Latency and Reliability-Aware Task Offloading and Resource Allocation for Mobile Edge Computing

Chen-Feng Liu*, Mehdi Bennis*, and H. Vincent Poor†
*Centre for Wireless Communications, University of Oulu, Finland
†Department of Electrical Engineering, Princeton University, NJ, USA

Abstract—While mobile edge computing (MEC) alleviates the computation and power limitations of mobile devices, additional latency is incurred when offloading tasks to remote MEC servers. In this work, the power-delay tradeoff in the context of task offloading is studied in a multi-user MEC scenario. In contrast with current system designs relying on average metrics (e.g., the average queue length and average latency), a novel network design is proposed in which latency and reliability constraints are taken into account. This is done by imposing a probabilistic constraint on users' task queue lengths and invoking results from *extreme value theory* to characterize the occurrence of low-probability events in terms of queue length (or queuing delay) violation. The problem is formulated as a computation and transmit power minimization subject to latency and reliability constraints, and solved using tools from Lyapunov stochastic optimization. Simulation results demonstrate the effectiveness of the proposed approach, while examining the power-delay tradeoff and required computational resources for various computation intensities.

Index Terms—5G, mobile edge computing, fog networking and computing, ultra-reliable low latency communications (URLLC), extreme value theory.

I. Introduction

Mobile edge/fog computing (MEC) is an architecture that distributes computation, communication, control, and storage at the network edge [1]–[3]. This work is motivated by recent advances in MEC (or fog computing) and surging traffic demands spurred by online video and Internet-of-things applications, including machine type and mission-critical communication (e.g., augmented/virtual reality (AR/VR) and drones). While today's communication networks have been engineered with a focus on boosting network capacity, little attention has been paid to latency and reliability performance. Indeed ultra-reliable and low latency communication (URLLC) is one of the most important enablers of 5G and is currently receiving significant attention in both academia and industry [4].

When executed at mobile devices, the performance and quality of experience of computation-intensive applications are significantly affected by the devices' limited computation capabilities. Additionally, intensive computations are energy-consuming which severely shortens the lifetime of battery-limited devices. To address the computation and energy issues, mobile devices can wirelessly offload their tasks to proximal MEC servers. On the flip side, offloading tasks incurs extra latency which cannot be overlooked and should be factored into the system design.

A. Related Work

The energy-delay tradeoff has received significant attention in MEC networks [5]–[7]. Kwak *et al.* studied an energy min- imization problem for local computation, task offloading, and wireless transmission [5]. Therein, a dynamic task offloading and resource allocation problem was solved using stochastic optimization. Taking into account energy consumption, the authors in [6] investigated the energy-delay tradeoff of an MEC system in which the user equipment (UE) is endowed with a multi-core central processing unit (CPU) to compute different applications simultaneously. In [5] and [6], one task type is offloaded to and computed at the single-core CPU MEC server. Furthermore, assuming a multi-core CPU server to compute different users' offloaded tasks in parallel, Mao *et al.* studied a multi-user task offloading and bandwidth allocation problem, in which the energy-delay tradeoff is investigated using the Lyapunov framework [7].

While considering the task queue length as a delay metric, the works in [5]–[7] only ensure a finite average queue length as time evolves, i.e., mean rate stability [8]. Nevertheless, relying on mean rate stability is insufficient for URLLC applications. Additionally, the authors in [7] assumed that the number of users is smaller than the number of server CPU cores, and hence users' offloaded tasks are computed immediately and simultaneously. However, if the number of task-offloading users is large, some offloaded tasks need to wait for the available computational resources, incurring latency. In this case, the waiting time for task computing at the server cannot be ignored and should be taken into account.

B. Our Contribution

In this work, resource-limited devices seek to offload their computations to nearby MEC servers with multiple CPU cores, while taking into account devices' and servers' computation capabilities, co-channel interference, queuing latency, and reliability constraints. The fundamental problem under consideration is cast as *a network-wide power minimization problem for task computation and offloading, subject to delay requirements and reliability constraints*. Enabling URLLC mandates a departure from expected utility-based approaches in which relying on average quantities (e.g., average throughput and mean response time) is no longer an option. Therefore, motivated by the aforementioned shortcomings, we explicitly model and characterize the higher-order statistics of UEs' task queue lengths and impose a probabilistic constraint on the delay/queue length deviation. Further, the queue length bound deviation is directly related to the URLLC constraint which we examine through the lens of *extreme value theory* [9]. Regarding the offloaded tasks at the servers, we take into account the waiting time for the available computational re-

sources and accordingly impose a probabilistic delay requirement. Subsequently, invoking tools from Lyapunov stochastic optimization, we propose a dynamic policy for local task execution, task offloading, and resource allocation subject to latency and reliability constraints. Numerical results highlight the effectiveness of the proposed framework while shedding light onto the power-delay tradeoff, and the scaling in terms of servers' computational resources according to the computation intensity.

II. SYSTEM MODEL

As shown in Fig. 1, we consider a set $\mathcal{U}$ of UEs with local computation capability in an MEC network. Due to devices' limited computation capabilities, a set $\mathcal{S}$ of MEC servers with N-core CPUs is deployed to help offload and compute UEs' tasks, where each UE wirelessly accesses multiple servers and dynamically offloads its tasks. The wireless channel gain between UE $i \in \mathcal{U}$ and server $j \in \mathcal{S}$ is denoted by h_{ij} which includes path loss and channel fading. Based on channel strength, we further assume that each UE $i \in \mathcal{U}$ accesses the set $\mathcal{S}_i$ of servers whose expected channel gains are higher than a threshold value h_i^{th}, i.e., $\mathcal{S}_i = \{j \in \mathcal{S} | \mathbb{E}[h_{ij}] \geq h_i^{\text{th}}\}$. For each server $j \in \mathcal{S}$, the set of accessing UEs is denoted by $\mathcal{U}_j$. Additionally, all channels experience block fading. We index the coherence block by $t \in \{0, 1, \cdots\}$, and each block is of unit time length for simplicity.

