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

(Invited Paper)

Yao Zhu[1], Yulin Hu[1*], Anke Schmeink[1] and James Gross[2]

[1]RWTH Aachen University, Email: $zhu|hu|schmeink$@isek.rwth-aachen.de

[2]KTH Royal Institute of Technology, Email: $james.gross$@ee.kth.se

Abstract

We consider a mobile edge computing network supporting low-latency and ultra reliable services. Task off-loading from the user to the edge server is operated under a truncated retransmission process, i.e., the allowed retransmission times are finite. For such network, we first characterize the end-to-end error probability and the total energy consumption. We subsequently provide a framework design allowing to determine the optimal number of allowed retransmissions and the blocklength for a single transmission/retransmission, where the objective is to minimize the expected total energy consumption while guaranteeing the end-to-end reliability. Via simulation, we confirm our analytical model and evaluated the system performance.

Keywords

edge computing, finite blocklength, offloading, retransmission

I. Introduction

Recently emerging mobile edge computing (MEC) technologies enable the flexible and rapid deployment of latency-critical closed-loop applications by offloading the computation process to the edge of networks [1]. This is motivated by much lower latencies compared to realizing the computation at the cloud. In addition to the low-latency aspect, the demand for ultra-reliability is another key concern in the design of MEC networks, e.g., to satisfy the requirement of online virtual reality gaming and vehicle edge computing applications. Moreover, note that achieving green communication has been one of the key strategies in the design of future networks. Hence, the future MEC networks are expected to guarantee reliable and low-latency services (including both communication and computation) for edge users, while consuming as little energy as possible.

Various energy-efficient offloading schemes have been studied for MEC networks [2]–[5]. In particular, [2] proposes a collaborative task execution algorithm with a partial offloading. The authors in [3] study a wireless powered scenario where the devices are charged by the MEC network and use the charged power to offload tasks. The work in [4] introduces a three-node partition offloading design by optimizing both the time allocation and power allocation. In addition, the trade-off between the communication energy cost and the computation energy cost in the task offloading is investigated in [5], following which the total energy consumption is minimized. However, the above works do not consider the critical applications that demand low latency. In particular, to model the communication, it is more accurate to incorporate finite blocklength (FBL) coding assumptions into the analysis when low-latency applications are considered [6].

Y. Hu is the corresponding author. This work was supported by the DFG research grant SCHM 2643/14.

In such FBL regime, the data transmission is no longer arbitrarily reliable, especially when the blocklength is short. To improve the reliability in the FBL regime, retransmission schemes are proposed in [7], [8] to improve the reliability of the communication with FBL codes. Reference [9] addresses the energy frame optimization problem by selecting the number of retransmissions and determining the frame length. However, the above works focus only on the retransmission process in the communications, i.e., the consideration on the impact of the computation process in the MEC networks is missing. In Particular, the end-to-end delay requirement of a task in a MEC network is fixed, which indicates that the total delay cost of the communication phase and the computation phase are limited. Generally, if the blocklength for each transmission/retransmission is relatively long, the allowed retransmission time is low. Moreover, if the transmission and retransmissions cost a relatively long time, then the remaining time for computation is short. Note that both retransmission times and computing time length significantly influences the total energy consumption of serving the task. It is interesting to have an optimal framework design to minimize the energy consumption while guaranteeing the reliability and latency requirement of the service. To the best of our knowledge, this problem has not been addressed in the literature.

We consider a MEC network with retransmission-supported task offloading in this work and aim at minimizing the expected total energy consumption. We propose a framework design by optimally allocating the time duration of the single (re)transmission. Furthermore, we determine the optimal number of maximal allowed retransmissions. To solve the optimization problem, we leverage the analytical results of the decomposed subproblems to reformulate the original problem as a mixed integer convex problem. Via simulation, we validate our analytical model and evaluate the impact of various setups on the network performance.

II. System Model

We consider a simple MEC network with a user equipment (UE) and a MEC server. Comprehensive computing tasks periodically generated at the UE, are required to be computed at the MEC server, e.g., the UE could be a state estimator which continuously reports time-sensitive information to the MEC server. The estimator is required to provide a state estimate within a fixed time bound T, i.e., the total cost of communication and computation time of a task is limited by T. In addition, due to the reliability requirement, the overall error probability of whole offloading needs to be lower than a threshold $\varepsilon_{\mathrm{tot,max}}$.

