Short-Packet Edge Computing Networks with Execution Uncertainty

Xiazhi Lai, Tuo Wu, Cunhua Pan, Lifeng Mai, and Arumugam Nallanathan, *Fellow, IEEE*

Abstract

Low-latency computational tasks in Internet-of-Things (IoT) networks require short-packet communications. In this paper, we consider a mobile edge computing (MEC) network under time division multiple access (TDMA)-based short-packet communications. Within the considering network, a mobile user partitions an urgent task into multiple sub-tasks and delegates portions of these sub-tasks to edge computing nodes (ECNs). However, the required computing resource varies randomly along with execution failure. Thus, we explore the execution uncertainty of the proposed MEC network, which holds broader implications across the MEC network. In order to minimize the probability of execution failure in computational tasks, we present an optimal solution that determines the sub-task lengths and the blocklengths for offloading. However, the complexity of the optimal solution increases due to the involvement of the Q function and incomplete Gamma function. Consequently, we develop a low-complexity algorithm that leverages alternating optimization and majorization-maximization (MM) methods, enabling efficient computation of semi-closed-form solutions. Furthermore, to reduce the computational complexity associated with sorting the offloading order of sub-tasks, we propose two sorting criteria based on the computing speeds of the ECNs and the channel gains of the transmission links, respectively. Numerical results have validated the effectiveness of the proposed algorithm and

(Corresponding authors: *Tuo Wu, Cunhua Pan, and Arumugam Nallanathan*).

X. Lai is with the School of Computer Science, Guangdong University of Education, Guangzhou 510220, Guangdong, China. (E-mail: xzlai@outlook.com).

T. Wu and A. Nallanathan are with the School of Electronic Engineering and Computer Science at Queen Mary University of London, London E1 4NS, U.K. (E-mail: {tuo.wu, a.nallanathan}@qmul.ac.uk).

C. Pan is with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China. (Email: cpan@seu.edu.cn).

L. Mai is with the School of Electronic and Information Technology, Sun Yat-sen University, Guangzhou 510006, Guangdong, China. (E-mail: mailf3@mail2.sysu.edu.cn).

criteria. The results also suggest that the proposed network achieves significant performance gains over the non-orthogonal multiple access (NOMA) and full offloading networks.

Index Terms

Internet-of-Things (IoT), short-packet, execution uncertainty, mobile edge computing (MEC)

I. Introductions

The proliferation of intelligent devices within Internet-of-Things (IoT) networks has revolutionized various applications, such as autonomous driving and tactile Internet, providing extensive services [1], [2]. However, these intelligent devices face limitations in terms of computational capacity, which hinders their ability to meet the growing demands of applications. To address this challenge, mobile edge computing (MEC) has emerged as a promising solution. MEC allows devices with limited computational capabilities to offload computationally intensive tasks to access nodes equipped with powerful servers, through wireless links [3].

In the context of MEC networks, minimizing latency and energy consumption while ensuring the successful execution of computations is a crucial concern. Given the constrained battery capacity of intelligent devices and the inherent uncertainty associated with wireless channels, substantial challenges arise. Therefore, it becomes imperative to devise strategies that can effectively reduce latency and energy consumption while maintaining high computation success rates. Hence, extensive research efforts have been dedicated to optimizing task offloading and computational resource allocation. These efforts aim to strike a balance between computational efficiency and resource utilization, considering factors such as task characteristics, network conditions, and user requirements. Several studies [4]–[6] have proposed algorithms and frameworks for efficient task offloading and resource allocation, aiming to achieve improved performance in terms of latency and energy consumption.

However, it's worth noting that the influence of communication resources represents an equally important role in the broader MEC research landscape. Consequently, a significant body of research has been dedicated to scrutinizing the ramifications of communication resources on overall system performance. In this vein, several studies [7]–[9] have specifically explored the effects of diverse communication techniques, including multi-access schemes and transmission protocols, on the overall performance of MEC networks. The objective of these studies is to optimize communication resources in order to improve the likelihood of successful task execution and facilitate dependable and efficient task offloading.

Overall, the field of MEC is a rapidly evolving area of research, addressing the challenges of low-latency computational tasks in IoT networks. Through the optimization of task offloading, resource allocation, and communication strategies, MEC offers promising solutions to meet the computational demands of IoT applications while ensuring efficient and reliable execution.

A. Literature

Short-packet communication scheme was proposed to provide ultra-low latency services for IoT networks. Compared with traditional communications with infinite-blocklength coding, block decoding errors may occur in short-packet communications due to the utilization of finite-blocklength coding. Therefore, plenty of work has been devoted to overcoming this drawback of short-packet communications [10]–[17]. Specifically, Sun *et al.* optimized the transmit power and data rate to maximize the effective throughput for a two-user non-orthogonal multiple access (NOMA) network with short-packet communications, and their work showed the superiority of NOMA to orthogonal multiple access (OMA) [12]. Furthermore, the authors of [14] extended to study the short-packet NOMA networks with multiple carriers, and solved the effective throughput maximization problem by using the block coordinate descent (BCD) and concave-convex procedure (CCCP) methods. Later, Chen *et al.* maximized the total effective throughput and minimized the transmission time for short-packet OMA networks, by alternately optimizing the block error rate (BLER) and blocklength for each device [15]. Moreover, Ren *et al.* devised joint power and blocklength allocation algorithms to minimize the BLER and maximize the secure sum rate for multi-user networks [16]–[18].

Besides, the deployment of short-packet communications in MEC networks boosts the development of low-latency applications in IoT scenarios, which raises research interest [19]–[23]. In specific, She *et al.* in [19] considered the queuing delay of tasks, and their works analyzed the BLER for orthogonal frequency division multiple access (OFDMA) MEC networks with short-packet communications and further optimized the allocation of tasks and sub-channels to minimize the BLER. To enhance the transmission efficiency while maintaining low latency, Liu *et al.* in [20] proposed to leverage the NOMA short-packet communications in MEC network, and the reliability of network has been improved by optimizing the channel use and power allocations factors. In [21], Zhou *et al.* managed to adjust the request rate for communication and computation, and jointly optimized the allocation of channel use, spectrum and computational resource, which breaks a balance between the effective amounts of information and the consumed energy. Moreover, to improve the freshness of data, Li *et al.* in [22] analyzed the average age of information for a short-packet edge computing network, and alternately optimized the offloading ratio of task and the blocklength with the aid of successive convex approximation (SCA) method. In addition, the utilization of hybrid automatic repeat request (HARQ) enhances the transmission reliability, and Zhu *et al.* in [23] exploited the benefit of HARQ and developed an optimal solution to minimize the energy consumption in short-packet MEC networks.

Most of the works in the field of MEC considered deterministic computational tasks with fixed workloads, and therefore, the required computational resources, e.g., central processing units (CPU) cycles, can be perfectly estimated and well allocated. However, information on computational resources and tasks becomes difficult to acquire with low-latency requirements, especially in IoT networks, which causes execution uncertainty [24]–[26]. To address this problem, the impact of execution uncertainty on the execution failure probability has been evaluated and modeled in [26]. Owing to the results in [26], Zhang *et al.* further optimized the computing speed to lower consumed energy with given execution failure probabilities in [27]. In practice, execution uncertainty may cause excessive energy consumption and result in power-consumption outages, and Yang *et al.* took notice of this phenomenon and analyzed the outage probability of power-consumption in multiple-input-multiple-output (MIMO) Gaussian channels [28].

However, it is not practical to propose short-packet communication without considering the occurrence of execution uncertainty. Specifically, in IoT applications, requires considering execution uncertainty due to the inherent complexity and variability of IoT devices. Recognizing execution uncertainty can significantly improve IoT system reliability. In the realm of MEC, the rate of successful computations becomes crucial. To address this, a new nonlinear function, including the gamma function, is developed to estimate the success probability of computations, taking into account factors like sub-task size and blocklength. This approach introduces fresh challenges for researchers. Additionally, while there is extensive research on execution uncertainty in standard MEC networks, the specific dynamics in short-packet MEC networks remain under-explored. This gap highlights the need for thorough research and analysis, which is the focus of this paper.

B. Contribution

The objective of this paper is to address the challenge of execution uncertainty in the context of short-packet MEC networks. Furthermore, the paper aims to mitigate the limitations posed by both short-packet communications and execution uncertainty within the constraints of latency. Specifically, we consider a short-packet MEC network, where a mobile terminal divides the computational task into multiple sub-tasks, and offloads parts of sub-tasks to edge computing nodes (ECNs). Moreover, the offloading and computing of the task are subject to latency constraints, and execution failure may occur due to the randomness of the number of CPU cycles required for each bit of task and the block decoding error. The main contributions of this paper are summarized as follows.

- Taking into account the impacts of communication and execution uncertainties, we formulate the execution failure probability minimization problem with the aim of optimizing the transmission time and the length of the sub-task for each ECN. In addition, an optimal solution for the execution failure minimization problem is provided.
- Moreover, since the objective function involves Q function and incomplete Gamma function, the execution failure minimization problem is complicated and the optimal solution is untraceable when the number of ECN is large. To cope with that, we develop a low-complexity alternating algorithm using the majorization-maximization (MM) method. Specifically, the second-order approximation based MM (MM-2) method is applied in the design of the algorithm, which provides semi-closed-form solutions and helps to obtain the results efficiently.
- As the time division multiple access (TDMA) scheme is considered, the mobile terminal offloads the sub-tasks to each ECN sequently, and we propose two sorting criteria for the sub-task offloading order, relying on the computing speeds of ECNs and the channel gains of transmission links, respectively.
- Numerical results show the effectiveness of the proposed alternating-MM algorithm and sorting criteria. Furthermore, the comparisons with NOMA networks and full offloading networks also demonstrate the superiority of the proposed TDMA network.

C. Structure

The rest of this paper is organized as follows. The system model and execution failure probability minimization problem are illustrated in Section II. To solve the problem, we provide an optimal solution in Section III and develop a low-complexity algorithm in Section IV. In Section V, two sorting criteria for task offloading are presented. We present the numerical results in Section VI and conclude our paper in Section VII.

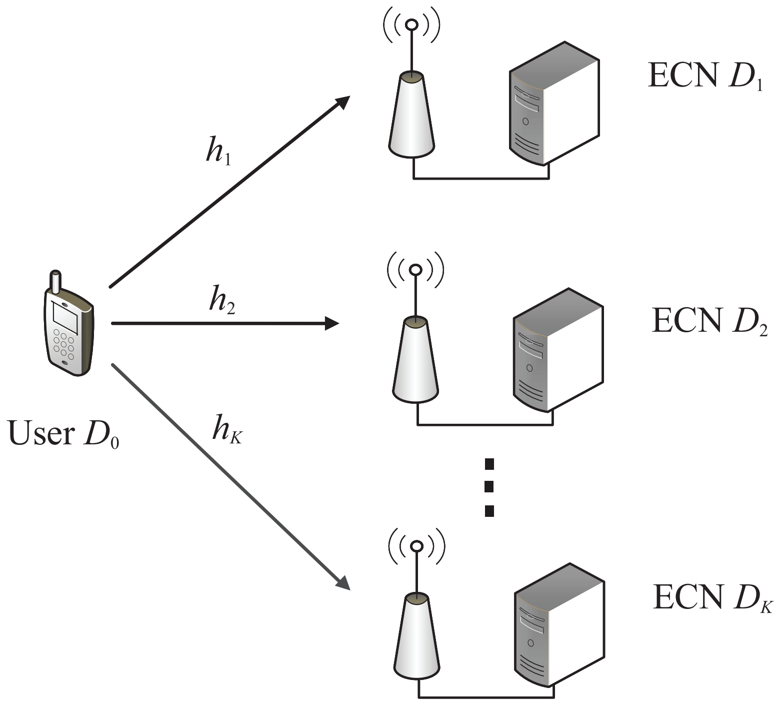

Fig. 1. Mobile edge computing network.

