Delay-Constrained Communication in Edge Computing Networks

(Invited Paper)

Yulin Hu and Anke Schmeink
Information Theory and Systematic Design of Communication Systems
Chair of Theoretical Information Technology, RWTH Aachen University, 52062 Aachen, Germany.
Email: $hu|schmeink$@umic.rwth-aachen.de

Abstract

In this paper, we consider an edge computing network supporting ultra-reliable low latency communications (URLLC) operating with finite blocklength codes. We derive the end-to-end (E2E) reliability, while both the delay violation probability and the decoding error probability are considered. In addition, we propose an optimal system design to minimize the E2E error probability by optimally setting the target decoding error probability for preserving blocklength/time for each transmission link. In particular, we solve the corresponding optimization problem by proving its convexity. Via simulations, we validate our analytical mode. In addition, we evaluate the considered network, and characterize the impact of delay constraint, target decoding error probability and packet size on the E2E reliability of the considered edge computing network.

Index Terms

edge computing, delay constraint, finite blocklength, URLLC.

I. Introduction

Future wireless networks are expected to support emerging traffic with diversified performance attributes, in order to promote the initiatives including massive machine-type communications (MTC) and Internet of Thing (IoT). Different from conventional throughput-oriented wireless communications, a range of new applications (such as industrial automation, augmented & virtual reality, and remote control) need to be supported ultra-reliable and low-latency communications (URLLC) [1], [2]. In particular, in such MTC URLLC networks, ultra-low end-to-end (E2E) latency may need to be as short as 1 milliseconds or even less.

On the other hand, it is known that edge computing introduces to future wireless MTC networks an efficient means of processing data in not the remote clouds or data centers but at local computing nodes [3]. In particular, processing data at edge computing nodes actually provides a near real-time service and also reduces the data packet size that needs to be sent to the remote receiver [4], [5]. Hence, deploying edge computing nodes for MTC devices exactly satisfies the demand of URLLC. However, prior studies on edge computing are conducted under the ideal assumption of communicating arbitrarily reliably at Shannon's channel capacity achieved when the coding blocklength grows with no bound. In other words, these results only hold in the so-called infinite blocklength (IBL) regime, i.e., code blocks have unbounded lengths. These results are likely also accurate for scenarios with finite but significantly long blocklengths. However, they do not reflect the performance in URLLC applications where the blocklengths are quite short due to the low latency requirements.

It is more accurate to incorporate finite blocklength (FBL) coding assumptions into the analysis when low-latency applications are considered, especially for the network supporting URLLC. In such FBL regime, the data transmission is no longer arbitrarily reliable. Especially when the blocklength is short, the error probability (due to noise) becomes significant even if the data coding rate is selected below the Shannon limit. Taking this into account, an accurate approximation of the achievable coding rate under the FBL assumption for an additive white Gaussian noise (AWGN) channel was derived in [6], [7] for a single-link transmission system. Subsequently, the initial work for AWGN channels was extended to quasi-static fading channels [8]–[10] as well as cooperative networks [11], [12], non-orthogonal multiple access networks [13], [14] and QoS-constrained networks [15], [16]. To the best of our knowledge, the FBL performance of an edge computing network has not been addressed. Nevertheless, recently in [17] the delay distribution of transmissions via an edge computing network is studied in the FBL regime, while the computing time is ignored. However, the general FBL reliability model and reliability-oriented design for edge computing networks is missing, especially when the delay constraint and the computing time cost are taken into account.

In this paper, we study a URLLC edge computing network operating with FBL code. The contributions of this paper can be further detailed as follows:

i. We derive the E2E reliability of the edge computing network, while both the delay violation probability and the decoding error probability are considered.
ii. The E2E reliability is maximized by optimally choosing the target decoding error probability. In particular the convexity of the considered optimal optimization problem is proved.
iii. Via simulations, we validate our analytical mode. In addition, we provide characterizations for the impact of the error probability, packet size and delay constraint on the FBL performance.

The remainder of the paper is organized as follows: In Section II, we describe the system model and briefly provide the background on the FBL regime. In Section III, we study the E2E delay-constrained error probability of the considered edge computing network. Then, we propose an optimal design in Section VI to minimize the E2E error probability. We provide our simulation results in Section IV and finally conclude the paper in Section V.

II. Preliminaries

In this section, we first describe the system model and subsequently briefly review the FBL performance model for a single-link transmission.

