Asymptotic Delay–Outage Analysis for Noise-Limited Wireless Networks with Caching, Computing, and Communications

Ming-Chun Lee, *Member, IEEE*, and Andreas F. Molisch, *Fellow, IEEE*

***Abstract*—Performance assessment and optimization for networks jointly performing caching, computing, and communication (3C) has recently drawn significant attention because many emerging applications require 3C functionality. However, studies in the literature mostly focus on the particular algorithms and setups of such networks, while the theoretical understanding and characterization of such networks has been less explored. To fill this gap, this paper conducts the asymptotic (scaling-law) analysis for the delay-outage tradeoff of noise-limited wireless edge networks with joint 3C. In particular, we derive closed-form expressions for the optimum outage probability as function of delay and other network parameters via first obtaining the outage probability expression and then deriving the optimal caching policy. We provide insights and interpretations based on the derived expressions. Computer simulations validate our analytical results and insights.**

I. Introduction

In the past years, numerous new mobile applications, e.g., ultra-high definition video services, augmented reality (AR), and virtual reality (VR), have emerged and led to a strong growth in the number of mobile devices. These new applications are expected to create an unprecedented increase of wireless traffic whose requirements are highly diverse, ranging from ultra-low latency to ultra-high data rate. To satisfy the resulting demands on wireless networks, new network architectures and novel solution technologies are needed [1].

However, these new applications not only demand for better data transmission rates but also for improved computing and caching capabilities to reduce the latency. To satisfy these requirements, mobile edge-computing and edge-caching have been deemed as two of the most promising technologies [2], [3], where edge-computing improves the performance by providing computational powers at the wireless edge without resorting to the cloud servers and edge-caching improves the network performance by exploiting the storages at the wireless edge, which brings the desired contents closer to users.

Noticing the benefits of edge-caching and edge-computing, numerous papers have been published that investigate one of those approaches. More recently, it became obvious that edge-caching and edge-computing need to be jointly considered as more and more applications require execution of computations whose input are large amounts of data. For example, in video services, video contents cached in the storages should be transcoded and delivered to the users for better user experiences. Another example is that when a user wants to use the machine learning-aided face recognition, this user needs to deliver the face image from the mobile device to an edge server for conducting the computational task via a series of well-trained neural networks (NNs). However, parameters of these NNs need to be stored somewhere at the wireless edge such that the edge server can fetch the parameters of NNs with low latency. The above examples clearly demonstrate that the network performance is determined jointly by how the caching, computing, and communication policies are designed and they are mutually coupled with one another. Since the joint use of caching, computing, and communication (3C) is getting more common for emerging applications, the joint 3C design has recently drawn significant attention [4]–[11].

However, to the best of our knowledge, studies in the literature, e.g., [4]–[11], were focusing on the practical design and implementation aspects of wireless edge networks with joint 3C. These investigations are indeed very important, but commonly lead to either very complicated solutions without closed-form expressions or even pure numerical solutions that could not be easily interpreted for obtaining insights. Furthermore, although there exist some papers investigating the theoretical aspects of either edge-caching or edge-computing, e.g., [12]–[17], it is non-trivial to extend their results to wireless edge networks jointly considering 3C. Based on the above observations, we aim to fill this gap by providing a theoretical analysis for wireless edge networks with joint 3C. Specifically, we aim to develop an asymptotic analysis that shows the basic dependence of performance on available link-rate, cache size, and computation resources. To the best of our knowledge, this paper is the first attempt to contribute to such an analysis of the wireless edge network with joint 3C.

In this paper, we consider a noise-limited wireless network, where the BSs are equipped with both computing units and storage for data and/or programs. We assume that to complete the tasks requested by users, the collaboration of caching, computing, and communications is required. As a result, given a latency requirement for completing the tasks, the network could fail to satisfy the requests of users when any part of

M.-C. Lee is with Institute of Communications, National Yang Ming Chiao Tung University and National Chiao Tung University, Hsinchu 30010, Taiwan. (email: mingchunlee@nycu.edu.tw)

A. F. Molisch is with Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA 90089, USA (email: molisch@usc.edu).

This work was supported in part by the National Science Foundation (NSF) under grant CNS-1816699, and by the Ministry of Science and Technology (MOST) of Taiwan under grants MOST 110-2222-E-A49-006 and MOST 110-2224-E-A49-001.