II. System Model and Problem Formulation

As depicted in Fig. 1, we consider a latency-limited edge computing network, where user D_0 attempts to execute the computational task of length $N_{\rm tol}$ bits and maximum latency γ_T s, with the aid of K ECNs $\{D_k|1 \leq k \leq K\}$. [1] Analogously to the partial offloading utilized in [5] and [6], the computational task is divided into $K+1$ sub-tasks, which are executed locally or offloaded to the ECNs. In specific, D_0 executes the sub-task of N_0 bits locally, and offloads the

[1] The work in this paper provides a preliminary study, which can be easily extended to multiuser networks, with techniques such as ECN selection, ECN allocation or power allocation [30].

sub-task of N_k bits to D_k with blocklength m_k channel uses, utilizing the TDMA scheme.[2] In addition, the computing speeds of D_0 and D_k are g_0 and g_k, respectively.

In practice, the number of CPU cycles required for each bit of task, i.e., η_k, may be difficult to estimate, while the statistical characteristic of η_k can be utilized to improve the system performance in the presence of execution uncertainty.[3] Taking into consideration this circumstance, we assume η_k varies randomly, and η_k can be modeled by a Gamma distribution as follows [26]–[28]

$$f_\eta(x) = \frac{1}{\beta\gamma(\kappa)} \left(\frac{x}{\beta}\right)^{\kappa-1} e^{-\frac{x}{\beta}}, \tag{1}$$

where κ is the shape parameter, β is the scale parameter and $\gamma(\kappa) = \int_0^\infty e^{-t} t^{\kappa-1} dt$ is the Gamma function. The estimation of distribution parameters depends on the nature of the application, e.g., the complexity of the algorithms [26]. It is assumed that each computational task consists of different files, thus $\{\eta_k | 1 \leq k \leq K\}$ are assumed to be approximately independently identically distributed [26]–[28]. We can easily find that the expectation of η_k equals to the product of κ and β, i.e., $\mathbf{E}(\eta_k) = \kappa\beta$. Therefore, with larger values of κ or β, the expected computation burden increases.

In the following, we illustrate the communication and computation models, and then detail the impacts of short-packet communication and execution uncertainty on the system performance.

A. Communication Model

Due to the use of TDMA, the offloadings of sub-tasks are arranged in chronological order. We assume the offloading of sub-task executed by D_k starts earlier than that executed by D_{k+1},

[2] For the OMA-based scheme, we see that the frequency division multiple access (FDMA)-based scheme costs the same transmission time for the offloading to each ECN. However, for the TDMA-based scheme, the ECN finishing the transmission earlier can start the computation of the task without waiting for the later ones. Therefore, the TDMA-based scheme can significantly reduce the latency compared with the FDMA-based scheme. Thus TDMA is considered in this paper.

[3] For instance, the data processing of acoustic streams or images, e.g. speech analysis or object recognition, poses different task burdens with various scenario, and it is easier to execute the task in a simple scenario with less distraction or noise [29]. However, we can estimate the statistical characteristic of η_k, e.g., the expectation and variance of η_k, through the collected data.

hence the total transmission latency for D_k is

$$L_{D,k} = \sum_{n=1}^{k} m_n T, \tag{2}$$

where $T = \dfrac{1}{B}$ and B is the channel bandwidth.

To meet the low-latency requirement, short-packet communications are utilized for offloading, and the successful offloading of sub-task cannot be guaranteed. Let P_S represent the transmission power, σ^2 represent the variance of the additive white Gaussian noise, and h_k represent the channel from D_0 to D_k. Hence, the received SNR for D_k is given by

$$\gamma_k = \frac{P_S |h_k|^2}{\sigma^2}, \tag{3}$$

and the corresponding successful transmission probability of the sub-task for D_k is given by

$$P_{D,k} \approx 1 - \Phi(\gamma_k, N_k, m_k), \tag{4}$$

where $\Phi(\gamma, N, m)$ represents the transmission error probability, which can be expressed as

$$\Phi(\gamma, N, m) = Q\left(\Lambda(\gamma, N, m)\right), \tag{5}$$

$\Lambda(\gamma, N, m)$ is defined as

$$\Lambda(\gamma, N, m) = (C(\gamma) - N/m) \cdot (V(\gamma)/m)^{-\frac{1}{2}}, \tag{6}$$

$C(\gamma) = \log_2(1+\gamma)$ denotes the channel capacity, $V(\gamma) = (\log_2 e)^2(1-(1+\gamma)^{-2})$ denotes the channel dispersion, and $Q(x) = \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_x^{\infty} e^{-\frac{t^2}{2}} dt$ is the Gaussian Q function. In this paper, we use normal approximation in (5) for the transmission error probability, and (5) becomes accurate in practice when the transmission error probability is higher than 10^{-7} [10]. In the context of long-packet MEC networks, the successful transmission probability for D_k becomes 1, if the transmission data rate, i.e., N_k/m_k, is smaller than channel capacity $C(\gamma_k)$.

B. Computation Model

Suppose the offloading of sub-task for D_k succeeds. In this case, D_k starts to execute the offloaded sub-task. The computing latency for D_k is

$$L_{C,k} = \frac{N_k \eta_k}{g_k}. \tag{7}$$

From (2) and (7), for given η_k, the total latency for D_k is

$$\begin{aligned} L_k &= L_{C,k} + L_{D,k} \\ &= \frac{N_k \eta_k}{g_k} + \sum_{n=1}^{k} m_n T, \end{aligned} \tag{8}$$

which consists of transmission latency $L_{D,k}$ and computing latency $L_{C,k}$. When the total latency for D_k does not exceed latency threshold γ_T, then D_k succeeds to finish its own sub-task. Due to the randomness of η_k, there exists the computation success probability for D_k, which is denoted as $P_{C,k}$. By using (1) and (8), $P_{C,k}$ can be formulated as

$$\begin{aligned} P_{C,k} &= \Pr\left(L_k \leq \gamma_T\right) \\ &= \Pr\left(\eta_k \leq \frac{g_k(\gamma_T - \sum_{n=1}^{k} m_n T)}{N_k}\right) \\ &= \gamma\left(\kappa, \frac{g_k(\gamma_T - \sum_{n=1}^{k} m_n T)}{N_k \beta}\right), \end{aligned} \tag{9}$$

where $\gamma(\kappa, x) = \frac{1}{\gamma(\kappa)} \int_0^x t^{\kappa-1} e^{-t} dt$ is the lower incomplete Gamma function. Also, $\gamma_T > \sum_{n=1}^{k} m_n T$ should hold to keep the number in incomplete Gamma function positive [4].

If the transmission and computing of sub-tasks for all ECNs and the local computing at N_0 are completed successfully, the execution of computational task succeeds. Following (4) and (9), the overall execution success probability of the computational task can be obtained as

$$P_{\text{sus}}(\mathbf{N}, \mathbf{m}) = \prod_{k=0}^{K} P_{D,k} P_{C,k}, \tag{10}$$

where $\mathbf{N} = [N_0, N_1, \cdots, N_K]$ and $\mathbf{m} = [m_1, m_2, \cdots, m_K]$. Accordingly, the execution failure probability is

$$P_{\text{fail}}(\mathbf{N}, \mathbf{m}) = 1 - P_{\text{sus}}(\mathbf{N}, \mathbf{m}). \tag{11}$$

[4] Other research has proposed different computation failure models, i.e., [31], whose computation failure is resulted from the shared computing resource. where failures arise from shared computing resources. However, this paper focuses on a more practical scenario, where task computation failures are attributed to the intrinsic nature of the content. Specifically, it deals with the variability in the number of CPU cycles required, which can fluctuate randomly.

C. Problem Formulation

In this paper, we aim to minimize the execution failure probability, or equivalently maximize the execution success probability, by optimizing the lengths of sub-tasks for the user and ECNs and the blocklengths for the offloading to ECNs. From (10) and using the monotonicity of the logarithm function, the execution failure probability minimization problem can be formulated as follows

$$\max_{\mathbf{N},\mathbf{m}} \quad \ln P_{\text{sus}}(\mathbf{N},\mathbf{m}) \tag{12a}$$

$$\text{s.t.} \quad \sum_{k=0}^{K} N_k \geq N_{\text{tol}}, \tag{12b}$$

$$N_k \leq N_{\text{tol}}, \quad k \in \mathcal{K}, \tag{12c}$$

$$\gamma_T - \sum_{n=1}^{k} m_n T > 0, \quad k \in \mathcal{K}, \tag{12d}$$

$$N_k, m_k \in \mathbb{N}, \quad k \in \mathcal{K}, \tag{12e}$$

where $\mathbb{N}$ is the set including all non-negative integers, $\mathcal{K} = \{0, 1, \cdots, K\}$. Constraint (12e) ensures the number of bits and the number of channel uses for each sub-task are non-negative integer, as we consider bit-wise level partial offloading and short-packet communications. It is readily to find that Problem (12) is non-convex due to the nonlinear functions in (12a) and the integer constraints in (12e), and the solution to Problem (12) is challenging to obtain. [5] In the sequent sections, we provide an optimal solution and a low-complexity alternating-MM algorithm to solve Problem (12).

III. Optimal Solution

In this section, we provide an optimal solution to Problem (12), based on the interior point method and K-dimension search. The detailed procedure is illustrated as follows.

By using the K-dimension search over $\mathbf{N}$, we can accordingly determine the optimal $\mathbf{m}$ by

[5] When considering the long-packet MEC network, we see that the execution failure probability minimization problem is convex, by setting $P_{D,k} = 1$ and $N_k/m_k \leq C(\gamma_k)$ for $k \in \mathcal{K}$ and relaxing constraint (12e).

solving the following problem

$$\begin{aligned} \max_{\mathbf{m}} \quad & \ln P_{\text{sus}}(\mathbf{N}, \mathbf{m}) \\ \text{s.t.} \quad & \text{(12d)}, \text{(12e)}. \end{aligned} \tag{13}$$

To find the optimal solution for Problem (13), we first relax the integer constraints in (12e) to $m_k \geq 0$, for $k \in \mathcal{K}$, and the relaxed Problem (13) is written as

$$\max_{\mathbf{m}} \quad \ln P_{\text{sus}}(\mathbf{N}, \mathbf{m}) \tag{14a}$$

$$\begin{aligned} \text{s.t.} \quad & \text{(12d)}, \\ & m_k \geq 0, \quad k \in \mathcal{K}. \end{aligned} \tag{14b}$$

According to the following theorems, we show that Problem (14) is convex.

Theorem 1: $\ln(1 - Q(x))$ is an increasing concave function.

Proof: See Appendix A. ■

Theorem 2: $\Lambda(\gamma, N, m)$ is an increasing concave function with respect to (w.r.t.) m.

