Energy Minimization of Mobile Edge Computing Networks with HARQ in the Finite Blocklength Regime

Yao Zhu, *Student Member, IEEE,* Yulin Hu*, *Senior Member, IEEE,* Anke Schmeink, *Senior Member, IEEE,* James Gross, *Senior Member, IEEE*

***Abstract*—We consider a mobile edge computing (MEC) network supporting low-latency, critical offloading workloads. The task offloading from the user to the server is operated under a truncated Hybrid Automatic Repeat reQuest (HARQ) process, i.e., we consider finite retransmission attempts. Both the HARQ type-I and type-II schemes are studied. For each scheme, we first characterize the total error probability and the total energy cost, while the impact of finite blocklength (FBL) on the stochastic retransmission behavior is considered. Following the characterizations, we are interested in optimal frameworks for each considered HARQ type, where the number of potential retransmission attempts is optimized together with the duration of each transmission, while the CPU frequency at the edge node is adjusted via voltage scaling. The objective is to minimize the total energy cost with error probability threshold. We show that the resulting stochastic optimization problems can be solved by means of convex optimization. We furthermore demonstrate that sharp minima exist among the energy consumption, underlying the importance of near-optimal parameter choice in the studied scenarios. Our results underline the importance of trading off communication and computational characteristics in delay-critical MEC setups with FBL codes.**

***Index Terms*—edge computing, finite blocklength, offloading, retransmission**

I. Introduction

Over the last few years, mobile edge computing (MEC) has received increasing attention from the research community. It is characterized by the provisioning of compute resources in the proximity of applications generating workloads [2], thus having advantages over cloud computing where compute resources are pooled in a scalable fashion and subsequently provided. However, due to the high degree of scalability involved in cloud computing, it usually cannot be realized with spatial proximity. This translates into the main drivers of MEC in contrast to cloud computing, which relates to lower access latencies, as well as bandwidth savings towards the backbone, paired with a different security/privacy profile, which makes the MEC more likely have a significant impact on public networked infrastructures over the next decade.

An area that has received less attention so far is the provisioning of latency-critical services over MEC infrastructures. From an application point-of-view, two main application classes stand out that are discussed in relation to MEC. On the one hand, closed-loop applications are seen as important application class with latency constraints. These applications are characterized by a sensor-controller-actuator data exchange, and are given for instance in industrial automation scenarios. However, this data exchange is also the essential principle in human-in-the-loop systems (like augmented reality, cognitive wearable assistance) [3] . On the other hand, analytics applications are seen as an attractive application class for future MEC systems. In contrast to closed-loop applications, these systems are based only on a provider-processor data forwarding and this do not comprise a feedback loop. Analytics applications can have a wide span of scenarios, for instance predictive maintenance is an important scenario in various contexts. Regarding latency constraints, online state-estimation is an important application case in particular if processes of higher dynamics are under consideration. In the following, we are mostly concerned with such analytics applications on the edge, demanding reliable latency-constrained services.

As with other applications, also in the latency-constrained case the central challenge in MEC remains the interplay between the communication and the computation. Efficient service provisioning is usually addressed through the optimization of energy-efficiency, while on the other hand latency constraints demand the consideration of specific models especially on the communication side. In the past, various energy-efficient offloading schemes have been studied on these issues. For example, the authors in [4] present offloading schemes in order to minimize the overall energy consumption in a delay-sensitive MEC system. Their main contribution relates to discussing when to offload compute tasks (in contrast to executing them locally) in combination with corresponding CPU frequency scheduling schemes executed at the edge node. This idea is extended in [5] to also take task partitioning into account. The authors propose an energy-efficient partial offloading scheme by introducing a task offloading ratio, based

Part of this paper (some of the materials in Section III characterizing the design for HARQ type-I) has been presented at IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Cannes, France, 2019 [1]. This work has been supported by the German Research Council (DFG) within the basic research project DFG SCHM 2643/16 and the Federal Ministry of Education and Research of Germany in the programme of "Souverän. Digital. Vernetzt." Joint project 6G-RIC, project identification number:16KISK028.

Y. Zhu, Y. Hu are with School of Electronic Information, Wuhan University, 430072 Wuhan, China, and with ISEK Research Area, RWTH Aachen University, D-52074 Aachen, Germany, (e-mail: zhu@rwth-aachen.de, hu@whu.edu.cn). Y. Hu is the corresponding author.

A. Schmeink is with the ISEK Research Area/Lab, RWTH Aachen University, 52074 Aachen, Germany (e-mail: anke.schmeink@rwth-aachen.de).

James Gross is with Electrical Engineering and Computer Science School of KTH Royal Institute of Technology, MALVINAS VAEG 10 (e-mail: james.gross@ee.kth.se).

on which the energy costs of computing parts locally vs. at the edge can be quantified and subsequently determined to minimize energy consumption while guaranteeing the stringent latency requirement. A further contribution in this direction is provided in [6], where the authors consider a relay-assisted offloading scenario, and discuss an optimal energy-efficient framework comprising of offloading via a relay to an edge node. All of these studies address the tradeoff between the communication and the computation in MEC systems, but without considering low-latency, critical offloading. Subsequently, data communication is modeled under the idealized assumption of being arbitrary reliable at Shannon's capacity, which strictly speaking holds only in the limit for code words of infinite length. We refer to this assumption in the following as the infinite blocklength (IBL) regime. By assuming the validity of the IBL regime also over finite code words (and hence finite time spans like communication slots), this has several simplifying consequences for data offloading to an edge node. Most importantly, given a certain data size of an offloadable task, and given a certain channel state of a wireless link, one can determine an exact (i.e. deterministic) time span which it takes to offload the task. However, this does not correspond to reality, where wireless data communication is always subject to an error probability, and the coding rate actually determines together with the time slot allocation the likelihood of such an error.

A more suitable modeling approach of the communication effects for latency-constrained MEC applications is the finite blocklength (FBL) regime [7]. In the FBL regime, reliability of communication becomes a probabilistic function, as transmission errors possibly occur even when setting the coding rate below the Shannon capacity. For a MEC offloading scheme this leads to the possibility of data loss and therefore possibly subsequent retransmission attempts via Hybird Automatic Repeat reQuest (HARQ) [8]–[11]. This more precise modeling leads to more realistic models of the the MEC network [12]–[14], while it also introduces additional costs in terms of a random offloading latency and associated transmission energy consumption. In the specific context of offloading in latency-critical scenarios, if the communications requires a relatively long time (for instance due to multiple transmission attempts) the edge node needs to finish the computation process more quickly to meet the end-to-end latency constraint. This requires the compute part to run with a higher CPU frequency and subsequently costs a higher energy consumption for the computation. Hence, aiming at the energy efficiency, there exists a clear tradeoff between the attempted reliability during the communication phase and the consequences for the computation phase. Intuitively, this tradeoff becomes more relevant the stricter the latency requirement is, in which case also the proper modeling by the FBL regime is essential. In related work on MEC offloading, these aspects are broadly ignored to date. Nevertheless, recent works in [15]–[17] consider the FBL impact in the MEC offloading design aiming at minimizing either the overall error probability or the energy consumption, while the lengths of the communication and computation phases are assumed to be constant. To the best of our knowledge, a complementary study addressing the fundamental tradeoff between the communication and computation phases in terms of time allocation for offloading time-critical tasks in MEC networks is still an open issue. More interestingly, when HARQ is leveraged for the communication phase with FBL codes, it can be considered as the joint tradeoff between the length of each single (re)transmission in the communication phase, the number of HARQ retransmissions, and the length of computation phase. Therefore, how to address such a three-way tradeoff should be carefully investigated.

In this work, we study a MEC network that can rely on either HARQ type-I or type-II schemes for securing the wireless communication. The communication phase is operated with FBL codes, while a delay-critical task is to be offloaded where the deadline relates to the joint task of communication and computation. A reliability constraint is given, with which the system is supposed to successfully complete the offloading task. For this setting, we consider the optimal time allocation between the communication and the computation phase in order to minimize the energy consumption while guaranteeing the reliability and latency requirements and taking the cost of NACK into account. For this general set-up we provide the following contributions:

- We characterize the communication reliability in the FBL regime and the total energy consumption of considered network under HARQ type-I and HARQ type-II schemes. Moreover, the error process and energy consumption of potential NACKs transmitted back from the edge node to the terminal are also taken into consideration.
- We provide an optimal framework design minimizing the expected total energy cost for both HARQ schemes via jointly allocating the blocklength of a single (re)transmission and determining the maximal allowed retransmission times. For HARQ type-I, we decompose the original problem. We for the first time rigorously prove convexity of error probability and energy cost within the region of interest. Based on these characterizations, we reformulate the original problem, resulting in an integer convex problem.
- For HARQ type-II, we provide a novel approach in a problem-splitting manner based on convexity analysis: First, we decompose the original problem to non-convex subproblems. Subsequently, by cutting the feasible duration of each subproblem into a set of intervals we prove that the subproblems are convex within each interval, i.e., each subproblem can be solved by comparing the locally optimal values of the subproblems over all feasible intervals. Following the approach, the original problem can be optimally solved.
- We extend the design to scenarios with a random queuing delay at the MEC server. We characterize the error probability and energy cost in such scenarios and prove that our designs including the proven convexity features are still valid.
- Simulation results demonstrate that the proposed designs resolve the essential tradeoff between the reliability and the energy consumption in the FBL regime. Furthermore, we show the necessity for investigating the considered

MEC scenarios in the FBL regime by illustrating the performance difference compared with the design under the IBL assumption.

The rest of the paper is organized as follows. Section II presents the system model. Section III characterizes the error probability and energy cost for HARQ type-I, and provides the optimal retransmission scheme design accordingly. Similarly, Section IV studies the HARQ type-II scheme and proposes an algorithm achieving an optimal design. We discuss the impact of queuing delay at the edge node in Section V. We provide numerical results in Section VI and conclude the work in Section VII.

II. System Model

We consider a straightforward MEC network with a user terminal (UE) communicating to an edge node which is connected to a base station. At the UE, the state information is generated periodically. It is supposed to be provided to the edge node which subsequently executes a time-critical computation task. For instance, the UE is assumed to be a sensor which continuously reports time-sensitive information to the MEC server which executes state estimation logic. An visualization of the considered system (also including the timing structure) is provided in Fig. 1.

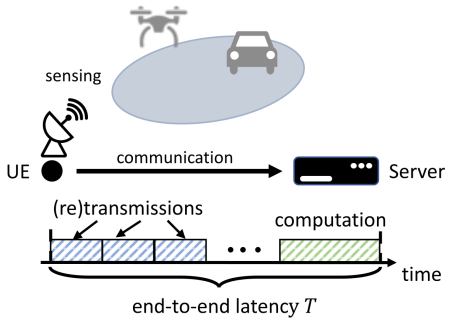

Fig. 1. Example of the considered system.

The entire task from the sensor generating until updating the state estimator is assumed to be of a random duration, which we denote by T. We consider delay-critical tasks, where the task duration T should be lower than a given threshold $T_{\max}$. This threshold could arise for instance from the high dynamicity of the observed device/object coupled with safety requirements. At the same time, the task also has a reliability requirement, i.e., the overall error probability of the whole offloading should be lower than a threshold $\varepsilon_{\text{tot,max}}$.

A. Communication Phase

In the communication phase, the UE transmits a local data packet (of a sensor reading) with a size of β bits via the wireless channel from the UE to the server. We consider a quasi-static channel fading model. Note that the task has a low-latency requirement such that $T_{\max}$ is quite short. Hence, a reasonable assumption is that the channel (including the pathloss and random fading) is constant during the whole end-to-end latency/duration of $T_{\max}$. Denote by z the gain of the fading, which is assumed to be perfectly known at the server side. Then, the signal-to-noise ratio (SNR) of the received packet at the server is $\gamma = \phi z P_{\text{ue}}/\sigma_S^2$, where P_{ue} is the transmit power of the UE, ϕ is the channel pathloss, and σ_S^2 represents the noise power. Moreover, the system utilizes HARQ retransmissions where the first attempt is the initial transmission and following transmissions are potential retransmissions. The blocklength (in symbols) of each attempt is $m \in \mathbb{Z}_+$, corresponding to a time length of $t = mT_S$, where T_S stands for the time length of one symbol. Due to the low-latency constraint, both t and m are required to be short. Hence, the transmission is possibly erroneous due to the impact of FBL codes [7]. We consider two typical HARQ schemes, namely HARQ type-I and type-II.

HARQ type-I scheme is usually applied in low-cost low-power IoT devices, due to its low complexity. In this work, we consider an energy efficient ARQ scheme, under which the receiver detects an error in transmission using the Cyclic Redundancy Check (CRC). If an error occurs, the erroneous packet will be simply discarded and a Negative Acknowledgement (NACK) with a fixed and small data size is sent to the UE within a fixed duration of t_{k}. This NACK might be subject to a decoding error at the UE in which case the UE interprets the payload as being correctly received by the base station, which adds to the overall error probability of the scheme. In addition, the transmit power of the NACK is denoted by P_{s}. After successfully decoding the NACK, the UE retransmits the entire data packet. This process repeats till the server successfully decodes the data (without sending ACK) or the maximal allowed transmission attempts N is reached. Denote by n the index of the possible transmission attempts up to N times, i.e., $n \in \mathcal{N} = \{1, ..., N\}$. So, $n = 1$ indicates the initial transmission and $N = 1$ represents the special case that no retransmission is allowed. Till n^{th} (re)transmission, nt transmission time has been spent, corresponding to nm symbols. Specially, it is also possible that the packet is not conveyed correctly after N attempts. In such case the communication phase ends immediately and no computation is carried out.