A. System model

We consider a simple edge computing network with a sensor node, an edge computing node and a control terminal/unit, as shown in Fig. 1-a. The entire system operates in a slotted fashion, where time is divided into frames. The frame structure is provided in Fig. 1-b, where each frame contains three phases. In the first phase with length m_1 (in symbols), the data generated/obtained at the sensor is transmitted to the edge computing node via an uplink (UL). Subsequently, in the second phase with length n (in symbols) the edge computing node processes the data (if the data is decoded correctly). We assume that the computing time cost n is randomly distributed with a probability density function (PDF) $f_N(n)$. Finally, in the last phase with length m_2 (in symbols) the edge computing nodes transmits the computed result, i.e., a control order, via a downlink (DL) to the control terminal. Denote the sum length of these three phases by M, we have $M = m_1 + n + m_2$. By T_{s} we denote a symbol length in time, then the time duration of a frame is expressed by $T_{\mathrm{f}} = MT_{\mathrm{s}}$. Different from a two-hop relaying network, the packets transmitted in the UL and DL of the considered edge computing network are different. In particular, these two packets are more likely to have different sizes. We express the sizes of the data packets transmitted via the UL and DL by D_1 and D_2. Hence, the data coding rates of the UL and DL are given by $\frac{D_1}{m_1}$ and $\frac{D_2}{m_2}$ in bits per symbol.

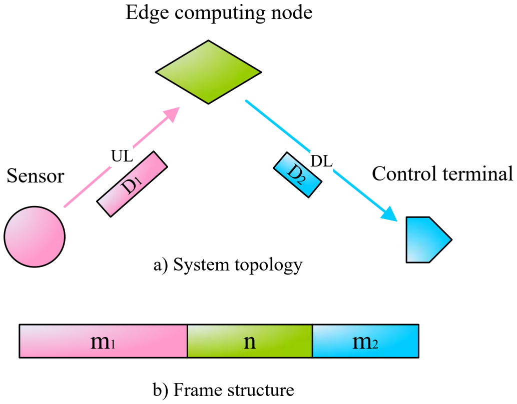

Fig. 1. The considered system topology and frame structure.

Channels are assumed to have random behaviors. In particular, we consider a Rayleigh quasi-static fading model, i.e., the channel fading remains the same within each frame and varies independently from one frame to the next. We express by z_1 and z_2 the random channel gain of the instantaneous UL channel and DL channel, receptively. Hence, the PDF of z_i, the gain due to Rayleigh fading, is given by the exponential distribution: $f(z) = e^{-z}$. Then, the instantaneous signal-to-noise ratios (SNR) of UL and DL are given by $\gamma_i = z_i \bar{\gamma}_i$, $i = 1, 2$, where $\bar{\gamma}_i$ is the average SNR of the link and is influenced by the corresponding path-loss, transmit power, noise and so on. Instantaneous channel state information (CSI) is assumed to be available at the sensor and edge computing node.

Finally and most importantly, the network is expected to support URLLC transmissions. In particular, we assume that the transmission reliability of the transmission via either the UL or the DL should be guaranteed, i.e., the target decoding error probability should be lower than a threshold $\varepsilon^* \leq 0.1$. According to the instantaneous CSI, the system preserves the lengths of m_1 and m_2 while guaranteeing the target decoding error probability. Moreover, the E2E transmission is required to be finished within a hard delay deadline S in symbols. In other words, the delay constraint is violated if the required E2E frame length is longer than the deadline, i.e., $M = m_1 + n + m_2 > S$.

B. Finite blocklength codes

The FBL performance of a single link transmission is analyzed in [6], [7] under an AWGN channel. In particular, with blocklength m, SNR γ and error probability ε the coding rate r (in bits per symbol) is shown to have the following asymptotic expression:

$$r = \mathcal{R}(\gamma, \varepsilon, m) \approx \log_2(1+\gamma) - \sqrt{\frac{V(\gamma)}{m}} Q^{-1}(\varepsilon), \quad (1)$$

where $V(\gamma) = \frac{\gamma(\gamma+2)}{(\gamma+1)^2}\log_2^2 e$ and $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt$.

Form (1), the (block) error probability is given by:

$$\varepsilon = \mathcal{P}(\gamma, r, m) \approx Q\left(\frac{\log_2(1+\gamma) - r}{\sqrt{V(\gamma)/m}}\right). \tag{2}$$

In this paper, we apply the above approximations for investigating the finite blocklength performance of the considered edge computing network. As these approximations have been shown to be accurate for a sufficiently large value of m [7], for simplicity we will employ them as the rate and error expressions in our analysis.