A. Traffic Model at the UE Side

Tasks arrive in a stochastic manner following an arbitrary probability distribution. We assume that most tasks' arrivals can be computed in one time slot while the large-size tasks are divided into small sub-tasks and computed in one time slot [6]. Hereafter, the terminology "*task*" refers to both the task which can be computed in one time slot and the divided sub-task. Each task can be computed at the UE or at the server, following the data-partition model [2]. We assume that each offloaded task is assigned to only one server. Moreover, each UE has a queue buffer to store the tasks' arrivals. Denoting the task queue length of UE $i \in \mathcal{U}$ in time slot t as $Q_i(t)$, the queue length (in the unit of bits) evolves as

$$Q_i(t+1) = \max\{Q_i(t) + A_i(t) - B_i(t), 0\}, \quad (1)$$

where the task arrivals $A_i(t)$ with mean value λ_i are independent and identically distributed over time. Additionally,

$$B_i(t) = \frac{f_i(t)}{L_i} + \sum_{j \in \mathcal{S}_i} R_{ij}(t) \quad (2)$$

is the task completion rate in time slot t in which

$$R_{ij}(t) = \frac{W}{|\mathcal{S}|} \log_2\left(1 + \frac{P_{ij}(t)h_{ij}(t)}{\frac{N_0 W}{|\mathcal{S}|} + \sum_{i' \in \mathcal{U}_j \setminus i} P_{i'j}(t)h_{i'j}(t)}\right) \quad (3)$$

is the transmission rate[1] for task offloading with transmit power $P_{ij}(t)$ from UE $i \in \mathcal{U}$ to server $j \in \mathcal{S}_i$, while $f_i(t)/L_i$ accounts for the local computation rate of UE i with CPU-cycle frequency $f_i(t)$. Here, L_i denotes the required CPU

[1] We assume reception in additive white Gaussian noise.
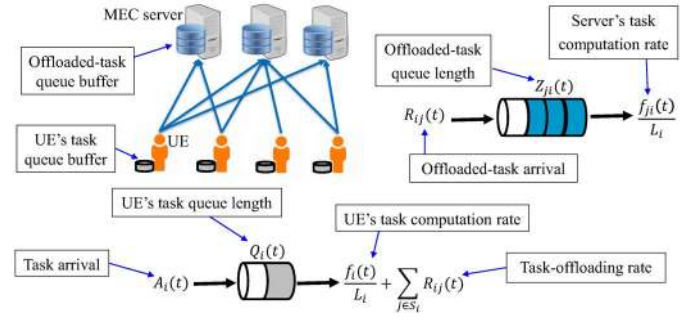

Figure 1. System model.

cycles per bit for computation, i.e., the processing density, which is application-dependent [5]. W and N_0 are the total system bandwidth and the power spectral density of the additive white Gaussian noise, respectively. Furthermore, the total bandwidth W is orthogonally and equally allocated to different servers whereas UEs share the allocated band of the accessed server for task offloading.

We note that given a CPU-cycle frequency $f_i(t)$, the power consumption of local computation is calculated as $\kappa[f_i(t)]^3$ in which the parameter κ depends on the hardware architecture [7], [10]. In order to minimize local computation power, the UE adopts the dynamic voltage and frequency scaling (DVFS) capability to adaptively adjust its CPU-cycle frequency [2], [10]. Thus, for local computational resource and transmit power allocation, we impose the following constraints for each UE $i \in \mathcal{U}$:

$$\begin{cases} 0 \leq f_i(t) \leq f_i^{\max}, \\ \sum_{j \in \mathcal{S}_i} P_{ij}(t) \leq P_i^{\max}, \\ P_{ij}(t) \geq 0, \ \forall j \in \mathcal{S}_i, \end{cases} \quad (4)$$

where $f_i^{\max}$ and $P_i^{\max}$ are the maximum local computation capability and transmit power budget, respectively.

B. Traffic Model at the Server Side

Each MEC server has multiple parallel buffers to process the UEs' offloaded tasks. In this regard, we denote the offloaded-task queue length (in units of bits) for UE $i \in \mathcal{U}_j$ at server $j \in \mathcal{S}$ in time slot t as $Z_{ji}(t)$. The evolution of $Z_{ji}(t)$ is given by

$$Z_{ji}(t+1) \leq \max\left\{Z_{ji}(t) + R_{ij}(t) - \frac{f_{ji}(t)}{L_i}, 0\right\}, \quad (5)$$

where $f_{ji}(t)$ is server j's allocated CPU-cycle frequency for UE i's offloaded tasks. Since the MEC server provides a faster computation capability, each CPU core is dedicated to at most one UE (i.e., its offloaded tasks) per time slot. Additionally, we assume that at each server, a given UE's offloaded tasks can only be computed by one CPU core for simplicity. Thus, each server $j \in \mathcal{S}$ schedules its computation and allocates resources based on the constraints

$$\begin{cases} \sum_{i \in \mathcal{U}_j} \{f_{ji}(t) > 0\} \leq N, \\ f_{ji}(t) \in \{0, f_j^{\max}\}, \ \forall i \in \mathcal{U}_j, \end{cases} \quad (6)$$

in which $f_j^{\max}$ is server j's computation capability of one CPU-core.

III. Latency Requirements and Reliability Constraints

In this work, the end-to-end delay is composed of the following elements: 1) queuing delay at the UE and the server, 2) computation delay at the UE and the server, and 3) wireless transmission delay for task offloading. According to Little's law, the average queuing delay is proportional to the average queue length [11]. Nevertheless, without considering the probability distribution of the queue length, relying merely on the average queue length fails to account for the low-latency and reliability constraints. To tackle this, we take into account the statistics of the UE's task queue length and impose a probabilistic constraint

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \Pr\left(Q_i(t) > d_i\right) \leq \epsilon_i, \ \forall i \in \mathcal{U}, \qquad (7)$$

on the task queue length, where d_i and $\epsilon_i \ll 1$ are the queue length bound and tolerable violation probability, respectively. Furthermore, the queue length/queuing delay bound violation also undermines the reliability of task computing. For example, if a finite-size queue buffer is over-loaded, the incoming tasks will be dropped.