Different from HARQ type-I, the system with HARQ type-II does not discard the previously erroneous packet(s) but combines them with the newly retransmitted one. In other words, instead of simply repeating the packet as in HARQ type-I, in HARQ type-II additional coding information is provided, assuming the receiver to store the previously sent coded bits. This type of HARQ is also known as Incremental Redundancy (IR). Recall that the UE only generates the information once in the beginning of the frame. Therefore, the retransmission contains the same data size as previous one that representing the same information. In other words, we consider a full HARQ-IR scheme, i.e., the coding rate $r = \frac{\beta}{m}$ of each (re)transmission is identical.

B. Computation Phase

The state estimator is initialized during the computation phase if a new sensor reading is successfully received at the server. To guarantee the stringent latency constraint, the server resources are reserved for the task-related computation of the

UE, i.e., the server is able to execute the task immediately after successfully decoding the data packet without waiting in the execution queue. In particular, it is assumed that the CPU frequency f per task can be adjusted via dynamic frequency and voltage scaling (DVFS) [22] to adopt the requirement of the current task. However, it can only be scaled up to a maximal available CPU frequency $f_{\max}$. we assume that the computational workload of each estimation task of c computation steps (in CPU cycles) is fixed. Let $t_c^{(n)}$ denote the execution time in the case that the n^{th} transmission attempt succeeds. Hence, CPU frequency is chosen according to $f^{(n)} = c/t_c^{(n)}, \forall n \in \mathcal{N}$. Note that the frequency cannot exceed the maximal available frequency, i.e., $0 \quad f^{(n)} \quad f_{\max}, \forall n \in \mathcal{N}$, must hold. For any $n \in \mathcal{N}$, by summing up of both communication and computation phase, the actual end-to-end latency, i.e., under the condition that the n^{th} (re)transmission succeeds, is given by

$$T = t_c^{(n)} + nt + (n \quad 1)t_{\text{k}} \quad T_{\max}, \forall n \in \mathcal{N}. \tag{1}$$

C. Problem Statement on Energy Minimization

We aim at providing the optimal retransmission design which minimizes the total energy consumption E_{tot} through determining m and N while guaranteeing that *the total error probability* ε_{tot} *is less than a given threshold* $\varepsilon_{\text{tot,max}}$. This results in the following optimization problem:

$$\begin{aligned}
\underset{m \in \mathbb{Z}_+, N \in \mathbb{Z}_+}{\text{minimize}} \quad & E_{\text{tot}} && \text{(2a)}\\
\text{subject to} \quad & T \quad T_{\max}, && \text{(2b)}\\
& 0 \quad f^{(n)} \quad f_{\max}, \ \forall n \in \mathcal{N}, && \text{(2c)}\\
& \varepsilon_{\text{tot}} \quad \varepsilon_{\text{tot,max}}. && \text{(2d)}
\end{aligned}$$

Recall that the computation phase follows immediately after the successful transmission attempt, if there is any, thus the fundamental tradeoff lies in the choice of m and N. On the one hand, the choice of m directly influences the error probability of each (re)transmission, and therefore leads to a stochastic retransmission behavior, while the optimal length of m is influenced by the reliability constraint (2d) under a given N. In general, larger m and N make the offloading more reliable, but also introduces higher energy cost for transmissions. On the other hand, in order to save the computation energy, the execution time $t_c^{(n)}$ is preferred to be as long as possible, i.e., requiring small m and N. In other words, there exists a tradeoff in term of the time allocation between the communication phase with stochastic behaviors and the computation phase which is influenced by such behaviors. For example, as illustrated in Fig. 2, for the special case of $N = 1$, m must be sufficiently long to guarantee the reliability within the one-shot transmission, which results in a fixed execution time (if the transmission succeeds). Meanwhile, if setting N too large, m must be very short to guarantee that the system can keep carrying out transmission attempts. At the same time, the length of the computation phase (execution time) is the remaining time after the random length of the communication phase.

It should be pointed out that the two HARQ schemes behave differently, and thus the corresponding system designs are addressed respectively in the following two sections.

III. Performance Characterizations and Optimal Design under HARQ Type-I

In this section, we aim at minimizing the (expected) energy consumption for the considered MEC network under HARQ type-I. We first characterize the FBL error probability and the energy consumption model. The optimal design will be addressed subsequently.

A. Total Error Probability in FBL Regime

The end-to-end error occurs if the data has not been successfully transmitted before the deadline. In particular, there are possibly multiple transmissions for each data packet. Recall that $n \in \mathcal{N}$ represents the index of a (re)transmission, i.e., $n = 1$ indicates the initial transmission and $n > 1$ indicates a retransmission. For the n^{th} (re)transmission, with fixed task data size β and blocklength m, the corresponding coding rate is given by $r = \frac{\beta}{m}$. Due to the FBL impact, the transmission is possible to be erroneous even if the coding rate is lower than Shannon capacity. We leverage the FBL model in [7] to characterize this decoding error probability for each (re)transmission, i.e., the (block) error probability of the n^{th} (re)transmission is

$$\varepsilon \quad Q\left(\sqrt{\tfrac{m}{V(\gamma)}}(C(\gamma) \quad r)\log_e 2\right), \tag{3}$$

where $C = \log_2(1+\gamma)$ is the Shannon capacity. In addition, V is the channel dispersion [21]. Under a complex AWGN channel, $V = 1 \quad (1+\gamma)^{-2}$. This closed-form expression is accurate when the blocklength is larger than 100 and the error probability is higher than 10^{-7} [19]. In fact, as shown in [20], the performance gap between (3) and the actual error probability is insignificant at an error probability of 10^{-7} with practical channel coding schemes.

On one hand, the (re)transmission can be erroneous. Denote by X_n the event that the n^{th} (re)transmission fails. In addition, denote by ε_n the error probability of the n^{th} (re)transmission, i.e., $\varepsilon_n = \mathbf{P}(X_n = 1)$. Note that the channel gain z is constant within T. Due to experiencing independent random noises, events $X_n, \ \forall n = 0, 1, ..., N$ are independent. Moreover, according to (3), the probabilities of these events are the same, i.e., the error probabilities of the initial transmission and retransmission are identical, as $\varepsilon = \varepsilon_n = \varepsilon_k = \mathbf{P}(X_n = 1), \ \forall (n,k) \in \mathcal{N}$. On the other hand, in this work, we consider a rather realistic assumption that the NACK might also be incorrectly detected at the UE due to the impact of FBL codes. Let Y_n denote the event that sending n^{th} NACK fails. Then, it also holds for error probabilities of detecting NACKs v_n that $\nu = \nu_n = \nu_k = \mathbf{P}(Y_n = 1), \ \forall (n,k) \in \{1, ..., N \quad 1\}$. Especially, although there is no $n = 0^{\text{th}}$ transmission or N^{th} NACK, for the convenience of notation, we define $\varepsilon_0 = \mathbf{P}(X_0 = 1) = 1$ and $\nu_N = \mathbf{P}(Y_N = 1) = 0$. Furthermore, recall that we consider a reliable scenario. To facilitate the derivation without losing the feature of error probability in the FBL regime, we treat the error probability ε_n as one if it violates a threshold $\varepsilon_{\max} < 0.1$. In the next section we

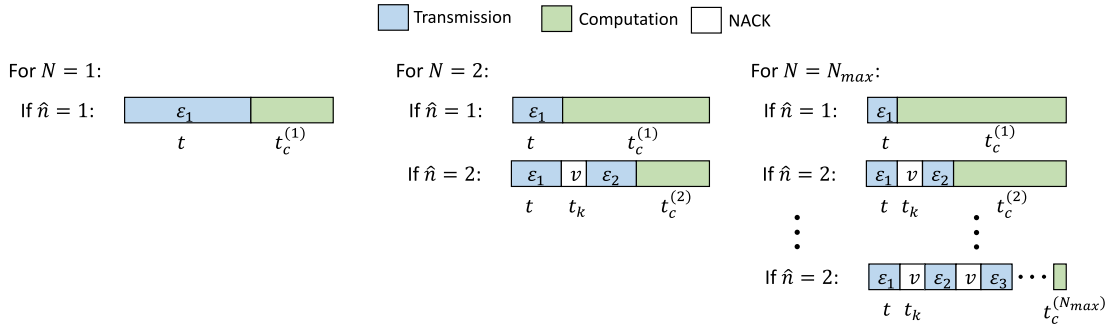

Fig. 2. Examples of possible retransmission schemes. In the figure, $\hat{n}$ denotes the number of (re)transmissions till the packet is successfully transmitted. Obviously, $N \quad \hat{n} \quad 1$ holds.

will show that if $\varepsilon_n > \varepsilon_{\max}$, i.e., the reliability constraint is violated, the considered problem becomes infeasible. In other words, the approximation treating ε_n as one in the infeasible domain facilities our analytical modeling but does not influence the achievable system performance.

Following the above notations, we derive the overall error probability by combining the errors of the data and the NACK transmission, while distinguishing different choices of N.

Case $N = 1$**:** No retransmissons are planned. Therefore, the error occurs only if the one-shot transmission fails, resulting in the total error probability $\varepsilon_{\mathsf{tot,I}}$ as:

$$\varepsilon_{\mathsf{tot,I}} = \varepsilon_1 = \varepsilon, \text{ for N = 1.} \tag{4}$$

Case $N > 1$**:** Firstly, the initial transmission ($n = 1$) is carried out and its error probability ($n = 1$) is ε. If the error occurs at $n = 1$ and the UE successfully decodes the NACK, the process of the next retransmission starts. In particular, the error at the server occurs at the n^{th} (re)transmission if one of the following two events happens: (A) the UE decodes the NACK of the n^{th} retransmission wrongly (the server receives nothing in the $(n + 1)^{\mathsf{th}}$ retransmission); (B) the NACK is decoded successfully but the $(n + 1)^{\mathsf{th}}$ retransmission fails.

Clearly, for the first joint event with 1^{st} transmission, we have $\mathbf{P}(X_1 \bigcap Y_1 = 1) = \varepsilon\nu$. Meanwhile, for the second joint event with 2^{nd} retransmission, we have $\mathbf{P}(Y_2 \bigcap X_1 \bigcap X_2 \bigcap \overline{Y}_1 = 1) = \varepsilon^2(1 \quad \nu)\nu$. Similarly, up to n^{th} retransmission, we have $\mathbf{P}(Y_n \bigcap_{i=1}^{n} {}^{1} X_i \ {}_{1}\overline{Y}_i = 1) = \sum_1^n \varepsilon^n(1 \quad \nu)^n \ {}^{1}\nu$. Moreover, $\mathbf{P}(X_{n+1} \bigcap_{i=1}^{n} X_i \overline{Y}_i = 1) = \varepsilon^n(1 \quad \nu)^n\varepsilon$ is the probability that all previous NACKs succeeded but all transmissions failed. As a result, the total error probability equals the probability of all the maximal allowed N transmission attempts being failed, which is given by

$$\varepsilon_{\mathsf{tot,I}} = \sum_{n=1}^{N} {}^{1} \varepsilon^n(1 \quad \nu)^n \ {}^{1}\nu + \varepsilon^N(1 \quad \nu)^N \ {}^{1}, \text{ for N} > 1. \tag{5}$$

B. Weighted Total Energy Cost

The total energy cost within end-to-end latency $T_{\max}$ consists of three parts: energy consumption of the UE $E_{\mathsf{t,I}}$, energy consumption at the server for transmitting the NACK $E_{\mathsf{k,I}}$ and the computation energy consumption at the server $E_{\mathsf{c,I}}$. Clearly, $E_{\mathsf{t,I}}$, $E_{\mathsf{k,I}}$ and $E_{\mathsf{c,I}}$ are influenced by the total transmission attempts n, which generally is a random variable in the range from 0 to N. In the following, we discuss the expected/average value of the three factors contributing to the energy consumption over the distribution of n. For the sake of clarity, in the rest of this paper, we use symbol E to indicate a deterministic energy consumption in the notation, while the expected energy consumption is represented by symbol E.