III. Delay-Constrained Error Probability of the Edge Computing Network

In this section, we characterize the impact of the transmission blocklength m_1 and m_2 and computing cost n on the overall error probability. The minimal blocklength cost (in terms of symbols) of reliably transmitting the data packet via each of the two links is subject to the channel SNR. In particular, for either the UL or the DL in each frame, based on the corresponding instantaneous SNR, the minimal blocklength cost satisfying the target decoding error probability ε^* can be determined. In particular, according to (2) the error probability of either the UL link or the DL link with packet size D_i, SNR γ_i and blocklength m_i is given by $\varepsilon_i = \mathcal{P}\left(\gamma_i, \frac{D_i}{m_i}, m_i\right)$, where the $\mathcal{P}$ function is defined in (2). Note that we consider a reliable transmission where the decoding error probabilities of UL and DL are lower than 0.1 and therefore are definitely lower than 0.5. Hence, in order to guarantee the target decoding error probability ε^*, the minimal blocklength (cost) m_i^* of each link is required to satisfy $\varepsilon^* = \mathcal{P}\left(\gamma_i, \frac{D_i}{m_i^*}, m_i^*\right)$. According to (2), we then have

$$m_i^* - \beta_i\sqrt{m_i^*} - \frac{D_i}{\log_2(1+\gamma_i)} = 0 \,, \tag{3}$$

where $\beta_i = Q^{-1}(\varepsilon^*)\frac{\log_2 e\sqrt{1-(1+\gamma_i)^{-2}}}{\log_2(1+\gamma_i)}$. Equation (3) is a quadratic equation with respect to $\sqrt{m_i^*}$. Therefore, we can obtain $\sqrt{m_i^*}$ and thus m_i^* by solving (3), where m_i^* represents the minimal blocklength for link i (recall that $i=1$ indicates the UL, while $i=2$ presents the DL). In particular, m_i^* is calculated by

$$m_i^* = \frac{D_i}{\log_2(1+\gamma_i)} + \frac{1}{2}\beta_i^2 + \beta_i\sqrt{\frac{D_i}{\log_2(1+\gamma_i)} + \left(\frac{\beta_i}{2}\right)^2} \,. \tag{4}$$

It is clear that m_i^* is a function of γ_i and β_i and that β_i is a function of γ_i. In other words, m_i^* is actually a function of γ_i. We denote this function by $g(\cdot)$, i.e., $m_i^* = g(\beta_i)$. Then, the corresponding inverse function is expressed as $\beta_i = g^{-1}(m_i^*)$. Based on the PDF of the channel gain $f_Z(z)$, the CDF of m_i^* is then given by

$$\begin{aligned} p_i &= F_{m_i^*}(m, \bar{\gamma}_i) \\ &= \int\limits_{z_i\in\Omega_i} f_Z(z_i)\,dz_i = \int\limits_0^{g^{-1}(m_i)/\bar{\gamma}_i} f_Z(z_i)\,dz_i \,, \end{aligned} \tag{5}$$

where $\Omega_i = \{z_i : m_i^*(z_i\bar{\gamma}_i) \le m\}$. Value p_i indicates the probability that the transmission of the i^{th} link can be finished within m symbols. Then, the PDF of m_i^* of either UL or DL with average channel gain $\bar{\gamma}_i$ is

$$f_{m_i^*}(m, \bar{\gamma}_i) = \frac{\partial F_{m_i^*}(m)}{\partial m} = \frac{p\left(\frac{g^{-1}(m)}{\bar{\gamma}_i}\right)}{\bar{\gamma}_i\frac{\partial g(g^{-1}(m))}{\partial m}} \,. \tag{6}$$

Then, the system determines the blocklength m_i, $i=1,2$ of the two links according to the minimal blocklength costs m_i^*, $i=1,2$. Note that the channels of the two hops are i.i.d., hence the blocklength cost $m_i, i{=}1,2$, are i.i.d.

Recall that the computing time cost n is also randomly distributed with a PDF $f_N(n)$. Then, the PDF of M is

$$f_M(m) = f_{m_1^*}(m, \bar{\gamma}_1) \otimes f_{m_2^*}(m, \bar{\gamma}_2) \otimes f_N(m) \,. \tag{7}$$

We obtain the CDF of M, which is given by

$$F_M(m) = \int\limits_{\tau=0}^{m} f_M(\tau)d\tau. \tag{8}$$

Here, we treat M as a continuous random variable for approximation and will show the appropriateness of the approximation via simulations in Section V.