The system operates in a time-slotted fashion, where time is divided into frames of length T equivalent to the "end-to-end" delay constraint of the estimation task. Each frame includes a communication phase and a computation phase. In the communication phase, the UE transmits the server the data packet of a task with a size of d bits via the wireless link from the UE to the server. The channel of the link is assumed to experience quasi-static fading, where the channel gain (including the pathloss) is constant within a frame and varies from one frame to the next. Denote by z the channel gain of the link, which is assumed to be perfectly known at the server. Then, the signal-to–noise ratio (SNR) of the received data at the server is given by $\gamma = \phi z P_{\mathrm{ue}}/\sigma_S^2$, where P_{ue} denotes the transmit power of the UE, ϕ is the channel pathloss and σ_S^2 represents the noise power.

Due to the impact of FBL, the transmission is possibly erroneous. A Negative Acknowledgement (NACK) with a fixed small data size of d_{NK} bits is sent to the UE within a fixed duration of t_{NK}. The transmit power of the NACK is denoted by P_{s}. The error probability of decoding/detecting the NACK at the UE is denoted by v. After successfully decoding the NACK, the UE retransmits the data packet till the server successfully decodes it or the maximal allowed retransmission attempts N is reached. Denote the length of a single transmission/retransmission by t and the time duration of one symbol by T_S. Therefore, the blocklength of the transmission is given by $m = \frac{t}{T_S}$. In addition, the time length of the communication phase in a frame is $(n+1)t$, corresponding to $\frac{(n+1)t}{T_S}$ symbols, where n is the number of retransmissions, i.e., $n \in \mathcal{N} = \{0, ..., N\}$. In particular, $n = 0$ represents the initial transmission and $N = 0$ indicates that no retransmission

is allowed.

To guarantee the stringent delay constraint, resources of the edge node are reserved for the task-related computation of the UE, i.e., the server with an adjustable CPU frequency f is able to execute the task immediately after successfully decoding the data packet. It is assumed that the server is able to adjust the frequency f per frame via the dynamic frequency and voltage scaling (DVFS) technique [10], [11] to adopt to the requirement of current task while the maximal available CPU frequency is $f_{\max}$. We assume in the following that there is a fixed computational load of each estimation task of c computation steps, while the execution time t_c depends on the chosen frequency of the processor. Hence, the frequency is chosen according to $f = c/t_c$ with $0 \leq f \leq f_{\max}$.

III. Characterizations of End-to-End Error Probability and Total Energy consumption

A. End-to-End Error Probability in FBL regime

With fixed task data size d and determined blocklength m of the n^{th} retransmission ($n = 0$ represents the initial transmission), the corresponding coding rate is $r = \frac{d}{m}$.

According to the FBL model in [6], the (block) error probability of the n^{th} retransmission is

$$\varepsilon = \mathcal{P}(\gamma, r, m) \approx Q\Big(\sqrt{\tfrac{m}{V(\gamma)}}(\mathcal{C}(\gamma) - r)\log_e 2\Big), \tag{1}$$

where $\mathcal{C} = \log_2(1+\gamma)$ is the Shannon capacity. In addition, V is the channel dispersion [12]. Under a complex AWGN channel, $V = 1 - (1+\gamma)^{-2}$.

Combining both the errors of data transmission and NACk decoding, we can determine the overall error probability in the following way.

Case $N = 0$: No retransmissons are planned. Therefore, the end-to-end error probability ε_{tot} is

