Joint Time and Power Allocation for NOMA-Assisted Low-Latency Mobile Edge Computing

Zhilin Liu[†‡], Yao Zhu[†‡*], Yulin Hu[†‡*], Peng Sun[¶], Ning Guo[†‡], and Anke Schmeink[‡]

[†]School of Electronic Information, Wuhan University, China, Email: *zhilin_liu|yulin.hu|ning.guo*@whu.edu.cn
[‡]INDA Institute, RWTH Aachen University, Germany, Email: *zhu|schmeink*@inda.rwth-aachen.de
[¶]Duke Kunshan University, KunShan, China, Email: *peng.sun568*@duke.edu

***Abstract*—In this paper, we study a mobile edge computing (MEC) network supporting latency-critical tasks. Data information generated at multiple devices are offloaded to and processed at the MEC node. Each service in the network is divided into two phases, i.e., a non-orthogonal multiple access (NOMA)-assisted communication phase and a MEC server computation phase, while the whole task offloading process is required to satisfy high-reliability and low-latency. We characterize the overall service error probability of the network, while taking into account the finite blocklength (FBL) impacts on both the communication and the queuing impacts on computation. Accordingly, a joint optimal design is introduced to minimize the overall service error probability by determining the phase lengths and transmit power at NOMA users. In particular, the formulated problem is nonconvex, for which a modified block coordinate descent method is proposed in order to decompose the problem into sub-problems which are characterized and solved efficiently. By means of simulations, we validate our analytical model and evaluate the considered network.**

***Index Terms*—mobile edge computing, NOMA, finite blocklength, resource allocation.**

I. Introduction

In future Internet of Things (IoT) networks, many devices are expected to monitor, sense, and generate enormous data that need to be processed timely for various smart industrial applications [1]–[4]. However, the computation-intensive and latency-critical tasks are unlikely to be handled by the resource constrained and non-rechargeable IoT devices themselves. To address this issue, mobile edge computing (MEC) is proposed, where the servers, e.g., base station or access point, deployed close to the users, could provide the computation services, i.e., significantly shortening the transmission time cost in comparison to a cloud computing. Despite of the enormous advantages of MEC, the key challenge in MEC remains to be the interplay between the communication and the computation with respect to latency and energy constraints [5]–[7].

On the one hand, due to the difference in the transmission power of multiple IoT devices as well as the random channel behavior, non-orthogonal multiple access (NOMA) is envisioned to be a promising solution that can further improve the performance of MEC systems [8]. Compared to conventional orthogonal multiple access (OMA) techniques, the use of NOMA can greatly improve the spectrum utilization by allowing users to share the radio resource block simultaneously in the power domain and thus, results in a higher network throughput and further reduce latency in the MEC network, especially there is a wide gap of the users' channel gains. Motivated by the benefit of NOMA in comparison to OMA, it has received an increasing attention from the research community. In [9], the authors proposed a NOMA-based optimization framework via jointly optimizing the user clustering, computing and communication resource allocation to improve the energy efficiency. In addition, the authors in [10] jointly consider a task offloading decision, local CPU frequency scheduling, power control, MEC computation resource and subchannel resource allocation to minimize the energy consumption of all users. However, so far it is still an open issue how the choices of communication and computation resource allocation influence the reliability of the NOMA-assisted MEC service, especially when an imperfect successive interference cancellation (SIC) in the NOMA process is considered.

In particular, most of the aforementioned studies are based on the ideal assumption of code words that transmissions are arbitrarily reliable at the Shannon capacity with infinite blocklength (IBL). This is actually overoptimistic for MEC services which usually require high reliability and low-latency. In fact, a more accurate model for low-latency MEC networks is often referred to as finite blocklength (FBL) model, with which the closed-form expression of a maximum allowable coding rate is provided by Polyanskiy *et al.* [11]. On this basis, FBL impacts have been studied in many wireless networks, such as quality-of-service (QoS) constrained downlink networks [12], multi-hop relaying networks [13] and frequency-selective fading channels [14]. Recently, in [15] the FBL performance of a NOMA-assisted MEC network is investigated.

In addition to the FBL of the communication phase of a MEC service, the time length/budget for computation is also limited. For the multi-user case, where the server provides computation power to multiple offloaded tasks, data may need to wait in a buffer before it can be processed. Hence, it is possible that the total computation time (including the

This work was supported by the seed-fund support program at the WHU-DKU joint research platform under Grant WHUDKUZZJJ202201, and in part by BMBF Germany in the program of "Souverän. Digital. Vernetzt." Joint Project 6G-RIC with project identification number 16KISK028.

*Y. Zhu and Y. Hu are corresponding authors.

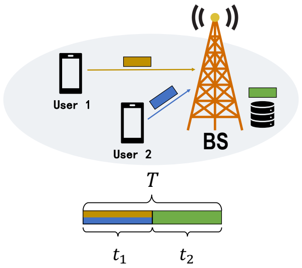

Fig. 1. System Model

waiting time) violates the targeted delay, which is unknown for the users. To this end, in [16], the authors investigated the queues attended by a single server system and the queue time distribution. In [17], the authors present results for the queue with Poisson arrivals. It is worth noting that the reliability of both communication and computation phase are related to their respective delay tolerances, since the total offloading process consists of wireless data transmission and task execution at the server sides. Therefore, it is significant to investigate the trade-off of the communication phase and computation phase in the perspective of overall service reliability.