Finally, based on the CDF of M, for given total E2E delay constraint S (in symbols), the probability that the E2E transmission can be finished within the delay constraint is given by

$$p_{\mathrm{d}} = F_M(S)\,. \tag{9}$$

Hence, the average delay violation probability (over fading) is given by $1 - p_{\mathrm{d}}$. Recall that the decoding error probability at either UL or DL is $\varepsilon^* \leq 10^{-1}$, then the decoding error probability of the E2E transmission is $2\varepsilon^* - (\varepsilon^*)^2 \approx 2\varepsilon^*$. Combining this decoding error probability with the average delay violation probability, the average overall E2E error probability can be finally obtained, which is given by

$$\begin{aligned}\varepsilon_{\mathrm{tot}} &= 1 - p_{\mathrm{d}} + 2\varepsilon^* - (1 - p_{\mathrm{d}}) \cdot 2\varepsilon^* \\ &= 1 - p_{\mathrm{d}} + 2\varepsilon^* p_{\mathrm{d}}\ .\end{aligned} \tag{10}$$

So far, we have derived the delay-constrained E2E reliability of the considered edge computing network. In the next section, we will follow the obtained model and propose an optimal design for the network.

IV. Optimal design for Maximizing the E2E Reliability

According to (10), the (average) E2E error probability $\varepsilon_{\mathrm{tot}}$ is actually a function of ε^* and p_{d}. Note that p_{d} is also influenced by ε^*. This actually allows us to minimize $\varepsilon_{\mathrm{tot}}$ by selecting the value of ε^*. The optimization problem can be expressed as

$$\min_{\varepsilon^*} \varepsilon_{\mathrm{tot}} \tag{11}$$

Next, we provide the following key proposition for solving the problem.

Proposition 1. *For the considered edge computing system operating with FBL codes, the average overall error probability $\varepsilon_{\mathrm{tot}}$ is convex in the target error probability ε^*.*

Proof: According to (10), we have $\frac{\partial \varepsilon_{\mathrm{tot}}}{\partial \varepsilon^*} = (2\varepsilon^* - 1)\frac{\partial p_{\mathrm{d}}}{\partial \varepsilon^*} + 2p_{\mathrm{d}}$ and $\frac{\partial^2 \varepsilon_{\mathrm{tot}}}{\partial^2 \varepsilon^*} = (2\varepsilon^* - 1)\frac{\partial^2 p_{\mathrm{d}}}{\partial^2 \varepsilon^*} + 2\frac{\partial p_{\mathrm{d}}}{\partial \varepsilon^*}$. In the following, we prove the proposition by showing $\frac{\partial^2 \varepsilon_{\mathrm{tot}}}{\partial^2 \varepsilon^*} \geq 0$.

According to (9) and (7), we have $\frac{\partial p_{\mathrm{d}}}{\partial \varepsilon^*} = \frac{\partial p_1}{\partial \varepsilon^*} \otimes f_{m_2}(S) \otimes f_N(S)$, where p_1 and p_2 are introduced in (5) as the probabilities that the UL (DL) can be finished within S symbols. We first show that the condition $\frac{\partial^2 p_1}{\partial^2 \varepsilon^*}(2\varepsilon^* - 1) + 2\frac{\partial p_1}{\partial \varepsilon^*} \geq 0$ holds in the considered system. According to the Rayleigh channel fading model, the probability that UL can be given by $p_1 = \int\limits_{\gamma_1^*/\bar{\gamma}_1}^{+\infty} e^{-z_1} dz_1 = \frac{e^{-\gamma_1^*/\bar{\gamma}_1}}{\bar{\gamma}_1}$, where γ_1^* is the SNR threshold for making the transmission reliable (guaranteeing ε^*). Obviously, γ_1^* is a function of ε^*. Then, the first and second derivatives of p_1 with respect to ε^* are given by $\frac{\partial p_1}{\partial \varepsilon^*} = -\frac{1}{\bar{\gamma}_1^2}\frac{\partial \gamma_1^*}{\partial \varepsilon^*} e^{\frac{-\gamma_1^*}{\bar{\gamma}_1}}$ and $\frac{\partial^2 p_1}{\partial^2 \varepsilon^*} = \frac{1}{\bar{\gamma}_1^2} e^{\frac{-\gamma_1^*}{\bar{\gamma}_1}} \left(\frac{1}{\bar{\gamma}_1}\left(\frac{\partial \gamma_1^*}{\partial \varepsilon^*}\right)^2 - \frac{\partial^2 \gamma_1^*}{\partial^2 \varepsilon^*}\right)$.