Let us further focus on $\bar{F}_{Q_i}(q_i) = \Pr\left(Q_i > q_i\right)$ in (7), i.e., the complementary cumulative distribution function (CCDF) of the UE's queue length, which reflects the queuing latency profile. If the monotonically decreasing CCDF decays faster while increasing q_i, the probability of having an extreme latency value is lower. Since the focus of this work is on extreme case events which happen with very low probabilities, i.e., $\Pr\left(Q_i > d_i\right) \ll 1$, we resort to principles of extreme value theory[2] to characterize the probability distribution of these extreme events and the extreme tails of distributions. Towards this goal, we first introduce a random variable Q whose cumulative distribution function (CDF) is denoted by $F_Q(q)$ and a threshold value d. Then, conditioned on $Q > d$, the conditional CDF of the excess value $X = Q - d > 0$ is given by

$$F_{X|Q>d}(x) = \Pr(Q - d \leq x | Q > d) = \frac{F_Q(x+d) - F_Q(d)}{1 - F_Q(d)}.$$

Theorem 1 (Pickands–Balkema–de Haan theorem for exceedances over thresholds [9]). *Consider the distribution of Q conditioned on exceeding some high threshold d. As the threshold d closely approaches $F_Q^{-1}(1)$, i.e., $d \to \sup\{q : F_Q(q) < 1\}$, the conditional CDF of the excess value $X = Q - d > 0$ is*

$$F_{X|Q>d}(x) \approx G(x; \sigma, \xi) = \begin{cases} 1 - \left(\max\left\{1 + \frac{\xi x}{\sigma}, 0\right\}\right)^{-1/\xi}, \\ \qquad\qquad\qquad\qquad when\ \xi \neq 0, \\ 1 - e^{-x/\sigma}, \ when\ \xi = 0. \end{cases}$$

Here, $G(x; \sigma, \xi)$ is the generalized Pareto distribution (GPD) whose mean and variance are $\sigma/(1 - \xi)$ and $\frac{\sigma^2}{(1-\xi)^2(1-2\xi)}$, respectively. Moreover, the characteristics of the GPD depend on the scale parameter $\sigma > 0$ and the shape parameter $\xi < 1/2$.

In other words, Theorem 1 shows that for sufficiently high threshold d, the distribution function of the excess may be approximated by the GPD. Now, let us define the conditional excess queue value (with respect to the threshold d_i in (7)) of each UE $i \in \mathcal{U}$ in time slot t as $X_i(t)|_{Q_i(t)>d_i} = Q_i(t) - d_i$. Then, after applying Theorem 1, we obtain the following approximations:

$$\mathbb{E}\left[X_i(t)|Q_i(t) > d_i\right] \approx \frac{\sigma_i}{1-\xi_i}, \qquad (8)$$

$$\mathrm{Var}\left(X_i(t)|Q_i(t) > d_i\right) \approx \frac{\sigma_i^2}{(1-\xi_i)^2(1-2\xi_i)}, \qquad (9)$$

with a scale parameter σ_i and a shape parameter ξ_i. Note that the smaller σ_i and ξ_i are, the smaller are the mean value and variance of the GPD. Hence, we impose thresholds on the scale parameter and the shape parameter, i.e., $\sigma_i \leq \sigma_i^{\mathrm{th}}$ and $\xi_i \leq \xi_i^{\mathrm{th}}$. Subsequently, applying the two parameter thresholds and $\mathrm{Var}(X_i) = \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2$ to (8) and (9), we consider the constraints for the time-averaged mean and second moment of the conditional excess queue value,

$$\bar{X}_i = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[X_i(t)|Q_i(t) > d_i\right] \leq \frac{\sigma_i^{\mathrm{th}}}{1-\xi_i^{\mathrm{th}}}, \qquad (10)$$

$$\bar{Y}_i = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[Y_i(t)|Q_i(t) > d_i\right] \leq \frac{2(\sigma_i^{\mathrm{th}})^2}{(1-\xi_i^{\mathrm{th}})(1-2\xi_i^{\mathrm{th}})}, \qquad (11)$$

where $Y_i(t) := [X_i(t)]^2$.

Likewise, the average queuing latency at the server is proportional to the ratio of the average queue length to the average offloading rate. Thus, analogous to (7), we consider the constraint

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \Pr\left(\frac{Z_{ji}(t)}{\tilde{R}_{ij}(t-1)} > d_{ji}\right) \leq \epsilon_{ji}, \ \forall i \in \mathcal{U}_j, j \in \mathcal{S}, \quad (12)$$

for the offloaded-task queue length, where $\tilde{R}_{ij}(t - 1) = \frac{1}{t} \sum_{\tau=0}^{t-1} R_{ij}(\tau)$ is the moving time-averaged offloading rate, d_{ji} is the latency bound, and $\epsilon_{ji} \ll 1$ is the tolerable violation probability. Additionally, for the offloaded-task queue length of UE $i \in \mathcal{U}_j$ at server $j \in \mathcal{S}$, we define the conditional exceedance in time slot t as $X_{ji}(t)|_{Z_{ji}(t)>\tilde{R}_{ij}(t-1)d_{ji}} = Z_{ji}(t) - \tilde{R}_{ij}(t-1)d_{ji}$ with respect to the threshold $\tilde{R}_{ij}(t-1)d_{ji}$. Similar to the task queue length at the user's side, we have the constraints for the conditional exceedance of the offloaded-task queue length, i.e., $\forall i \in \mathcal{U}_j, j \in \mathcal{S}$,

$$\bar{X}_{ji} = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[X_{ji}(t)|Z_{ji}(t) > \tilde{R}_{ij}(t-1)d_{ji}\right] \leq \frac{\sigma_{ji}^{\mathrm{th}}}{1-\xi_{ji}^{\mathrm{th}}}, \qquad (13)$$

$$\bar{Y}_{ji} = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[Y_{ji}(t)|Z_{ji}(t) > \tilde{R}_{ij}(t-1)d_{ji}\right]$$
$$\leq \frac{2(\sigma_{ji}^{\mathrm{th}})^2}{(1-\xi_{ji}^{\mathrm{th}})(1-2\xi_{ji}^{\mathrm{th}})}, \qquad (14)$$

with $Y_{ji}(t) := [X_{ji}(t)]^2$ and the thresholds $\sigma_{ji}^{\mathrm{th}}$ and ξ_{ji}^{th} for the approximated GPD.