In this paper, we study the joint blocklength (in symbols) and transmit power allocation for the considered MEC network, where uplink-NOMA is carried out for data uploading and the queue behavior of the MEC server is taken into account. Following the characterization of the total error probability, we formulate an optimization problem that minimizes such error probability. To address the issue of nonconvexity, we decompose the original problem into two sub-problems and address them independently. In those sub-problems, on one hand, we reduce the dimension of the variables via exploiting the interplay between communication and computation phases with respect to the energy constraint. On the other hand, *for the first time*, we analyze the joint convexity of decoding error probability with respect to the transmit power of NOMA users, and hence, provide the convexity of the corresponding sub-problem. These sub-problems are solved iteratively, and finally the (nearly) optimal solution of the original problem is obtained.

The rest of the paper is organized as follows: The system model is introduced in Section II. The reliability performance of the considered network is derived in Section III. In Section IV we propose our reliability-oriented design. We provide simulation results in Section V. Finally, we conclude our work in Section VI.

II. System model

We consider a MEC network, where two local users, e.g., sensor nodes, upload the local information to the base station mounted with a server in an uplink NOMA manner via a wireless link. For example, the applications relating to industrial processes, which need to process their collected data (e.g. video) from the sensors reliably and in real-time, in order for their mission to proceed safely. The server computes a task after receiving the local information belonging to the same task from both users as the prerequisite input. The computation results are for real-time control. Therefore, it has a stringent delay requirement. In other words, the whole offloading process, which consists of both communication and computation phases, is demanded to be accomplished before a given deadline of $T_{\max}$. Furthermore, we denote by t_1 the time duration of the communication phase for both users [18]. And t_2 is denoted as the time duration of the computation phase, as shown in Fig. 1. It must holds that $T = t_1 + t_2 \leq T_{\max}$. Let T_{s} be the duration of one symbol. Then, the total available blocklength can be written as $M_{\max} = \frac{T_{\max}}{T_{\mathrm{s}}}$. Similarly, we have $m_1 = \frac{t_1}{T_{\mathrm{s}}}$ and $m_2 = \frac{t_2}{T_{\mathrm{s}}}$ as the corresponding blocklength of each phase, as well as the blocklength constraint $m_1 + m_2 \leq M_{\max}$.

Note that the NOMA scheme is carried out for the data transmissions from the users to the server in the communication phase, i.e., user 1 and user 2 offload the data with size of d_1 and d_2, respectively, via shared wireless channels. We denote by p_1 and p_2 the transmit power of user 1 and user 2, respectively. Since the total energy of the system is constrained, the energy consumption should fulfill $m_1 T_{\mathrm{s}}(p_1+p_2) \leq E_{\max}$, where $E_{\max}$ is the maximal allowed energy consumption and the energy of computation phase is fixed. We assume that the channels experience block-fading, i.e., the channel state is constant within the frame, but may vary in the next. Furthermore, the channels from different users to the server are assumed to be independent. In particular, let z_1 and z_2 be the channel power gains with path-loss for user 1 and user 2, respectively. Then, after the transmissions, the server receives the signal y as follows:

$$y = \sqrt{z_1 p_1} x_1 + \sqrt{z_2 p_2} x_2 + w, \tag{1}$$

where $w \sim (0, \sigma^2)$ represents the additive white Gaussian noise (AWGN) with mean zero and variance σ^2. In addition, x_1 and x_2 are the users' transmitted signals. Note that the server always decodes the strong signal first. Therefore, for the sake of clarity, we refer to the stronger user as user 2, i.e., $z_2 p_2 \geq z_1 p_1$. After receiving signal y, the server first attempts to decode the signal x_2 from the stronger user based on the signal-to-interference-plus-noise ratio (SINR) expressed as

$$\gamma_{2|1} = \frac{z_2 p_2}{z_1 p_1 + \sigma^2} \approx \frac{z_2 p_2}{z_1 p_1}. \tag{2}$$

The approximation is based on the assumption that compared to interference the noise is low enough to be ignored. Once user 2's signal has been successfully decoded by the server, it leverages SIC to extract signal x_2 from y and decodes user 1's signal with the signal-to-noise ratio (SNR) (without interference) given by $\gamma_{1|1} = \frac{z_1 p_1}{\sigma^2}$. Note that communications based on FBL codes may result in errors. Therefore, the assumption of SIC always to be reliable does not hold. As a result, the server has to decode the weak signal while the interference is preserved, i.e., the SINR is given by $\gamma_{1|2} = \frac{z_1 p_1}{z_2 p_2}$. Recall that the task requires the data from both user as mandatory input. The computation can only start after both packets being successfully decoded. Then, the task is put into the server's queue waiting to be proceeded. Clearly, the actual duration of the computation phase depends on the waiting time and the required workloads. Therefore, it is possible that it exceeds t_2. We investigate the possibility of such events to characterize the error probability in the next section.

III. End-to-End Reliability Characterization

In this section, we first characterize the error probability of the task offloading of communication operating with FBL codes. Subsequently, the computation error probability is characterized by considering the queue time distribution. Finally, we conduct the expression of the end-to-end error probability of the considered NOMA-assisted MEC network.