the caching, computing, and communication is insufficient, leading to the occurrences of the outage. We then analyze the outage probability as a function of the latency requirement and the 3C network parameters. Specifically, we first derive the expression for the outage probability. Then, based on the derived expression, we obtain an analytical expression for the optimal caching policy that can minimize the outage probability and, based on this, the optimal outage probability. We then provide insights and interpretations using the derived outage probability expression. Furthermore, with some reformulation, the minimum achievable latency can be expressed as the function of the outage probability and the 3C network parameters, leading to the fundamental interpretation of the overall latency as the sum of effective transmission delay and computing delay. Finally, computer simulations validate our analysis and insights.

II. Network Model

In this paper, we consider using BSs to serve users. We assume no data communication is possible between BSs and no cloud server is available for the BSs. We then consider that the caching and computing are implemented at the BSs only and users cannot provide caching and computing resources. We assume users in the network have tasks that require the collaborations of caching, computing, and communications and assume that to complete a task requested by a user, the following steps are required: (i) input data upload from the user to the BS; (ii) auxiliary dataset retrieval from the storage of the BS; (iii) computation for converting the data to the necessary content for completing the task; and (iv) final content delivery to the user. Such task process model is general and can be applied to many practical applications, e.g., AR/VR and face recognition. We assume there are M tasks to request, and thus the library has M different auxiliary datasets corresponding to the tasks. We assume different datasets have the same size for simplicity and assume different datasets are used for completing different tasks. Thus, a user requesting task m needs to be associated with the BS having dataset m in its storage. We assume a BS can cache S datasets. We adopt the Poisson point process (PPP) for users and BSs, where the density of the BSs is λ. We assume a noise-limited network, where each user can obtain a fixed amount of communication and computational resource from the connected BS and the interference between users and BSs can be ignored.

We assume the received signal power between a typical user located at origin $(0,0)$ and a BS located at $\mathbf{x} = (x_1, x_2)$ is given by $P|h_{\mathbf{x}}|^2\|\mathbf{x}\|^{-\alpha}$, where P is the transmit power; $h_{\mathbf{x}}$ is the small-scale fading coefficient; and α is the pathloss factor. We denote Φ as the set of BSs in the network, and denote Φ_m as the set of BSs that cache dataset m. We assume the association follows the largest received power principle in which the typical user is associated with the BS which has the largest received power among the BSs that cache the required dataset m. Therefore, the received power when requesting task m at the associated BS is:

$$\max_{\mathbf{x}\in\Phi_m} P|h_{\mathbf{x}}|^2\|\mathbf{x}\|^{-\alpha}. \tag{1}$$

We consider the randomized caching policy [13], where $p_c(m)$ is the probability for a BS to cache dataset m and $\sum_{m=1}^{M} p_c(m) = S$. As a result, the density of Φ_m is $\lambda p_c(m)$. We assume the channel is invariant in a timeslot with duration D. We assume the Nakagami-m_{D} small-scale fading, where $m_{\mathrm{D}} \geq \frac{1}{2}$ and $m_{\mathrm{D}} = 1$ corresponds to the Rayleigh fading. By the analysis in [14], when the required rate for conducting the transmission between the associated BS in Φ_m and the typical user is ρ_m, we can obtain the successful transmission probability as:

$$\mathbb{P}\left[R_m \geq \rho_m\right] = 1 - \exp\left(-\kappa p_c(m)\left(\frac{\eta}{2^{\rho_m}-1}\right)^{\delta}\right), \tag{2}$$

where R_m is the link-capacity (spectral efficiency), $\kappa = \pi\lambda\frac{\Gamma(\delta+m_{\mathrm{D}})}{m_{\mathrm{D}}^{\delta}\Gamma(m_{\mathrm{D}})}$, $\delta = \frac{2}{\alpha}$, $\eta = \frac{P}{\sigma_n^2}$, and σ_n^2 is the noise power. Now, we assume that the required latency for completing the task is D; thus the channel is invariant during the implementation of the task. Suppose that the number of bits to upload for a task is F^{U}; the number of bits to download for a task is F^{D}; and the number of cycles to compute for a task is $\nu^{\mathrm{U}}F^{\mathrm{U}} + \nu^{\mathrm{D}}F^{\mathrm{D}}$, where ν^{U} and ν^{D} are the computational scaling parameters. Then, the probability to successfully complete task m within a latency requirement D is given as:

$$\begin{aligned}\mathbb{P}\left[d_m \leq D\right] &= \mathbb{P}\left[\frac{F^{\mathrm{U}}}{BR_m} + \frac{F^{\mathrm{D}}}{BR_m} + \frac{\nu^{\mathrm{U}}F^{\mathrm{U}} + \nu^{\mathrm{D}}F^{\mathrm{D}}}{E_{\mathrm{c}}} \leq D\right] \\ &= \mathbb{P}\left[R_m \geq \frac{1}{B}\frac{F^{\mathrm{U}} + F^{\mathrm{D}}}{D - \frac{\nu^{\mathrm{U}}F^{\mathrm{U}} + \nu^{\mathrm{D}}F^{\mathrm{D}}}{E_{\mathrm{c}}}}\right],\end{aligned} \tag{3}$$

where d_m is the latency for completing task m; B and E_{c} are the bandwidth and computing power allocated to a user, respectively. We assume $D - \frac{\nu^{\mathrm{U}}F^{\mathrm{U}} + \nu^{\mathrm{D}}F^{\mathrm{D}}}{E_{\mathrm{c}}} > 0$ for simplicity; otherwise, the task can never be successfully completed. It follows from (2) and (3) that the probability of successfully completing task m is:

$$\begin{aligned}&\mathbb{P}\left[d_m \leq D\right] \\ &= 1 - \exp\left(-\kappa p_c(m)\left(\frac{\eta}{2^{\left(\frac{1}{B}\frac{F^{\mathrm{U}}+F^{\mathrm{D}}}{D-\frac{\nu^{\mathrm{U}}F^{\mathrm{U}}+\nu^{\mathrm{D}}F^{\mathrm{D}}}{E_{\mathrm{c}}}}\right)}-1}\right)^{\delta}\right).\end{aligned} \tag{4}$$

We denote the probability for the typical user to request task m as $P_r(m)$ and assume that the requesting distribution is modeled by a Zipf distribution given as

$$P_r(m;\gamma) = \frac{(m)^{-\gamma}}{\sum_{f=1}^{M}(f)^{-\gamma}} = \frac{m^{-\gamma}}{H(1,M,\gamma)}, \tag{5}$$

where γ is the Zipf factor and $H(a,b,\gamma) := \sum_{f=a}^{b}(f)^{-\gamma}$. We focus on the case that $\gamma < 1$ in this paper, and the case that $\gamma > 1$ will be considered in our journal version [18].

$$P_s = \sum_{m=1}^{M} P_r(m)\mathbb{P}\left[d_m \le D\right] = 1 - \left[\sum_{m=1}^{M} P_r(m)\exp\left(-\kappa p_c(m)\left(\frac{\eta}{2^{\left(\frac{1}{B}\frac{F^{\mathrm{U}}+F^{\mathrm{D}}}{D-\frac{\nu^{\mathrm{U}}F^{\mathrm{U}}+\nu^{\mathrm{D}}F^{\mathrm{D}}}{\bar{E}_{\mathrm{c}}}}\right)}-1}\right)^{\delta}\right)\right]. \quad (6)$$

By using (4) and (5), we obtain the successful probability for completing a task given as in (6) on the top of next page. Hence, the outage probability is:

$$P_o = 1 - P_s = \sum_{m=1}^{M} P_r(m)\exp\left(-\kappa p_c(m)\left(\frac{\eta}{2^{\left(\frac{1}{B}\frac{F^{\mathrm{U}}+F^{\mathrm{D}}}{D-\frac{\nu^{\mathrm{U}}F^{\mathrm{U}}+\nu^{\mathrm{D}}F^{\mathrm{D}}}{\bar{E}_{\mathrm{c}}}}\right)}-1}\right)^{\delta}\right). \quad (7)$$

By letting $M \to \infty$ and $S \to \infty$ (i.e., the library and the cache size of the BSs go to infinity), we then use (7) to conduct our asymptotic analysis in the next section.[1]

III. Delay–Outage Analysis

In this section, we start the analysis with finding the optimal caching policy that minimizes the outage probability. Then, the optimal delay-outage tradeoff would be obtained. Finally, insights derived from the analysis are provided.