$$\varepsilon_{\text{tot}} = \varepsilon, \text{ for N} = 0. \tag{2}$$

Case $N \geq 1$: Firstly, the error probability of the initial transmission ($n = 0$) is ε. If the error occurs at $n = 0$ and the UE successfully decodes the NACK, the process of the 1^{st} retransmission starts. In particular, the error at the server occurs at the n^{th} retransmission if one of the following two events happen: (A). the UE decodes the NACK of the $(n-1)^{\text{th}}$ retransmission wrongly (the server receives nothing in the n^{th} retransmission); (B). the NACK is decoded successfully but the n^{th} retransmission fails. We denote by $P_v(n)$ the error probability of the n^{th} retransmission resulting from the failure of decoding NACK at the UE and by $P_\varepsilon(n)$ the part resulting from the data transmission error. Clearly, we have $P_v(1) = \varepsilon v$ and $P_\varepsilon(1) = \varepsilon^2(1-v)$. Similarly, up to n^{th} retransmission, we have $P_v(n) = \sum_1^n \varepsilon^n (1-v)^{n-1} v$. $P_\varepsilon(n)$ is the probability that all previous NACKs succeeded but all transmissions failed, i.e., $P_\varepsilon(n) = \varepsilon^n (1-v)^n \varepsilon$. As a result, the end-to-end error probability for N maximal allowed retransmission attempts is given by

$$\begin{aligned} \varepsilon_{\text{tot}} &= P_v(N) + P_\varepsilon(N) \\ &= \sum_{n=1}^{N} \varepsilon^n (1-v)^{(n-1)} v + \varepsilon^{(N+1)} (1-v)^N, \text{ for N} \geq 1. \end{aligned} \tag{3}$$

B. Total Energy Consumption

The total energy consumption E_{tot} in a frame consists of three parts: energy consumption of the UE E_{t}, energy consumption at the server for transmitting NACK E_{k} and the computation energy cost at the server E_{c}. Clearly, E_{t}, E_{k}, E_{c} are influenced by the total retransmission 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.

1) Energy consumption of data (re)transmission: The expected energy consumption of the UE $\bar{E}_{\rm t}$ depends on the error probability of NACK and the maximal number of retransmission attempts. Clearly, the expected energy consumption of either the initial transmission or a retransmission is given by $E_{\rm t,0} = tP_{ue} + E'_{\rm s}$, where $E'_{\rm s}$ is the constant energy consumption at the server for receiving a task (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 $(n+1)^{\rm th}$ retransmission depends on the error probability of the $n^{\rm th}$ retransmission and the reliability of the $n^{\rm th}$ NACK. Hence, we have

$$\begin{aligned}\bar{E}_{\rm t} &= E_{\rm t,0} + \varepsilon(1-v)E_{\rm t,0} + ... + \varepsilon^N(1-v)^N E_{\rm t,0} \\ &= \sum\nolimits_{n=0}^{N} \varepsilon^n (1-v)^n E_{\rm t,0}.\end{aligned} \tag{4}$$

2) Energy consumption for sending NACK: Clearly, the energy cost for sending a NACK is given by $E_{\rm k,0} = t_{\rm NK}P_s + E'_{\rm ue}$, where $E'_{\rm ue}$ 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., $E_{\rm k} = 0$. 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 $\varepsilon E_{\rm k,0}$. Moreover, the second NACK occurs if the first two (re)transmissions fail while the previous NACK is detected successfully, i.e., with probability $\varepsilon E_{\rm k,0}$. Similarly, the probability of $n^{\rm th}$ NACK is $\varepsilon^{n+1}(1-v)^n$. Hence, the expected energy consumption $\bar{E}_{\rm k}$ in a frame for sending all NACKs is

$$\begin{aligned}\bar{E}_{\rm k} &= \varepsilon E_{\rm k,0} + \varepsilon^2(1-v)E_{\rm k,0} ... + \varepsilon^{N+1}(1-v)^N E_{\rm k,0} \\ &= \sum\nolimits_{n=0}^{N} \varepsilon^{n+1}(1-v)^n E_{\rm k,0}.\end{aligned} \tag{5}$$

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 non-linear energy consumption model of computation introduced in [14], given by

$$E_{\rm c} = \kappa c f^2 = \kappa c^3 t_c^{-2}, \tag{6}$$

where κ is a constant related to the hardware architecture.