A. Reliability of FBL Communication

Note that the blocklength is considered as finite due to the demands of low latency. Therefore, the Shannon capacity is longer accurate to characterize the FBL performance directly. Therefore, the work in [11] investigated the relationship between transmission rate and error probability and provided a closed-form expression of the (block) error probability:

$$\varepsilon = \mathcal{P}(\gamma, \tfrac{d}{m}, m) \approx Q\Big(\sqrt{\tfrac{m}{V(\gamma)}}(\mathcal{C}(\gamma) - \tfrac{d}{m}) \ln 2\Big), \tag{3}$$

where $\mathcal{C}(\gamma) = \log_2(1+\gamma)$ is the Shannon capacity and $V(\gamma)$ is the channel dispersion. It is worth noting that for a complex AWGN channel it holds that $V(\gamma) = 1 - (1+\gamma)^{-2}$ [11]. In addition, $Q(x)$ is the Q-function defined as $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$.

Recall that the signal of user 2 is always decoded with interference according to (2). Then, we can characterize its error probability as $\varepsilon_{2|1} = \mathcal{P}(\gamma_{2|1}, d_2, m_1)$. Assuming that the signal of user 2 is decoded correctly, in this condition, the decoding error probability of user 1 can be expressed as $\varepsilon_{1|1} = \mathcal{P}(\gamma_{1|1}, d_1, m_1)$. Otherwise, i.e., if the previous decoding fails, the error probability of user 1 is given by $\varepsilon_{1|2} = \mathcal{P}(\gamma_{1|2}, d_1, m_1)$.

Note that without SIC, the decoding will fail with a high probability since $\gamma_{1|2}$ is generally less than $\gamma_{1|1}$, and is considerably less than $\gamma_{2|1}$. Based on this, for the probability of decoding failure it holds that $\varepsilon_{1|2} \approx 1$. In the light of this, the overall decoding error probability for user 1 is expressed as follows:

$$\varepsilon_{1-1} = (1 - \varepsilon_{2|1})\varepsilon_{1|1} + \varepsilon_{2|1}\varepsilon_{1|2} \approx \varepsilon_{2|1} + \varepsilon_{1|1}. \tag{4}$$

Since the server first attempts to decode the signal from the stronger user, i.e., user 2, so the error probability of decoding user 2's signal can be obtained directly, i.e.,$\varepsilon_{1-2} = \varepsilon_{2|1}$. To summarise, the error probability of the communication phase is given by

$$\begin{aligned} \varepsilon_1 &= 1 - (1 - \varepsilon_{1-1})(1 - \varepsilon_{1-2}) \\ &= \varepsilon_{1-1} + \varepsilon_{1-2} - \varepsilon_{1-1}\varepsilon_{1-2}. \end{aligned} \tag{5}$$

B. Reliability Model in the Computation Phase

Then, we characterize the computation error of the MEC server in the computation phase of time length t_2. D is used to denote the computing time. We assume that the server follows the first-come-first-serve (FCFS) principle, i.e., tasks from the two users are managed through a queue in a FCFS manner. In general, when tasks arrive at the MEC server, due to the limited computing capability of the MEC server, they cannot be executed immediately, so they may wait in the queue for further service. Therefore, the computing time at the MEC server has two parts: task execution time and the latency of queue (delay in waiting time in the queue buffer). We consider the offloading task follows data-partition model [19]. We denote by c the workload of the server, computation power is denoted by f and W is the queuing latency, which is determined by the computation power f and the queue length c_w, i.e., $W = \frac{c_w}{f}$. The computing time D can be given by

$$D = \frac{c}{f} + W, \tag{6}$$

If the computing time exceeds the maximum allowable time, a computation delay violation error will occur. We use $Pr\left(D \geq t_2\right)$ to express the probability that the computation time exceeds t_2. As a result, the probability of computation error at the server is given by

$$\varepsilon_2 = \Pr\left(D \geq t_2\right). \tag{7}$$

Since the workloads and CPU-frequency of the MEC server is usually definite, the distribution of the computing time is related to the distribution of the waiting time W, i.e.,

$$\varepsilon_2 = \Pr\left(D \geq t_2\right) = \Pr\left(W \geq t_2 - \frac{c}{f}\right). \tag{8}$$

Note that the waiting time W is non-negative, for more accurate analysis, a modified delay tolerance is introduced in the computation phase, given by $\hat{t}_2 = \max\left\{t_2 - \frac{c}{f}, 0\right\}$, for which we have $\varepsilon_2 = \Pr\left(W \geq \hat{t}_2\right)$.

The arrival of the tasks at the queue of the server will be processed following a Poisson process with an arrival rate λ [20] since we assume that the server follows FCFS principle. So the queue length c_w of the server obey the Poisson distribution. According to Little's Law, the queuing latency W also obeys the Poisson distribution [21] and in linear correlation with c_w. Moreover, the execution time $\frac{c}{f}$ is determined, i.e., the computing time also follows Poisson distribution. Therefore, with $\lambda\hat{t}_2$ given, the error probability of computation phase can be given by

$$\varepsilon_2 = (1 - F_W\left(\hat{t}_2\right)) = e^{-\lambda\hat{t}_2}. \tag{9}$$

Where $F_W(x)$ is the complementary cumulative distribution function (CCDF) of the queue delay W.