[2]Extreme value theory is a powerful and robust framework to study the tail behavior of a distribution. Extreme value theory also provides statistical models for the computation of extreme risk measures [9].

$$Q_i^{(X)}(t+1) = \max\left\{Q_i^{(X)}(t) + \left(X_i(t+1) - \frac{\sigma_i^{\text{th}}}{1-\xi_i^{\text{th}}}\right) \times \ \{Q_i(t+1) > d_i\}, 0\right\}, \tag{18}$$

$$Q_i^{(Y)}(t+1) = \max\left\{Q_i^{(Y)}(t) + \left(Y_i(t+1) - \frac{2(\sigma_i^{\text{th}})^2}{(1-\xi_i^{\text{th}})(1-2\xi_i^{\text{th}})}\right) \times \ \{Q_i(t+1) > d_i\}, 0\right\}, \tag{19}$$

$$Q_{ji}^{(X)}(t+1) = \max\left\{Q_{ji}^{(X)}(t) + \left(X_{ji}(t+1) - \frac{\sigma_{ji}^{\text{th}}}{1-\xi_{ji}^{\text{th}}}\right) \times \ \{Z_{ji}(t+1) > \tilde{R}_{ji}(t)d_{ji}\}, 0\right\}, \tag{20}$$

$$Q_{ji}^{(Y)}(t+1) = \max\left\{Q_{ji}^{(Y)}(t) + \left(Y_{ji}(t+1) - \frac{2(\sigma_{ji}^{\text{th}})^2}{(1-\xi_{ji}^{\text{th}})(1-2\xi_{ji}^{\text{th}})}\right) \times \ \{Z_{ji}(t+1) > \tilde{R}_{ji}(t)d_{ji}\}, 0\right\}, \tag{21}$$

$$Q_i^{(Q)}(t+1) = \max\left\{Q_i^{(Q)}(t) + \ \{Q_i(t+1) > d_i\} - \epsilon_i, 0\right\}, \tag{22}$$

$$Q_{ji}^{(Z)}(t+1) = \max\left\{Q_{ji}^{(Z)}(t) + \ \{Z_{ji}(t+1) > \tilde{R}_{ji}(t)d_{ji}\} - \epsilon_{ji}, 0\right\}. \tag{23}$$

At the UE, the computation delay and transmission delay for task offloading are inversely proportional to $f_i(t)$ and $\sum_{j\in\mathcal{S}_i} R_{ij}(t)$, respectively, as per (2). To decrease the computation delay, the UE should allocate more local CPU-cycle frequency while computing tasks locally. Analogously, more transmit power is allocated in order to decrease transmission delay while offloading the task. Since the queue length can also be further reduced by allocating higher CPU-cycle frequency or/and transmit power, both delays are implicitly taken into account in the queue length constraints (7), (10), and (11). On the other hand, allocating more computation capability or/and transmit power depletes the UE's battery quickly. Therefore, the power-delay tradeoff is crucial and will be investigated in the numerical results section. Regarding the computation delay at the server, this is negligible since one CPU-core with better computation capability is dedicated for one UE's offloaded tasks. Nevertheless, since queuing delay at the server with the requirements (12)-(14) is considered, the server needs to schedule computation and allocate CPU cores when the number of accessing UEs is larger than number of the server's CPU cores. Taking into account the latency requirements and reliability constraints, we propose a dynamic task offloading and resource allocation approach that minimizes the UEs' power consumption.

IV. LATENCY AND RELIABILITY-AWARE TASK OFFLOADING AND RESOURCE ALLOCATION

Denoting the network-wide power allocation vector and the computational resource allocation vector of server $j \in \mathcal{S}$ as $\mathbf{P}(t) = (P_{ij}(t), i \in \mathcal{U}, j \in \mathcal{S}_i)$ and $\mathbf{f}_j(t) = (f_{ji}(t), i \in \mathcal{U}_j, j \in \mathcal{S})$, respectively, our studied optimization problem is formulated as follows:

$$\underset{\mathbf{P}(t),f_i(t),\mathbf{f}_j(t)}{\text{minimize}} \quad \sum_{i\in\mathcal{U}} \bar{P}_i \tag{15}$$

$$\text{subject to} \quad \text{(4) and (6) for resource allocation,}$$
$$\text{(7) and (12) for delay bound violation,}$$
$$\text{(10), (11), (13), and (14) for the GPD,}$$

where $\bar{P}_i = \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1}\left(\kappa[f_i(t)]^3 + \sum_{j\in\mathcal{S}_i} P_{ij}(t)\right)$ is UE i's long-term time average power consumption. To find the optimal task offloading and computational resource allocation

policy for problem (15), we resort to techniques from Lyapunov stochastic optimization.

A. Lyapunov Optimization Framework

We first rewrite (7) and (12) as

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\ \{Q_i(t) > d_i\}\big] \leq \epsilon_i, \tag{16}$$