Noting that the computation proceeds immediately, once the input data is received and occupies the rest of the frame, the computation time t_c depends on the number of retransmissions n. In particular, the duration of communication phase (including data transmission and NACK transmission) is $(n+1)t + nt_{\rm NK}$. Then, the remaining time for computation is given by $t_c^{(n)} \leq T - (n+1)t - nt_{\rm NK}$. Since E_c is a monotonic increasing function of t_c, the equality should always hold to minimize the energy consumption. Denote by $E_c^{(n)}$ the computation energy consumption in the case that the server decodes the task data successfully with $n+1$ times transmission attempts ($n=0$ represents the initial transmission). Then, we have

$$E_{\rm c}^{(n)} = \kappa c^3 \tfrac{1}{(T-(n+1)t-nt_{\rm NK})^2}. \tag{7}$$

Clearly, the probability of $n=0$ is $1-\varepsilon$. In addition, the $n^{\rm th}$ retransmission happens when the first n attempts (initial transmission and $n-1$ retransmissions) fail and the corresponding n times NACKs are incorrectly decoded while the $n^{\rm th}$ retransmission is successful. Hence, the probability that $n^{\rm th}$ retransmission happens and is successful, is given by $\varepsilon^n(1-v)^n(1-\varepsilon)$. Therefore, the expected energy consumption for computation is

$$
\begin{aligned}
\bar{E}_{\mathrm{c}} &= (1-\varepsilon)E_{\mathrm{c}}^{(0)} + \varepsilon(1-v)(1-\varepsilon)E_{\mathrm{c}}^{(1)} + ... \\
&\qquad + \varepsilon^N (1-v)^N (1-\varepsilon) E_{\mathrm{c}}^{(N)} \\
&= E_c^{(0)} - \varepsilon^{N+1}(1-v)^N E_c^{(N)} \\
&\quad + \sum\nolimits_{n=1}^{N} \varepsilon^n (1-v)^{n-1} \left((1-v) E_c^{(n)} - E_c^{(n-1)} \right) \\
&\approx E_c^{(0)} + \sum\nolimits_{n=1}^{N} \varepsilon^n (1-v)^{n-1} \left(E_c^{(n)} - E_c^{(n-1)} \right),
\end{aligned} \tag{8}
$$

where the approximation in the last step is tight due to the following reason: Note that we consider ultra reliable scenarios, i.e., $\varepsilon \ll 1$ and $v \ll 1$ hold. Hence, we have $E_c^{(0)} + \varepsilon^N (1-v)^N E^{(N)} \gg \varepsilon^{N+1}(1-v)^N E^{(N)}$

So far, we have derived the expected energy consumptions for task transmission, NACK and computation. Combining these results, the expected total energy consumption (within a frame) $\bar{E}_{\mathrm{tot}}$ can be written as

$$
\bar{E}_{\mathrm{tot}} = \bar{E}_{\mathrm{t}} + \bar{E}_{\mathrm{k}} + \bar{E}_{\mathrm{c}}. \tag{9}
$$

IV. Optimal Framework Design

In this section, we provide a framework design for optimally determining the time duration of a single transmission/retransmission t and the maximal allowed retransmission times N.

A. Problem Statement

Our objective is to minimize expected total energy consumption $\bar{E}_{\mathrm{tot}}$ while guaranteeing the given reliability requirements. In particular, the server should have sufficient time t_c to finish the task within the duration of the frame even in the worst-case scenario, where the task data is received after N attempts of transmission and retransmissions. In addition, recall that the overall error probability needs to be lower than $\varepsilon_{\mathrm{tot,max}}$.[1] Therefore, the problem is formulated by

$$
\begin{aligned}
\underset{t,\,N}{\text{minimize}} \quad & \bar{E}_{\mathrm{tot}} && \text{(10a)} \\
\text{subject to} \quad & t_c^{(n)} + (n+1)t + nt_{\mathrm{NK}} = T,\ \forall n \in \mathcal{N}, && \text{(10b)} \\
& \frac{c}{f_{\max}} + (N+1)t - N t_{\mathrm{NK}} \le T, && \text{(10c)} \\
& \varepsilon_{\mathrm{tot}} \le \varepsilon_{\mathrm{tot,max}}, && \text{(10d)} \\
& N \in \mathbb{Z}. && \text{(10e)}
\end{aligned}
$$

B. Optimal Solution

In this subsection, we solve problem (10) in the following way. We firstly decompose the original problem (10) into $N_{\max}$ subproblems, where $N_{\max}$ is the maximal value of N which is feasible for the original problem. In addition, we derive $N_{\max}$ which limits the total number of subproblems. Moreover, we characterize the subproblems and based on that, we reformulate the original problem to be a solvable integer convex problem.