C. End-to-End Error Probability

Note, that the whole service process includes two phases: a communication phase and a computation phase. Therefore, the service can complete successfully only when there is no communication error and computation error occur on the server. We denote by ε_o the end-to-end error probability, which can be expressed as

$$\varepsilon_o = \varepsilon_1 + \varepsilon_2 - \varepsilon_1\varepsilon_2 \approx 2\varepsilon_{2|1} + \varepsilon_{1|1} + \varepsilon_2. \tag{10}$$

The approximation tightly holds, since the product of ε_1 and ε_2 is significantly lower than ε_1 and ε_2. Similarly, the product of ε_{1-1} and ε_{1-2} is much smaller than ε_{1-1} and ε_{1-2}.

IV. Framework Optimization

A. Problem Formulation

We aim at minimizing ε_o by jointly allocating the transmit power $\mathbf{p} = \{p_1, p_2\}$ and the blocklength assigned to each phase $\mathbf{m} = \{m_1, m_2\}$, i.e., the time phases $\mathbf{t} = \{t_1, t_2\}$ Meanwhile, the total energy consumption constraint of users and the total blocklength limitation should also be taken into consideration. Then, the original optimization problem can be formulated as

$$\min_{\mathbf{m},\mathbf{p}} \quad \varepsilon_o \tag{11a}$$

$$\text{s.t.} \quad m_1 T_{\mathrm{s}}(p_1 + p_2) \leq E_{\max}, \tag{11b}$$

$$m_1 + m_2 \leq M_{\max}. \tag{11c}$$

Note that (11c) is the constraint of total blocklength, which can be expressed as $t_1 + t_2 = (m_1 + m_2)T_s \leq T_{\max}$ in time domain.

B. Optimal solution of problem (11)

However, directly solving Problem (11) seems intractable since it is a non-convex problem due to the non-convex objective function as well as the energy constraint (11b) that involves variable multiplications. To this end, we leverage an interactive search method similar to the block coordinate descent (BCD) optimizer [22].

In particular, in the k-th ($k = 1, 2, 3, ...$) iteration, we optimize either blocklength $\mathbf{m}$ or transmit power $\mathbf{p}$ while fixing the other. Then, we update the variable in the $(k+1)$-th iteration with the obtained solution in the previous k-th iteration. The iteration keeps repeating until the stop criteria is fulfilled, i.e., the gap between two iterations is lower than the threshold. Therefore, in what follows, we investigate the two sub-problems derived via decomposing the original problem (11). The optimal solutions of those sub-problems are provided via our analytical finding. Finally, we present our proposed interactive algorithm to solve Problem (11).

First, we consider a sub-problem of the original problem (11) by fixing the power $\mathbf{p}$ as $\mathbf{p}^{(k-1)}$ at k-th interaction, denoted as $P1^{(k)}$. Hence, $P1^{(k)}$ can be expressed as

$$\min_{\mathbf{m}} \quad \varepsilon_o \tag{12a}$$

$$\text{s.t.} \quad \mathbf{p} = \mathbf{p}^{(k-1)}, \text{ (11b) and (11c).} \tag{12b}$$

We provide the following key lemma to address the problem

Lemma 1. *The objective function of problem (12) is jointly convex in* $\mathbf{m}$*.*

Proof. To prove the joint-convexity of ε_o in $\mathbf{m}$, we show the Hessian matrix of ε_o as follows:

$$\mathbf{H} = \begin{pmatrix} \frac{\partial^2 \varepsilon_o}{\partial m_1^2} & \frac{\partial^2 \varepsilon_o}{\partial m_1 \partial m_2} \\ \frac{\partial^2 \varepsilon_o}{\partial m_2 \partial m_1} & \frac{\partial^2 \varepsilon_o}{\partial m_2^2} \end{pmatrix}. \tag{13}$$

Recall that the overall error probability of the system can be given by $\varepsilon_o = 2\varepsilon_{2|1} + \varepsilon_{1|1} + \varepsilon_2$. Therefore, it is convex if every component is also convex. First, $\varepsilon_{2|1}$, $\varepsilon_{1|1}$ only depends on m_1 and ε_2 only depends on m_2, i.e.,

$$\frac{\partial^2 \varepsilon_o}{\partial m_1 \partial m_2} = \frac{\partial^2 \varepsilon_o}{\partial m_2 \partial m_1} = 0. \tag{14}$$

Next, the upper-left element of $\mathbf{H}$, i.e., the second derivative of ε_o can be decomposed as

$$\frac{\partial^2 \varepsilon_o}{\partial m_1^2} = 2\frac{\partial^2 \varepsilon_{2|1}}{\partial m_1^2} + \frac{\partial^2 \varepsilon_{1|1}}{\partial m_1^2} + \frac{\partial^2 \varepsilon_2}{\partial m_1^2}. \tag{15}$$

The second derivative of $\varepsilon_{2|1}$ and the second derivative of $\varepsilon_{1|1}$ can be demonstrated to be non-negative in [23].