Proof: See Appendix B. ■

Theorem 3: $\ln P_{C,k}$ is a concave function of $\mathbf{m}$.

Proof: See Appendix C. ■

From Theorems 1-2 and (4), it is trivial to find that $\ln P_{D,k}$ is a concave function of $\mathbf{m}$. In addition, utilizing Theorem 3, we can further conclude that $\ln P_{\text{sus}}$ is a concave function of $\mathbf{m}$. As such, Problem (14) is a convex optimization problem, and we can obtain the optimal $\tilde{\mathbf{m}}(\mathbf{N}) = [\tilde{m}_1(\mathbf{N}), \tilde{m}_2(\mathbf{N}), \cdots, \tilde{m}_K(\mathbf{N})]$ for Problem (14) via the interior point method in [35]. Furthermore, using the exhausting search over the rounding integers of $\tilde{m}_k(\mathbf{N})$, we can obtain the optimal solution to Problem (13), as follows

$$\mathbf{m}^*(\mathbf{N}) = \arg \max_{\substack{m_k \in \mathcal{B}_k,\ k \in \mathcal{K} \\ \text{(12d),(12e)}}} \ln P_{\text{sus}}(\mathbf{N}, \mathbf{m}), \tag{15}$$

where $\mathcal{B}_k = \{\lfloor \tilde{m}_k(\mathbf{N}) \rfloor, \lceil \tilde{m}_k(\mathbf{N}) \rceil\}$ is the set containing the rounding integers of $\tilde{m}_k(\mathbf{N})$, $\lfloor x \rfloor$ denotes the floor function, and $\lceil x \rceil$ denotes the ceiling function. Moreover, we can easily see that the intersection of $m_k \in \mathcal{B}_k$, for $k \in \mathcal{K}$, and (12d) includes all the possible optimal solutions to Problem (13), since $\ln P_{\text{sus}}$ is a concave function of $\mathbf{m}$.

Furthermore, using the K-dimension search over $\mathbf{N}$ and comparing the values of $\ln P_{\text{sus}}$ with $\mathbf{m} = \mathbf{m}^*(\mathbf{N})$, the optimal $\mathbf{N}^*$ for problem (12) can be obtained as

$$\mathbf{N}^* = \arg \max_{(12c),(12e)} \ln P_{\text{sus}}(\mathbf{N}, \mathbf{m}^*(\mathbf{N})). \tag{16}$$

In this way, the optimal solution to Problem (12) can be obtained by the K-dimension search over $\mathbf{N}$ and interior point method.

Algorithm 1 Proposed Optimal Solution

```
Using K-dimension search to list all possible compositions of $\mathbf{N}$ in $[\mathbf{N}_1, \mathbf{N}_2, \cdots, \mathbf{N}_L]$.
Set $l = 0$, $\mathbf{N}^* = \mathbf{N}_1$.
while $l \leq L$ do
    Set $\mathbf{N} = \mathbf{N}_l$.
    Solve Problem (14) and obtain $\tilde{\mathbf{m}}(\mathbf{N})$.
    Solve Problem (15) and obtain $\mathbf{m}^*(\mathbf{N}_l)$ using exhausting search method.
    if $\ln P_{\text{sus}}(\mathbf{N}_l, \mathbf{m}^*(\mathbf{N}_l)) > \ln P_{\text{sus}}(\mathbf{N}^*, \mathbf{m}^*(\mathbf{N}^*))$ then
        Set $\mathbf{N}^* = \mathbf{N}_l$.
    end if
    $l := l + 1$.
end while
```

Computational Complexity Analysis: For the K-dimension search over $\mathbf{N}$ with $\sum_{k=0}^{K} N_k = N_{\text{tol}}$, the computational complexity is $\mathcal{O}\left(N_{\text{tol}}^{K-1}\right)$. Moreover, for the optimization of $\mathbf{m}$, from [32], the computational complexity is $\mathcal{O}\left(K^{3.5} \cdot \log_2(1/\epsilon)\right)$, where ϵ is the numerical accuracy. Hence we can conclude that the overall computational complexity for the optimal solution is

$$\mathcal{O}\left(N_{\text{tol}}^{K-1} K^{3.5} \cdot \log_2(1/\epsilon)\right). \tag{17}$$

We see that the proposed optimal solution is suitable for small value of K, e.g. $K \leq 3$.

IV. Alternating-MM Algorithm

The computational complexity of the proposed optimal solution becomes obnoxious when K is large, therefore we develop a low-complexity algorithm in this section. Analogously to the works in [22] and [33], we leverage the alternating optimization method in [34] and optimize $\mathbf{N}$ and $\mathbf{m}$ alternately, where the integer constrainta in (12e) are relaxed to $m_k, N_k \geq 0$, for $k \in \mathcal{K}$. In specific, given $\mathbf{N}$, we solve Problem (14) in Section III. With fixed $\mathbf{m}$, we solve the following problem

$$\begin{aligned} \max_{\mathbf{N}} \quad & \ln P_{\text{sus}}(\mathbf{N}, \mathbf{m}) && \text{(18a)} \\ \text{s.t.} \quad & \text{(12b)}, \text{(12c)}, \\ & N_k \geq 0, \quad k \in \mathcal{K}. && \text{(18b)} \end{aligned}$$

However, Problem (18) is still non-convex owing to the involvement of non-linear functions, i.e., $Q(x)$ and $\gamma(x, y)$. Leveraging the MM method in [36], [37] and the second-order approximation, we develop an MM-2 method to solve Problem (18). The core idea of the MM method is to find a sequence of surrogate functions of the original objective function, and solve the problem with surrogate functions iteratively [36], [37]. To construct valid surrogate functions for Problem (18), we propose the following theorems.

Theorem 4: The second-order derivative of $Q(x)$ function is lower-bounded by

$$Q''(x) \geq -(2\pi e)^{-\frac{1}{2}}. \tag{19}$$

Proof: See Appendix E. ■

Theorem 5: The second-order derivative of $\ln \gamma\left(\kappa, \psi t^{-1}\right)$ w.r.t. $t \geq 0$ is lower-bounded by

$$\frac{\partial^2 \ln \gamma\left(\kappa, \psi t^{-1}\right)}{\partial t^2} \geq B_L(\psi, \kappa), \tag{20}$$

where $B_L(\psi, \kappa)$ is given in (70).

Proof: See Appendix F. ■

Based on Theorem 4, we apply the second-order Taylor series expansion at $\hat{N}_k$ and obtain the following inequality

$$P_{D,k} \geq \hat{P}_{D,k}(\hat{N}_k) = q_{k,1} N_k^2 + q_{k,2} N_k + q_{k,3}, \tag{21}$$

where

$$q_{k,1} = -\left(8\pi V(\gamma_k) m_k e\right)^{-\frac{1}{2}}, \tag{22}$$

$$q_{k,2} = \left(\hat{N}_k e^{-\frac{1}{2}} - e^{-\Lambda^2(\gamma_k,\hat{N}_k,m_k)/2}\right) \cdot \left(2\pi V(\gamma_k) m_k\right)^{-\frac{1}{2}}, \tag{23}$$

$$\begin{aligned} q_{k.3} =& 1 - \Phi(\gamma_k, \hat{N}_k, m_k) + \hat{N}_k^2 q_{k,1} \\ &+ \hat{N}_k e^{-\Lambda^2(\gamma_k,\hat{N}_k,m_k)/2} \left(2\pi V(\gamma_k) m_k\right)^{-1}. \end{aligned} \tag{24}$$

As we can see, $\hat{P}_{D,k}(\hat{N}_k)$ is a concave function of N_k.

From Theorem 5 and using the second-order Taylor series expansion at $\hat{N}_k$, we can obtain the following inequality

$$\ln P_{C,k} \geq \hat{P}_{C,k}(\hat{N}_k) = s_{k,1} N_k^2 + s_{k,2} N_k + s_{k,3}, \tag{25}$$

where

$$\psi_k = \frac{g_k(\gamma_T - \sum_{n=1}^{k} m_n T)}{\beta}, \tag{26}$$

$$s_{k,1} = \frac{B_L(\psi_k, \kappa)}{2}, \tag{27}$$

$$s_{k,2} = \hat{N}_k B_L(\psi_k) - f_\eta\Big(\frac{\beta\psi_k}{\hat{N}_k}\Big)\frac{\beta\psi_k}{\hat{N}_k^2}\gamma^{-1}\Big(\kappa, \frac{\psi_k}{\hat{N}_k}\Big), \tag{28}$$

$$s_{k,3} = \ln\gamma\Big(\kappa, \frac{\psi_k}{\hat{N}_k}\Big) + f_\eta\Big(\frac{\beta\psi_k}{\hat{N}_k}\Big)\frac{\beta\psi_k}{\hat{N}_k^2}\gamma^{-1}\Big(\kappa, \frac{\psi_k}{\hat{N}_k}\Big) + \hat{N}_k^2 s_{k,1}. \tag{29}$$

Accordingly, $\hat{P}_{C,k}(\hat{N}_k)$ is a concave function of N_k, since $s_{k,1}$ is non-positive from (27) and (70).

In the $(l+1)$-th step of the proposed MM-2 method, we replace $\ln P_{C,k}P_{D,k}$ with $\hat{P}_{C,k}(N_{k,l}) + \ln \hat{P}_{D,k}(N_{k,l})$ and obtain a valid surrogate function for Problem (18), then solve the following problem

$$\max_{\mathbf{N}} \quad \sum_{k=0}^{K} \hat{P}_{C,k}(N_{k,l}) + \ln \hat{P}_{D,k}(N_{k,l}) \tag{30a}$$

$$\text{s.t.} \quad \text{(12b)},$$

$$N_{L,k} \leq N_k \leq N_{U,k}, \quad k \in \mathcal{K}, \tag{30b}$$

where $\mathbf{N}_l = [N_{0,l}, N_{1,l}, \cdots, N_{K,l}]$ is the optimal solution to Problem (30) in the l-th step of the proposed MM-2 method. In addition, we set

$$N_{L,k} = \max(0, N_{a,k}), \quad N_{U,k} = \min(N_{\text{tol}}, N_{b,k}), \tag{31}$$

where

$$N_{a,k}=\big((q_{k,2}^2-4q_{k,1}q_{k,3})^{\frac{1}{2}}-q_{k,2}\big)\cdot\big(2q_{k,1}\big)^{-1}, \tag{32}$$

$$N_{b,k}=-\big((q_{k,2}^2-4q_{k,1}q_{k,3})^{\frac{1}{2}}+q_{k,2}\big)\cdot\big(2q_{k,1}\big)^{-1}, \tag{33}$$

are the solutions to $\hat{P}_{D,k}(N_{k,l})=0$. Constraints in (30b) ensure the numbers in logarithm function are positive.

As the objective function of Problem (30) is a concave function of $\mathbf{N}$, from [35], we can rewrite (30) into the following equivalent problem

$$\begin{aligned}&\max_{\mathbf{N},\mu\geq 0} \quad \sum_{k=0}^{K}\hat{P}_{C,k}(N_{k,l})+\ln\hat{P}_{D,k}(N_{k,l})+\mu N_k-\mu N_{\text{tol}}\\&\text{s.t.} \quad \text{(30b)},\end{aligned} \tag{34}$$

where μ is the dual variable.