1) Energy Consumption of Data Transmission: Due to the randomness of the (re)transmission process, i.e., number of (re)transmissions till the packet being successfully sent is random, the exact energy consumption is unknown before operating the (re)transmissions. On the other hand, the expected energy consumption can be characterized according to the stochastic retransmission behavior. In particular, the expected energy consumption of the UE $E_{\mathsf{t,I}}$ depends on the error probability of NACK and the maximal number of transmission attempts N. Clearly, the expected energy consumption of either the initial transmission or a retransmission is given by $E_{\mathsf{t,0}} = tP_{\mathsf{ue}} + E_{\mathsf{s}}^0$, where E_{s}^0 is the constant energy consumption at the server for receiving a packet (with a given data size). Note that the server sends a NACK if the received packet is incorrectly decoded, while the corresponding retransmission occurs if the NACK is successfully decoded. Moreover, the initial transmission is always carried out regardless of N. Therefore, the expected energy consumption of the $(n + 1)^{\mathsf{th}}$ retransmission depends on the error probability of the n^{th} retransmission and the reliability of the n^{th} NACK. Hence, we have:

$$\begin{aligned} E_{\mathsf{t,I}} &= E_{\mathsf{t,0}} + \varepsilon(1 \quad \nu)E_{\mathsf{t,0}} + \ldots + \varepsilon^N \ {}^{1}(1 \quad \nu)^N \ {}^{2}E_{\mathsf{t,0}} \\ &= \sum\nolimits_{n=1}^{N} {}^{1} \varepsilon^n(1 \quad \nu)^n \ {}^{1}E_{\mathsf{t,0}}. \end{aligned} \tag{6}$$

2) Energy Consumption for Sending NACK: Clearly, the energy cost for sending a NACK is given by $E_{\mathsf{k,0}} = t_{\mathsf{k}}P_s + E_{\mathsf{ue}}^0$, where E_{ue}^0 is the constant energy consumption at the UE for receiving a NACK. If the initial transmission succeeds, no NACK needs to be sent, i.e., $E_{\mathsf{k,I}} = 0$. The probability that the first NACK occurs, equals to the error probability of the initial transmission. Hence, the expected energy consumption of the first NACK is $\varepsilon E_{\mathsf{k,0}}$. Moreover, the second NACK occurs if the first two (re)transmissions fail while the previous NACK is detected successfully, i.e., with probability $\varepsilon E_{\mathsf{k,0}}$. Similarly, the probability of n^{th} NACK is $\varepsilon^{n+1}(1 \quad \upsilon)^n$. Hence, the expected energy consumption $E_{\mathsf{k,I}}$ in N possible

(re)transmissions for sending all NACKs is

$$E_{\mathsf{k,l}} = \varepsilon E_{\mathsf{k},0} + \varepsilon^2(1-v)E_{\mathsf{k},0} + \cdots + \varepsilon^N(1-v)^{N-1}E_{\mathsf{k},0} = \sum_{n=1}^{N}\varepsilon^n(1-v)^{n-1}E_{\mathsf{k},0}. \tag{7}$$

3) Computation Energy Consumption: The energy consumption of computation is generally proportional to the workloads and the CPU frequency. In this paper, we adopt the nonlinear energy consumption model of computation introduced in [24], given by $E_{\mathsf{c,l}} = \kappa c f^2 = \kappa c^3 t_c^{-2}$, where κ is a constant related to the hardware architecture and t_c is the computation time.

Noting that the computation proceeds immediately, once the input data is received and the computation phase occupies the rest of the time slots, the computation time t_c depends on the number of proceeded transmissions n. In particular, the duration of communication phase (including data transmission and NACK transmission) is $nt+(n-1)t_{\mathsf{k}}$. Then, the remaining time for computation after n transmission attempts is given by $t_c^{(n)} = T_{\max} - nt - (n-1)t_{\mathsf{k}}$. Since $E_{\mathsf{c,l}}$ is a monotonically decreasing function with respect of t_c, the equality should always hold to minimize the energy consumption. Denote by $E_{\mathsf{c,l}}^{(n)}$ the computation energy consumption in the case that the server decodes the task data successfully in n^{th} transmission attempt ($n=1$ represents the initial transmission). Then, we have:

$$E_{\mathsf{c,l}}^{(n)} = \kappa c^3 \frac{1}{(T_{\max} - nt - (n-1)t_{\mathsf{k}})^2}. \tag{8}$$

Specially, we define $E_c^{(0)} = 0$, which indicates if no packet is transmitted, the computation will also not be carried out.

It is possible that the first transmission is successful and the computation starts immediately after the $n=1^{\mathsf{st}}$ transmission with the possibility of $1-\varepsilon$. Moreover, for $n>1$ if the previous $n-1$ transmission attempts fail (while the $n-1$ times NACKs are correct) and the n^{th} attempt successes, the corresponding probability is given by $\varepsilon^{n-1}(1-\nu)^{n-1}(1-\varepsilon)$, i.e., with this probability the computation starts after $n>1$ (re)transmissions. Combining the above two cases, the expected energy consumption for computation is

$$\begin{aligned} E_{\mathsf{c,l}} &= (1-\varepsilon)E_{\mathsf{c}}^{(1)} + \varepsilon(1-v)(1-\varepsilon)E_{\mathsf{c}}^{(2)} + \ldots + \varepsilon^N(1-v)^N(1-\varepsilon)E_{\mathsf{c}}^{(N)} \\ &\approx E_c^{(1)} + \sum_{n=1}^{N-1}\varepsilon^{n+1}(1-v)^n\left(E_c^{(n+1)} - E_c^{(n)}\right), \end{aligned} \tag{9}$$

where the approximation in the last step is tight due to the fact that we consider ultra-reliable scenarios, i.e., $\varepsilon \ll 1$ and $v \ll 1$ hold, and thus having $E_c^{(1)} + \varepsilon^{N-1}(1-v)^N E^{(N)} \approx \varepsilon^N(1-v)^N E^{(N)}$.

The expression can be also interpreted as follows: The surplus from $E_c^{(n+1)} - E_c^{(n)}$ is the extra computational energy the system has to consume if n^{th} (re)transmission fails and the computation has to be delayed after next retransmission including sending both NACK and data, which leads to shorter available computation time. However, the next retransmission does not guarantee the success of decoding. Therefore, the total computational energy consumption is the sum of all such surpluses up to $N-1$ (re)transmissions (if N^{th} retransmission fails, no computation will be carried out, i.e., no energy consumption.).

So far, we have derived the expected energy consumption for task transmission, NACK and computation. To further represent the different capacity of energy from different source, we consider a weighted total energy cost (within T) E_{tot} instead of the absolute value of energy consumption, which can be written as

$$E_{\mathsf{tot,l}} = \alpha_t E_{\mathsf{t,l}} + \alpha_k E_{\mathsf{k,l}} + \alpha_c E_{\mathsf{c,l}}, \tag{10}$$

where α_{t}, α_{k} and α_{c} are the non-negative weights for the energy consumption of (re)transmissions, sending NACKs and computation, respectively. Then, $E_{\mathsf{tot,l}}$ takes the energy intensity[1] from different sources into account.

C. Optimal Retransmission Scheme Design under HARQ Type-I

Following the above characterizations, we provide in this subsection a retransmission scheme design for the considered network under HARQ type-I by optimally determining the blocklength of a single (re)transmission m and the maximal allowed transmission attempts N.

1) Problem Formulation: Our objective is to minimize expected total energy cost $E_{\mathsf{tot,l}}$ while guaranteeing the given reliability requirements. In particular, the server should have sufficient time t_c to compute the task within the end-to-end latency constraint $T_{\max}$ even in the worst-case scenario, where the task data is received just after N transmission attempts. Moreover, recall that the overall error probability needs to be lower than $\varepsilon_{\mathsf{tot,max}}$ [2]. Furthermore, it is trivial to show that the equality should be hold for the inequality constraint (2b), in order to obtain the optimal solution. Combing with the CPU frequency constraint (11c), the original problem in (2) can be reformulated as

$$\underset{m\in\mathbb{Z}_+, N\in\mathbb{Z}_+}{\text{minimize}} \quad E_{\mathsf{tot,l}} \tag{11a}$$

$$\text{subject to} \quad t_c^{(n)} + nt + (n-1)t_{\mathsf{k}} = T_{\max}, \forall n \in \mathcal{N}, \tag{11b}$$

$$\frac{c}{f_{\max}} + Nt + (N-1)t_{\mathsf{k}} \le T_{\max}, \tag{11c}$$

$$\varepsilon_{\mathsf{tot,l}} \le \varepsilon_{\mathsf{tot,max}} \tag{11d}$$

2) Optimal Solution: We solve Problem (11) in the following three steps: We firstly determine the maximal possible transmission attempts $N_{\max}$, which is the upper bound of N that are feasible for the original problem. Next, we decompose the original problem given in (11) into $N_{\max}$ subproblems, i.e., corresponding to each feasible value of $N \in \{1,2,\ldots,N_{\max}\}$. Moreover, we characterize the subproblems and based on that,

[1]Note that the UE is possible to be battery-enabled and the MEC server is usually connected to the grid. Hence, the UE is possible to be more sensitive to the energy cost. This is the motivation of introducing α_{t}, α_{k} and α_{c}.

[2]Note that to support a reliable transmission, the link SNR cannot be extremely low. Hence, the extreme low SNR cases with $\gamma < \gamma_{th} < 0$dB are out of scope in this design, i.e., operating the system with such low SNR means just wasting the energy.

we reformulate the original problem to be a solvable integer convex problem. The details of the three steps are given as follows:

Step 1: Decomposition of Problem (11)**:** Since N is a positive integer and upper-bounded by $N_{\max}$, there exists $N_{\max}$ possible outcomes with respect to the retransmission events of the frame. For a given $N \in \{0, 1, .., N_{\max}\}$, we have the following subproblem:

$$\begin{aligned} \underset{m}{\text{minimize}} \quad & E_{\text{tot,I}} && \text{(12a)} \\ \text{subject to} \quad & \text{(11b), (11c) and (11d).} \end{aligned}$$

Without constraints, $N_{\max}$ is an unbounded integer resulting in infinite subproblems. However, the maximal number of retransmissions is restricted due to the limited server computation power $f_{\max}$. By combining the constraints (11b) and (11c), we obtain an upper bound for $N_{\max}$:

$$N_{\max} \leq \left\lfloor \frac{T_{\max} - \frac{c}{f_{\max}} - t}{t + t_{\text{k}}} \right\rfloor, \tag{13}$$

where $\lfloor \cdot \rfloor$ is the floor function.

Step 2: Determining Optimal Solution of Subproblem (12)**:** For a given N, we first relax the integer constraint to $m > 0$. Subsequently, we have the following key lemma to handle Subproblem (12).

Lemma 1. *The total error probability $\varepsilon_{\text{tot,I}}$ is convex in the relaxed blocklength $m > 0$.*

Proof. Since ν and t_{k} are fixed, we can obtain $t_c^{(n)}$ as a function of t, according to (11b):

$$t_c^{(n)} = \max\{T_{\max} - nt - (n-1)t_{\text{k}}, 0\}. \tag{14}$$

To show the convexity of $\varepsilon_{\text{tot,I}}$ in m, we show necessary conditions for the second derivative. For $N = 1$, we have $\frac{\partial^2 \varepsilon_{\text{tot,I}}}{\partial m^2} = \frac{\partial^2 \varepsilon}{\partial m^2}$. In addition, for $N \geq 2$, we have

$$\begin{aligned} \frac{\partial^2 \varepsilon_{\text{tot,I}}}{\partial m^2} = \frac{\partial^2 \varepsilon}{\partial m^2}\nu + \sum_{n=2}^{N} n \Bigg(& (n-1)\varepsilon^{n-2}\left(\frac{\partial \varepsilon}{\partial m}\right)^2 \\ & + \varepsilon^{n-1}\frac{\partial^2 \varepsilon}{\partial m^2}\Bigg) + N(N+1)\varepsilon^{N-1}(1-\nu)^N \frac{\partial^2 \varepsilon}{\partial m^2}. \end{aligned} \tag{15}$$

As shown in our previous work [23], $\frac{\partial^2 \varepsilon}{\partial m^2} \geq 0$ holds. Hence, the overall error probability $\varepsilon_{\text{tot,I}}$ is convex in m for both the cases of $N = 1$ and $N \geq 2$. □

Lemma 1 implies that the exponent of error probability $\varepsilon^n, \forall n \in \mathbb{N}$, is convex in m. This characterization can be utilized in any FBL scenarios that involve ε^n, as well as the polynomial $\sum \varepsilon^n$, e.g., for average age-of-information (AoI) minimization [25]. Moreover, it also indicates that constraint (11d) actually results in a convex feasible set for m in Subproblem (12). Note that the other constraints are linear. Subproblem (12) is convex if the objective $E_{\text{tot,I}}$ is also convex in m, which is addressed in the following lemma:

Lemma 2. *The total energy cost $E_{\text{tot,I}}$ is convex in the relaxed blocklength $m > 0$.*

Proof. Recall that $E_{\text{tot,I}}$ consists of three parts, i.e., $E_{\text{tot,I}} = \alpha_{\text{t}} E_{\text{t,I}} + \alpha_{\text{k}} E_{\text{k,I}} + \alpha_{\text{c}} E_{\text{c,I}}$. In the following, we prove the convexity of each part respectively.