Notice that ε_2 is independent of m_1, so $\frac{\partial^2 \varepsilon_2}{\partial m_1^2} = 0$. As a result, we have

$$\frac{\partial^2 \varepsilon_o}{\partial m_1^2} = 2\frac{\partial^2 \varepsilon_{2|1}}{\partial m_1^2} + \frac{\partial^2 \varepsilon_{1|1}}{\partial m_1^2} + \frac{\partial^2 \varepsilon_2}{\partial m_1^2} \geq 0. \tag{16}$$

And $\frac{\partial^2 \varepsilon_o}{\partial m_2^2}$ can be expressed as

$$\frac{\partial^2 \varepsilon_o}{\partial m_2^2} = \lambda^2 T_s^2 e^{\left(-\frac{\lambda^2 \eta^2\left(m_2 - \frac{c}{fT_s}\right)T_s c}{f} \cdot m_2\right)} \cdot \eta^2 \left(m_2 - \frac{c}{fT_s}\right) \\ \geq 0. \tag{17}$$

where $\eta(x)$ is the unit step function, i.e., $\eta(x) = 1$ if $x \geq 0$, and $\eta(x) = 0$ otherwise.

As a result, we have

$$\det(\mathrm{H}) = \frac{\partial^2 \varepsilon_o}{\partial m_1^2}\frac{\partial^2 \varepsilon_o}{\partial m_2^2} - \left(\frac{\partial^2 \varepsilon_o}{\partial m_1 \partial m_2}\right)^2 \geq 0. \tag{18}$$

Thus, according to (16) and (18), the objective function of problem (12) is jointly convex in $\mathbf{m}$. ■

Note that the energy constraint (11b) is affine. According to Lemma 1, Problem (12) is convex, and the optimal solution can be efficiently obtained via standard convex programming methods. We denote the corresponding optimal solution as $\mathbf{m}^{(k)}$.

Next, we fix $\mathbf{m} = \mathbf{m}^{(k)}$ and formulate another sub-problem by optimizing $\mathbf{p}$ in the (k)-th iteration. Hence, the power allocation sub-problem $P2^{(k)}$ can be expressed as

$$\min_{\mathbf{p}} \quad \varepsilon_o \tag{19a}$$

$$\text{s.t.} \quad \mathbf{m} = \mathbf{m}^{(k)}, \text{ (11b) and (11c).} \tag{19b}$$

However, Problem (19) is still nonconvex since the energy constraint (11b) is not convex. To tackle this problem, we have the following lemma by investigating its optimal condition:

Lemma 2. *The optimal solutions to Problem (19), denoted by* $p_1{}^*$ *and* $p_2{}^*$*, hold that* $p_1^* + p_2^* = \frac{E_{\max}}{m_1^{(k)} T_s}$.

Proof. We prove this lemma by contradiction. In particular, assume there is an optimal solution $\mathbf{p}' = \{p_1', p_2'\}$ that holds $p_1' + p_2' + \alpha = \frac{E_{\max}}{m_1^{(k)} T_s}$, where $\alpha > 0$. Then, for any other solutions $\mathbf{p}$, it must hold $\varepsilon(\mathbf{p}) \leq \varepsilon(\mathbf{p}')$ due to the optimality. However, we can always construct another solution $(p_1'' = p_1', p_2'' = p_2' + \alpha) \in \left\{p_1, p_2 \mid p_1 + p_2 = \frac{E_{\max}}{m_1^{(k)} T_s}\right\}$, which is also feasible. It is trivial to show that ε_0 is decreasing in p_2. Since $p_2'' = p_2' + \alpha > p_2'$, we can conclude $\varepsilon_o(p_1'', p_2'') < \varepsilon_o(p_1', p_2')$. In other words, the assumption of $\mathbf{p}'$ being optimal is contradicted. This completes the proof. ■

$$\begin{aligned}\frac{\partial^2 \varepsilon_o}{\partial p_1^2} &= 2\frac{\partial^2 \varepsilon_{2|1}}{\partial p_1^2} + \frac{\partial^2 \varepsilon_{1|1}}{\partial p_1^2} + \frac{\partial^2 \varepsilon_{2}}{\partial p_1^2} \;=\; 2\left(\frac{\partial \varepsilon_{2|1}}{\partial \gamma_{2|1}}\frac{\partial^2 \gamma_{2|1}}{\partial\ p_1^2} + \frac{\partial^2 \varepsilon_{2|1}}{\partial\ {\gamma_{2|1}}^2}\left(\frac{\partial \gamma_{2|1}}{\partial\ p_1}\right)^2\right) + \frac{\partial^2 \varepsilon_{1|1}}{\partial \gamma_{1|1}^2}\frac{z_1^2}{\sigma^4} + 0\\ &\geq\; (\frac{2}{\sqrt{2\pi}}\frac{2\left(E_{\max}/\left(m_1 T_s\right)\right).}{z_1 \left(p_1\right)^3}\frac{\partial \omega_{2|1}}{\partial \gamma_{2|1}}) \cdot (\omega_{2|1} z_1 \frac{\partial \omega_{2|1}}{\partial \gamma_{2|1}} \underbrace{\frac{\left(E_{\max}/\left(m_1 T_s\right)\right)}{z_1 \left(p_1\right)^2}}_{>1} - 2z_2) + \frac{\partial^2 \varepsilon_{1|1}}{\partial \gamma_{1|1}^2}\frac{z_1^2}{\sigma^4}\\ &\overset{z_2 \geq z_1}{\geq}\; \frac{2}{\sqrt{2\pi}}\frac{2z_2\left(E_{\max}/\left(m_1 T_s\right)\right).}{z_1 \left(p_1\right)^3} \cdot \underbrace{(\omega_{2|1}\frac{\partial \omega_{2|1}}{\partial \gamma_{2|1}}\frac{\left(E_{\max}/\left(m_1 T_s\right)\right)}{z_1 \left(p_1\right)^2} - 2)}_{\geq 0} + \frac{\partial^2 \varepsilon_{1|1}}{\partial \gamma_{1|1}^2}\frac{z_1^2}{\sigma^4} \;\geq\; 0.\end{aligned} \tag{20}$$