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\ \{Z_{ji}(t) > \tilde{R}_{ji}(t-1)d_{ji}\}\big] \leq \epsilon_{ji}, \tag{17}$$

respectively, where $\{\cdot\}$ is the indicator function. In order to ensure all time-averaged constraints (10), (11), (13), (14), (16), and (17), we respectively introduce the corresponding virtual queues (18)–(23) with the dynamics shown on the top of this page. Then, letting $\mathbf{Q}_u(t) := \big(Q_i^{(X)}(t), Q_i^{(Y)}(t), Q_i^{(Q)}(t), i \in \mathcal{U}\big)$ and $\mathbf{Q}_s(t) := \big(Q_{ji}^{(X)}(t), Q_{ji}^{(Y)}(t), Q_{ji}^{(Z)}(t), i \in \mathcal{U}_j, j \in \mathcal{S}\big)$ denote the combined queue vectors for notational simplicity, we express the conditional Lyapunov drift-plus-penalty for slot t as

$$\mathbb{E}\Big[\mathcal{L}(\mathbf{Q}(t+1)) - \mathcal{L}(\mathbf{Q}(t)) + \sum_{i\in\mathcal{U}} VP_i(t)\Big|\mathbf{Q}_u(t), \mathbf{Q}_s(t)\Big], \tag{24}$$

where $\mathcal{L}(\mathbf{Q}(t)) = \frac{1}{2}\sum_{i\in\mathcal{U}}\big[\big(Q_i^{(X)}(t)\big)^2 + \big(Q_i^{(Y)}(t)\big)^2 + \big(Q_i^{(Q)}(t)\big)^2\big] + \frac{1}{2}\sum_{j\in\mathcal{S}}\sum_{i\in\mathcal{U}_j}\big[\big(Q_{ji}^{(X)}(t)\big)^2 + \big(Q_{ji}^{(Y)}(t)\big)^2 + \big(Q_{ji}^{(Z)}(t)\big)^2\big]$ is the Lyapunov function. $V \geq 0$ is a parameter that trades off power consumption and end-to-end latency. Subsequently, substituting $(\max\{x, 0\})^2 \leq x^2$ and all physical and virtual queue dynamics into (24), we obtain

$$(24) \leq C + \mathbb{E}\Big[-\sum_{i\in\mathcal{U}}\Big[B_i(t)\big(Q_i^{(X)}(t) + Q_i(t) + A_i(t)$$
$$+ 2Q_i^{(Y)}(t)\big(Q_i(t) + A_i(t)\big) + 2\big(Q_i(t) + A_i(t)\big)^3\big)$$
$$+ Q_i^{(Q)}(t)\Big] \times \ \{\max\{Q_i(t) + A_i(t) - B_i(t), 0\} > d_i\}$$
$$+ \sum_{j\in\mathcal{S}}\sum_{i\in\mathcal{U}_j}\Big[\Big(R_{ij}(t) - \frac{f_{ji}(t)}{L_i}\Big)\big(Q_{ji}^{(X)}(t) + Z_{ji}(t)$$
$$+ 2Q_{ji}^{(Y)}(t)Z_{ji}(t) + 2[Z_{ji}(t)]^3\big) + Q_{ji}^{(Z)}(t)\Big]$$
$$\times \ \Big\{\max\Big\{Z_{ji}(t) + R_{ij}(t) - \frac{f_{ji}(t)}{L_i}, 0\Big\} > \tilde{R}_{ji}(t)d_{ji}\Big\}$$
$$+ \sum_{i\in\mathcal{U}} V\Big(\kappa[f_i(t)]^3 + \sum_{j\in\mathcal{S}_i} P_{ij}(t)\Big)\Big|\mathbf{Q}_u(t), \mathbf{Q}_s(t)\Big]. \tag{25}$$

Due to space limitations, we omit the details of the constant C which does not affect the system performance. Note that the solution to problem (15) can be obtained by minimizing the upper bound in (25) [8]. To this end, we have three optimization problems to be solved in each time slot: two optimization problems for local computation and task offloading at the UE side while the other problem is at the server side.

B. Task Computation and Offloading at the UE Side

For local computation, UE $i \in \mathcal{U}$ allocates the CPU-cycle frequency by solving

$$\underset{0 \le f_i(t) \le f_i^{\max}}{\text{minimize}} \quad V\kappa[f_i(t)]^3 - a_i(t)f_i(t)/L_i \quad (26)$$

with $a_i(t) = Q_i^{(Q)}(t) + Q_i(t) + A_i(t) + [Q_i^{(X)}(t) + Q_i(t) + A_i(t) + 2Q_i^{(Y)}(t)(Q_i(t) + A_i(t)) + 2(Q_i(t) + A_i(t))^3] \times \{Q_i(t) + A_i(t) > d_i\}$. The optimal solution to (26), i.e., $f_i^*(t) = \min\left\{\sqrt{\frac{a_i(t)}{3V\kappa L_i}}, f_i^{\max}\right\}$, can be obtained by differentiation.

The second problem for task offloading is given as follows:

$$\underset{\mathbf{P}(t)}{\text{minimize}} \quad \sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{S}_i}\left[VP_{ij}(t) + \left(b_{ji}(t) - a_i(t)\right)R_{ij}(t)\right] \quad (27a)$$

$$\text{subject to} \quad \sum_{j\in\mathcal{S}_i} P_{ij}(t) \le P_i^{\max}, \ \forall i \in \mathcal{U}, \quad (27b)$$

$$P_{ij}(t) \ge 0, \ \forall i \in \mathcal{U}, j \in \mathcal{S}_i, \quad (27c)$$

which jointly decides all UEs' task offloading policies. Here, $b_{ji}(t) = Q_{ji}^{(Z)}(t) + Z_{ji}(t) + [Q_{ji}^{(X)}(t) + Z_{ji}(t) + 2Q_{ji}^{(Y)}(t)Z_{ji}(t) + 2[Z_{ji}(t)]^3] \times \{Z_{ji}(t) + R_i^{\max} > \tilde{R}_{ji}(t-1)d_{ji}\}$, and $R_i^{\max}$ is UE i's maximum offloading rate. Solving (27) without a central controller, each UE requires other UEs' channel state information (CSI) and queue state information (QSI) as per (3) and (27a), which is impractical and incurs high overhead, especially in dense networks. In order to alleviate this problem, we decompose the summation (over all UEs) in the objective (27a) and reformulate (27) as $|\mathcal{U}|$ sub-problems, i.e., $\forall i \in \mathcal{U}$,

$$\underset{P_{ij}}{\text{minimize}} \quad \sum_{j\in\mathcal{S}_i} VP_{ij} + \sum_{j\in\mathcal{S}_i}\left(b_{ji} - a_i\right)$$
$$\times \mathbb{E}_{I_{ij}}\left[\frac{W}{|\mathcal{S}|}\log_2\left(1 + \frac{|\mathcal{S}|P_{ij}h_{ij}}{N_0W + |\mathcal{S}|I_{ij}}\right)\right] \quad (28a)$$