We start with $E_{\text{t,I}}$ and have

$$\begin{aligned} \frac{\partial^2 E_{\text{t,I}}}{\partial m^2} = P_{\text{ue}}\Bigg(& (1-v)A \\ & + \sum_{n=2}^{N} (1-v)^n n(n-1)\varepsilon^{n-2}\left(\frac{\partial \varepsilon}{\partial m}\right)^2 \\ & + n\varepsilon^{n-1}(1-v)^n A\Bigg), \end{aligned} \tag{16}$$

where $A = \frac{\partial^2 \varepsilon}{\partial m^2} t + 2\frac{\partial \varepsilon}{\partial m}$. Clearly, $\frac{\partial^2 E_{\text{t,I}}}{\partial m^2} \geq 0$ if $A \geq 0$. Note that $V \geq 1$, $m \geq 1$ and $t = mT_S$. Hence, we have

$$\begin{aligned} A = & \left(\frac{\partial^2 \varepsilon}{\partial m^2} t + 2\frac{\partial \varepsilon}{\partial m}\right) \\ & \sqrt{\frac{m}{V}}\left(\frac{(C-\beta/m)(C+\beta/m)^2}{4Vm^2} - \frac{3C+\frac{\beta}{m}}{4m^2}\right) \\ & \frac{B}{m^3}, \end{aligned} \tag{17}$$

where $B = C^3 m^3 + (C^2\beta - 3C)m^2 - (C\beta^2 - 3\beta)m - \beta^3$ is a third degree polynomial with the greatest root $m = \frac{\beta}{C}$. Since $\varepsilon < \varepsilon_{\max} \leq 1$, it holds $C > \frac{\beta}{m}$ for the transmission. In other words, the polynomial B is always positive (negative) when the first derivative of B is positive (negative) in the feasible regime. We thus have

$$\begin{aligned} \frac{\partial B}{\partial m} = & 2C^2\beta m - C\beta^2 + 3\beta + 3Cm(C^2 m - 3) \\ & 2C\beta^2 - C\beta^2 + 3\beta + 3Cm(C^2 m - 3) \\ & 0. \end{aligned} \tag{18}$$

Hence, $B \geq 0$ holds. According to (17), $A \geq 0$ also holds. As a result, $\frac{\partial^2 E_{\text{t,I}}}{\partial m^2} \geq 0$, i.e., $E_{\text{t,I}}$ is convex in m.

Secondly, for $E_{\text{k,I}}$ we have

$$\frac{\partial^2 E_{\text{k,I}}}{\partial m^2} = \frac{\partial^2 \varepsilon}{\partial m^2} E_{\text{k,I}} + \sum_{n=1}^{N} n(n-1)\varepsilon^{n-2}\left(\frac{\partial \varepsilon}{\partial m^2}\right)^2 + n\varepsilon^{n-1}\frac{\partial^2 \varepsilon}{\partial m^2}$$

As shown in [23], $\frac{\partial^2 \varepsilon}{\partial m^2} \geq 0$ holds. It is clear that $\frac{\partial^2 E_{\text{k,I}}}{\partial m^2} \geq 0$, which proves the convex of $E_{\text{k,I}}$ to m.

$$\begin{aligned} \frac{\partial^2 E_{\text{c,I}}}{\partial m^2} = & \frac{\partial^2 E_{\text{c,I}}^{(0)}}{\partial m^2} \\ & + \sum_{n=1}^{N}\Bigg[\left(n(n-1)\varepsilon^{n-2}\left(\tfrac{\partial \varepsilon}{\partial m}\right)^2 + n\varepsilon^{n-1}\tfrac{\partial^2 \varepsilon}{\partial m^2}\right) D_1^{(n)} \\ & 2n\varepsilon^{n-1}\frac{\partial \varepsilon}{\partial m} D_2^{(n)} + \varepsilon^n D_3^{(n)}\Bigg], \end{aligned} \tag{19}$$

Finally, we study the convexity of $E_{\text{c,I}}$ regarding m. The second order derivative of $E_{\text{c,I}}$ to m is given in (19) on the top of next page, where $D_1^{(n)} = E_{\text{c,I}}^{(n)} - E_{\text{c,I}}^{(n-1)}$, $D_2^{(n)} = \frac{\partial E_{\text{c,I}}^{(n)}}{\partial m} - \frac{\partial E_{\text{c,I}}^{(n-1)}}{\partial m}$ and $D_3^{(n)} = \frac{\partial^2 E_{\text{c,I}}^{(n)}}{\partial m^2} - \frac{\partial^2 E_{\text{c,I}}^{(n-1)}}{\partial m^2}$. As proven previously, ε is a convex and monotonically decreasing function with respect to n, i.e., $\frac{\partial \varepsilon}{\partial m} < 0$ and $\frac{\partial^2 \varepsilon}{\partial m^2} \geq 0$. In particular, we have $\frac{\partial^2 E_{\text{c,I}}^{(0)}}{\partial m^2} = 6(T - mT_S)^{-4} \geq 0$. Therefore,

all the terms besides $D_i^{(n)}, \forall i \in \{1,2,3\}$ in (19) are non-negative, i.e., to determine the convexity of $E_{\mathsf{c,I}}$ is to determine the sign of $D_i^{(n)}$.

For $D_1^{(n)}$, we have

$$D_1^{(n)} = \frac{1}{(T - nt - (n-1)t_{\mathsf{k}})^2} - \frac{1}{(T - (n-1)t - (n-2)t_{\mathsf{k}})^2} - \frac{1}{(T - nt - (n-1)t_{\mathsf{k}})^2} + \frac{1}{(T - nt - (n-1)t_{\mathsf{k}})^2} = 0. \tag{20}$$

Similarly, it is easy to show that $D_2^{(n)} \geq 0$ and $D_3^{(n)} \geq 0$ also hold by exploiting $n+1 \geq n$ to carry out the inequality chains. As a result, we have $\frac{\partial^2 E_{\mathsf{c,I}}}{\partial m^2} \geq 0$.

So far, we have proven that $E_{\mathsf{t,I}}$, $E_{\mathsf{k,I}}$ and $E_{\mathsf{c,I}}$ are convex in m. As α_{t}, α_{k}, α_{c} are non-negative, $E_{\mathsf{tot,I}} = \alpha_{\mathsf{t}} E_{\mathsf{t,I}} + \alpha_{\mathsf{k}} E_{\mathsf{k,I}} + \alpha_{\mathsf{c}} E_{\mathsf{c,I}}$ is also convex in m. □

Lemma 2 implies that the objective function of Problem (13) is also convex. Moreover, since the proof of Lemma 2 is conducted by proving the convexity of $E_{\mathsf{t,I}}$, $E_{\mathsf{k,I}}$ and $E_{\mathsf{c,I}}$, respectively, the value of α does not influence the validity of Lemma 2. In other words, the energy cost of the pure communication or computation phase are also convex in m.

Step 3: Reformulation of the Original Problem (11)**:** According to Lemma 2 and the upper bounds $N_{\max}$, we can reformulate the original problem as

$$\underset{m \in [0,N]}{\text{minimize}} \quad E_{\mathsf{tot,I}} \tag{21a}$$

$$\text{subject to} \quad N \leq \frac{T - \frac{c}{f_{\max}} t}{t + t_{\mathsf{k}}} c, \tag{21b}$$

(11b), (11c) and (11d).

With Lemma 1 and 2, the objective function and all constraints are either affine or convex. Hence, Problem (21) is a mixed integer convex problem with a relaxed of m, which can be handled by solving $N_{\max}$ convex subproblems and comparing the resulted $N_{\max}$ optimal values. This approach has a computational complexity of $\mathcal{O}(N_{\max})$ [26]. Denote by m^* the relaxed optimal solution of m and by N^* the global optimal solution of N. The integer solution of m that achieving the global optimum is obtained by comparing all possible integer neighbours, i.e.,

$$m^* = \underset{m \in \{\lfloor m^* \rfloor, \lceil m^* \rceil\}}{\arg\max} E_{\mathsf{tot,I}}(N^*) \tag{22}$$

Remark: As a comparison, if we consider the IBL regime (where transmissions are arbitrarily reliable at Shannon's capacity), the decoding error probability, denoted by $\varepsilon_{\mathsf{IBL}}$, does not depend on the blocklength as long as $r(m) < C$. Then, the solution to Problem (11) becomes quite straightforward. In particular, we have $m_{\mathsf{IBL}} = \lceil \frac{k}{C} \rceil$ and $N_{\mathsf{IBL}} = 1$, i.e., without retransmission due to the AWGN channel. However, it should be pointed out that under the FBL regime such an IBL-optimal allocation becomes inaccurate when the blocklength of the transmissions is short, which we show later via numerical simulation.

IV. Performance Characterization and Optimal Retransmission Scheme Design under HARQ Type-II

In this section, we study the network performance under HARQ type-II. We first investigate the error probability by exploiting the performance bound of HARQ type-II. Based on the error probability, we characterize the energy consumption. Similar to the design of HARQ type-I, we provide the optimal design and propose an algorithm accordingly.

A. Total Error Probability in FBL Regime

Recall that failed decoded packets under HARQ type-II are kept at the receiver, which will be combined with the packet from the next retransmission to enhance the reliability until reaching the maximal allowed transmission attempts or being successfully decoded. Therefore, the effective blocklength after the n^{th} (re)transmission $m_{(n)}$ is the sum of all previous transmissions, i.e., $m_{(n)} = nm$. We denote by $\varepsilon_{(n)} = \mathbf{P}(\bigcup^{n} X_n Y_n = 1)$ the joint error probability of all the n (re)transmissions. Since $m_{(n)} \leq m_{(k)}$, where $n \leq k \in \mathbb{N}$, it always holds that $X_n Y_n \geq X_k Y_k$ if the blocklength is infinite. In the FBL regime, the exact expression of the total error probability seems however intractable. To address this issue, we adopt an approximation[3] applied in [9], where $\varepsilon_{(n)}$ is approximated by the probability that error occurs up n^{th} (re)transmission ε_n with blocklength $m_{(n)}$. The error probability of n^{th} (re)transmission ε_n up to previous $n-1$ failed transmissions is given by [9]:

$$\varepsilon_n = Q\left(\sqrt{\frac{m_{(n)}}{V(\gamma)}}(C(\gamma) - \beta/m_{(n)})\right). \tag{23}$$

As a result, the total error probability with HARQ type-II is approximated as

$$\varepsilon_{\mathsf{tot,II}} = \varepsilon_{(N)} \approx \begin{cases} \varepsilon_N & \text{if } N = 1, \\ (1-v)\varepsilon_N + v^N & \text{otherwise.} \end{cases} \tag{24}$$

The approximation is accurate for any $v \leq \varepsilon_N$ and the performance mismatch between the approximation and the exact expression is up-bounded by $\sum_{n=1}^{N-1} v^n$ while it coincides with (23) if we ignore the impact of NACK.

B. Weighted Total Energy Cost

Similar to HARQ type-I, the weighted total energy cost with HARQ type-II, denoted by $E_{\mathsf{tot,II}}$, is decomposed into three parts: the energy consumption of data transmissions $E_{\mathsf{t,II}}$, the energy consumption for transmitting NACKs $E_{\mathsf{k,II}}$ and the energy consumption of the computation $E_{\mathsf{c,II}}$. For $N > 1$, the energy consumption of a single (re)transmission remains the same as $E_{\mathsf{t},0}$. Therefore, the expected energy consumption of data (re)transmission with HARQ type-II is:

$$E_{\mathsf{t,II}} = \sum_{n=1}^{N} \varepsilon_{(n-1)} E_{\mathsf{t},0}. \tag{25}$$

[3]Although block-fading channels are considered in [9], it has been shown via numerical evaluation in [12] that the approximation provided in [9] regarding HARQ Type-II is tight also for AWGN channels (with constant channel gains).

Especially, we define $\varepsilon_{(0)} = 1$ to facilitate the theoretical analysis in modelling the energy cost. In addition, the expected energy cost for transmitting NACKs is given by:

$$E_{\mathsf{k,II}} = \sum_{n=1}^{N-1} \varepsilon_{(n-1)} E_{\mathsf{k},0}. \tag{26}$$

Regarding the computation, it will be carried out once the data transmission is successful. Since the computational energy consumption depends on the correctness of the (re)transmission, i.e., it can be calculated by averaging over all possible (re)transmission correctness combinations, which can be written as:

$$\begin{aligned} E_{\mathsf{c,II}} = &\sum_{n=1}^{N} \left(\varepsilon_{(n-1)} - \varepsilon_{(n)}\right) E_c^{(n)} \\ &\sum_{n=1}^{N} \varepsilon_{(n-1)} \left(E_c^{(n)} - E_c^{(n-1)}\right). \end{aligned} \tag{27}$$

C. Optimal Retransmission Scheme Design under HARQ Type-II

In this section, we provide the optimal retransmission scheme for the considered network under HARQ type-II. In particular, we minimize the total energy cost by determining the optimal blocklength of each (re)transmission m and the maximal allowed retransmission attempts N.

1) Problem Formulation: The objective of the design here is also minimizing the weighted total energy cost $E_{\mathsf{tot,II}}$, as introduced in (10). Under HARQ type-II, it is expressed as:

$$E_{\mathsf{tot,II}} = \alpha_{\mathsf{t}} E_{\mathsf{t,II}} + \alpha_{\mathsf{k}} E_{\mathsf{k,II}} + \alpha_{\mathsf{c}} E_{\mathsf{c,II}}. \tag{28}$$

Similar to Problem (21), the problem for the design of HARQ type-II scheme is formulated as

$$\begin{aligned} \underset{m,N}{\text{minimize}} \quad & E_{\mathsf{tot,II}} && \text{(29a)} \\ \text{subject to} \quad & \text{(11b), (11c), (11d) and (21b).} \end{aligned}$$

2) Decomposing Problem (29)*:* Following the methodology of handling Problem (11), we solve Problem (29) by decomposing it into $N_{\max}$ subproblems with given N. In particular, for a given N, the subproblem is expressed as:

$$\begin{aligned} \underset{m}{\text{minimize}} \quad & E_{\mathsf{tot,II}} && \text{(30a)} \\ \text{subject to} \quad & \text{(11b), (11c), (11d) and (21b).} \end{aligned}$$

We address the subproblem in the following way. First, we relax the interger constraint as $m \geq 0$. Subsequently, we have the following key lemma, which provides the convexity of total error probability under HARQ type-II $\varepsilon_{\mathsf{tot,II}}$.