By exploiting Lemma 2, we can substitute p_2 with $\frac{E_{\max}}{m_1 T_S} - p_1$. Problem (19) can be further reformulated as

$$\min_{p_1} \quad \varepsilon_o \tag{21a}$$

$$\text{s.t.} \quad p_2 = \frac{E_{\max}}{m_1 T_S} - p_1, \tag{21b}$$

$$\mathbf{m} = \mathbf{m}^{(k)}. \tag{21c}$$

We have the following lemma to solve problem (21), denoted as $\hat{P}2^{(k)}$:

Lemma 3. *Problem (21) is convex.*

Proof. It is obvious that all constraints of Problem (21) are affine. Therefore, in order to prove the convexity of Problem (21), we only need to focus on the convexity of the objective function ε_o. For the overall error probability ε_o ,the second derivative with respect to p_1 is given as (20) on the top of the page.

As a result, ε_o is convex in p_1. In summary, both the objective function and constraints are convex, i.e., Problem (21) is a convex optimization problem. ■

According to Lemma 3, similar to the previous sub-problem, Problem (21) is also convex. And the convexity of the two sub-problems still holds in multi-user scenarios. Thus, we can obtain the optimal power allocation by efficiently solving it via convex programming. This solution is served as the fixed transmission power $\mathbf{p}^{(k)}$. This process can be iterated in the next, where we solve problem $P1^{(k+1)}$ by fixing $\mathbf{p} = \mathbf{p}^{(k)}$. This iteration will continue until the stop criteria fulfills, i.e., $\varepsilon_{(k)} - \varepsilon_{(k-1)} < \theta$, where $\theta > 0$ is a given threshold depending on the pre-defined resolution. Finally, we take $\mathbf{m}^* = \mathbf{m}^{(k)}$ and $\mathbf{p}^* = \mathbf{p}^{(k)}$ as the solutions for original problem. Specially, to initialize the iteration, we set ${p_1}^{(0)}$ and ${p_2}^{(0)}$ to $p_1^{(0)} = p_2^{(0)} = \frac{E_{\max}}{M_{\max} T_s}$ as the initial points. It is worth to mention that the optimization of the powers can also be done entirely separately with iterative search algorithm but the optimal solution can not be obtained and the complexity of the algorithm is high. However, the presented block coordinate descent algorithm is simple to construct and is able to achieve a nearly optimal solution with the complexity of $\mathcal{O}\left((N)^2 \ln(1/\delta)\right)$, where δ is the solution accuracy. The convergence of the algorithm is ensured by the convexity of problem (12) and problem (19) and the convergence speed is related to the complexity and solution accuracy [22].

V. Numerical Results

In this section, numerical simulations are deployed to evaluate our analytical characterizations and the modified BCD algorithm. The maximum allowable time $T_{\max}$ is set as 0.1s. Since we set the duration of one symbol as $T_s = 0.25$ms, the corresponding total available blocklength is $M_{\max} = 400$ symbols. The maximum energy limitation of $E_{\max} = 400$J for the reliability-oriented scheme. For the offloading via the wireless links, noise power $\sigma^2 = 0.001W$, the two channels are assumed to experience independent and identically distributed block Rayleigh fading and we set a unit average channel gain (including path-loss). Assuming that the data size is $d_1 = d_2 = 375$bits in each time frame. The arriving tasks follows the Poisson arrivals with rate $\lambda = 3$ M cycles/s. For MEC server execution, we set the CPU-frequency of the server as $f = 3$ GHz and unless otherwise specified, the workloads of the server is set as $c_o = 24$Mcycles. Exhaustive search is set as the performance bound of our proposed algorithm.

We start with Fig. 2 to investigate the system's performance as the data size increases. The reliability of the system decreases with the increasing data size regardless of different settings of blocklength and energy constraint. We can see that the performance of our algorithm nearly matches the one of exhaustive search, which shows the advantage of our proposed design. It is obvious that higher energy consumption or blocklength budget correspond to lower error probability. However, when the total energy constraint is relatively loose, the reliability of the system does not decline significantly with the data size improves. Fig. 2 shows the trade-off of blocklength and energy as well as the packet size with consideration of system's reliability, which is instructive for us to design the system in practical situations, especially when the resources are limited.