To proceed, we use the water-filling method and decompose Problem (34) into multiple convex sub-problems. With fixed μ, the objective function is separable regarding N_k, therefore we can respectively solve

$$\begin{aligned}&\max_{N_k} \quad \ln\hat{P}_{D,k}(N_{k,l})+\hat{P}_{C,k}(N_{k,l})+\mu N_k\\&\text{s.t.} \quad \text{(30b)},\end{aligned} \tag{35}$$

with optimal solution $N_{k,l+1}(\mu)$. To lower the computational complexity, we have the following theorem.

Theorem 6: The closed-form solution to Problem (35) can be determined with Cardano's formula in [38].

Proof: See Appendix G. ■

Furthermore, since the second-order derivative of $\ln\hat{P}_{D,k}(N_{k,l})+\hat{P}_{C,k}(N_{k,l})$ w.r.t. N_k is negative, $N_{k,l+1}(\mu)$ is a non-decreasing function of μ. Therefore, we use the bisection search method to find the optimal $\bar{\mu}$ for Problem (34), which satisfies

$$\sum_{k=0}^{K}N_{k,l+1}(\bar{\mu})=N_{\text{tol}}. \tag{36}$$

In the next step of the MM-2 method, we set $N_{k,l+1}=N_{k,l+1}(\bar{\mu})$. In this way, we solve Problem (18) with locally optimum $\bar{N}_k$, for $k\in\mathcal{K}$.

Remark 1: As a comparison, we can leverage the first-order approximation based MM (MM-1) method, which has been widely used in the field of short-packet communications to deal with the difference-of-convex (DC) problems [11], [14], [15]. Based on Theorem 1 and Lemma 1, we can use the MM-1 method to solve Problem (18) in a manner analog to the proposed MM-2 method. However, the MM-1 method for Problem (18) requires a double bisection search in each step to obtain the optimal $N_{k,l+1}(\mu)$ and μ^*. Different from the MM-1 method, the developed MM-2 method provides semi-closed-form solutions and requires one bisection search only in each step.

Algorithm and Convergency: To summarize, we provide the following Algorithm 2 to explain the procedure of the developed alternating-MM algorithm for Problem (15) with relaxed $m_k, N_k \geq 0$, for $k \in \mathcal{K}$. The convergency of the developed alternating-MM algorithm can be proved in a way similar to the work in [22], which can be summarized as follows.

In the t-th step of the alternating optimization method, we have

$$\begin{aligned} \ln P_{\text{sus}}(\mathbf{N}^t, \mathbf{m}^t) &\overset{(a)}{\leq} \ln P_{\text{sus}}(\mathbf{N}^t, \mathbf{m}^{(t+1)}) \\ &\overset{(b)}{\leq} \ln P_{\text{sus}}(\mathbf{N}^{(t+1)}, \mathbf{m}^{(t+1)}), \end{aligned} \tag{37}$$

where step (a) corresponds to Problem (14) with given $\mathbf{N}^t$, and step (b) corresponds to Problem (18) with given $\mathbf{m}^t$. Step (a) hold since Problem (14) is convex and the objective value of Problems (14) is non-decreasing. Moreover, step (b) holds since for Problem (18) we have

$$\begin{aligned} \ln P_{\text{sus}}(\mathbf{N}_0, \mathbf{m}) &= \sum_{k=0}^{K} \hat{P}_{C,k}(N_{k,0}) + \ln \hat{P}_{D,k}(N_{k,0}) \\ &\leq \sum_{k=0}^{K} \hat{P}_{C,k}(N_{k,l}) + \ln \hat{P}_{D,k}(N_{k,l}) \\ &\leq \ln P_{\text{sus}}(\mathbf{N}_l, \mathbf{m}), \end{aligned} \tag{38}$$

that is the objective value of Problem (18) is non-decreasing. We see that (37) indicates the objective value of Problems (14) and (18) is non-decreasing after each iteration in alternating optimization method, hence the convergency of Algorithm 2 is proved.

Finally, applying the exhausting search with the rounding integers of $\tilde{m}_k$ and $\bar{N}_k$ for $k \in \mathcal{K}$, we solve Problem (12).

Algorithm 2 Proposed Alternating-MM Algorithm

```
Initialize $\mathbf{m}^{(0)}$, $\mathbf{N}^{(0)}$, $t = 1$.
repeat
    Set $\mathbf{N} = \mathbf{N}^{(t-1)}$ and solve Problem (14) with $\mathbf{m}^{(t)}$.
    Set $\mathbf{N}_0 = [N_{1,0}, N_{2,0}, \cdots, N_{K,0}] = \mathbf{N}^{(t-1)}$, $l = 0$, $\mathbf{m} = \mathbf{m}^{(t)}$.
    repeat
        Solve Problem (18) and obtain $\mathbf{N}_{l+1}$ using MM-1 or MM-2 methods.
        $l := l + 1$.
    until $\mathbf{N}_l$ converges.
    Set $\mathbf{N}^{(t)} = \mathbf{N}_l$.
    $t := t + 1$.
until Converge.
Output ($\mathbf{m}^t$, $\mathbf{N}^t$) as ($\tilde{\mathbf{m}}$, $\bar{\mathbf{N}}$).
```

Computational Complexity Analysis: For the optimization of m_k, from [32], the computational complexity is $\mathcal{O}\left(K^{3.5} \cdot \log_2(1/\epsilon)\right)$, where ϵ is the numerical accuracy. For the optimization of N_k, the computational complexity of the MM-2 method is $\mathcal{O}\left(L(K+1) \cdot \log_2(1/\epsilon)\right)$, where L_2 is the number of iterations for the convergence of the MM-2 method. Therefore, the computational complexity of the proposed alternating-MM-2 algorithm is

$$\mathcal{O}\left(L\left(\left(L_2 \cdot (K+1) + K^{3.5}\right) \cdot \log_2(1/\epsilon)\right)\right). \tag{39}$$

However, for the alternating-MM-1 algorithm, the computational complexity is

$$\mathcal{O}\left(L\left(L_1(K+1)\log_2(1/\epsilon) + K^{3.5}\right) \cdot \log_2(1/\epsilon)\right), \tag{40}$$

where L_1 is the number of iterations for the convergence of the MM-1 method. The results in (39) and (40) indicate that the computational complexity of the alternating-MM-2 algorithm is significantly lower than that of the alternating-MM-1 algorithm.

V. Sorting Criterion for Sub-task Offloading Order

The TDMA scheme is employed for sub-task offloading in this paper, making the order of offloading a crucial consideration in the proposed system and its optimization. Accordingly, we introduce two sorting criterion for determining the order of sub-task offloading in this section. It is evident that the total number of potential sorting orders is $K!$, an exponential increase with the growth of ECNs' number K, which can become infeasible for larger K values. To address this computational complexity, we propose two sorting criteria for order arrangement.

1) Criterion 1: The sub-task offloading order is determined in a descending manner w.r.t. computing speed g_k, i.e., $g_1 > g_2, ..., g_K$. Due to the diversity of the computational capacity of ECNs, the performance of criterion 1 is no worse than the ECN selection based on the computational capacities of ECNs.

2) Criterion 2: The sub-task offloading order is determined in a descending manner w.r.t. channel gain $|h_k|^2$, i.e., $|h_1| > |h_2|, ..., |h_K|$. Due to the diversity of the transmission links, the performance of criterion 2 is no worse than the ECN selection based on channel gain.

VI. Numerical Results

In this section, simulation results are carried out to validate the effectiveness of the proposed algorithms and criteria. In the simulation, the shape and scale parameters of the distribution of η_k are set to $\kappa = 30$ and $\beta = 300$, respectively. In addition, we consider the NB-IoT scenarios in [39] and [40], and set the channel bandwidth to $B = 100$ KHz and the corresponding time duration for each channel use is $T = 10^{-5}$ s. The computation speeds for D_0 and D_k are set to $g_0 = 0.5$ GHz and $g_k = 2$ GHz, respectively. The maximum latency is $\gamma_T = 10^{-2}$ s. If not specified, the normalized channel gains from user to ECNs are ordered as $|h_k|^2 = 1.2 - 0.2k$, and we select the first K ECNs for simulation. For instance, when $K = 2$, we have $|h_1|^2 = 1$ and $|h_2|^2 = 0.8$.

Fig. 2 illustrates the convergence behavior of the proposed alternating-MM algorithm, which has been employed to optimize the computational task offloading in the considered scenario. The experimental setup involves specific parameters, including $N_{\text{tol}} = 2500$bits, $K = \{2, 3\}$ and $P_S/\sigma^2 = \{15, 20\}$ dB, which are indicative of the system's characteristics and performance evaluation metrics. To provide a comprehensive analysis, the results obtained from the optimal solution, alternating-MM-1 algorithm, and alternating-MM-2 algorithm are presented for comparative purposes. In the legend, the proposed optimal solution, alternating-MM-1, and alternating-MM-2 algorithms are denoted as ”OP,” ”AO-MM-1,” and ”AO-MM-2,” respectively. The convergence analysis reveals that both the alternating-MM-1 and alternating-MM-2 algorithms achieve convergence within approximately ten iterations, leading to results that are identical to those obtained from the proposed optimal solution. This convergence behavior showcases the effectiveness and reliability of the proposed algorithms in determining the optimal offloading strategy and resource allocation scheme. It is worth noting that the computational complexity of the alternating-MM-2 algorithm is significantly reduced compared to the alternating-MM-1 algorithm. This advantage arises from the fact that the MM-1 method necessitates a double bisection method at each MM step, while the MM-2 method only requires a single bisection search. The reduced computational complexity of the alternating-MM-2 algorithm contributes to its computational efficiency and practical feasibility in real-world implementations.

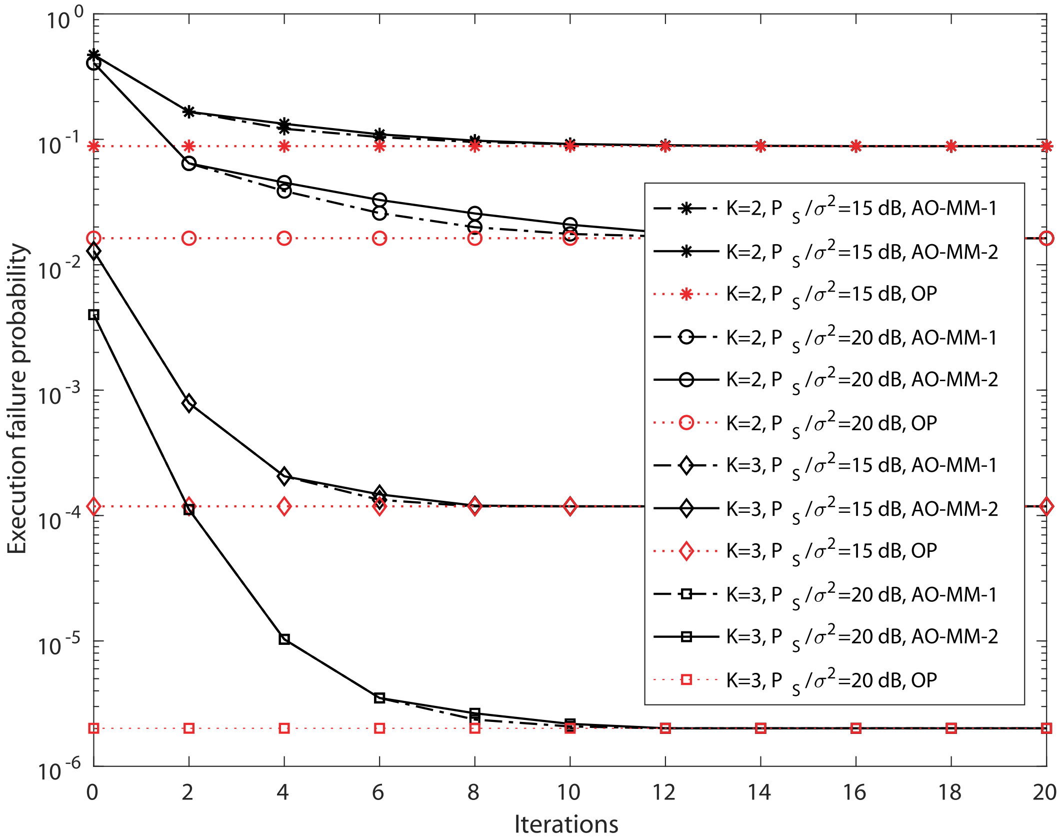

Fig. 2. Execution failure probabilities versus the number of iterations, where $K = 3$.