Lemma 3. *The total error probability with HARQ type-II, denoted by $\varepsilon_{\mathsf{tot,II}}$, is convex in the relaxed blocklength $m \geq 0$.*

Proof. We show the convexity of $\varepsilon_{\mathsf{tot,II}}$ in m by showing that second derivative of that is positive. First, for $N > 1$, we have

$$\frac{\partial^2 \varepsilon_{\mathsf{tot,II}}}{\partial m^2} = (1-v) N^2 \frac{\partial^2 \varepsilon_N}{\partial m^2} \geq 0. \tag{31}$$

Since the second derivative of both cases with and without HARQ type-II in Lemma 1 for $N = 1$ coincide, we can also show that it is non-negative. Hence the overall error probability with HARQ type-II $\varepsilon_{\mathsf{tot,II}}$ is convex in m. □

Next, we characterize the objective of Problem (30). Different from the HARQ type-I case where $E_{\mathsf{tot,I}}$ is convex in m, the expected energy cost $E_{\mathsf{tot,II}}$ of HARQ type-II (the same as its relaxation to m) is a non-convex non-concave function with respect to m. To tackle this issue, we first relax the constraint as $m \geq 0$. Then, we divide the feasible blocklength of m into N intervals. The interval of **dom** $\mathbf{f_n}$, where $n \in \mathcal{N}$, is defined by $[m_{\mathsf{cut},N-n}, m_{\mathsf{cut},N-n+1})$, which can be obtained as follows:

$$m_{\mathsf{cut},n} = \begin{cases} 0 & \text{if } n = 0, \\ \frac{T}{T_S} & \text{if } n = N, \\ \varepsilon_n^{-1}(\varepsilon_{\max}) & \text{otherwise.} \end{cases} \tag{32}$$

where $\varepsilon_n^{-1}(\cdot)$ is the inverse function of ε_n and $\varepsilon_{\max}$ is the error probability threshold of single (re)transmission. $m_{\mathsf{cut},n}$ represents the blocklength of the transmission after the n^{th} (re)transmission, which exactly satisfies the equality of (11d). Since $\varepsilon_n(m)$ is a monotonic function with respect to m, $\varepsilon_n^{-1}(\varepsilon_{\max})$ is also unique for each n. Note that we treat/approximate ε_n being higher than $\varepsilon_{\max}$ as one. Therefore, in **dom** $\mathbf{f_n}$, it holds that $\varepsilon_k = 1$, if $k < n$ and $\varepsilon_n = 1$. As a result, we can further decompose the Problem (30) with the given N into another N subproblems. For each **dom** $\mathbf{f_n}$, such subproblem is given by:

$$\begin{aligned} \underset{m \geq 0}{\text{minimize}} \quad & E_{\mathsf{tot,II}} && \text{(33a)} \\ \text{subject to} \quad & m_{\mathsf{cut},N-n} \leq m < m_{\mathsf{cut},N-n+1}, && \text{(33b)} \\ & \text{(21b), (11b), (11c) and (11d).} \end{aligned}$$

We hence have the following lemma:

Lemma 4. *The expected energy consumption under HARQ type-II $E_{\mathsf{tot,II}}$ is convex in the relaxation of blocklength $m \geq 0$ within each* ***dom*** $\mathbf{f_n}$, $n \in \mathcal{N}$.

Proof. Recall that $E_{\mathsf{tot,II}}$ consists of three parts, namely $E_{\mathsf{t,II}}$, $E_{\mathsf{k,II}}$, $E_{\mathsf{c,II}}$. As the energy consumption from different sources are weighted with different factors, the convexity of $E_{\mathsf{tot,II}}$ can only be ensured if all the three parts are convex. To this end, in following, we consider an arbitrary $n \in \mathcal{N}$ and we prove the convexity of each part within **dom** $\mathbf{f_n}$, respectively. Note that we consider $\varepsilon_{n^0} = 1$ if a given threshold $\varepsilon_{\max}$ is not fulfilled. Therefore, it holds that $\varepsilon_{n^0} \geq \varepsilon_k$, if $n^0 < k$. For the energy consumption of sending NACKs $E_{\mathsf{k,II}}$, we have

$$\frac{\partial^2 E_{\mathsf{k,II}}}{\partial m^2} = \sum_{n=n^0+1}^{N-1} n^2 (1-v)^{n-1} \frac{\partial^2 \varepsilon_{n-1}}{\partial m_{n-1}^2} E_{\mathsf{k},0} \geq 0. \tag{34}$$

Next, for the energy consumption of data-(re)transmission

$E_{\mathrm{t,II}}$ we have:

$$\frac{\partial^2 E_{t,\mathrm{II}}}{\partial m^2} = \sum_{n=1}^{N} (1 \quad v)^n P_{\mathrm{ue}} \left(\frac{\partial^2 \varepsilon_{n\ 1}}{\partial m^2} m + 2 \frac{\partial \varepsilon_{n\ 1}}{\partial m} \right) = \sum_{n=n^0+1}^{N\ 1} K_n \left(\frac{\partial W_{n\ 1}}{\partial m} G_{n\ 1} \quad \frac{\partial^2 W_{n\ 1}}{\partial m^2} \right),$$

where $K_n = (1 \quad v)^n P_{\mathrm{ue}} t \exp\left(\frac{W_{n\ 1}}{2} \right)$ and $G_{n\ 1} = \frac{C^2 m_{(n\ 1)}^2 \quad \beta^2 \quad 2Vm_{(n\ 1)}^2}{2Vm}$. By investigate the root of the numerator, we have $m_{n\ 1}^{+} = \frac{C^2\beta^2+V^2+V}{C^2} \quad \frac{\beta}{C} + \frac{2V}{C} \quad \frac{\beta}{C} \quad m$, where $m_{n\ 1}^{+}$ is the only positive root of the numerator. The approximation holds due to that the blocklength is always greater than 1 in the FBL regime and $\frac{2V}{C}$ is smaller than 1 for the reliable transmission. It implies that for any m in **dom** $\mathbf{f_n}$, we have $G_{n\ 1} \quad 0$. Furthermore, we also have $K_n \quad 0$ and $\frac{\partial^2 W_{n\ 1}}{\partial m^2} = \sqrt{\frac{m_{(n)}}{V}} \left(C + \frac{3\beta}{m_{(n)}} \right) \frac{1}{4m_{(n)}^3} \quad 0$. Hence, $E_{\mathrm{t,II}}$ is convex in m.

Finally, for the energy consumption of computation $E_{\mathrm{c,II}}$, we have:

$$\frac{\partial^2 E_{\mathrm{c,II}}}{\partial m^2} = \frac{\partial^2 E_{\mathrm{c,I}}^{(0)}}{\partial m^2} + \sum_{n=1}^{n^0} D_3^{(n)} + \sum_{n=n^0+1}^{N\ 1} \left[\left(\frac{\partial^2 \varepsilon_n}{\partial m^2} \right) D_1^{(n)} \quad 2 \frac{\partial \varepsilon_n}{\partial t} D_2^{(n)} + \varepsilon_n D_3^{(n)} \right], \tag{35}$$

As we showed in Lemma 2, $D_1^{(n)}$, $D_2^{(n)}$ and $D_3^{(n)}$ are all non-negative for any $(n+1)t \quad T \quad nt_{\mathsf{k}}$. Hence, $E_{\mathrm{c,II}}$ is convex in m. We have showed all three parts of $E_{\mathrm{tot,II}}$ are convex in m within any **dom** $\mathbf{f_n}$. As a result, $E_{\mathrm{tot,II}}$ is also convex in m within any **dom** $\mathbf{f_n}$. □

In fact, those analytical findings can be observed implicitly in the numerical simulations in [12],[14]. More interestingly, the convexity does not hold outside of each **dom** $\mathbf{f}_n$. In other words, this characterization cannot be directly utilized with original problem. Therefore, we propose a novel approach to solve this problem as follows.

3) Determining the Optimal Solution of (30)*:* Note that Lemma 4 only holds within the targeted interval instead of the whole feasible set of m. We need to calculate the optimal solution of Subproblem (33) within each of the N feasible intervals. Then, by comparing the N optimal values of Subproblem (33), we obtain the global optimal solution from Subproblem (30).

Combining with the decomposition process to the original problem, we introduce an algorithm to obtain the optimal solution to the original Problem (29), where the flow of the algorithm is described as follows. First of all, we choose an initial $N \quad N_{\max}$ and determine $m_{\mathrm{cut},n}$, where $n \quad N$, which is calculated by the inverse function of ε_n with respect to the error probability threshold $\varepsilon_{\max}$. Then, we solve Subproblem (33) for a given N according to Lemma 4 in each **dom** $\mathbf{f_n}$. We denote by $E_{n,N}$ the optimal result and $m_{n,N}$ the corresponding solution. Subsequently, we compare all N optimal results and select the minimal one as the optimal solution of Subproblem (33) for the given N, i.e., $E_{\mathrm{II},N} = \min_n E_{n,II}$ and $m_{\mathrm{II},N} = m_{n\ ,N}$, where $n = \arg\min_n E_{n,N}$. Note that it is possible that there is no feasible solution for the optimization problem. To address this, we set $E_{\mathrm{II},n}$ to an extreme high value E_{1} and $m_{n\ ,N} = 0$ if the problem is infeasible. Next, we compare total $N_{\max}$ results from the subproblems and select again the minimal result as the optimal of original problem, i.e., $E_{\mathrm{tot,II}} = \min_N E_{\mathrm{II},N}$ and $m = m_N$, where $N = \arg\min_N E_{\mathrm{II},N}$. Finally, we obtain the optimal solution m via comparing the integer neighbours of m, i.e.,

$$m = \arg\max_{m2\,fbm\ c,dm\ eg} E_{\mathrm{tot,II}}(N), \tag{36}$$

where m is the optimal blocklength of a single (re)transmission and N is the optimal number of allowed transmission attempts. We provide the flow of the algorithm as Algorithm 1. The algorithm solves essentially maximal $N_{\max}^2$ convex problems and sorts the results, which leads to the computational complexity $O\left(N_{\max}^2 \log(N_{\max}^2)\right)$.

Algorithm 1 Algorithm for Joint Design under HARQ Type-II

1: **for** $N = 1 : N_{\max}$ **do**
2: **for** $n = 1 : N$ **do**
3: calculate $m_{\mathrm{cut},N\ n}$ and $m_{\mathrm{cut},N\ n+1}$ according to (32).
4: Let **dom** $\mathbf{f_n} = [m_{\mathrm{cut},N\ n},\ m_{\mathrm{cut},N\ n+1})$.
5: solve $\min_m E_{\mathrm{tot,II}}$ according to Lemma 4, denote the solution by $m_{n,N}$ and result by $E_{n,N}$.
6: **end for**
7: Let $E_N = \min_n E_{n,N}$ and corresponding $m_N = m_{n\ ,N}$, where $n = \arg\min_n E_{n,N}$.
8: **end for**
9: Let $E_{\mathrm{tot,II}} = \min_n E_N$ and corresponding $m = m_N$, where $N = \arg\min_N E_N$.
10: Calculate m according to (36).
11: **return** $E_{\mathrm{tot,II}} = E_{\mathrm{tot,II}}(N, m)$ as optimal result, m and N as optimal solutions.

V. Extension to Scenarios with Random Queuing Delay at The MEC Server

Recall that we consider a MEC scenario, where the offloaded task from UE is continuously generated and its results are vital for the system, e.g., state estimation logic. Therefore, the server reserves a slice of computational resources (e.g., dedicated virtual machine) for the tasks while the rest of the resources are used to provide computation services for other UEs with lesser important tasks. However, in some scenarios, the computation resource is shared among UEs, instead of dedicated to a single UE. In particular, assume that the server is only able to provide the computational service for one task at the time and the rest in the queue follows the first-come first-serve (FCFS) policy, i.e., the offloaded task is only executed if all previous tasks are finished. We denote by t_{q} the random waiting time of the offloaded task in the server's queue (queuing delay). In addition, to describe the randomenss, its probability density function (PDF) and cumulative distribution function (CDF) are denoted by $f_{t_{\mathsf{q}}}(\ )$ and $F_{t_{\mathsf{q}}}(\ )$, respectively. Then, the end-to-end latency requirement in (1), i.e., the

task execution should be finished before the deadline, can be rewritten as:

$$T = t_{\mathsf{q}} + t_c^{(n)} + nt + (n \quad 1)t_{\mathsf{k}} \quad T_{\max}, \forall n \in \mathcal{N}. \quad (37)$$

In the following, we investigate the impacts of involving this t_{q} in our analytical model and design in the previous sections, including the computation error model, the energy cost model, and more impotently on the convexity of the modified optimization problem of the design.