$$\text{subject to} \quad \sum_{j\in\mathcal{S}_i} P_{ij} \le P_i^{\max}, \quad (28b)$$

$$P_{ij} \ge 0, \ \forall j \in \mathcal{S}_i, \quad (28c)$$

which is solved at UE i in a decentralized way. Here, the expectation is with respect to the estimated distribution of the aggregate interference $I_{ij} = \sum_{i'\in\mathcal{U}_j\setminus i} P_{i'j}h_{i'j}$, and we suppress the time index t for simplicity. In the decomposed problem (28), each UE decides on the transmit power for task offloading without requiring other UEs' CSI and QSI. Subsequently, applying the Karush-Kuhn-Tucker (KKT) conditions to the convex optimization problem (28), UE i offloads tasks to server j with the optimal transmit power $P_{ij}^* > 0$ satisfying $\mathbb{E}_{I_{ij}}\left[\frac{(a_i - b_{ji})Wh_{ij}}{(N_0W + I_{ij}|\mathcal{S}| + P_{ij}^*h_{ij}|\mathcal{S}|)\ln 2}\right] = V + \gamma$

Algorithm 1 Computational resource allocation at the server.
1: Initialize $n = 1$ and $\tilde{\mathcal{U}} = \mathcal{U}_j$.
2: **while** $n \le N$ and $\tilde{\mathcal{U}} \ne \emptyset$ **do**
3:      $i^* = \text{argmax}_{i\in\tilde{\mathcal{U}}}\left\{b_{ji}(t)/L_i\right\}$.
4:      $f_{ji^*}(t) = f_j^{\max}$.
5:      Update $n \leftarrow n + 1$ and $\tilde{\mathcal{U}} = \tilde{\mathcal{U}} \setminus i^*$.
6: **end while**

Table I
SIMULATION PARAMETERS [5], [7], [12], [13].

Parameter	Value	Parameter	Value
N_0	-174 dBm/Hz	W	10 MHz
κ	10^{-27} Watt $\cdot$ s^3/cycle^3	$P^{\max}$	20 dBm
L_i	$\{737.5, 1760, 2640, 8250\}$	$f_i^{\max}$	10^9 cycle/s
λ_i	$\{0.1, 0.2, \cdots, 1.5\}$ Mbps	$f_j^{\max}$	10^{10} cycle/s
d_i	$4\lambda_i$ (bit)	ϵ_i	0.01
σ_i^{th}	$4\lambda_i$ (bit)	ξ_i^{th}	0.3
d_{ji}	20 sec	ϵ_{ji}	0.01
σ_{ji}^{th}	$4\tilde{R}_{ij}(\infty)$ (bit)	ξ_{ji}^{th}	0.3

if $\mathbb{E}_{I_{ij}}\left[\frac{(a_i - b_{ji})Wh_{ij}}{(N_0W + I_{ij}|\mathcal{S}|)\ln 2}\right] > V + \gamma$. Otherwise, $P_{ij}^* = 0$. The Lagrange multiplier γ is 0 if $\sum_{j\in\mathcal{S}_i} P_{ij}^* < P_i^{\max}$. When $\gamma > 0$, $\sum_{j\in\mathcal{S}_i} P_{ij}^* = P_i^{\max}$.

C. Computational Resource Allocation at the Server Side

At the server side, each MEC server $j \in \mathcal{S}$ allocates computational resources by solving the following optimization problem:

$$\underset{f_{ji}(t)}{\text{maximize}} \quad \sum_{i\in\mathcal{U}_j} b_{ji}(t)f_{ji}(t)/L_i \quad (29a)$$

$$\text{subject to} \quad \sum_{i\in\mathcal{U}_j}\{f_{ji}(t) > 0\} \le N, \quad (29b)$$

$$f_{ji}(t) \in \{0, f_j^{\max}\}, \ \forall i \in \mathcal{U}_j. \quad (29c)$$

To maximize (29a), the server equally dedicates its N CPU cores to the UEs with the N largest values of $b_{ji}(t)/L_i$. The solution to problem (29) is detailed in Algorithm 1.

After computing and offloading the tasks in time slot t, the UEs update the queue length in (1), (18), (19), and (22) while the MEC servers update (5), (20), (21), and (23), for the next slot $t + 1$. Moreover, based on the transmission rate $R_{ij}(t)$, the UE empirically estimates the statistics of I_{ij} for slot $t + 1$ as per $\hat{\text{Pr}}(\tilde{I}_{ij}; t+1) = \frac{\{I_{ij}(t) = \tilde{I}_{ij}\}}{t+2} + \frac{(t+1)\hat{\text{Pr}}(\tilde{I}_{ij}; t)}{t+2}$.

V. NUMERICAL RESULTS

We consider an MEC architecture in which four servers, each with 9 CPU cores, are uniformly deployed in a 100×100 m^2 indoor area. Additionally, 36 UEs are randomly distributed, where each UE associates with the nearest server, i.e., $|\mathcal{S}_i| = 1, \forall i \in \mathcal{U}$. For task offloading, we assume that the transmission frequency is 5.8 GHz with the path loss model $24\log x + 20\log 5.8 + 60$ (dB) [14], where x in meters is the distance between the transmitter and receiver. Further, all wireless channels experience Rayleigh fading with unit variance, with a coherence time of 40 ms [15]. Moreover, we
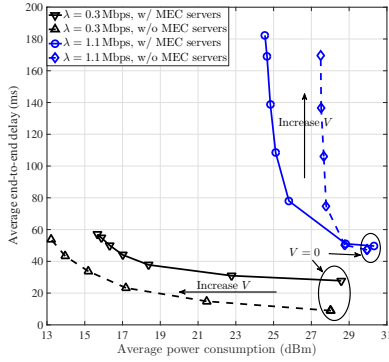
(a) $L = 737.5$ cycle/bit.
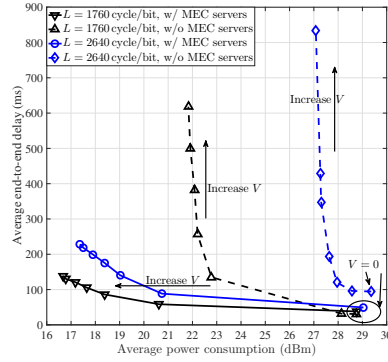
(b) $\lambda = 0.3$ Mbps.