1) Decomposition of problem (10)*:* 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:

$$
\begin{aligned}
\underset{t}{\text{minimize}} \quad & \bar{E}_{\mathrm{tot}} && \text{(11a)} \\
\text{subject to} \quad & \text{(10b), (10c) and (10d)} && \text{(11b)}
\end{aligned}
$$

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

$$\frac{\partial^2 \bar{E}_{\rm c}}{\partial t^2} = \frac{\partial^2 E_{\rm c}^{(0)}}{\partial t^2} + \sum_{n=1}^{N} \Big[\Big(n(n-1)\varepsilon^{n-2}\big(\tfrac{\partial \varepsilon}{\partial t}\big)^2 + n\varepsilon^{n-1}\tfrac{\partial^2 \varepsilon}{\partial t^2} \Big) D_1^{(n)} - 2n\varepsilon^{n-1}\tfrac{\partial \varepsilon}{\partial t} D_2^{(n)} + \varepsilon^n D_3^{(n)} \Big] , \tag{13}$$

2) Upper Bounds of $N_{\max}$: Without the constraints, $N_{\max}$ is an unbounded integer, resulting in infinite subproblems. However, the maximal number of retransmissions is restricted due to the limited computation power of the server. In particular, by combining the constraints (10b) and (10c), we we obtain an upper bound for $N_{\max}$

$$N_{\max} \leq \lfloor \frac{T - \frac{c}{f_{\max}} - t}{t + t_{\rm NK}} \rfloor, \tag{12}$$

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

3) Optimal Solution of subproblem (11)*:* For a given N, we have the following lemma to handle the subproblem.

Lemma 1. *The total error probability $\varepsilon_{\rm tot}$ is convex in the time length of a single transmission/retransmission t.*

Proof: Since v and $t_{\rm NK}$ are fixed, we are able to obtain $t_c^{(n)}$ as an expression of t, according to (10b):

$$t_c^{(n)} = \max\{T - (n+1)t - nt_{\rm NK}, 0\}. \tag{14}$$

To show the convexity of $\varepsilon_{\rm tot}$ in t, we show necessary conditions for the second derivative. For $N = 0$, we have $\frac{\partial^2 \varepsilon_{\rm tot}}{\partial t^2} = \frac{\partial^2 \varepsilon}{\partial t^2}$. In addition, for $N \geq 1$, we have

$$\begin{aligned} \frac{\partial^2 \varepsilon_{\rm tot}}{\partial t^2} = \frac{\partial^2 \varepsilon}{\partial t^2} v + \sum_{n=2}^{N} n\Big((n-1)\varepsilon^{n-2}\big(\tfrac{\partial \varepsilon}{\partial t}\big)^2 + \varepsilon^{n-1}\tfrac{\partial^2 \varepsilon}{\partial t^2} \Big) \\ + N(N+1)\varepsilon^{N-1}(1-v)^N \tfrac{\partial^2 \varepsilon}{\partial t^2}. \end{aligned} \tag{15}$$

As shown in our previous work [13], $\frac{\partial^2 \varepsilon}{\partial t^2} \geq 0$ holds. Hence, the overall error probability $\varepsilon_{\rm tot}$ is convex in t for both the cases of $N = 0$ and $N \geq 1$. ■

Lemma 2 indicates that Constraint (10d) actually results in a convex feasible set for t in subproblem (11). Note that the other constraints are linear. Subproblem (11) is convex if the objective $\bar{E}_{\rm tot}$ is convex in t, which is addressed in the following lemma.

Lemma 2. *The total energy consumption $\bar{E}_{\rm tot}$ is convex in t.*

Proof: Recall that $\bar{E}_{\rm tot}$ consists of three parts, i.e., $\bar{E}_{\rm tot} = \bar{E}_{\rm t} + \bar{E}_{\rm k} + \bar{E}_{\rm c}$. In the following, we prove the convexity of each part respectively.