Recall that $\varepsilon^* = \mathcal{P}\left(\gamma_1, \frac{D_1}{m_1^*}, m_1^*\right)$. Then, we have $Q^{-1}(\varepsilon) = \frac{\sqrt{m_1}}{\log_2 e} \frac{\log_2(1+\gamma_1) - \frac{D_1}{m_1}}{\sqrt{\frac{{\gamma_1}^2 + 2\gamma_1}{(1+\gamma_1)^2}}}$. Hence, the first derivative of $Q^{-1}(\varepsilon^*)$ with respect to ε^* is given by

$$\dot{Q}^{-1}(\varepsilon^*) = \frac{\sqrt{m_1}}{\log_2 e} \frac{\log_2 e - \frac{1}{(\gamma_1^2 + 2\gamma_1)}\left(\log_2(1+\gamma_1) - \frac{D_1}{m_1}\right)}{\sqrt{\gamma_1^2 + 2\gamma_1}} \frac{\partial \gamma_1^*}{\partial \varepsilon^*}\ . \tag{12}$$

On the other hand, according to the definition of the Q-function, we have another expression of this derivative

$$\dot{Q}^{-1}(\varepsilon^*) \;=\; -\sqrt{2\pi} e^{\frac{\left(Q^{-1}(\varepsilon^*)\right)^2}{2}} < 0\,. \tag{13}$$

In other words, the right side of (12) is negative. Thus, we have $1 - \frac{1}{(\gamma_1^2 + 2\gamma_1)}\left(\log_2(1+\gamma_1) - \frac{D_1}{m_1}\right) > 0$ as $\gamma_1^2 + 2\gamma_1 > \log_2(1+\gamma_1) > \log_2(1+\gamma_1) - \frac{D_1}{m_1}$ for $\gamma_1 > 0$. Hence, $\frac{\partial \gamma_1^*}{\partial \varepsilon^*} < 0$ holds. Moreover, regarding the second order derivative, the following relationship holds:

$$\begin{aligned}\bar{\gamma}_1 \frac{\partial \gamma_1^*}{\partial \varepsilon^*} &= \bar{\gamma}_1 \frac{-\sqrt{2\pi} e^{\frac{\left(Q^{-1}(\varepsilon^*)\right)^2}{2}}}{\frac{\sqrt{m_1}}{\log_2 e} \frac{1 - \frac{1}{(\gamma_1^2 + 2\gamma_1)}\left(\log_2(1+\gamma_1) - \frac{D_1}{m_1}\right)}{\sqrt{\gamma_1^2 + 2\gamma_1}}} \\ &< -2\bar{\gamma}_1 \sqrt{\frac{\gamma_1^2 + 2\gamma_1}{m_1}} \cdot e^{\frac{m_1}{2}(1+\gamma_1)^2 \left(\frac{\log_2(1+\gamma_1) - \frac{D_1}{m_1}}{\log_2 e \sqrt{\gamma_1^2 + 2\gamma_1}}\right)^2} \\ &\ll -2\ .\end{aligned} \tag{14}$$

Based on (12) and (13), we obtain the following two expressions for the second derivative of $Q^{-1}(\varepsilon^*)$:

$$
\begin{aligned}
\ddot{Q}^{-1}(\varepsilon^*) = & \frac{\sqrt{m}_1}{\log_2 e} \frac{\log_2 e - \frac{1}{(\gamma_1^2+2\gamma_1)} \left(\log_2(1+\gamma_1) - \frac{D_1}{m_1}\right)}{\sqrt{\gamma_1^2+2\gamma_1}} \frac{\partial^2 \gamma_1^*}{\partial^2 \varepsilon^*} \\
& - \frac{\sqrt{m}_1}{\log_2 e} \frac{\log_2 e - \frac{1}{(\gamma_1^2+2\gamma_1)} \left(\log_2(1+\gamma_1) - \frac{D_1}{m_1}\right)}{(\gamma_1^2+2\gamma_1)^{\frac{3}{2}}} \left(\frac{\partial \gamma_1^*}{\partial \varepsilon^*}\right)^2, \\
\ddot{Q}^{-1}(\varepsilon^*) = & 2\pi Q^{-1}(\varepsilon^*) e^{\left(Q^{-1}(\varepsilon^*)\right)^2}, \quad .
\end{aligned}
$$

where $\ddot{Q}^{-1}(\varepsilon^*) > 0$ due to the fact that $Q^{-1}(\varepsilon^*) > 0$ for all the $\varepsilon^* \leq 0.1 < 0.5$. Recall that $1 - \frac{1}{(\gamma_1^2+2\gamma_1)} \left(\log_2(1+\gamma_1) - \frac{D_1}{m_1}\right) > 0$, thus it is easy to conclude $\frac{\partial^2 \gamma_1^*}{\partial^2 \varepsilon^*} < 0$ from the above two expressions.