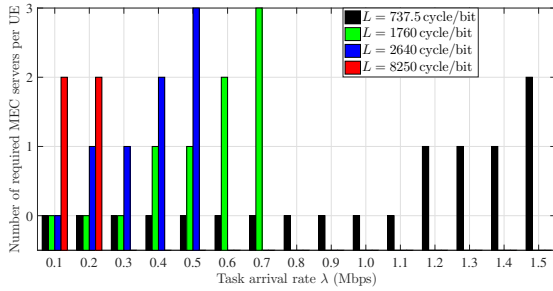Figure 2. Tradeoff between a UE's average power consumption and end-to-end delay.

Figure 3. Average end-to-end delay versus number of accessed MEC servers per UE.

Figure 4. Number of required MEC servers per UE for various processing densities and task arrival rates.

consider Poisson task arrivals. The remaining parameters for all UEs $i \in \mathcal{U}$ and servers $j \in \mathcal{S}$ are listed in Table I.[3]

We first show the tradeoff between a UE's average power consumption and end-to-end delay in Fig. 2. By varying the non-negative Lyapunov tradeoff parameter V, we obtain the tradeoff curve in which a small V induces a low end-to-end delay at the cost of higher power consumption. In contrast, a large V asymptotically minimizes the power consumption at the expense of a higher end-to-end latency. Given the light processing density and low task arrival rate, e.g., $L = 737.5$ cycle/bit and $\lambda = 0.3$ Mbps, the UE's computation capability is sufficient to execute all tasks locally. In this case, the UE's computation rate is higher than the transmission rate for offloading. Thus, consuming power for task offloading instead of computing locally incurs a worse delay performance. As the processing density increases, or tasks have a higher arrival rate, the UE is unable to execute tasks with limited computation capability. Here, consuming power for task offloading achieves a higher task completion rate while the server provides a faster computation capability. The benefit is more prominent when the processing density is more intense. To summarize, the MEC architecture achieves better power-delay tradeoff performance when UEs have high-computation tasks or higher task arrival rates.

[3]The four considered processing densities, L_i (cycle/bit), correspond to the applications of the English Wikipedia main page, 6-queen puzzle, 400-frame video game, and 7-queen puzzle, respectively [5], [13].
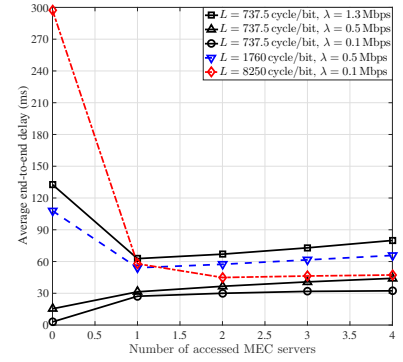
Subsequently, we focus on the performance of the average end-to-end delay and the statistics of the queue length in which the Lyapunov parameter is set to zero ($V = 0$ gives the lowest average end-to-end delay and queue length [8]). In Fig. 3, we first show the average end-to-end delay versus the number of UE accessed servers, i.e., $|\mathcal{S}_i|$, in which each UE $i \in \mathcal{U}$ accesses $|\mathcal{S}_i|$-nearest servers. As discussed in Fig. 2, the MEC-centric approach outperforms the local computation one without MEC, i.e., $|\mathcal{S}_i| = 0$, in terms of average end-to-end delay when the UE has high computation tasks or higher task arrival rate. Let us further consider the two settings, 1) $L = 737.5$ cycle/bit and $\lambda = 1.3$ Mbps, and 2) $L = 1760$ cycle/bit and $\lambda = 0.5$ Mbps, in Fig. 3. Although the UE can leverage more computational resources by accessing MEC servers, the multiple UE scenario incurs severe interference as well as longer waiting times. Therefore, accessing multiple servers does not decrease the end-to-end delay in these two settings. Nevertheless, if the processing is intense, the benefit of having access to multiple servers' resources outweighs these shortcomings. In Fig. 3, we can find that $|\mathcal{S}_i| = 2$ achieves the lowest average end-to-end delay for $L = 8250$ cycles/bit. Based on the results in Fig. 3, Fig. 4 further shows the number of required MEC servers per UE to achieve the lowest average end-to-end delay given that the task traffic demand with a given processing density can be satisfied (shown by a rectangular bar) in the simulated MEC architecture or local computation scenario.

Further considering the settings (of Fig. 3) in which the MEC-centric approach is superior to the local computation one, we vary the threshold δ of the probabilistic queuing delay (accounting for both the UE and server) requirement, $\Pr(\text{Queuing delay} > \delta) \leq 10^{-2}$, and plot in Fig. 5 the exact delay bound violation probability, i.e., reliability. As shown in Figs. 3 and 5, in addition to the average end-to-end delay, the MEC architecture achieves better reliability performance. In other words, even though there is an extra queuing delay at the server side, offloading tasks to the MEC servers reduces the waiting time for task execution when UEs have high-computation tasks or higher task arrival rates. Corresponding to the results in Fig. 3, the reliability enhancement is more
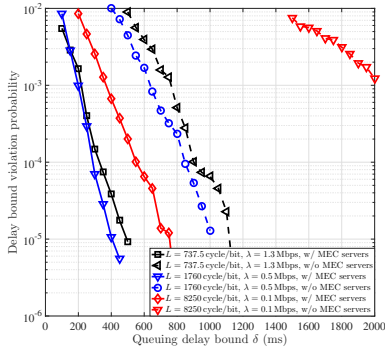
Figure 5. Delay bound violation probability versus queuing delay bound.
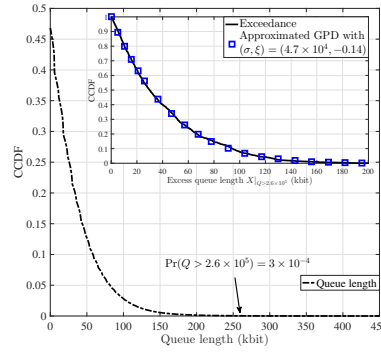