We start with $\bar{E}_{\rm t}$ and have

$$\frac{\partial^2 \bar{E}_{\rm t}}{\partial t^2} = \frac{P_{tx}}{T_S} \begin{pmatrix} (1-v)A \\ + \sum_{n=2}^{N} (1-v)^n n(n-1)\varepsilon^{n-2}\big(\tfrac{\partial \varepsilon}{\partial m}\big)^2 t \\ + n\varepsilon^{n-1}(1-v)^n A \end{pmatrix}, \tag{16}$$

where $A = \frac{\partial^2 \varepsilon}{\partial m^2} t + 2\frac{\partial \varepsilon}{\partial m}$. Clearly, $\frac{\partial^2 \bar{E}_{\rm t}}{\partial t^2} \geq 0$ if $A \geq 0$.

Note that $V \leq 1$, $m \geq 1$ and $m = \frac{t}{T_S}$. Hence, we have

$$\begin{aligned} A &= \frac{1}{T_S}\left(\frac{\partial^2 \varepsilon}{\partial m^2} t + 2\frac{\partial \varepsilon}{\partial m} \right) \\ &= \sqrt{\frac{m}{V}} \left(\frac{(C - k/m)(C + k/m)^2}{4Vm^2} - \frac{3C + \frac{k}{m}}{4m^2} \right) \geq \frac{B}{m^3}, \end{aligned} \tag{17}$$

where $B = C^3 m^3 + (C^2 k - 3C)m^2 - (Ck^2 - 3k)m - k^3$ is a third degree polynomial with the greatest root $m = \frac{k}{C}$.

Since $\varepsilon < \varepsilon_{\max} \ll 1$, it holds $C > \frac{k}{m}$ 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

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

Hence, $B \geq 0$ holds. According to (17), $A \geq 0$ also holds. As a result, $\frac{\partial^2 \bar{E}_{\mathrm{t}}}{\partial t^2} \geq 0$, i.e., $\bar{E}_{\mathrm{t}}$ is convex in t.

Secondly, for $\bar{E}_{\mathrm{k}}$ we have

$$\frac{\partial^2 \bar{E}_{\mathrm{k}}}{\partial t^2} = \frac{\partial^2 \varepsilon}{\partial t^2} E_{\mathrm{k}} + \sum_{n=1}^{N} n(n-1)\varepsilon^{n-2}\left(\frac{\partial \varepsilon}{\partial t^2}\right)^2 + n\varepsilon^{n-1}\frac{\partial^2 \varepsilon}{\partial t^2}$$

As shown in [13], $\frac{\partial^2 \varepsilon}{\partial t^2} \geq 0$ holds. It is clear that $\frac{\partial^2 \bar{E}_{\mathrm{k}}}{\partial t^2} \geq 0$, which proves that $\bar{E}_{\mathrm{k}}$ is convex in t.

Finally, we study the convexity of $\bar{E}_{\mathrm{c}}$ regarding t. The second order derivative of $\bar{E}_{\mathrm{c}}$ to t is given in (13), where $D_1^{(n)} = E_{\mathrm{c}}^{(n)} - E_{\mathrm{c}}^{(n-1)}$, $D_2^{(n)} = \frac{\partial E_{\mathrm{c}}^{(n)}}{\partial t} - \frac{\partial E_{\mathrm{c}}^{(n-1)}}{\partial t}$ and $D_3^{(n)} = \frac{\partial^2 E_{\mathrm{c}}^{(n)}}{\partial t^2} - \frac{\partial^2 E_{\mathrm{c}}^{(n-1)}}{\partial t^2}$. As proven previously, ε is a convex and monotonically decreasing function with respect to t, i.e., $\frac{\partial \varepsilon}{\partial t} < 0$ and $\frac{\partial^2 \varepsilon}{\partial t^2} \geq 0$. In particular, we have $\frac{\partial^2 E_{\mathrm{c}}^{(0)}}{\partial t^2} = 6(T-t)^{-4} \geq 0$. Therefore, all the terms besides $D_i^{(n)}, \forall i \in \{1,2,3\}$ in (13) are non-negative, i.e., to determine the convexity of $\bar{E}_{\mathrm{c}}$ is to determine the sign of $D_i^{(n)}$.

