Globecom Workshops*, 2017.

[39] J. Liu and Q. Zhang, “Offloading schemes in mobile edge computing for ultra-reliable low latency communications,” *IEEE Access*, vol. 6, pp. 12 825–12 837, 2018.

[40] S.-W. Ko, K. Han, and K. Huang, “Wireless networks for mobile edge computing: Spatial modelling and latency analysis,” *IEEE Trans. Wireless Commun., early access*, 2018.

[41] H. A. Omar, W. Zhuang, A. Abdrabou, and L. Li, “Performance evaluation of VeMAC supporting safety applications in vehicular networks,” *IEEE Trans. Emerg. Topics Comput.*, vol. 1, no. 1, pp. 69–83, Aug. 2013.

[42] S. A. Ashraf, I. Aktas, E. Eriksson, K. W. Helmersson, and J. Ansari, “Ultra-reliable and low-latency communication for wireless factory automation: From LTE to 5G,” in *Proc. IEEE Emerg. Tech. and Factory Automation (ETFA)*, 2016.

[43] Z. Hou, C. She, Y. Li, T. Q. Quek, and B. Vucetic, “Burstiness-aware bandwidth reservation for ultra-reliable and low-latency communications in tactile internet,” *IEEE J. Select. Areas Commun.,*, vol. 36, no. 11, pp. 2401–2410, Nov. 2018.

[44] K. Sriram and W. Whitt, “Characterizing superposition arrival processes in packet multiplexers for voice and data,” *IEEE J. Sel. Areas Commun.*, vol. 4, no. 6, pp. 833–846, Sep. 1986.

[45] G. Durisi, T. Koch, and P. Popovski, “Toward massive, ultra-reliable, and low-latency wireless communication with short packets,” *Proc. IEEE*, vol. 104, no. 9, pp. 1711–1726, Aug. 2016.

[46] K. Cheng, Y. Teng, W. Sun, A. Liu, and X. Wang, “Energy-efficient joint offloading and wireless resource allocation strategy in multi-mec server systems,” in *Proc. IEEE ICC*, 2018.

[47] C. You, Y. Zeng, R. Zhang, and K. Huang, “Asynchronous mobile-edge computation offloading: Energy-efficient resource management,” 2018. [Online]. Available: https://arxiv.org/pdf/1801.03668.pdf

[48] X. Chen, H. Zhang, C. Wu, S. Mao, Y. Ji, and M. Bennis, “Performance optimization in mobile-edge computing via deep reinforcement learning,” 2018. [Online]. Available: https://arxiv.org/pdf/1804.00514v1.pdf

[49] A. Gravey, J.-R. Louvion, and P. Boyer, “On the Geo/D/1 and Geo/D/1/n queues,” *Performance Evaluation*, vol. 11, no. 2, pp. 117–125, Jul. 1990.

[50] J. Östman, G. Durisi, E. G. Ström, J. Li, H. Sahlin, and G. Liva, “Low-latency ultra-reliable 5G communications: Finite-blocklength bounds and coding schemes,” in *Int. ITG Conf. on Syst., Commun. and Coding*, Feb. 2017.

[51] C. She, Z. Chen, C. Yang, T. Q. S. Quek, Y. Li, and B. Vucetic, “Improving network availability of ultra-reliable and low-latency communications with multi-connectivity,” *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5482–5496, Nov. 2018.

[52] D. Feng, C. She, K. Ying *et al.*, “Toward ultra-reliable low-latency communications: Typical scenarios, possible solutions, and open issues,” *IEEE Veh. Tech. Mag., early access*, 2019.

[53] C. She, C. Yang, and L. Liu, “Energy-efficient resource allocation for MIMO-OFDM systems serving random sources with statistical QoS requirement,” *IEEE Trans. Commun.*, vol. 63, no. 11, pp. 4125–4141, Nov. 2015.

[54] S. Boyd and L. Vandanberghe, *Convex Optimization*. Cambridge Univ. Press, 2004.

[55] 3GPP, *LTE; Evolved Universal Terrestrial Radio Access*. ETSI TR 36.931 v9.0.0, May 2011.