Figs. 3 presents a comprehensive comparison among the proposed TDMA network, the synchronous TDMA network and the NOMA-assisted MEC network described in [20]. In this comparison, all networks are equipped with two ECNs, and the parameter P_S/σ^2 is set to 15

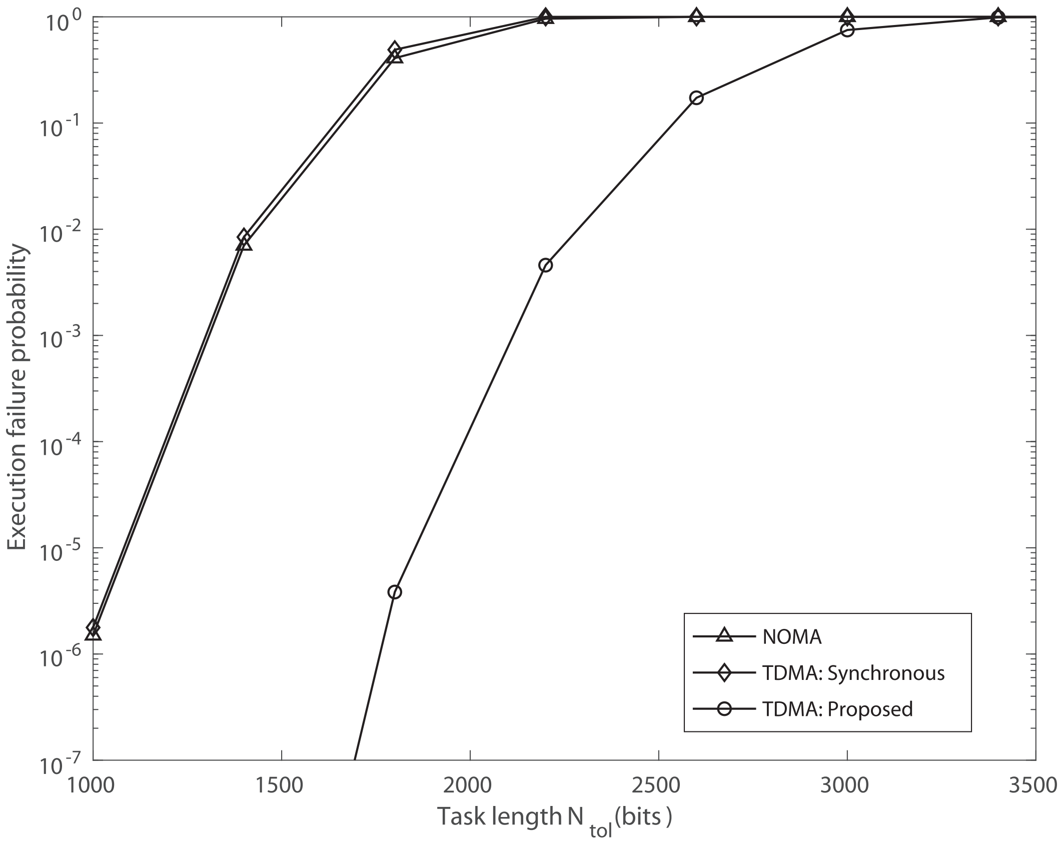

Fig. 3. Comparison on the performance between NOMA, FDMA and TDMA, where task length N_{tol}varies from 1000 to 3500 bits, $K = 2$ and $P_S/\sigma^2 = 15$ dB.

dB, which characterizes the power-to-noise ratio. In the synchronous TDMA network, both ECNs start to compute synchronously, and the optimization results can be easily obtained based on the works in this paper. To optimize the performance of the NOMA network with two ECNs, a systematic optimization process is employed, which involves alternating optimization of power allocation factors, sub-task lengths for each ECN, and the number of channel uses for offloading. Theorems 1-3 and Lemma 1 provide theoretical foundations for efficiently obtaining locally optimal solutions through the utilization of bisection search methods. Analyzing the results depicted in Fig. 3, it becomes evident that the proposed TDMA network outperforms the synchronous TDMA network and the NOMA network in terms of execution failure probability. This superiority can be attributed to the inherent characteristics of the proposed TDMA scheme, which enables the earlier offloaded ECN to initiate the computation of the task without being constrained by the subsequent transmissions from other ECNs. Conversely, in the NOMA scheme, both ECNs commence task offloading and task computation simultaneously. Consequently, by assigning a larger proportion of the task to the earlier offloaded ECN, the proposed TDMA network achieves a lower probability of execution failure compared to the NOMA network. Similar reason results in higher execution failure probability of the synchronous TDMA network than the proposed TDMA network. This comparative analysis substantiates the effectiveness of the proposed TDMA network and highlights its advantages over the NOMA network in mitigating the risk of execution failure. The ability to allocate tasks more efficiently to the earlier offloaded ECN contributes to the enhanced performance of the TDMA network in terms of execution success probability, ultimately promoting reliable and efficient task execution in mobile edge computing scenarios.

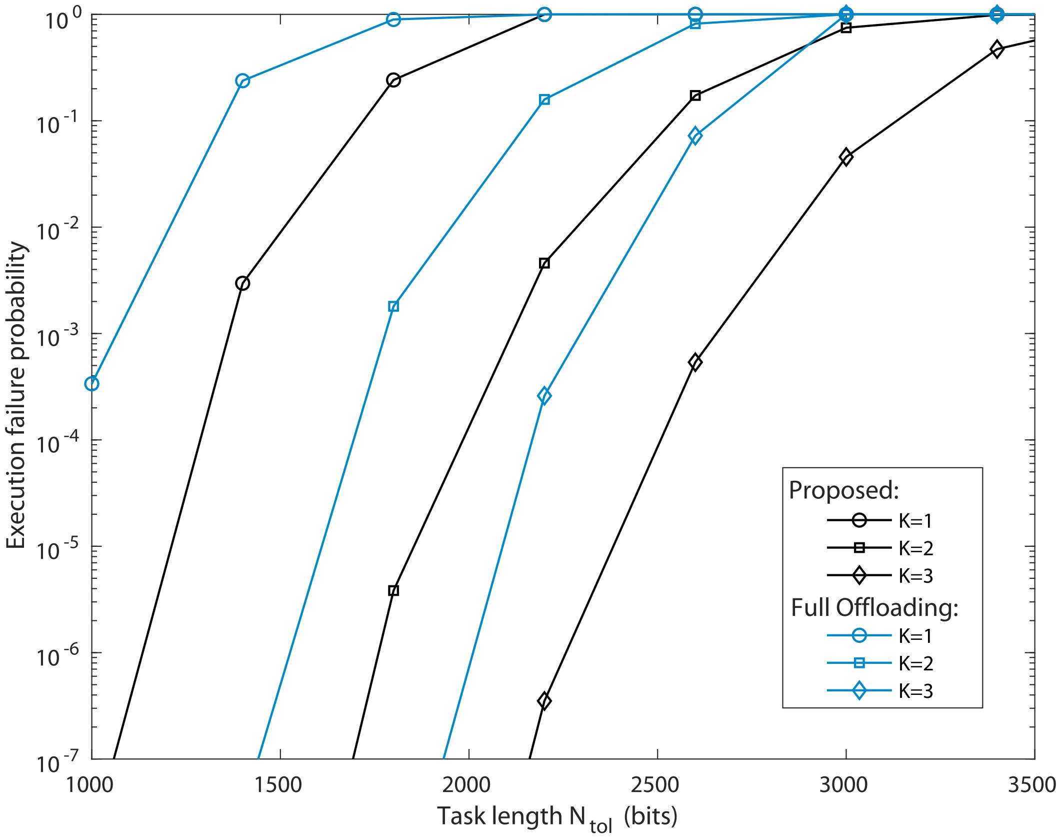

Fig. 4. Comparison on the performance with different values of task lengths N_{tol} and numbers of users K, where $P_S/\sigma^2 = 15$ dB.

In Figs 4–6, we compare the execution failure probabilities of the proposed network and the

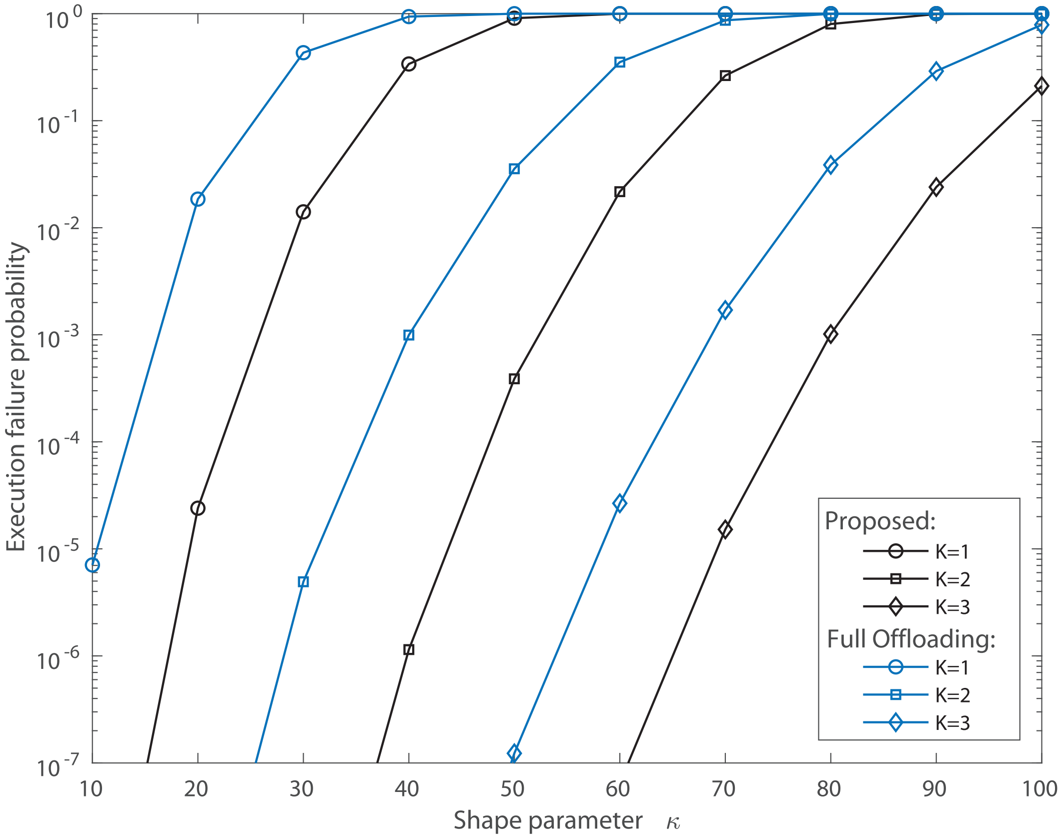

Fig. 5. Comparison on the performance with different values of shape parameter κ and numbers of users K, where $P_S/\sigma^2 = 15$ dB.

full offloading network. In specific, Fig. 4 illustrates the impact of the task length, denoted as N_{tol}, on the execution failure probability in the system. The simulation settings are such that P_S/σ^2 is set to 15 dB, and the number of ECNs, denoted as K, varies from 1 to 3. Analyzing the results depicted in Fig. 4, several observations can be made. Firstly, as the task length increases, the system performance in terms of execution failure probability deteriorates. This can be attributed to the increased complexity and resource requirements associated with longer tasks, which make them more susceptible to failures. Furthermore, it is evident that as the number of ECNs (K) increases, the system performance is significantly enhanced. The presence of multiple ECNs allows for task parallelization and distributed computation, which improves the overall reliability and efficiency of task execution.