Further, we obtain $\frac{\partial^2 p_1}{\partial^2 \varepsilon^*}(2\varepsilon^* - 1) + 2\frac{\partial p_1}{\partial \varepsilon^*} = \frac{1}{\bar{\gamma}_1^2} e^{-\frac{\gamma_1^*}{\bar{\gamma}_1}} \left(\frac{1}{\bar{\gamma}_1}\left(\frac{\partial \gamma_1^*}{\partial \varepsilon^*}\right)^2 - \frac{\partial^2 \gamma_1^*}{\partial^2 \varepsilon^*}\right)(2\varepsilon^* - 1) - \frac{2}{\bar{\gamma}_1^2}\frac{\partial \gamma_1^*}{\partial \varepsilon^*} e^{-\frac{\gamma_1^*}{\bar{\gamma}_1}}$. Note that $\frac{\partial \gamma_1^*}{\partial \varepsilon^*} < 0$. Hence, we have that $\frac{\partial^2 p_1}{\partial^2 \varepsilon^*}(2\varepsilon^* - 1) + 2\frac{\partial p_1}{\partial \varepsilon^*}$ is larger than $\frac{1}{\bar{\gamma}_1^3} e^{-\frac{\gamma_1^*}{\bar{\gamma}_1}} \left(\left(\frac{\partial \gamma_1^*}{\partial \varepsilon^*}\right)^2 - 2\bar{\gamma}_1 \frac{\partial \gamma_1^*}{\partial \varepsilon^*}\right)$, which is further lager than a positive value $\frac{1}{\bar{\gamma}_1^3} e^{-\frac{\gamma_1^*}{\bar{\gamma}_1}} \left(\left(\frac{\partial \gamma_1^*}{\partial \varepsilon^*}\right)^2 + 4\right)$ as $\bar{\gamma}_1 \frac{\partial \gamma_1^*}{\partial \varepsilon^*} \ll -2$ according to (14). Therefore, the condition $\frac{\partial^2 p_1}{\partial^2 \varepsilon^*}(2\varepsilon^* - 1) + 2\frac{\partial p_1}{\partial \varepsilon^*} \geq 0$ holds.

Finally, we have

$$
\begin{aligned}
\frac{\partial^2 \varepsilon_{\text{tot}}}{\partial^2 \varepsilon^*} &= (2\varepsilon^* - 1)\frac{\partial^2 p_d}{\partial^2 \varepsilon^*} + 2\frac{\partial p_d}{\partial \varepsilon^*} \\
&= (2\varepsilon^* - 1)\frac{\partial^2 p_1}{\partial^2 \varepsilon^*} \otimes f_{m_2}(S) \otimes f_N(S) \\
&\quad + 2 f_{m_1}(S) \otimes f_{m_2}(S) \otimes f_N(S) \\
&= \left(\frac{\partial^2 p_1}{\partial^2 \varepsilon^*}(2\varepsilon^* - 1) + 2\frac{\partial p_1}{\partial \varepsilon^*}\right) \otimes f_{m_2}(S) \otimes f_N(S).
\end{aligned}
$$

Note that the PDFs $f_{m_2}(S)$ and $f_N(S)$ are non-negative. $\frac{\partial^2 \varepsilon_{\text{tot}}}{\partial^2 \varepsilon^*} \geq 0$ as condition $\frac{\partial^2 p_1}{\partial^2 \varepsilon^*}(2\varepsilon^* - 1) + 2\frac{\partial p_1}{\partial \varepsilon^*} \geq 0$ holds. Hence, $\frac{\partial^2 \varepsilon_{\text{tot}}}{\partial^2 \varepsilon^*} \geq 0$ and ε_{tot} is convex in ε^*.

■

Therefore, the E2E delay-constrained reliability (combining both the delay violation probability and the decoding error probabilities) of the considered edge computing network can be optimized by choosing an appropriate target decoding error probability, i.e., the optimization problem in (11) can be solved efficiently [18].

V. NUMERICAL RESULTS

In this section, we resort to Monte Carlo simulations to confirm the accuracy of our analytical model and evaluate the network performance. In the simulations, we consider the following parameterization. First, we set the average SNRs of the two links to 20 dB, while 15 dB is also considered in Fig. 2. In addition, without being specifically noted, the default setups for delay constraint and the packet sizes are S = 2000 symbols, D_1 = 400 bits and D_2 = 40 bits. Moreover, we consider a Poisson distributed computing time cost n (in symbols) with an average value $\bar{n}$, while the default setup of $\bar{n}$ is $\bar{n} = 100$ symbols.