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

$$\begin{aligned}D_1^{(n)} &= \frac{1}{(T-(n+1)t-nt_{\mathrm{NK}})^2} - \frac{1}{(T-nt-(n-1)t_{\mathrm{NK}})^2} \\ &\geq \frac{1}{(T-nt-(n-1)t_{\mathrm{NK}})^2} - \frac{1}{(T-nt-(n-1)t_{\mathrm{NK}})^2} = 0.\end{aligned} \tag{19}$$

Similarly, we can show that $D_2^{(n)} \geq 0$ and $D_1^{(n)} \geq 0$ also hold by exploiting $n+1 \geq n$ to carry out the inequality chains. As a result, we have $\frac{\partial^2 \bar{E}_{\mathrm{c}}}{\partial t^2} \geq 0$.

So far, we have proven that $\bar{E}_{\mathrm{t}}$, $\bar{E}_{\mathrm{k}}$ and $\bar{E}_{\mathrm{c}}$ are convex in t. As a result, $E_{\mathrm{tot}} = \bar{E}_{\mathrm{t}} + \bar{E}_{\mathrm{k}} + \bar{E}_{\mathrm{c}}$ is also convex in t. ■

4) Reformulation of the original problem (10) *:* According to Lemma 1 and the upper bounds of $N_{\max}$, we can reformulate the original problem as

$$\underset{t,\,N}{\text{minimize}} \quad E_{\mathrm{tot}} \tag{20a}$$

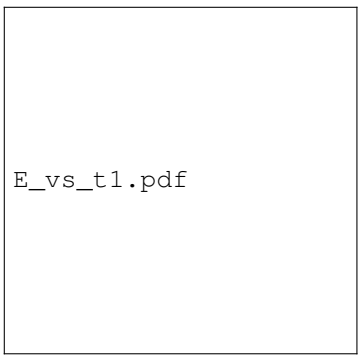
$$\text{subject to} \quad N \leq \left\lfloor \frac{T - \frac{c}{f_{\max}} - t}{t + t_{\mathrm{NK}}} \right\rfloor, \tag{20b}$$

$$\text{(10b), (10c) and (10d).} \tag{20c}$$

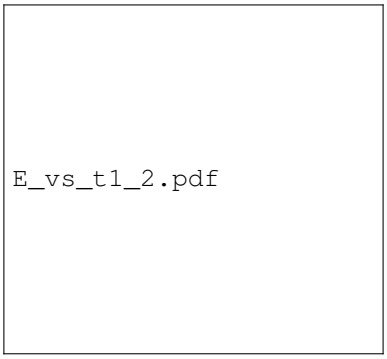
With Lemma 1 and Lemma 2, the objective function and all constraints are either convex or linear. Hence problem (20) is a mixed integer convex problem, which can be solved efficiently [15].

V. Numerical Results

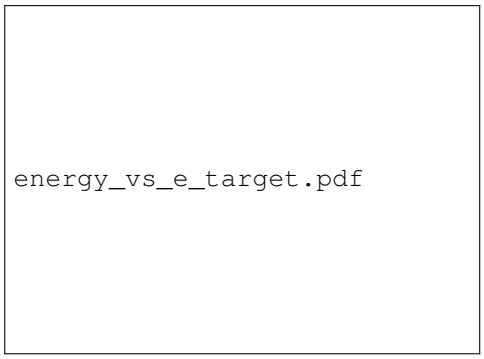
In this section, we provide our numerical results obtained via Monte Carlo simulation. In the simulation, we consider the following parameter setups: First, the data size of a task d is set to 1200 bits. In addition, the distance of the transmission is set to 10 m, while adopting the path-loss model in [16], given by $PL = 17.0 + 40.0\log_{10}(r)$ with 2.4 GHz carrier frequency, where $r = 15$ m is the distance between the UE and the server. Moreover, we set the length of the frame to $T = 50$ ms and the symbol length to $T_s = 0.04$ ms. Furthermore, we set the bandwidth to $B = 5$ Mhz, transmit power to $P_{tx} = P_k = 20$ dBm and noise power to $N = -174$ dBm. Furthermore, we set $t_k = 3$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. Finally, we



(a) $\bar{E}_{\mathrm{tot}}$ vs.t with different N



(b) $\bar{E}_{\mathrm{tot}}$ vs. t with variant T and r



(c) optimal E^*_{tot} and optimal t^* vs. reliability constraint $\varepsilon_{\mathrm{tot,max}}$

Fig. 1. Simulation results.

consider for the maximal total error probability as $\varepsilon_{\mathrm{tot,max}} = 0.001$.