Fig. 5 and Fig. 6 demonstrate the effects of the shape parameter, κ, and the scale parameter, β, on the system's execution failure probability. Specifically, in Fig. 5, κ ranges from 10 to 100

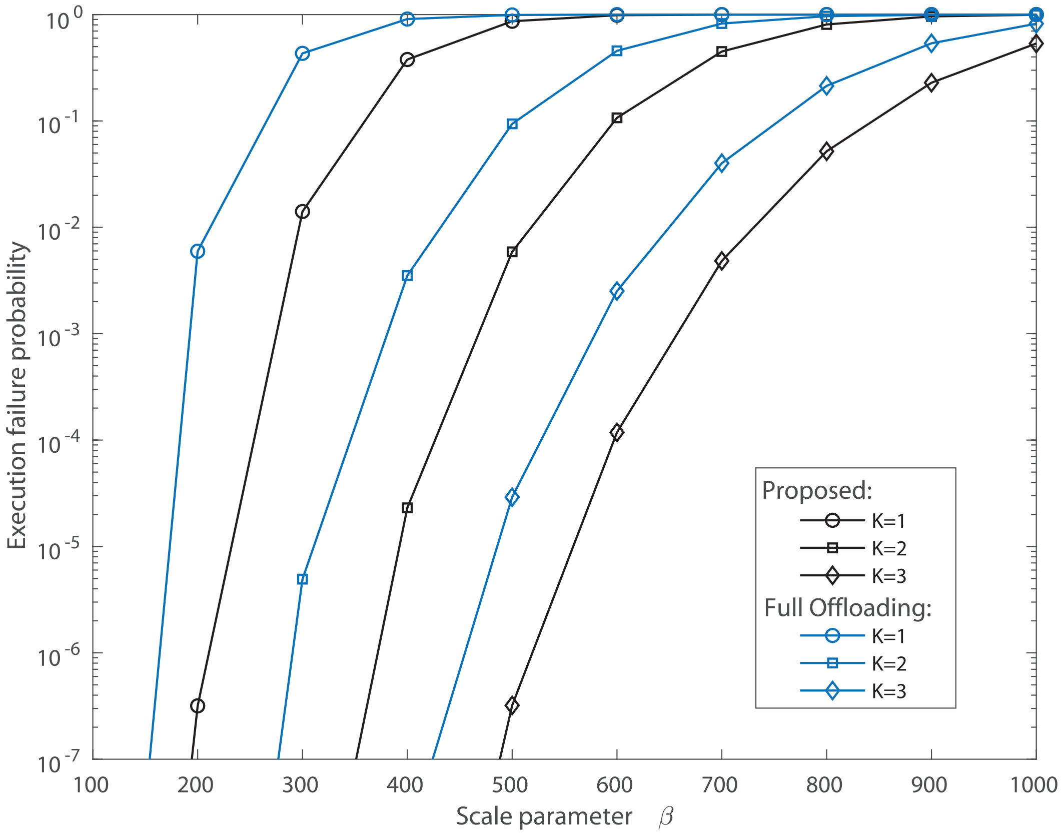

Fig. 6. Comparison on the performance with different values of scale parameter β and numbers of users K, where $P_S/\sigma^2 = 15$ dB.

with a constant β of 300, while in Fig. 6, β varies from 100 to 1000, maintaining κ at 300. As the expected number of CPU cycles required per bit of task, $\mathbf{E}(\eta_k) = \kappa\beta$, increases, higher values of κ or β lead to greater computational burden and, consequently, an elevated execution failure probability. The outcomes depicted in Fig. 5 and Fig. 6 corroborate this analysis. Crucially, the network model proposed in this study offers a marked improvement in execution failure probabilities compared to networks relying solely on full offloading. The reduced failure rates in the proposed network underscore its efficacy in addressing the inherent risks of full offloading, where tasks are entirely transferred to the ECNs. This improvement is largely due to the optimized distribution of sub-tasks and resources within the proposed framework, which effectively balances offloading with local processing, thereby enhancing overall system performance.

Fig. 7 demonstrates the execution failure probabilities for the proposed sorting criteria, with different numbers of ECNs K and task lengths N_{tol}, where $P_S/\sigma^2 = 15$ dB. For comparison, the

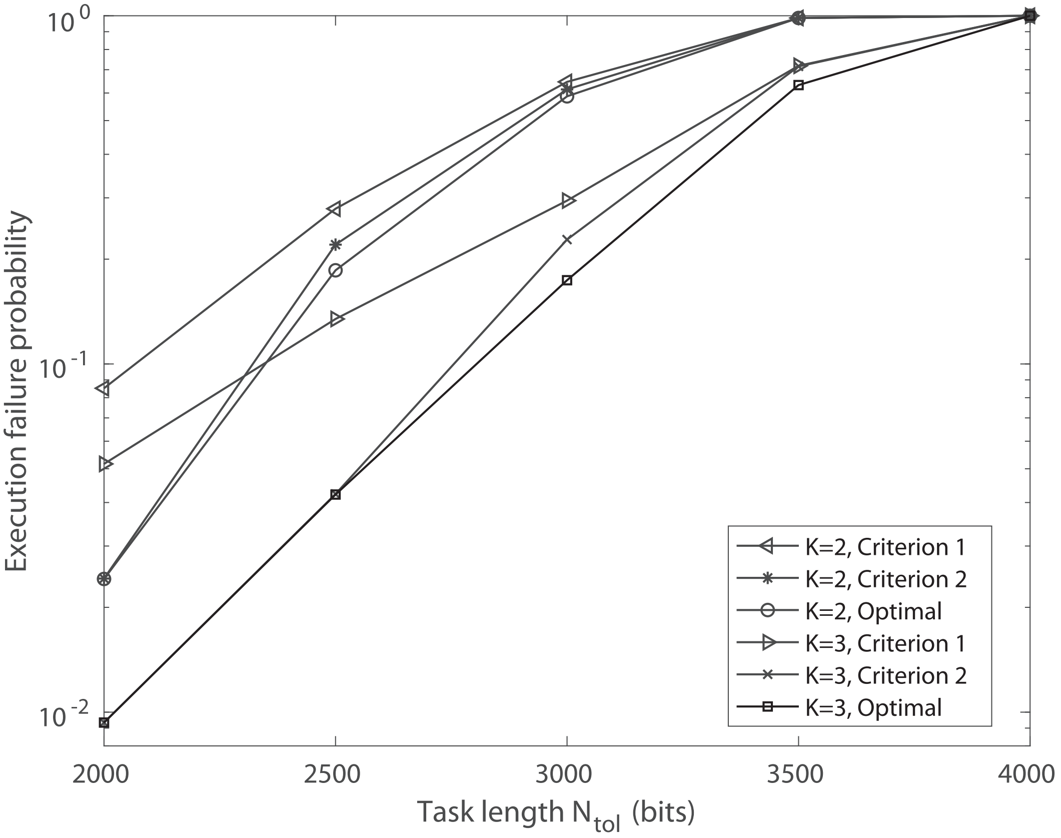

Fig. 7. Comparison on the performance between Criterion 1 and Criterion 2, where $P_S/\sigma^2 = 15$ dB.

results of the optimal offloading orders are also provided and denoted as “Optimal” in the legend. The results are obtained by averaging 100 realizations, where we set the channel parameters for each transmission link with $h_k \sim \mathcal{CN}(0,1)$ and the computing speeds for each ECN with $g_k \sim \mathcal{U}(1,5)$ GHz. As evident from the figure, Criterion 2 exhibits superior performance compared to Criterion 1. Moreover, the results achieved using Criterion 2 closely approach those attained by the optimal offloading order. The underlying reason for this phenomenon lies in the fact that Criterion 2 facilitates a more significant offloading of the task to the ECNs. This approach offers substantial benefits in terms of system performance, particularly when the computational capability of the mobile user is limited. Conversely, Criterion 1 does not consider the channel gains of the transmission links, and as a result, the execution failure probability increases significantly under unfavorable transmission conditions.

VII. Conclusion

In this paper, we studied a MEC network with short-packet communications, where the success execution of the computational task was degenerated due to the uncertainties from communication and computation. The execution failure probability has been improved by jointly optimizing the blocklength and the length of sub-tasks for each ECN. We also developed a low-complexity algorithm based on the alternating optimization method and the MM method. Moreover, two sorting criteria for sub-task offloading order were proposed to lower the implementation complexity, depending on the computational speeds and transmission links respectively. Numerical results were given to validate the effectiveness of the proposed algorithm and criteria. In addition, the results also showed the superiority of the proposed network over NOMA and full offloading networks.

Appendix A
Proof of Theorem 1

Note that $\ln(1-Q(x)) = \ln Q(-x)$. Therefore, we can compute the first-order derivative of $\ln Q(-x)$ as

$$\frac{\partial \ln Q(-x)}{\partial x} = \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}}{Q(-x)} > 0, \tag{41}$$

and the second-order derivative of $\ln Q(-x)$ as

$$\frac{\partial^2 \ln Q(-x)}{\partial x^2} = \frac{\partial^2 \ln Q(x)}{\partial x^2}, \tag{42}$$

which indicates that the convexities of $\ln Q(-x)$ and $\ln Q(x)$ are identical. From [35], we see that $\ln Q(x)$ is a concave function w.r.t. x. Thus based on (41) and (42), we can conclude that $\ln(1-Q(x))$ is an increasing concave function of x.

Appendix B
Proof of Theorem 2

The first-order and second-order derivatives of $\Lambda(\gamma, N, m)$ w.r.t. m are respectively given as

$$\frac{\partial \Lambda(\gamma, N, m)}{dm} = \frac{C(\gamma) m^{-1/2} + N m^{-3/2}}{2\sqrt{V(\gamma)}} > 0, \tag{43}$$

$$\frac{\partial^2 \Lambda(\gamma, N, m)}{dm^2} = -\frac{C(\gamma) m^{-3/2} + 3N m^{-5/2}}{4\sqrt{V(\gamma)}} < 0. \tag{44}$$

As we see from (43) and (44) that $\Lambda(\gamma, N, m)$ is an increasing concave function w.r.t. m.