Then, in Fig. 3, we show the impact of another important parameter, the noise power σ^2 on the reliability of the system. We can see that with the increase of noise power, the reliability of the system decreases and the decreasing speed gradually increases. At the same time, the performance difference between our proposed algorithm and exhaustive search is more and more obvious, since the ignorance of noise power when we characterize SINR. Furthermore, although the channel gains we set differ greatly, the system reliability gap is not obvious. This is due to one of the advantages of the NOMA technique exploiting the channel differences between users. In practical scenarios, even if the channel condition of one user occasionally becomes very poor, the interference to

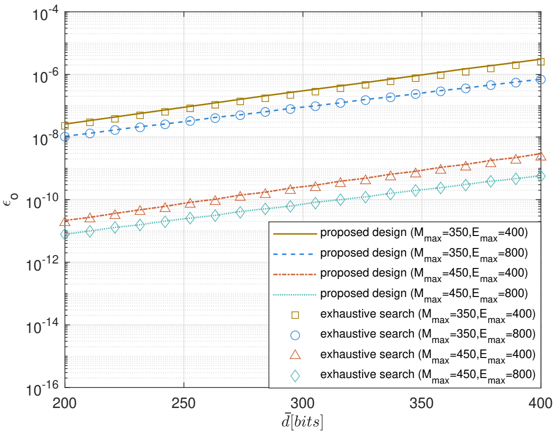

Fig. 2. The comparison of system's reliability with the change of data size under different setups of total blocklength and energy constraint. The performance of our proposed design and exhaustive search are evaluated.

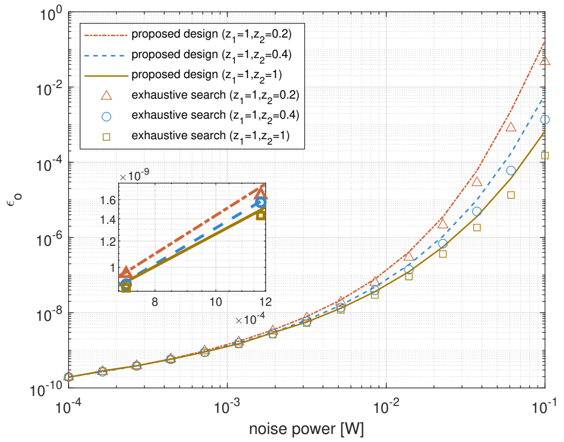

Fig. 3. Overall error probability against noise power with different setups of channel gain. Evaluations of the performance between our proposed design and exhaustive search.

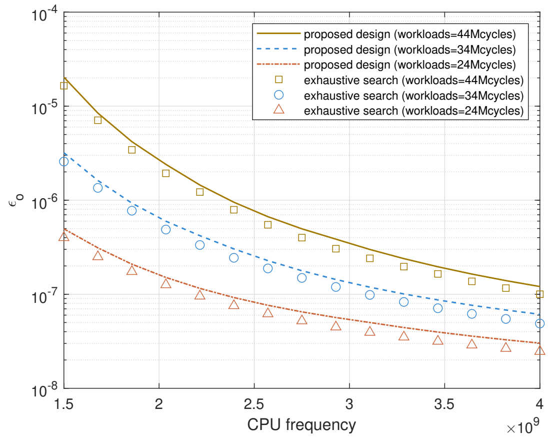

Fig. 4. The influence of the CPU frequency on the system reliability with various workloads. A comparison of our proposed design and exhaustive search is shown.

the system's reliability is limited compared to a traditional OMA system.

Finally, we focus on the influence of CPU frequency on the system's overall error probability in Fig. 4. On the one hand, as the CPU frequency becomes higher, the error probability decreases gradually. And when the CPU frequency is limited, the reliability of the system varies greatly with different setups of workloads. In this case, the error probability of the system mainly depends on the error probability in the computation phase and when the workload is large, the execution time increases and the computation delay violation error is more likely to occur. On the other hand, when the CPU frequency is relatively large, the system's performance bottleneck depends on communication.

VI. Conclusion

In this paper, we developed a reliability-optimal design for a NOMA-assisted MEC system, where the blocklength for the communication and computation phases and the transmit power of each user were jointly optimized to minimize the error probability. In order to solve the highly coupled non-convex problem, we apply the BCD method to decompose the original problem into a series of solvable problems and the convexity of those problems were further proved. We evaluated our analytical models and confirmed the advantages of the proposed design via numerical simulations, which also show that the proposed design nearly achieved the performance of exhaustive search. The results of this approach give a series of guidelines for a practical system design also showing the effects of various setups. The results of our work will be used to facilitate the studies on multi-user scenarios in our future work.

VII. Acknowledgement

This work was supported in part by the China National Key Research and Development Program under Grant 2021YFB2900301, in part by BMBF Germany in the program of "Souverän. Digital. Vernetzt." Joint Project 6G-ANNA with project identification number 16KISK097.

References

[1] X. Krasniqi and E. Hajrizi, "Use of IoT Technology to Drive the Automotive Industry from Connected to Full Autonomous Vehicles," *IFAC-PapersOnLine*, vol. 49, no. 29, pp. 269–274, 2016, 17th IFAC Conference on International Stability, Technology and Culture TECIS 2016.