A. Computation Error

First, having such a random queuing delay, the computation is no longer always a reliable process. In other words, the delay-violation errors possibly occur, when the random queuing delay makes the rest of time (till the delay constraint) to be insufficient for the task execution. To facilitate the notation, we further define:

$$t_{\mathsf{r}}^{(n)} = T_{\max} \quad \frac{\kappa c^2}{f_{\max}} \quad nt \quad (n \quad 1)t_{\mathsf{k}}, \forall n \in \mathcal{N}, \quad (38)$$

as the maixmal remaining computation time after n^{th} transmission attempt without violating the end-to-end delay requirement. Therefore, the computation error probability after n^{th} successful transmission attempt $\varepsilon_{\mathsf{comp}}^{(n)}$ is given by:

$$\varepsilon_{\mathsf{comp}}^{(n)} = \mathbf{P}\left(t_{\mathsf{q}} > t_{\mathsf{r}}^{(n)}\right) = 1 \quad F_{t_{\mathsf{q}}}\left(t_{\mathsf{r}}^{(n)}\right), \forall n \in \mathcal{N}. \quad (39)$$

Without loss of generality, t_{q} could follow any possible random distribution. Since the task is mission-critical, the system demands an extremely low computation error, i.e., low end-to-end delay violation probability. Therefore, it should fulfill that $1 \quad \varepsilon_{\mathsf{comp}}^{(N)} \quad \varepsilon_{\mathsf{comp}}^{(N \quad 1)} \quad \varepsilon_{\mathsf{comp}}^{(1)}$. In other words, it is more likely to violate the deadline if we spend more time resources for transmission attempts. In such cases, the computation error after N^{th} transmission attempt always dominates other occurrences while keeping a low error probability, i.e., $\varepsilon_{\mathsf{comp}} \quad \varepsilon_{\mathsf{comp}}^{(N)} \quad 1$. Moreover, we adopt the extreme value theory (EVT) model[4] [29], which is able to characterize the tail distribution of ε. In particular, we denote $X = \max\{t_{\mathsf{r}}^{(N)} \quad t_{\mathsf{th}}, 0\}$ as the exceedance of waiting time tolerance, where t_{th} is a sufficiently high threshold. Then, according to [28], if the threshold t_{th} closely approaches $F_{t_{\mathsf{q}}}{}^{1}(1)$, the conditional CDF of the exceedance X is given by:

$$F_{X|t_{\mathsf{r}}>d}(x) = \mathbf{P}(t_{\mathsf{r}} \quad d \quad x|t_{\mathsf{r}} > t_{\mathsf{th}}) \quad G(x;\sigma,\xi) = \begin{cases} e^{\quad x/\sigma}, & \text{if } \xi = 0, \\ 1 \quad \left(1 + \frac{\xi x}{\sigma}\right)^{\quad \frac{1}{\xi}}, & \text{otherwise}, \end{cases} \quad (40)$$

where $G(x;\sigma,\xi)$ is the generalized Pareto distribution, which can be characterized by its parameters σ and ξ. Based on above characterizations, the computation error probability with maximal remaining time $t_{\mathsf{r}}^{(N)}$ and the threshold t_{th} can be written as:

$$\varepsilon_{\mathsf{comp}} = (1 \quad F_{t_{\mathsf{q}}}(d))\left(1 \quad G(\max\{t_{\mathsf{r}}^{(N)} \quad t_{\mathsf{th}}, 0\};\sigma,\xi)\right). \quad (41)$$

[4]Compared to the pioneering work [28], which leverages EVT to characterize the MEC queuing behavior, we further assume that the distribution is unbounded, i.e., $\xi > \frac{1}{2}$. In other word, there is always a possibility the queue may be overloaded.

Importantly, the validity of above expression does not depend on any specific task distribution model [29].

Unlike the originally considered scenario, where only FBL communications contribute to the error performance, the task in the current system could fail either due to the data lost during transmission or the computation duration violating the end-to-end delay requirement. Therefore, the total error probability for current system is as follows:

$$\varepsilon_{\mathsf{tot,q}} = \varepsilon_{\mathsf{comp}} + \varepsilon_{\mathsf{comm}} \quad \varepsilon_{\mathsf{comp}}\varepsilon_{\mathsf{comm}} \quad \varepsilon_{\mathsf{comp}} + \varepsilon_{\mathsf{comm}}. \quad (42)$$

Specially, the communication error probability is defined according to (5) or (25) depending on the HARQ type, respectively.

B. Energy Consumption

The CPU frequency has to adopt the queuing delay as long as it does not exceed $f_{\max}$. Therefore, the computation energy consumption in the case that the server decodes the data successfully in the n^{th} transmission attempt is given by:

$$E_{\mathsf{c,q}}^{(n)} = \int_0^{t_{\mathsf{r}}^{(n)}} \frac{\kappa c^3}{(T_{\max} \quad nt \quad (n \quad 1)t_{\mathsf{k}} \quad \tau)^2} f_{t_{\mathsf{q}}}(\tau)\mathrm{d}\tau, \forall n \in \mathcal{N}. \quad (43)$$

In this way, the expected energy consumption for computation $E_{\mathsf{c,I}}$ and $E_{\mathsf{c,II}}$ in (10) and (28) are still valid without any reformulation while the energy consumption of transmission and feedback are also not influenced by t_{q}. It should be pointed out that if we let t_{q} be constant or zero, our original system model also holds via replacing $T_{\max}$ with $T_{\max} = T_{\max} \quad t_{\mathsf{q}}$.

C. Convexity Revisiting

However, the above modifications on the energy model and error model may also influence the convexity of the optimization problem. In particular, the convexity characterizations of the total error probability (provided in Lemma 1 and Lemma 3) and the total energy cost (provided in Lemma 2 and Lemma 4) should be revisited. In fact, based on the analytical results of Lemma 1 and Lemma 3, we can also show the convexity of the (modified) total error probability in this queuing impact case with the following corollary:

Corollary 1. *The total error probability with the consideration of the queuing at the MEC node is still convex within the feasible set of (13) and the interval of* ***dom*** $\mathbf{f_n}$ *of (34), respectively.*

Proof. The total error probability $\varepsilon_{\mathsf{tot,q}}$ is a sum of communication error probability $\varepsilon_{\mathsf{comm}}$ and computation error probability $\varepsilon_{\mathsf{comp}}$. Since we have already proven that $\varepsilon_{\mathsf{comm}}$ is convex within the feasible set of (13) and the interval of **dom** $\mathbf{f_n}$ of (34), respectively, we only need to investigate the convexity of $\varepsilon_{\mathsf{comp}}$. Assuming $t_{\mathsf{r}} \quad t_{\mathsf{th}}$, we can show that the second derivative:

$$\frac{\partial^2 \varepsilon_{\mathsf{comp}}}{\partial m^2} = (1 \quad F_{D_k}(d))\frac{\partial^2 G(t_{\mathsf{r}} \quad t_{\mathsf{th}};\sigma,\xi)}{\partial m^2} = (1 \quad F_{D_k}(d))\frac{(1+\xi)}{T_S^2\sigma^2}\left(1 + \frac{\xi(t_{\mathsf{r}} \quad t_{\mathsf{th}})}{\sigma}\right)^{\quad \frac{2+\xi}{\xi}} \quad 0. \quad (44)$$

Moreover, it is trivial to show that $\frac{\partial^2 \varepsilon_{\mathsf{comp}}}{\partial m^2} = 0$ if $t_{\mathsf{r}} < t_{\mathsf{th}}$. As a result, $\varepsilon_{\mathsf{comp}}$ is convex in m regardless of HARQ types,

and hence the total error probability $\varepsilon_{\mathsf{tot,q}}$ is convex within the feasible set of (13) and the interval of **dom** $\mathbf{f_n}$ of (34), respectively. □

Similarly, we also characterize the convexity of total energy cost with the current computation model as follows:

Corollary 2. *The total energy cost E_{tot} with the consideration of queuing delay is still convex within the feasible set of (13) and the interval of* ***dom*** $\mathbf{f_n}$ *of (34), respectively.*

Proof. Recall that t_{q} only influences the energy consumption of computation while the convex features of the rest of energy costs remain. Specially, according to Lemma 2 and Lemma 4, we can determine the convexity of total energy cost via determine the sign of $D_1^{(n)} = E_{\mathsf{c}}^{(n)} \quad E_{\mathsf{c}}^{(n\ 1)}$, $\forall n \quad N$. Therefore, we have:

$$
\begin{aligned}
D_1^{(n)} = & \int_0^{t_{\mathsf{r}}^{(n)}} \frac{f_{t_{\mathsf{q}}}(\tau)}{(} T_{\max} \quad nt \quad (n \quad 1)t_{\mathsf{k}})^2 \\
& \tau \mathrm{d}\tau \quad \int_0^{t_{\mathsf{r}}^{(n\ 1)}} \frac{f_{t_{\mathsf{q}}}(\tau)}{(T_{\max} \quad (n \quad 1)t \quad (n \quad 2)t_{\mathsf{k}})^2} \mathrm{d}\tau \\
& \int_0^{t_{\mathsf{r}}^{(n)}} \Bigg(\frac{1}{(T_{\max} \quad nt \quad (n \quad 1)t_{\mathsf{k}})^2} \\
& \frac{1}{(T_{\max} \quad (n \quad 1)t \quad (n \quad 2)t_{\mathsf{k}})^2} \Bigg) f_{t_{\mathsf{q}}}(\tau) \mathrm{d}\tau \\
& \int_0^{t_{\mathsf{r}}^{(n)}} \Bigg(\frac{1}{(T_{\max} \quad nt \quad (n \quad 1)t_{\mathsf{k}})^2} \\
& \frac{1}{(T_{\max} \quad (n)t \quad (n \quad 1)t_{\mathsf{k}})^2} \Bigg) f_{t_{\mathsf{q}}}(\tau) \mathrm{d}\tau = 0.
\end{aligned}
\tag{45}
$$

Hence, the total energy cost E_{tot} is convex within the considered interval. □

According to above corollaries, Lemma 1-4 still hold after introducing the queuing delay in the server's queue t_{q}, as well as the computation error probability $\varepsilon_{\mathsf{comp}}$ for both HARQ type-I and type-II. In other words, although the optimal results may vary, they can still be obtained via our proposed approaches in section III and IV without affecting our analytical findings.

VI. Numerical Evaluation

In this section, we provide our numerical evaluations obtained via Monte Carlo simulations to validate our analytical model and evaluate the considered network.

A. Parameter Setup

We consider the following parameter setups: First, the data size of a task is set to $\beta = 240$ bits. We assume a distance of $d = 50$ m between the UE and the base station, while adopting the NLOS path-loss model in [27], given by $\phi = 17.0 + 40.0 \log_{10}(d)$ with 2.4 GHz carrier frequency. Moreover, we set the end-to-end latency to $T = 60$ ms and the symbol length to $T_S = 0.025$ ms. Furthermore, we set the bandwidth to $B = 5$ MHz, transmit power to $P_{\mathsf{ue}} = P_{\mathsf{k}} = 20$ dBm and noise power to $N = \quad 174$ dBm. Furthermore, we set $t_{\mathsf{k}} = 3$ ms for NACK. For the computation, we set the maximal CPU frequency to $f_{\max} = 3.5$ GHz and total required workload to $c = 20$ Mcycles. We also set the weight factors equally, i.e., $\alpha_{\mathsf{t}} = \alpha_{\mathsf{k}} = \alpha_{\mathsf{c}} = 1$. Finally, we consider an ultra-reliable scenario, where the maximal allowed total error probability is $\varepsilon_{\mathsf{tot,max}} = 0.00001$. For the scenario with queuing delay , we consider the queuing delay follows the exponential distribution with average delay of 3ms.

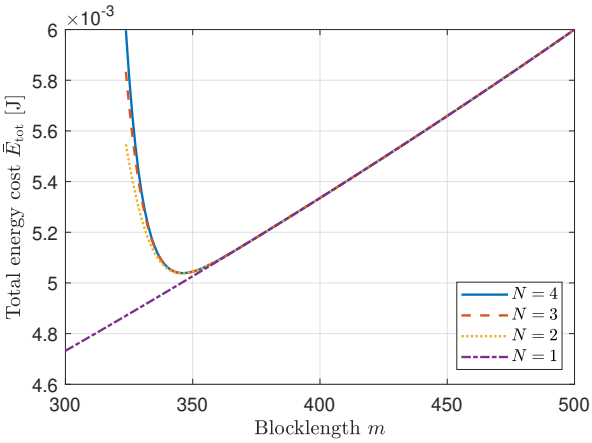

Fig. 3. Total energy cost $E_{\mathsf{tot,I}}$ versus blocklength m under number of (re)transmission $N = 1, 2, 3, 4$.

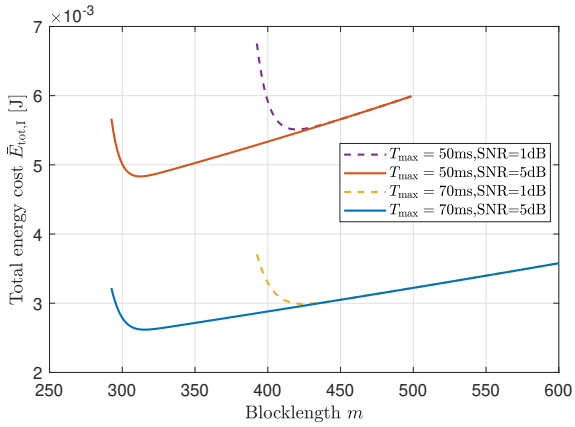

Fig. 4. Total energy cost $E_{\mathsf{tot,I}}$ versus blocklength m with variant setups of end-to-end latency $T_{\max}$ and SNR.