APPENDIX C
PROOF OF THEOREM 3

From Theorem 2, we find that for $l > k$ or $d > k$, there exists

$$\frac{\partial^2 \ln P_{C,k}}{\partial m_l \partial m_d} = 0, \tag{45}$$

and when $l < k$ and $d < k$, there exists

$$\frac{\partial^2 \ln P_{C,k}}{\partial m_l \partial m_d} = \frac{\partial^2 \ln \gamma(\kappa, r_k)}{\partial r_k^2} \left(\frac{g_k T}{N_k \beta} \right)^2, \tag{46}$$

where

$$r_k = \frac{g_k(\gamma_T - \sum_{n=1}^{k} m_n T)}{N_k \beta}. \tag{47}$$

Therefore, from (45)-(47) the Hessian matrix of $\ln P_{C,k}$ w.r.t. $\mathbf{m}$ can be expressed as

$$\frac{\partial^2 \ln P_{C,k}}{\partial \mathbf{m}^2} = \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \tag{48}$$

where $\mathbf{A}$ is a $k \times k$ matrix, and all elements in $\mathbf{A}$ equal to $\frac{\partial^2 \ln \gamma(\kappa, r_k)}{\partial r_k^2} \left(\frac{g_k T}{N_k \beta} \right)^2$. It is trivial to find that the eigenvalues of the Hessian matrix of $\ln P_{C,k}$ w.r.t. $\mathbf{m}$ are equal to $k \frac{\partial^2 \ln \gamma(\kappa, r_k)}{\partial r_k^2} \left(\frac{g_k T}{N_k \beta} \right)^2$ or 0. Furthermore, we also have the following theorem.

Lemma 1: $\ln \gamma(\kappa, x)$ is a non-decreasing concave function w.r.t. x.

Proof: See Appendix D. ■

From Lemma 1, we know that the eigenvalues of the Hessian matrix of $\ln P_{C,k}$ w.r.t. $\mathbf{m}$ are non-positive, thus $\ln P_{C,k}$ is a concave function of $\mathbf{m}$ [35].

APPENDIX D
PROOF OF LEMMA 1

The first-order derivative of $\ln \gamma(\kappa, x)$ w.r.t. x is

$$\frac{\partial \ln \gamma(\kappa, x)}{\partial x} = \frac{\beta f_\eta(\beta x)}{\gamma(\kappa, x)}, \tag{49}$$

which is non-negative with $x \geq 0$, and thus $\ln \gamma(\kappa, x)$ is a non-decreasing function of x. Further, the second-order derivative of $\ln \gamma(\kappa, x)$ w.r.t. x is

$$\frac{\partial^2 \ln \gamma(\kappa, x)}{\partial x^2} = \frac{\gamma(\kappa, x) \beta \frac{\partial f_\eta(\beta x)}{\partial x} - \beta^2 f_\eta^2(\beta x)}{\gamma^2(\kappa, x)} \tag{50}$$

$$= \frac{\frac{\zeta}{\gamma(\kappa)} x^{\kappa-2} e^{-x}}{\gamma^2(\kappa, x)}, \tag{51}$$

where

$$\frac{\partial f_\eta(\beta x)}{\partial x} = \frac{x^{\kappa-2}(\kappa-1-x)}{\beta\gamma(\kappa)} e^{-x}, \tag{52}$$

$$\zeta = \gamma(\kappa,x)(\kappa-1-x) - \frac{x^\kappa}{\gamma(\kappa)} e^{-\frac{x}{\beta}}. \tag{53}$$

We see from (52) that $\frac{\partial^2 \ln\gamma(\kappa,x)}{\partial x^2} \leq 0$ holds when ζ is negative. The first-order derivative of ζ w.r.t. x is

$$\frac{\partial\zeta}{\partial x} = -\beta f_\eta(\beta x) - \gamma(\kappa,x) \leq 0. \tag{54}$$

which is non-positive when $x \geq 0$. From (53) and (54), we have $\zeta \leq 0$ and $\frac{\partial^2 \ln\gamma(\kappa,x)}{\partial x^2} \leq 0$ when $x \geq 0$. Thus, $\ln\gamma(\kappa,x)$ is a non-decreasing concave function of x, and the proof is completed.

Appendix E
Proof of Theorem 4

The second-order derivative of $Q(x)$ w.r.t. x can be calculated as

$$Q''(x) = \frac{x}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}. \tag{55}$$

It is difficult to determine the monotonicity of the second-order derivative of $Q(x)$ to x in (55), thus we turn to compute and set the third-order derivative of $Q(x)$ w.r.t. x to 0, i.e.,

$$\frac{\partial^3 Q(x)}{\partial x^3} = \frac{1-x^2}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} = 0. \tag{56}$$

and the solutions to (56) is $x = \pm 1$. Therefore, $Q''(x)$ is a decreasing function in the regions of $(-\infty,-1)$ and $(1,\infty)$, and an increasing function in the region of $(-1,1)$. Thus, the minimum values of $Q''(x)$ can be obtained by setting $x=-1$ or $x=+\infty$.

Moreover, from L'Hospital's rule, we have

$$\lim_{x\to-1} Q''(x) = -2\pi e^{-\frac{1}{2}} < 0, \tag{57}$$

and

$$\lim_{x\to+\infty} Q''(x) = 0, \tag{58}$$

thus the minimum value of $Q''(x)$ is $-2\pi e^{-\frac{1}{2}}$. Therefore, we can prove that $Q''(x)$ is lower-bounded by

$$Q''(x) \geq -2\pi e^{-\frac{1}{2}}. \tag{59}$$

Appendix F
Proof of Theorem 5

Firstly, we can calculate the second-order derivative of $\ln \gamma\left(\kappa, \psi t^{-1}\right)$ w.r.t. t as

$$\frac{\partial^2 \ln \gamma\left(\kappa, \psi t^{-1}\right)}{\partial t^2}=\frac{e^{-v} v^{k+2}}{\psi^2 \gamma(\kappa, v)}\left(-\frac{e^{-v} v^k}{\gamma(\kappa, v)}+\kappa+1-v\right), \tag{60}$$

where $v=\psi t^{-1}$ and $v \geq 0$ since $t \geq 0$. It is trivial to verify that $\frac{e^{-v} v^k}{\gamma(\kappa, v)}$ is a decreasing function of v, thus we have

$$\begin{aligned}\frac{\partial \ln \gamma\left(\kappa, \psi t^{-1}\right)}{\partial t} &\geq \frac{e^{-v} v^{k+2}}{\psi^2 \gamma(\kappa, v)}\left(-\frac{e^{-v} v^k}{\gamma(\kappa, v)}\Big|_{v=0}+\kappa+1-v\right)\\ &=\frac{e^{-v} v^{k+2}}{\psi^2 \gamma(\kappa, v)}(\rho-v),\end{aligned} \tag{61}$$

where $\rho=(1-\gamma(\kappa))(\kappa+1)$.

In the following, we provide a proper lower bound of $\frac{e^{-v} v^{k+2}}{\psi^2 \gamma(\kappa, v)}(\rho-v)$. Assume there exists a number M which satisfies

$$\frac{e^{-v} v^{k+2}}{\psi^2 \gamma(\kappa, v)}(\rho-v) \geq M, \tag{62}$$

and we also define

$$U(v)=e^{-v} v^{k+2}(\rho-v)-M \psi^2 \gamma(\kappa, v), \tag{63}$$

then (62) holds if $U(v) \geq 0$. Also, since $U(0)=0$, if $U^{\prime}(v) \geq 0$, then (62) holds. Thus, we turn to find M that satisfies $U^{\prime}(v)>0$, and obtain a proper lower bound of the second-order derivative of $\ln \gamma\left(\kappa, \psi t^{-1}\right)$ w.r.t. t.

To proceed, we calculate the first-order derivative of $U(v)$ as

$$U^{\prime}(v)=e^{-v} v^{\kappa-1} G(v), \tag{64}$$

where

$$G(v)=v^4-(\kappa+3+\rho) v^3+(\kappa+2) \rho v^2-M \psi^2 \gamma(\kappa). \tag{65}$$

The first-order derivative of $G(v)$ is

$$G^{\prime}(v)=v\left(4 v^2-3(\kappa+3+\rho) v+2 \rho(\kappa+1)\right), \tag{66}$$

which is a cubic function of v, and we can determine the minimum value of $G(v)$ with $v \geq 0$. Define

$$\Delta = 3(\kappa + 3 + \rho) - 32\rho(\kappa + 2), \tag{67}$$

then if $\Delta < 0$, $G'(v) \geq 0$ holds for $v \geq 0$, which indicates $U(v) \geq 0$ with $M \leq 0$. If $\Delta \geq 0$, we define

$$v_1 = \frac{3(\kappa + 3 + \rho) - \sqrt{\Delta}}{8}, \tag{68}$$

$$v_2 = \frac{3(\kappa + 3 + \rho) + \sqrt{\Delta}}{8}, \tag{69}$$

with $0 \leq v_1 \leq v_2$, and the minimum value of $G(v)$ can be obtained when $v = 0$ or $v = v_2$. Therefore, we can find M satisfying $\min\left(G(0), G(v_2)\right) \geq 0$ from (65), and $U(v) \geq 0$ holds.

In conclusion, we have $\dfrac{\partial^2 \ln \gamma\left(\kappa, \psi t^{-1}\right)}{\partial t^2} \geq B_L(\psi, \kappa)$, and $B_L(\psi, \kappa)$ is given by

$$B_L(\psi, \kappa) = \begin{cases} 0, & \Delta < 0 \\ \dfrac{v_2^4 - (\kappa + 3 + \rho)v_2^3 + (\kappa + 2)\rho v_2^2}{\psi^2 \gamma(\kappa)}, & \Delta \geq 0 \end{cases}. \tag{70}$$

Appendix G

Proof of Theorem 6

Taking the first-order derivative of the objective function of Problem (35) w.r.t. N_k to 0, we have

$$\frac{2q_{k,1}N_k + q_{k,2}}{q_{k,1}N_k^2 + q_{k,2}N_k + q_{k,3}} + 2s_{k,1}N_k + s_{k,2} + \mu = 0, \tag{71}$$

and (71) can be rewritten as the following cubic equation

$$aN_k^3 + bN_k^2 + cN_k + d = 0, \tag{72}$$

where

$$a = 2q_{k,1}s_{k,1}, \tag{73}$$

$$b = \mu q_{k,1} + 2q_{k,2}s_{k,1} + q_{k,1}s_{k,2}, \tag{74}$$

$$c = \mu q_{k,2} + q_{k,2}s_{k,2} + 2q_{k,3}s_{k,1} + 2q_{k,1}, \tag{75}$$

$$d = \mu q_{k,3} + q_{k,2}s_{k,3} + q_{k,2}. \tag{76}$$

Applying Cardano's formula in [38], the roots to equation (72) are given by

$$\epsilon_1 = -\frac{b}{3a} + z_1 + z_2, \tag{77}$$

$$\epsilon_2 = -\frac{b}{3a} + \omega z_1 + \omega z_2, \tag{78}$$

$$\epsilon_3 = -\frac{b}{3a} + \omega^2 z_1 + \omega^2 z_2, \tag{79}$$

where

$$\omega = \frac{-1+\sqrt{3}i}{2}, \tag{80}$$

$$p = \frac{3ac - b^2}{3a^2}, \tag{81}$$

$$q = \frac{27a^2 d - 9abc + 2b^2}{27a^3}, \tag{82}$$

$$z_1 = \left(-\frac{q}{2} + \left(\frac{q^2}{4} + \frac{p^3}{27}\right)^{\frac{1}{2}}\right)^{\frac{1}{3}}, \tag{83}$$

$$z_2 = \left(-\frac{q}{2} - \left(\frac{q^2}{4} + \frac{p^3}{27}\right)^{\frac{1}{2}}\right)^{\frac{1}{3}}. \tag{84}$$

There are three closed-form solutions to (72). Since the objective function of Problem (35) is concave, there is at most one solution to (72) that satisfies (30b). If none of the solutions to (72) satisfies (30b), we can substitute $N_{L,k}$ and $N_{U,k}$ into the objective function of (35), and the one with larger value is the solution to (35) .