To start with, we study the impact of the target decoding error probability on the average E2E overall error probability. The results are provided in Fig. 2 where different average SNR setups are considered. It can be observed that the overall error probabilities are convex in the target decoding error probability, which confirms the analytical results in Proposition 1. Moreover, although the packet size of the UL is set to be significantly higher than the one of the DL, it makes no big difference by setting the average SNR of either the UL or DL a bit higher, i.e., increasing the SNRs at different links actually introduces a similar improvement on the E2E reliability performance.

Next, we evaluate the impact on the packet sizes of UL and DL transmissions on the optimal E2E error probability (achieved by choosing the optimal target decoding error probability). The results are provided in Fig. 3. As expected, the overall error probabilities are increasing in the packet sizes. Moreover, the E2E transmission is more reliable when the control information D_2 is less or the delay constraint is loose.

Finally, we investigate the relationship between the available E2E reliability and the computing cost n. When the delay constraint is loose, the impact of n is not considerable if the average computing cost is less half of the delay constraint. However, when the delay constraint is stringent, the impact of computing cost becomes significant, i.e., the slop of the curve with $S = 1000$ symbols is quite steep. This indicates a guideline to improve the E2E reliability of ultra low latency service by deploying more powerful computing nodes or providing higher priority at the computing node than other services with relatively longer loose latency constraint.

VI. CONCLUSION

In this paper, we have investigated the E2E reliability performance of an edge computing network with quasi-static fading channels. The E2E error probability is derived in the FBL regime. Moreover, we have proposed an optimal design to minimize the E2E error probability by choosing the optimal target decoding error probability. Via numerical analysis, first we have validated our analytical model. We have also observed that the E2E reliabilities are convex in the target decoding error probability and are increasing in the packet sizes and computing time cost. Future work will focus on extending the current model to scenarios with multiple computing nodes and multiple control terminals.

REFERENCES

[1] C. She, C. Yang and T. Q. S. Quek, "Radio resource management for ultra-reliable and low-latency communications," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 72-78, Jun. 2017.

[2] Y. Hu, M. C. Gursoy and A. Schmeink, "Relaying-Enabled Ultra-Reliable Low Latency Communications in 5G", *IEEE Network*, , vol. 32, no. 2, pp. 62-68, Mar.-Apr. 2018.

[3] N. Abbas *et.al.*, "Mobile Edge Computing: A Survey," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 450-465, Feb. 2018.

[4] W. Yu *et.al.*, "A Survey on the Edge Computing for the Internet of Things," *IEEE Access*, vol. PP, no. 99, pp. 1-1.

[5] F. Wang *et.al.*, "Joint Offloading and Computing Optimization in Wireless Powered Mobile-Edge Computing Systems," in *IEEE Trans. Wireless. Commn.*, vol. 17, no. 3, pp. 1784-97, Mar. 2018.

[6] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Dispersion of gaussian channels,"in *IEEE International Symposium on Information Theory (ISIT)*, pp. 2204–2208, 2009.

[7] 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.

[8] W. Yang *et.al.* “Quasi-static multiple-antenna fading channels at finite blocklength,” *IEEE Trans. Inf. Theory*, vol. 60, no. 7, Jul. 2014.

[9] G. Ozcan and M. C. Gursoy, “Throughput of cognitive radio systems with finite blocklength codes,” *IEEE J. Sel. Areas Commun.*, vol. 31, no. 11, pp. 2541-2554, Nov. 2013.

[10] S. Xu *et.al.*, "Energy-efficient packet scheduling with finite blocklength codes: Convexity analysis and efficient algorithms," *IEEE Trans. Wireless. Commn.*, vol.15, no.8, pp.5527-40, Aug. 2016.

[11] Y. Hu, J. Gross and A. Schmeink, “On the capacity of relaying with finite blocklength,” *IEEE Trans. Veh. Technol.*, vol. 65, no. 3, pp. 1790-1794, Mar. 2016.

[12] Y. Hu, J. Gross and A. Schmeink, “On the performance advantage of relaying under the finite blocklength regime,” *IEEE Commun.Letters*, vol. 19, no. 5, pp. 779–782, May 2015.

[13] X. Sun *et.al.*, "Short-Packet Downlink Transmission with Non-Orthogonal Multiple Access," *IEEE Trans. Wireless. Commn.*, Early Access Apr. 2018

[14] Y. Hu, M. C. Gursoy and A. Schmeink, “Efficient Transmission Schemes for Low-Latency Networks: NOMA vs. Relaying", in *Proc. IEEE PIMRC*, Montreal, QC, Oct. 2017. (Best Paper Award).