B. HARQ type-I

We start with the numerical results under HARQ type-I. First, we evaluate the impact of blocklength m on the energy cost while considering different setups of N for HARQ type-I. As shown in Fig. 3, the total energy cost $E_{\mathsf{tot,I}}$ is convex in m for each setup of N, which confirms our analytical results. In addition, it is shown that boosting N increases the energy consumption. However, as N grows, the increment of $E_{\mathsf{tot,I}}$ becomes smaller.

Secondly, by considering different setups of end-to-end latency constraint $T_{\max}$ and SNR, we show the total energy cost $E_{\mathsf{tot},I}$ versus blocklength m in Fig. 4. It can be observed that all curves are also convex in m. In addition, the energy costs (under different setups of $T_{\max}$ and SNR) are highly different when blocklength m is short. In fact, with long m, the transmission is reliable, and thus the energy cost

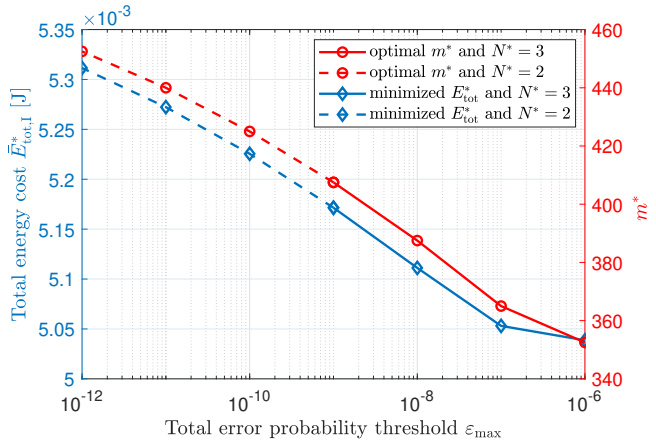

Fig. 5. Optimal energy const $E_{\mathsf{tot,I}}$ and optimal blocklength m vs. error probability threshold $\varepsilon_{\mathsf{tot,max}}$. The optimal maximal allowed transmission times N is indicated with different line type.

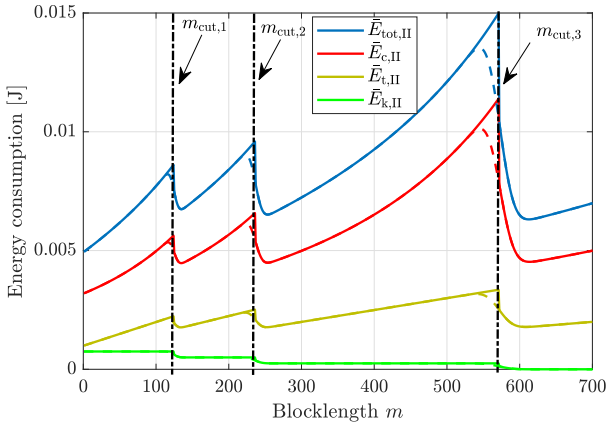

Fig. 6. Total energy cost $E_{\mathsf{tot,II}}$ and the energy consumption for data-transmission $E_{\mathsf{t,II}}$, for sending NACK $E_{\mathsf{k,II}}$, for computation $E_{\mathsf{c,II}}$ versus blocklength m with $N = 4$. $m_{\mathsf{cut},n}$ is indicated with vertical line.

increases sub-linearly along with m. However, note that the blocklength is not always feasible due to the computation time constraint (11c) and error probability constraint (11d). Hence, the feasible values of m are therefore restricted within a convex set. It is also worth to mention that this feasible set of m shrinks while decreasing T_{max} or SNR. In addition, we observe that the convexity of the curves with a relatively shorter T_{max} is relatively sharper than the one with a longer T_{max}, i.e., the accuracy of the optimal solution of m is more important for the considered system under a more strict latency constraint. As the error probability is a monotonically decreasing function with respect to both SNR and blocklength, a longer transmission distance, i.e., corresponding to a lower average channel SNR, significantly increases the energy consumption.

Thirdly, we plot the minimized total energy cost $E_{\mathsf{tot,I}}$ and corresponding optimal blocklength m versus the target error probability $\varepsilon_{\mathsf{tot,max}}$ in Fig. 5. In addition, the optimal solution of allowed transmission attempts N is also shown in the plot. The figure reveals that for stringent $\varepsilon_{\mathsf{tot,max}}$, it requires a sufficiently long blocklength m , resulting in a higher energy cost. Moreover, the dash line implies the optimal $N = 2$

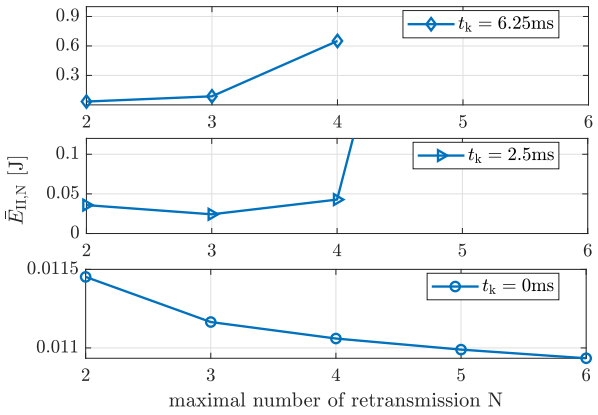

Fig. 7. Optimal energy cost versus maximal number of (re)transmission N at different NACK-transmission lengths t_{k}.

and the solid line represents the optimal $N = 3$. It can be intuitively interpreted from the perspective of the computing energy consumption: if the target error probability is high, the optimum is located at the short transmission duration, with which both the computation energy consumption and the communication energy consumption are low. To compromise the relatively higher error probability caused by the short blocklength, the system has to offer more retransmission attempts. On the other hand, if the target error probability is low, for the given channel quality, the system is expected to have a longer blocklength to guarantee the reliability. Meanwhile, to keep energy consumption low, the length of the computation phase cannot be too short. As a result, the allowed retransmission attempts are reduced.

C. HARQ type-II

Next, we move on to provide numerical results for HARQ type-II. Fig. 6 depicts the impact of blocklength m to the total energy cost $E_{\mathsf{tot,II}}$ as well as all energy consumption components $E_{\mathsf{t,II}}$, $E_{\mathsf{k,II}}$ and $E_{\mathsf{c,II}}$, which corresponds to Fig. 3 under HARQ type-I. We also set $N = 4$ in the figure. To show the influence of the approximated error probability in (23) to the system, we plot the expected energy consumption with approximation in solid line and without approximation in dash line, respectively. Regardless of the approximation, the energy consumption is non-convex and non-concave with respect to m. However, we can observe the energy cost shows convexity within each domain $\mathbf{dom\ f_n}$ after approximation including the energy consumption from all sources, which confirms the Lemma 4. In addition, the gap between dash and solid line becomes high if m is closer to the boundaries of domain, where the error probability of n^{th} (re)transmission increases significantly. Meanwhile, the gap shrinks while the error probability of n^{th} (re)transmission decreases. It implies the approximation is accurate in the low error probability cases, which are the scenarios we consider.

Subsequently, we plot the optimal energy cost $E_{\mathsf{II},N}$ by solving Subproblem (33) versus maximal number of transmission N for different transmission duration of sending NACK t_{k} in Fig. 7. To obtain the optimal solution, we have to compare all E_N and determine the minimum according

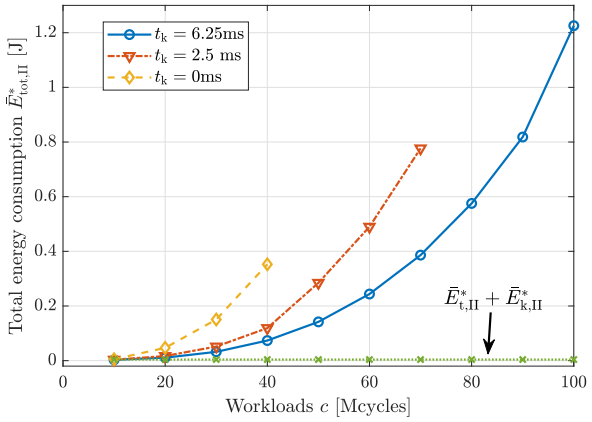

Fig. 8. Optimal total energy cost $E_{\mathsf{tot,II}}$ versus workloads c with variable length of NACK-transmission t_{k}.

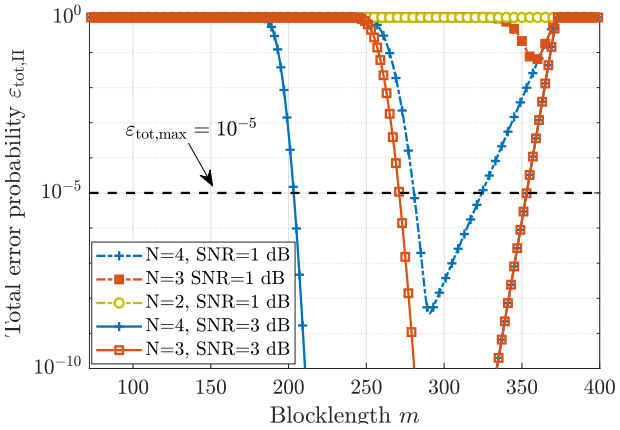

Fig. 9. Total error probability $\varepsilon_{\mathsf{tot,II}}$ versus blocklength m with different setups of N and SNR.

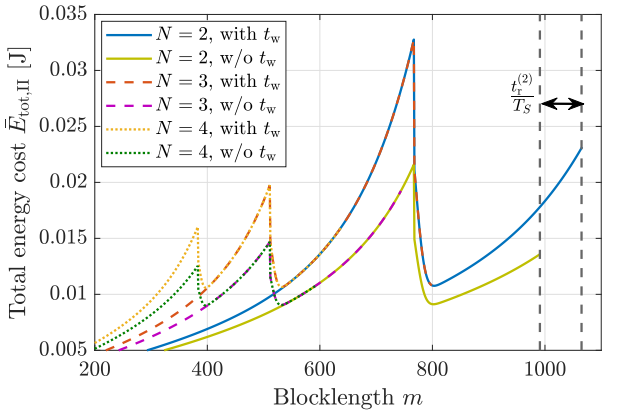

Fig. 10. Total energy cost $E_{\mathsf{tot,II}}$ versus blocklength m with and without consideration of waiting time t_{q}.

to Algorithm 1. Since each retransmission introduces extra energy consumption into the system, small N is preferred if t_{k} is high. In our simulation, there exists no feasible solution for (33) when N is greater than 4 as shown in upper subfigure of Fig. 7. On the other hand, if t_{k} is small, the more retransmission attempts give the better system performance, since it leads to low energy consumption in each retransmission. The lower subfigure shows the extreme case of $t_{\mathsf{k}} = 0$, which implies that sending NACK takes zero time duration. In fact, if we consider an ideal scenario of HARQ type-II neglecting of all cost for sending NACK (including energy consumption of decoding), increasing N always leads to a better performance, since erroneous transmissions have no negative influence to energy consumption and the packets are kept in the receiver anyways. The optimal solution is therefore to choose as many transmissions as possible with the smallest blocklength for transmission, i.e., $N = \frac{T}{T_S}$ and $m = 1$. Nevertheless, it implies that the cost for sending NACK cannot be neglected in the practical system. To obtain the optimal solution, all possible choices of N have to be evaluated and compared accordingly. For instance, in the middle subfigure, the minimum is located at neither boundaries but $N = 3$.

Next, we plot the optimal total energy cost $E_{\mathsf{tot,II}}$ versus workloads c with variant length of NACK-transmission t_{k}, as shown in Fig. 8. We can observe that the increase of workloads c also increases the energy cost regardless of t_{k}. Note that the maximal CPU frequency of the server is limited by f_{max}. Therefore, if c or/and t_{k} is too high, the system is unable to provide the computation service within the end-to-end latency requirement T_{max} while fulfilling the error probability constraints. As a result, increasing c also eventually leads to infeasibility of the original Problem (29). By solely comparing the performance with fixed c, we observe that increasing t_{k} also forces the server to execute the computational task with higher CPU frequency, in order to finish it in time. Recall that the objective function is actually an utility with the sum of weighted. Therefore, to demonstrate the performance difference between only considering communication cost and our objective, we also show the optimal energy cost $E_{\mathsf{t,II}} + E_{\mathsf{k,II}}$. Clearly, since the communication cost is not influenced by the computation, it is directly corresponding to the minimal (expected) blocklength cost that satisfies the error probability constraint. Therefore, increasing workload c does not has any impact to communication energy cost.

D. Impact of The Queuing Delay

Although we have proven that our analytical results and proposed approaches can be extended to the scenario taking queuing delay into account seamlessly, the queuing delay t_{q} and computation error $\varepsilon_{\mathsf{comp}}$ still bring significant impact on the system performance, which should be carefully evaluated as follows. We demonstrate the results for HARQ type-II, where the performance of HARQ type-I is omitted to avoid repetition.