References

[1] G. Qiao, S. Leng, K. Zhang, and Y. He, "Collaborative task offloading in vehicular edge multi-access networks," *IEEE Commun. Mag.*, vol. 56, no. 8, pp. 48–54, Aug. 2018.

[2] Z. Xiang, F. Gabriel, E. Urbano, G. T. Nguyen, M. Reisslein, and F. H. P. Fitzek, "Reducing latency in virtual machines: Enabling tactile internet for human-machine co-working," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 5, pp. 1098–1116, May 2019.

[3] J. Chen and X. Ran, "Deep learning with edge computing: A review," *Proc. IEEE*, vol. 107, no. 8, pp. 1655–1674, Aug. 2019.

[4] X. Meng, W. Wang, Y. Wang, V. K. N. Lau, and Z. Zhang, "Closed-form delay-optimal computation offloading in mobile edge computing systems," *IEEE Trans. Wirel. Commun.*, vol. 18, no. 10, pp. 4653–4667, Oct. 2019.

[5] L. Qian, Y. Wu, F. Jiang, N. Yu, W. Lu, and B. Lin, "NOMA assisted multi-task multi-access mobile edge computing via deep reinforcement learning for industrial internet of things," *IEEE Trans. Ind. Informatics*, vol. 17, no. 8, pp. 5688–5698, Aug. 2021.

[6] L. Qian, Y. Wu, N. Yu, D. Wang, F. Jiang, and W. Jia, “Energy-efficient multi-access mobile edge computing with secrecy provisioning,” *IEEE Trans. Mobile Comput.*, vol. 22, no. 1, pp. 237–252, Jan. 2023.

[7] Y. Zuo, S. Jin, S. Zhang, Y. Han, and K. Wong, “Delay-limited computation offloading for MEC-assisted mobile blockchain networks,” *IEEE Trans. Commun.*, vol. 69, no. 12, pp. 8569–8584, Dec. 2021.

[8] Y. Gu, Y. Yao, C. Li, B. Xia, D. Xu, and C. Zhang, “Modeling and analysis of stochastic mobile-edge computing wireless networks,” *IEEE Internet Things J.*, vol. 8, no. 18, pp. 14051–14065, 15 Sept. 2021.

[9] X. Lai, L. Fan, X. Lei, Y. Deng, G. K. Karagiannidis, and N. Nallanathan, “Secure mobile edge computing networks in the presence of multiple eavesdroppers,” *IEEE Trans. Commun.*, vol. 70, no. 1, pp. 500–513, Jan. 2022.

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

[11] H. Ren, C. Pan, Y. Deng, M. Elkashlan, and A. Nallanathan, “Joint pilot and payload power allocation for massive-MIMO-enabled URLLC IIoT networks,” *IEEE J. Sel. Areas Commun.*, vol. 38, no. 5, pp. 816–830, May 2020.

[12] X. Sun, S. Yan, N. Yang, Z. Ding, C. Shen, and Z. Zhong, “Short packet downlink transmission with non-orthogonal multiple access,” *IEEE Trans. Wirel. Commun.*, vol. 17, no. 7, pp. 4550–4564, Jul. 2018.

[13] J. Yao, Q. Zhang, and J. Qin, “Joint decoding in downlink NOMA systems with finite blocklength transmissions for ultrareliable low-latency tasks,” *IEEE Internet Things J.*, vol. 9, no. 18, pp. 17705–17713, Sept. 2022.

[14] J. Chen, L. Zhang, Y.-C. Liang, and S. Ma, “Optimal Resource allocation for multicarrier NOMA in short packet communications,” *IEEE Trans. Veh. Technol.*, vol. 69, no. 2, pp. 2141–2156, Feb. 2020.

[15] J. Chen, L. Zhang, Y.-C. Liang, X. Kang, and R. Zhang, “Resource allocation for wireless-powered IoT networks with short packet communication,” *IEEE Trans. Wirel. Commun.*, vol. 18, no. 2, pp. 1447–1461, Feb. 2019.

[16] H. Ren, C. Pan, Y. Deng, M. Elkashlan, and A. Nallanathan, “Joint power and blocklength optimization for URLLC in a factory automation scenario,” *IEEE Trans. Wirel. Commun.*, vol. 19, no. 3, pp. 1786–1801, March 2020.

[17] H. Ren, C. Pan, Y. Deng, M. Elkashlan, and A. Nallanathan, “Resource allocation for secure URLLC in mission-critical IoT scenarios,” *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5793–5807, Sept. 2020.

[18] W. R. Ghanem, V. Jamali and R. Schober, “Optimal Resource Allocation for Multi-User OFDMA-URLLC MEC Systems,” *IEEE Open Journal of the Communications Society*, vol. 3, pp. 2005-2023, 2022.

[19] C. She, Y. Duan, G. Zhao, T. Q. S. Quek, Y. Li, and B. Vucetic, “Cross-layer design for mission-mritical IoT in mobile edge computing systems,” *IEEE Internet Things J.*, vol. 6, no. 6, pp. 9360–9374, Dec. 2019.

[20] Z. Liu, Y. Zhu, Y. Hu, P. Sun, A. Schmeink “Reliability-oriented design framework in NOMA-assisted mobile edge computing,” *IEEE Access*, vol. 10, pp. 103598-103609, 2022.

[21] Y. Zhou, F. R. Yu, J. Chen, and B. He, “Joint resource allocation for ultra-reliable and low-Latency radio access networks with edge computing,” *IEEE Trans. Wirel. Commun.*, vol.22, no. 1, pp. 444–460, Jan. 2022.

[22] J. Li, J. Tang, and Z. Liu, “On the data freshness for industrial Internet of things with mobile-edge computing,” *IEEE Internet Things J.*, vol. 9, no. 15, pp. 13542–13554, Aug. 2022.

[23] Y. Zhu, Y. Hu, A. Schmeink, and M. C. Gursoy, “Energy minimization of mobile edge computing networks with HARQ in the finite blocklength regime,” *IEEE Trans. Wirel. Commun.*, vol.21, no. 9, pp. 7105–7120, Sept. 2022.

[24] Z. Zhou, H. Liao, X. Zhao, B. Ai, and M. Guizani, “Reliable task offloading for vehicular fog computing under information asymmetry and information uncertainty,” *IEEE Trans. Veh. Technol.*, vol. 68, no. 9, pp. 8322–8335, Sept. 2019.

[25] M. Lombardi, M. Milano, and L. Benini, “Robust scheduling of task graphs under execution time uncertainty,” *IEEE Trans. Comput*, vol. 62, no. 1, Jan. 2013.

[26] J. R. Lorch and A. J. Smith, “Improving dynamic voltage scaling algorithms with PACE,” in *Proc. ACM Sigmetrics Performance Evaluation Review*, 2001, pp. 50–61.

[27] W. Zhang, Y. Wen, K. Guan, D. Kilper, H. Luo, and D. O. Wu, “Energy-optimal mobile cloud computing under stochastic wireless channel,” *IEEE Trans. Wirel. Commun.*, vol. 12, no. 9, pp. 4569–4581, Sept. 2013.

[28] J. Yang, X. Ge, J. Thompson, and H. Gharavi, “Power-consumption outage in beyond fifth generation mobile communication systems,” *IEEE Trans. Wirel. Commun.*, vol. 20, no. 2, pp. 897–910, Feb. 2021.

[29] G. Muhammad, M. F. Alhamid, M. Alsulaiman, and B. Gupta, “Edge computing with cloud for voice disorder assessment and treatment,” *IEEE Commun. Mag.*, vol. 56, no. 4, pp. 60–65, April 2023.

[30] L. Qin, H. Lu, and F. Wu, “When the user-centric network meets mobile edge computing: Challenges and optimization,” *IEEE Commun. Mag.*, vol. 61, no. 1, pp. 114–120, Jan. 2023.

[31] P. A. Apostolopoulos, E. E. Tsiropoulou and S. Papavassiliou, “Risk-aware data offloading in multi-server multi-access edge computing environment,” *IEEE/ACM Trans. Netw.*, vol. 28, no. 3, pp. 1405–1418, June 2020.

[32] I. Polik and T. Terlaky, “Interior point methods for nonlinear optimization,” in *Nonlinear Optimization*, G. Di Pillo and F. Schoen, Eds., 1st ed. New York, NY, USA: Springer, 2010, ch. 4.

[33] Y. Hu, Y. Zhu, M. C. Gursoy, and A. Schmeink, “SWIPT-enabled relaying in IoT network operating with finite blocklength codes,” *IEEE J. Sel. Areas Commun.*, vol. 37, no. 1, pp. 74–88, Jan. 2019.

[34] J. C. Bezdek and R. J. Hathaway, “Some notes on alternating optimization,” in *Proc. Adv. Soft Comput. (AFSS)*, Calcutta, India, Feb. 2002, pp. 288–300.

[35] S. Boyd and L. Vandenberghe, Convex Optimization. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[36] Y. Sun, P. Babu, and D. P. Palomar, “Majorization-minimization algorithms in signal processing, communications, and machine learning,” *IEEE Trans. Signal Process.*, vol. 65, no. 3, pp. 794-816, Feb. 2017.

[37] J. Rubio, A. Pascual-Iserte, D. P. Palomar, and A. Goldsmith, “Joint optimization of power and data transfer in multiuser MIMO systems,” *IEEE Trans. Signal Process.*, vol. 65, no. 1, pp. 212-227, Jan. 2017.

[38] M. Abramowitz and I. A. Stegun, Handbook of mathematical functions with formulas, graphs, and mathematical tables, New York: Dover, 1972.

[39] A. Adhikary, X. Lin, Y. Wang, “Performance Evaluation of NB-IoT Coverage,” in *IEEE Vehicular Technology Conference (VTC-Fall)*, Montreal, Canada, 2016, pp. 1-5.

[40] J. Lee, Y. Kim, Y. Kwak, J. Zhang, A. Papasakellariou, T. Novlan, C Sun, and Y. Li, “LTE-advanced in 3GPP Rel-13/14: An evolution toward 5G,” *IEEE Commun. Mag.*, vol. 54, no. 3, pp. 36-42, March 2016.