[15] M. C. Gursoy, “Throughput analysis of buffer-constrained wireless systems in the finite blocklength regime,” *EURASIP J. Wireless Commun.*, vol. 2013:290, Dec. 2013.

[16] Y. Hu, A. Schmeink and J. Gross “Blocklength-limited performance of relaying under quasi-static Rayleigh channels,” *IEEE Trans. Wireless. Commn.*, vol. 15, no. 7, pp. 4548 - 4558, Jul. 2016.

[17] S. Schiessl *et.al.*, "Finite Length Coding in Edge Computing Scenarios," *21th ITG WSA*, Berlin, Germany, 2017

[18] S. Boyd and L. Vandenberghe, Convex optimization. New York, NY, USA: Cambridge Univ. Press, 2004

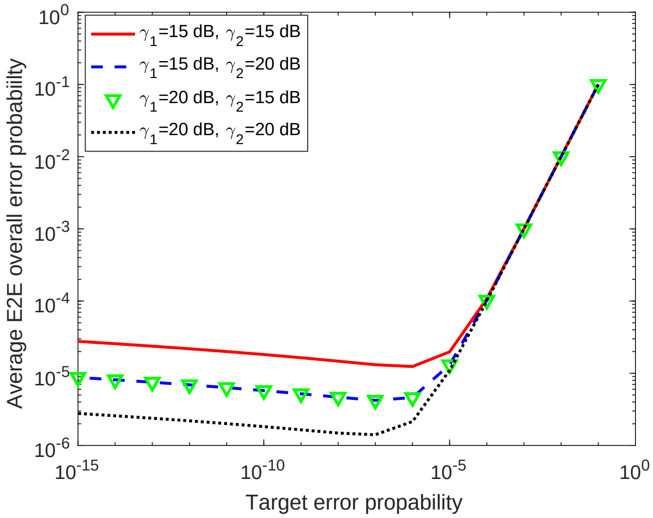

Fig. 2. The E2E reliability performance at different average SNR setups. In the simulation, we set delay constraint to $S = 2000$ symbols and computing cost to $n = 100$ symbols.

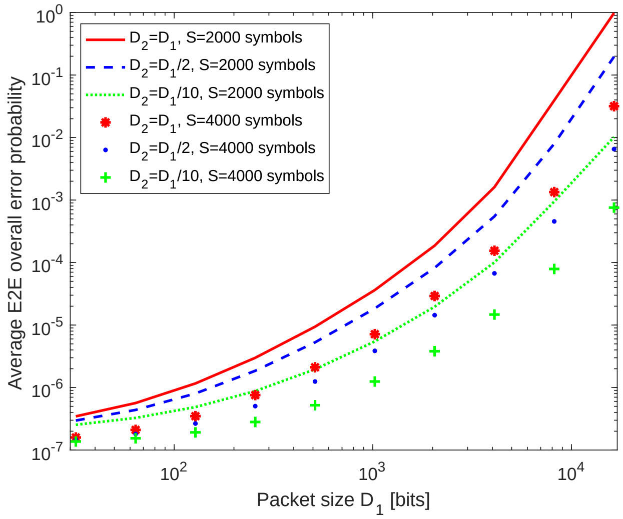

Fig. 3. The impact of packet sizes on the reliability. In the simulation, we set computing cost to $n = 100$ symbols.

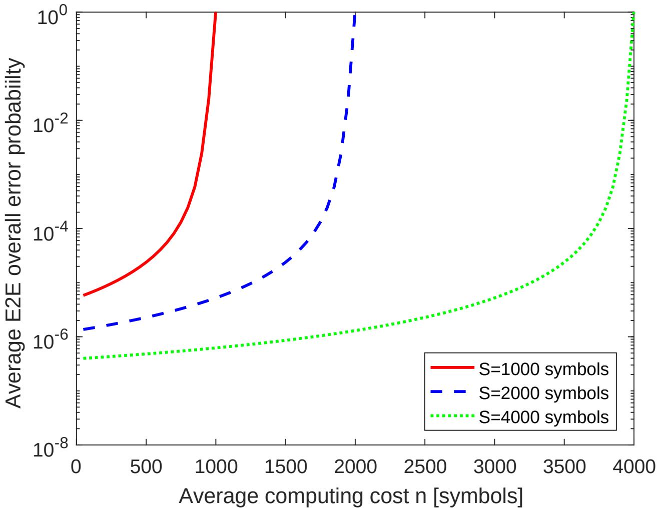

Fig. 4. The impact of average computing cost on the reliability. In the simulation, we set SNRs to 20 dB.