First, we evaluate the impact of queuing delay on the performance of total error probability $\varepsilon_{\mathsf{tot,II}}$ with different setups of N and SNRs in Fig. 9. As expected, the monotonicity of total error probability $\varepsilon_{\mathsf{tot,II}}$ no longer holds when we take the influence of computation error into account. However, we can still observe the convexity, which confirms the analytical results of Corollary 1. In particular, after introducing the queuing delay, higher m improves the reliability of transmissions while leading to a shorter remaining computation time with fixed end-to-end delay. In other words, there exists a trade-off

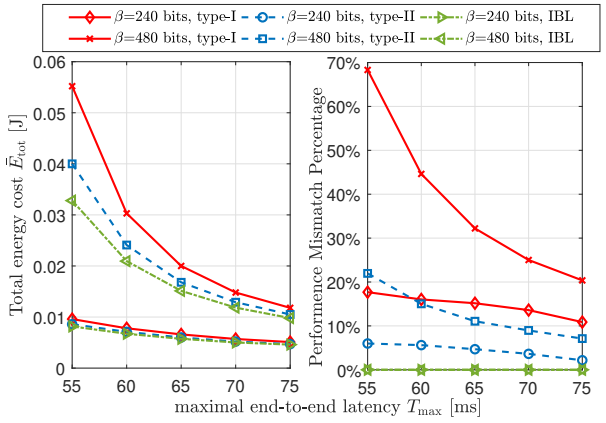

Fig. 11. Performance comparison between the proposed design in FBL regime and the one in the IBL regime.

between the communication phase and the computation phase in the perspective of total error probability. due to the error probability constraint $\varepsilon_{\mathsf{tot}} \quad \varepsilon_{\mathsf{tot,max}}$, it also implies that the reliability requirement becomes more stringent. As shown in the figure, for the setups with low SNR and low retransmission attempts, the problem may be infeasible.

Next, we further investigate the influence of impact of the queuing delay on the energy cost in Fig. 10. It presents the total energy cost $E_{\mathsf{tot,II}}$ versus the blocklength m with and without consideration of queuing delay t_{q}. As shown in the plot, the number of convex sub-domain, i.e., the number of local minima $E_{\mathsf{tot,II}}$, depends on N. We can observe that the curves in both cases have similar structure while the convexity within each domain **dom** $\mathbf{f}_n$ remains. However, the interval of **dom** $\mathbf{f}_n$, i.e., the values of $m_{\mathsf{cut},n}$, varies due to the impact of $\varepsilon_{\mathsf{comp}}$ on the total error probability. In particular, to achieve the same error probability, it requires more blocklength to compromise the negative influence of $\varepsilon_{\mathsf{comp}}$. Therefore, under current setup, $m_{\mathsf{cut},n}$ is shifted slightly. On the other hand, to adopt the random computation waiting time, CPU frequency has to be increased accordingly to avoid violation of end-to-end delay requirement $T \quad T_{\max}$. Therefore, the energy cost also increases significantly, as it is proportional to f^2.

E. Performance Comparison

Finally, in Fig. 11 we compare the performance of HARQ type-I and type-II while varying the end-to-end latency requirement $T_{\max}$. In addition, the performance of the design in the infinite blocklength (IBL) regime (ignoring the FBL impact in the design) is also provided as the benchmark. Note that the data transmission is error-free at the Shannon's capacity in the IBL regime, i.e., the data is transmitted only once and no HARQ is applied. In total two groups of results are provided in two sub-plots of Fig. 11. First, the exact total energy cost performance is shown in the left sub-plot. With the above results, it confirms that the total energy cost E_{tot} decreases while increasing $T_{\max}$ or decreasing β for both types. In addition, As expected, HARQ type-II outperforms type-I, especially when $T_{\max}$ is short. On the other hand, the gap between the two types of HARQ shrinks when a long $T_{\max}$ is available. With different setups of β, we observe that $E_{\mathsf{tot,I}}$ changes dramatically due to the extra energy cost from either increasing blocklength or adding additional retransmission attempts to compromise the increased transmission rate. In such cases, the performances of both types are even close, where the advantage of HARQ type-II may not surpass the implementation complexity introduced comparing to HARQ type-I.

Another observation from the left sub-plot is that both designs perform differently than the one ignoring the FBL impact. We further show in the right sub-plot the performance mismatch percentage of the proposed design in comparison to the IBL ones, i.e., $\frac{E_{\mathsf{tot,FBL}} \quad E_{\mathsf{tot,IBL}}}{E_{\mathsf{tot,IBL}}}$ 100%. The results show more clearly the significant inaccuracy of the design ignoring the FBL impact, especially when the the end-to-end latency is short, or the packet size is large. Moreover, it should be pointed out that if we check the FBL performance of the optimal solution to the design in the IBL regime, it could not satisfy the reliability constraints. In particular, the optimal solution of the IBL design, which sets the coding rate to the Shannon capacity, results in a significantly high error probability in the low-latency transmission with short blocklength according to (3). Hence, the necessity of our design for a state estimator with latency-critical estimation tasks, i.e., leveraging the transmission error model to characterize the FBL impact and applying the HARQ to improve the reliability, is highly confirmed.

VII. Conclusion

In this paper, we studied a MEC network with HARQ schemes in the FBL regime. For both HARQ type-I and HARQ type-II, we provided corresponding optimal retransmission scheme designs by optimally joint allocating the blocklength of a single (re)transmission and determining the maximal allowed transmission attempts, while the objective is to minimize the expected total energy cost for both HARQ schemes. In particular, when the network operates under HARQ type-I, we address the original problem of the retransmission scheme design by decomposing the original problem and characterizing the obtained subproblems. Following the characterizations, we reformulated the original problem to be a solvable integer convex problem. In addition, for retransmission scheme design under HARQ type-II, we solve the problem in a split manner. In particular, after decomposing the original problem into non-convex subproblems, we cut the whole feasible duration of each subproblem into a set of intervals and proved the subproblem to be convex within each interval. An algorithm addressing the flow of solving original problem for HARQ type-II is provided. Furthermore, we showed that our proposed approaches can also be extended with the scenarios that take the impact of queuing delay into account.

Via simulations, we confirmed our analytical model and evaluated the system performance. Moreover, we learnt that the target error probability influences significantly not only in the energy consumption of the system but also the maximal transmission attempts. Finally, when comparing the results of the two types of HARQ, the HARQ type-II outperforms the HARQ type-I, which is as expected since the complexity

of implementing HARQ type-II is significantly higher than HARQ type-I. From this point of view, our results provide guidelines for the designs of MEC networks with both low-power low-cost sensors (where HARQ type-I fits well) and smart sensors (where HARQ type-II is more preferred). Finally, the significant performance gain of our designs in comparison to the ones ignoring the FBL impact is observed, which confirms the necessity of a special design for the considered network in the FBL regime.

Finally, it should be pointed out that the proposed approaches in this work have a high extensibility. Although the considered problem is studied based on a MEC scenario, our approaches can be applied to facilitate the designs with the similar problem structure (blocklength resource allocation with HARQ scheme), while the objective can be extended to effective throughput or expected age-of-information. In addition, the designs can also be extended to multiple hop relaying systems or energy harvesting-enabled systems. Moreover, instead of unifying the blocklength of all transmission attempts, one can extend this work to adjusting the blocklength of current attempt according to the results of previous attempts, i.e, via dynamic programming techniques.

REFERENCES

[1] Y. Zhu, Y. Hu, A. Schmeink, J. Gross, "Energy Minimization of Mobile Edge Computing Networks with Finite Retransmissions in the Finite Blocklength Regime", *IEEE SPAWC 2019*, Cannes, France, 2019.

[2] Y. Mao, C. You, J. Zhang, K. Huang and K. B. Letaief, "A Survey on Mobile Edge Computing: The Communication Perspective," in *IEEE Commun. Surv. Tut.*, vol. 19, no. 4, pp. 2322-2358, 2017.

[3] P. Porambage, J. Okwuibe, M. Liyanage, M. Ylianttila and T. Taleb, "Survey on Multi-Access Edge Computing for Internet of Things Realization," in *IEEE Commun. Surv. Tut.*, vol. 20, no. 4, pp. 2961-2991, Fourthquarter 2018.

[4] X. Lyu, H. Tian, W. Ni, Y. Zhang, P. Zhang, and R. P. Liu, "Energy-Efficient Admission of Delay-Sensitive Tasks for Mobile Edge Computing," in *IEEE Trans. Commun.*, vol. 66, no. 6, pp. 2603-2616, June 2018.

[5] M. Zhao et al., "Energy-Aware Offloading in Time-Sensitive Networks with Mobile Edge Computing," arxiv:2003.12719 [cs], Mar.2020.

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

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

[8] Z. Shi, S. Ma, H. ElSawy, G. Yang and M. Alouini, "Cooperative HARQ-Assisted NOMA Scheme in Large-Scale D2D Networks," in *IEEE Trans. Commun.*, vol. 66, no. 9, pp. 4286-4302, Sept. 2018.

[9] B. Makki, T. Svensson and M. Zorzi, "Finite Block-Length Analysis of the Incremental Redundancy HARQ," in *IEEE Wireless Commun. Lett.*, vol. 3, no. 5, pp. 529-532, Oct. 2014.

[10] B. Makki, T. Svensson, G. Caire and M. Zorzi, "Fast HARQ Over Finite Blocklength Codes: A Technique for Low-Latency Reliable Communication," in *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 194-209, Jan. 2019.

[11] Q. He, Y. Zhu, P. Zheng, Y. Hu and A. Schmeink, "Multi-Device Low-Latency IoT Networks with Blind Retransmissions in the Finite Blocklength Regime," *in IEEE Trans. Veh. Technol., early access*.

[12] A. Anand and G. de Veciana, "Resource Allocation and HARQ Optimization for URLLC Traffic in 5G Wireless Networks," in *IEEE J. Sel. Areas Commun.*, vol. 36, no. 11, pp. 2411-2421, Nov. 2018.

[13] A. Avranas, M. Kountouris and P. Ciblat, "Energy-Latency tradeoff in Ultra-Reliable Low-Latency Communication With Retransmissions," in *IEEE J. Sel. Areas Commun.*, vol. 36, no. 11, pp. 2475-2485, Nov. 2018.

[14] T. Koketsu Rodrigues, J. Liu and N. Kato, "Offloading Decision for Mobile Multi-Access Edge Computing in a Multi-Tiered 6G Network," *in IEEE Trans. Emerg. Topics Comput.,* early access.

[15] C. She, Y. Duan, G. Zhao, T. Q. S. Quek, Y. Li and B. Vucetic, "Cross-Layer Design for Mission-Critical IoT in Mobile Edge Computing Systems," in *IEEE IoT Journal*, vol. 6, no. 6, pp. 9360-9374, Dec. 2019.

[16] R. Dong, C. She, W. Hardjawana, Y. Li and B. Vucetic, "Improving Energy Efficiency of Ultra-Reliable Low-Latency and Delay Tolerant Services in Mobile Edge Computing Systems," in *IEEE ICC Workshops*, Shanghai, China, 2019, pp. 1-6.

[17] R. Dong, C. She, W. Hardjawana, Y. Li and B. Vucetic, "Deep Learning for Hybrid 5G Services in Mobile Edge Computing Systems: Learn From a Digital Twin," in *IEEE Trans. Wireless Commun.*, vol. 18, no. 10, pp. 4692-4707, Oct. 2019.

[18] W. Zhang, Y. Wen, and D. O. Wu, "Collaborative task execution in mobile cloud computing under a stochastic wireless channel," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 81–93, Jan. 2015.

[19] A. Lancho, J. Ostman, G. Durisi, T. Koch, and G. Vazquez-Vilar, "Saddlepoint approximations for short-packet wireless communications," *IEEE Trans. Wirel. Commun.*, vol. 19, no. 7, pp. 4831-4846, Jul. 2020.

[20] M. Shirvanimoghaddam *et al.*, "Short block-length codes for ultra-reliable low latency communications," *IEEE Commun. Mag.* , vol. 57, no. 2, pp. 130–137, Feb. 2019.

[21] W. Yang, G. Durisi, T. Koch and Y. Polyanskiy, "Quasi-Static Multiple-Antenna Fading Channels at Finite Blocklength," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4232-4265, July 2014.

[22] Y. Mao, J. Zhang and K. B. Letaief, "Dynamic Computation Offloading for Mobile-Edge Computing With Energy Harvesting Devices," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3590-3605, Dec. 2016.

[23] Y. Hu, Y. Zhu, M. C. Gursoy and A. Schmeink, "SWIPT-Enabled Relaying in IoT Networks Operating With Finite Blocklength Codes," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 1, pp. 74-88, Jan. 2019.

[24] 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. Wireless Commun.*, vol. 12, no. 9, pp. 4569–4581, Sep. 2013.

[25] Y. Zhu, X. Yuan, B. Han, Y. Hu and A. Schmeink, "Average Age-of-Information Minimization in EH-enabled Low-Latency IoT Networks," in *ICC 2021*, Montreal, 2021, pp. 1-6.

[26] Lubin, M., Yamangil, E., Bent, R. *et al.* , "Polyhedral approximation in mixed-integer convex optimization" *Math. Program.*, vol. 172, no. 1, pp.139-168, Nov. 2018.

[27] Y. Corre, J. Stephan and Y. Lostanlen, "Indoor-to-outdoor path-loss models for femtocell predictions," in Proc. *IEEE PIMRC*, Toronto, ON, 2011, pp. 824-828.

[28] C. Liu, M. Bennis, M. Debbah and H. V. Poor, "Dynamic Task Offloading and Resource Allocation for Ultra-Reliable Low-Latency Edge Computing," *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4132-4150, June 2019.

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