[2] A. Albahri, J. K. Alwan, Z. K. Taha, S. F. Ismail, R. A. Hamid, A. Zaidan, O. Albahri, B. Zaidan, A. Alamoodi, and M. Alsalem, "IoT-based telemedicine for disease prevention and health promotion: State-of-the-Art," *Journal of Network and Computer Applications*, vol. 173, p. 102873, 2021.

[3] D. Yan-e, "Design of Intelligent Agriculture Management Information System Based on IoT," in *2011 Fourth International Conference on Intelligent Computation Technology and Automation*, vol. 1, 2011, pp. 1045–1049.

[4] X. You, H. Yin, and H. Wu, "On 6G and wide-area IoT," *Chinese Journal on Internet of Things*, vol. 4, no. 1, p. 3, 2020.

[5] G. Cui, X. Li, L. Xu, and W. Wang, "Latency and Energy Optimization for MEC Enhanced SAT-IoT Networks," *IEEE Access*, vol. 8, pp. 55 915–55 926, 2020.

[6] M. Qin, N. Cheng, Z. Jing, T. Yang, W. Xu, Q. Yang, and R. R. Rao, "Service-Oriented Energy-Latency Tradeoff for IoT Task Partial Offloading in MEC-Enhanced Multi-RAT Networks," *IEEE Internet of Things Journal*, vol. 8, no. 3, pp. 1896–1907, 2021.

[7] C. Zheng, S. Liu, Y. Huang, and L. Yang, "MEC-Enabled Wireless VR Video Service: A Learning-Based Mixed Strategy for Energy-Latency Tradeoff," in *2020 IEEE Wireless Communications and Networking Conference (WCNC)*, 2020, pp. 1–6.

[8] S. Kekki, W. Featherstone, Y. Fang, P. Kuure, A. Li, A. Ranjan, D. Purkayastha, F. Jiangping, D. Frydman, G. Verin *et al.*, "MEC in 5G Networks," *ETSI white paper*, vol. 28, no. 28, pp. 1–28, 2018.

[9] A. Kiani and N. Ansari, "Edge Computing Aware NOMA for 5G Networks," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1299–1306, 2018.

[10] C. Xu, G. Zheng, and X. Zhao, "Energy-Minimization Task Offloading and Resource Allocation for Mobile Edge Computing in NOMA Heterogeneous Networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 16 001–16 016, 2020.

[11] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel Coding Rate in the Finite Blocklength Regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307 – 2359, 2010, achievability;Converse;Finite blocklength regimes;Noisy channel;Shannon theory;. [Online]. Available: http://dx.doi.org/10.1109/TIT.2010.2043769

[12] Y. Hu, M. Ozmen, M. C. Gursoy, and A. Schmeink, "Optimal Power Allocation for QoS-Constrained Downlink Multi-User Networks in the Finite Blocklength Regime," *IEEE Transactions on Wireless Communications*, vol. 17, no. 9, pp. 5827–5840, 2018.

[13] F. Du, Y. Hu, L. Qiu, and A. Schmeink, "Finite Blocklength Performance of Multi-Hop Relaying Networks," in *2016 International Symposium on Wireless Communication Systems (ISWCS)*. IEEE, 2016, pp. 466–470.

[14] Y. Wu, D. Qiao, and H. Qian, "Efficient Bandwidth Allocation for URLLC in Frequency-Selective Fading Channels," in *GLOBECOM 2020-2020 IEEE Global Communications Conference*. IEEE, 2020, pp. 1–6.

[15] Y. Yang, Y. Hu, and M. C. Gursoy, "Energy Efficiency Analysis in RIS-aided MEC Networks with Finite Blocklength Codes," in *2022 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2022, pp. 423–428.

[16] L. Takács, "Two Queues Attended By A Single Server," *Operations Research*, vol. 16, no. 3, pp. 639–650, 1968.

[17] N. Prabhu, "Some Results for the Queue with Poisson Arrivals," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 22, no. 1, pp. 104–107, 1960.

[18] F. Fang, Y. Xu, Z. Ding, C. Shen, M. Peng, and G. K. Karagiannidis, "Optimal resource allocation for delay minimization in noma-mec networks," *IEEE Transactions on Communications*, vol. 68, no. 12, pp. 7867–7881, 2020.

[19] X. Cao, F. Wang, J. Xu, R. Zhang, and S. Cui, "Joint Computation and Communication Cooperation for Energy-Efficient Mobile Edge Computing," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 4188–4200, 2019.

[20] K. Sriram and W. Whitt, "Characterizing Superposition Arrival Processes in Packet Multiplexers for Voice and Data," *IEEE Journal on Selected Areas in Communications*, vol. 4, no. 6, pp. 833–846, 1986.

[21] "A distributional form of little's law," *Operations Research Letters*, vol. 7, no. 5, pp. 223–227, 1988. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0167637788900351

[22] P. Tseng, "Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization," *Journal of optimization theory and applications*, vol. 109, no. 3, pp. 475–494, 2001.

[23] Y. Zhu, Y. Hu, X. Yuan, M. C. Gursoy, and A. Schmeink, "Joint Convexity of Error Probability to Blocklength and Transmit Power in the Finite Blocklength Regime," 2021, [Online]. Available: https://www.isek.rwth-aachen.de/2021_Yao_JointConvexity_letter.pdf.