Figure 6. Tail distributions of a UE's task queue length, queue length exceedances over threshold, and the approximated GPD of exceedances.
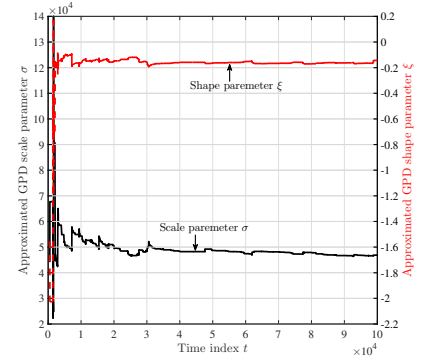

Figure 7. Convergence of the approximated GPD scale and shape parameters of exceedances.

prominent while the average end-to-end delay is reduced, e.g., the case $L = 8250$ cycles/bit.

Finally, let us investigate the statistics of the queue length. Due to space limitations, we only plot the UE's task queue length but note that the following illustration is applicable to the offloaded-task queue length at the server. Let us consider $\lambda_i = 1.3$ Mbps, $L_i = 737.5$ cycle/bit, $d_i = 2.6 \times 10^5$ bit, and $d_{ji} = 0.2$ sec, $\forall i \in \mathcal{U}, j \in \mathcal{S}$. Fig. 6 shows the tail distribution, i.e., CCDF, of the UE's queue length in which we not only ensure the probabilistic constraint on the queue length but also achieve a CCDF value, i.e., $\Pr(Q > 2.6 \times 10^5) = 3 \times 10^{-4}$, approaching zero. Applying Theorem 1 to the conditional excess queue value $X|_{Q>2.6\times10^5} = Q - 2.6 \times 10^5$, we show the tail distributions of the conditional excess queue value and the approximated GPD which coincide with each other. In Fig. 7, we show the convergence of the scale and shape parameters of the approximated GPD. Once convergence is achieved, we can estimate the statistics of the extreme queue length/queuing delay which enables us to proactively deal with the occurrence of extreme events.

VI. CONCLUSIONS

In this work, we have studied an URLLC-enabled MEC architecture with multiple users and servers, in which the high-order statistics of latency and reliability are taken into account. To this end, we have imposed a probabilistic constraint on the queue length (or queuing delay) violation and invoked extreme value theory to deal with low-probability (extreme) events. The studied problem has been cast as a computation and transmit power minimization problem subject to latency requirements and reliability constraints. Subsequently, utilizing Lyapunov stochastic optimization, we have proposed a dynamic latency and reliability-aware policy for task computation, task offloading, and resource allocation. Numerical results have shown that the considered MEC network achieves a better power-delay tradeoff for intense computation requirements and higher task arrival rates.

REFERENCES

[1] M. Chiang, S. Ha, C.-L. I, F. Risso, and T. Zhang, "Clarifying fog computing and networking: 10 questions and answers," *IEEE Commun. Mag.*, vol. 55, no. 4, pp. 18–20, Apr. 2017.

[2] 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.*, to be published.

[3] M. S. ElBamby, M. Bennis, and W. Saad, "Proactive edge computing in latency-constrained fog networks," in *Proc. European Conf. Netw. Commun.*, Jun. 2017, pp. 1–6.

[4] B. Soret, P. Mogensen, K. I. Pedersen, and M. C. Aguayo-Torres, "Fundamental tradeoffs among reliability, latency and throughput in cellular networks," in *Proc. IEEE Global Commun. Conf. Workshops*, Dec. 2014, pp. 1391–1396.

[5] J. Kwak, Y. Kim, J. Lee, and S. Chong, "Dream: Dynamic resource and task allocation for energy minimization in mobile cloud systems," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 12, pp. 2510–2523, Dec. 2015.

[6] Z. Jiang and S. Mao, "Energy delay tradeoff in cloud offloading for multi-core mobile devices," *IEEE Access*, vol. 3, pp. 2306–2316, 2015.

[7] Y. Mao, J. Zhang, S. H. Song, and K. B. Letaief, "Power-delay tradeoff in multi-user mobile-edge computing systems," in *Proc. IEEE Global Commun. Conf.*, Dec. 2016, pp. 1–6.

[8] M. J. Neely, *Stochastic Network Optimization with Application to Communication and Queueing Systems.* Morgan and Claypool Publishers, Jun. 2010.

[9] S. Coles, *An Introduction to Statistical Modeling of Extreme Values.* Springer, 2001.

[10] T. D. Burd and R. W. Brodersen, "Processor design for portable systems," *J. VLSI Signal Process. Syst.*, vol. 13, no. 2-3, pp. 203–221, Aug. 1996.

[11] J. D. C. Little, "A proof for the queuing formula: $L = \lambda W$," *Operations Research*, vol. 9, no. 3, pp. 383–387, 1961.

[12] A. Al-Shuwaili and O. Simeone, "Energy-efficient resource allocation for mobile edge computing-based augmented reality applications," *IEEE Wireless Commun. Lett.*, vol. 6, no. 3, pp. 398–401, Jun. 2017.

[13] A. P. Miettinen and J. K. Nurminen, "Energy efficiency of mobile clients in cloud computing," in *Proc. 2nd USENIX Conf. on Hot Topics in Cloud Computing*, Jun. 2010.

[14] Radio Communications Sector of ITU, "P.1238-8: Propagation data and prediction methods for the planning of indoor radiocommunication systems and radio local area networks in the frequency range 300 MHz to 100 GHz," ITU-R, Tech. Rep., Jul. 2015.

[15] Z. Pi and F. Khan, "System design and network architecture for a millimeter-wave mobile broadband (MMB) system," in *Proc. 34th IEEE Sarnoff Symp.*, May 2011, pp. 1–6.