To start with, we evaluate the impact of t on the energy consumption while considering different setups of N. As shown in Fig. 1(a), the overall energy consumption $\bar{E}_{\mathrm{tot}}$ is convex in t for each setup of N, which confirms our analytical model. In addition, it is shown that boosting N increases the energy consumption. However, as N grows, the increment of $\bar{E}_{\mathrm{tot}}$ becomes smaller.

Secondly, by considering different setups of T and r, the convexity of $\bar{E}_{\mathrm{tot}}$ in t is further confirmed in Fig. 1(b) where curves of $r = 15$ m have only feasible values (due to the reliability constraint) when $t > 20$ ms. In addition to the convexity, we observe that a relatively shorter T, i.e., a more strict delay constraint, enhances the sharpness of the convexity. Moreover, a longer transmission distance, i.e., corresponding to a lower average channel SNR, significantly increases the energy consumption.

Finally, in Fig. 1(c), we present the minimized total energy consumption $\bar{E}^*_{\mathrm{tot}}$ and corresponding optimal transmission duration solution l^* versus the target error probability $\varepsilon_{\max}$. In addition, the optimal solution of allowed transmission attempts N^* is also shown in the plot. The figure reveals that for stringent $\varepsilon_{\mathrm{tot,max}}$, it requires a sufficiently long transmission duration t^*, resulting in a higher energy consumption. Moreover, the dash line implies the optimal $N^* = 2$ 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 system prefers a short transmission duration, such that 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 offers more retransmission attempts. On the other hand, if the target error probability is low, for given channel quality, the system is excepted to have a longer transmission duration to guarantee the reliability. Meanwhile, to maintain a low energy consumption, the length of the computation phase cannot be too short. As a result, the allowed retransmission attempts are reduced.

VI. Conclusion

We studied an edge computing network with a retransmission process. We provided an optimal framework design by optimally allocating the time duration of a single (re)transmission and determining the maximal allowed retransmission times, while the objective is to minimize the expected total energy consumption. By decomposing the original problem and characterizing the obtained subproblems, we gain insights on the design. Following these insights, we reformulated the original problem to be a solvable integer convex problem. Via simulation, we confirmed our analytical model and evaluated the system performance.

References

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

[2] 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.

[3] C. You, K. Huang, and H. Chae, “Energy efficient mobile cloud computing powered by wireless energy transfer,” IEEE J. Sel. Areas Commun., vol. 34, no. 5, pp. 1757–1771, May 2016.

[4] X. Cao, F. Wang, J. Xu, R. Zhang and S. Cui, "Joint Computation and Communication Cooperation for Energy-Efficient Mobile Edge Computing," *IEEE IoT Journal* (Early Access).

[5] Y. Yu, J. Zhang, and K. B. Letaief, “Joint subcarrier and CPU time allocation for mobile edge computing,” *IEEE GLOBECOM*, Washington, Dec. 2016.

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

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

[8] B. Makki, T. Svensson and M. Zorzi, "Wireless Energy and Information Transmission Using Feedback: Infinite and Finite Block-Length Analysis," *IEEE Trans. Commun.*, vol. 64, no. 12, pp. 5304-5318, Dec. 2016.

[9] A. Avranas, M. Kountouris, and P. Ciblat, “Energy-latency tradeoff in ultra-reliable low-latency communication with retransmissions.” [Online]. Available: https://arxiv.org/pdf/1805.01332.pdf

[10] G. Semeraro, et al, "Energy-efficient processor design using multiple clock domains with dynamic voltage and frequency scaling," *HPCA*, Cambridge, MA, USA, 2002, pp. 29-40.

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

[12] C. Chen, J. Yan, N. Lu, Y. Wang, X. Yang and X. Guan, "Ubiquitous monitoring for industrial cyber-physical systems over relay-assisted wireless sensor networks," *IEEE Trans. on Emerging Topics in Computing*, vol. 3, no. 3, pp. 352-362, Sept. 2015.

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

[14] 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.

[15] 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.

[16] 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.