A. Main Results

To simplify the notation, we define

$$\kappa' = \kappa\left(\frac{\eta}{2^{\left(\frac{1}{B}\frac{F^{\mathrm{U}}+F^{\mathrm{D}}}{D-\frac{\nu^{\mathrm{U}}F^{\mathrm{U}}+\nu^{\mathrm{D}}F^{\mathrm{D}}}{\bar{E}_{\mathrm{c}}}}\right)}-1}\right)^{\delta}. \quad (8)$$

It then follows that we can express the outage probability as:

$$P_o = \sum_{m=1}^{M} P_r(m)\exp\left(-\kappa' p_c(m)\right). \quad (9)$$

Using (9), we first provide proposition 1 which provides the general expression for the optimal caching policy:

Proposition 1: The optimal caching policy that minimizes the outage probability P_o is given as:

$$P_c^*(m) = \min\left(1, \left[\left(\log\frac{\kappa' P_r(m)}{\zeta}\right)^{\frac{1}{\kappa'}}\right]^{+}\right), \quad (10)$$

where $P_c(m)^*$ is caching probability for dataset m, ζ is the Lagrangian multiplier such that $\sum_{m=1}^{M} P_c^*(m) = S$, and $[a]^+ = \max(a, 0)$.

Proof. See Appendix A of [19]. □

We then denote $m_1^* \ge 0$ as the smallest index such that $P_c^*(m_1^*+1) < 1$ and m_2^* as the smallest index such that $P_c^*(m_2^*+1) = 0$. It follows that according to the regimes of m_1^* and m_2^*, we need to split the discussion into three regimes: (i) $0 \le m_1^* < m_2^* < M$; (ii) $m_1^* \le 0 < m_2^* \le M$; and (iii) $0 < m_1^* < M \le m_2^*$. In the following, we present the theorems that respectively characterize the optimum policies of the above regimes:

Theorem 1: Let $M \to \infty$ and $S \to \infty$. Denote $m_1^* \ge 0$ as the smallest index such that $P_c^*(m_1^*+1) < 1$ and m_2^* as the smallest index such that $P_c^*(m_2^*+1) = 0$. Assume $m_2^* < M$ is a large number as $M \to \infty$. The caching distribution $P_c^*(\cdot)$ that minimizes the outage probability p_o is as follows:

$$\begin{aligned} P_c^*(f) &= 1, && f = 1, ..., m_1^* \\ P_c^*(f) &= \log\left(\frac{z_f}{\nu}\right), && f = m_1^*+1, ..., m_2^* \\ P_c^*(f) &= 0, && f = m_2^*+1, ..., M \end{aligned} \quad (11)$$

where $m_1^* + \sum_{f=m_1^*+1}^{m_2^*} \log\left(\frac{z_f}{\nu}\right) = S$, $z_f = (P_r(f))^{\frac{1}{\kappa'}}$, and

$$m_1^* = c_1 S; \quad m_2^* = c_2 S, \quad (12)$$

where $c_1 = \frac{1}{\frac{\gamma}{\kappa'}\left(e^{\frac{\kappa'}{\gamma}}-1\right)}$ and $c_2 = \frac{e^{\frac{\kappa'}{\gamma}}}{\frac{\gamma}{\kappa'}\left(e^{\frac{\kappa'}{\gamma}}-1\right)}$.

Proof. See Appendix B of [19]. □

Theorem 2: Let $M \to \infty$ and $S \to \infty$. Suppose $P_c^*(f) < 1, \forall f$ is satisfied. Then, we denote m^* as the smallest index such that $P_c^*(m^*+1) = 0$. The caching distribution $P_c^*(\cdot)$ that minimizes the outage probability p_o is as follows:

$$P_c^*(f) = \left[\log\left(\frac{z_f}{\nu}\right)\right]^{+}, \quad f = 1, ..., M, \quad (13)$$

where $\sum_{f=1}^{m^*} \log\left(\frac{z_f}{\nu}\right) = S$, $z_f = (P_r(f))^{\frac{1}{\kappa'}}$, and

$$m^* = \min\left(\frac{S\kappa'}{\gamma}, M\right). \quad (14)$$

Proof. See Appendix C of [19]. □

Theorem 3: Let $M \to \infty$ and $S \to \infty$. Denote $m_1^* > 0$ as the index such that $P_c^*(m_1^*+1) < 1$ and assume $P_c^*(M) > 0$. Let $C_2 = \frac{S}{M}$ and let $0 < C_1 \le 1$ be the solution of the following equality: $C_1 - \log(C_1) = \frac{\kappa'}{\gamma}(1 - C_2) + 1$. Then, the caching distribution $P_c^*(\cdot)$ that minimizes the outage probability p_o is as follows:

$$\begin{aligned} P_c^*(f) &= 1, && f = 1, ..., m_1^* \\ P_c^*(f) &= \log\left(\frac{z_f}{\nu}\right), && f = m_1^*+1, ..., M \end{aligned} \quad (15)$$

where $m_1^* + \sum_{f=m_1^*+1}^{M} \log\left(\frac{z_f}{\nu}\right) = S$, $z_f = (P_r(f))^{\frac{1}{\kappa'}}$, and

$$m_1^* = C_1 M. \quad (16)$$

[1] Note that we cannot let the density λ go to infinity for the analysis because this would break the basic assumption in stochastic geometry that $\mathbb{E}_{\mathbf{x}}[|h_{\mathbf{x}}|^2\|\mathbf{x}\|^{-\alpha}] < \infty$.

$$P_o^* = 1 - \left[(c_2)^{1-\gamma} - (c_1)^{1-\gamma}e^{-\kappa'} - (1-\gamma)e^{\gamma}(c_2-c_1)(c_2)^{-\gamma}\left(\frac{c_2}{c_1}\right)^{\frac{-\gamma c_1}{c_2-c_1}} e^{\frac{-(1-c_1)\kappa'}{c_2-c_1}}\right]\left(\frac{S}{M}\right)^{1-\gamma} = 1 - \Theta\left(\left(\frac{S}{M}\right)^{1-\gamma}\right), \tag{17}$$

Proof. See Appendix D of [19]. □

Theorems 1, 2, and 3 analytically describe the optimal caching policies for different regimes.[2] Based on them, we can have the following theorems, namely, Theorems 4, 5, and 6, which characterize the outage probability as a function of the delay requirement and network parameters:

Theorem 4: Let $M \to \infty$ and $S \to \infty$. Consider $\gamma < 1$. Suppose the caching policy is given by Theorem 1. Then, the optimal (minimum) achievable outage probability can be expressed as in (17) on the top of this page, where

$$c_1 = \frac{1}{\frac{\gamma}{\kappa'}\left(e^{\frac{\kappa'}{\gamma}} - 1\right)}; \quad c_2 = \frac{e^{\frac{\kappa'}{\gamma}}}{\frac{\gamma}{\kappa'}\left(e^{\frac{\kappa'}{\gamma}} - 1\right)}. \tag{18}$$

Proof. See Appendix E of [19]. □

Theorem 5: Let $M \to \infty$ and $S \to \infty$. Consider $\gamma < 1$. Suppose the caching policy is given by Theorem 2. Then, the optimal (minimum) achievable outage probability is:

$$P_o^* = (1-\gamma)e^{\gamma}e^{\frac{-S\kappa'}{M}}. \tag{19}$$

Proof. See Appendix F of [19]. □

Theorem 6: Let $M \to \infty$ and $S \to \infty$. Consider $\gamma < 1$. Suppose the caching policy is given by Theorem 3. Then, the optimal (minimum) achievable outage probability is:

$$P_o^* = \left[(1-\gamma)e^{\gamma}(1-C_1)(C_1)^{\frac{\gamma C_1}{1-C_1}}\right] e^{\frac{-\kappa'(C_2-C_1)}{1-C_1}} + e^{-\kappa'}(C_1)^{1-\gamma}, \tag{20}$$

where C_1 and C_2 are given according to Theorem 3.

Proof. See Appendix G of [19]. □

Since the expression in Theorem 6 might not provide clear insight, we conduct additional approximation to derive the more insightful expression for the outage probability characterized by Theorem 6, leading to Corollary 6.1:

Corollary 6.1: Let $M \to \infty$ and $S \to \infty$. Consider $\gamma < 1$. Suppose the caching policy is given by Theorem 3. Assume C_2 is small. Then, the optimal (minimum) achievable outage probability in Theorem 6 can be approximated as:

$$P_o^* \approx (1-\gamma)e^{\gamma}e^{\frac{-S\kappa'}{M}} + (C_1)^{1-\gamma}e^{-\kappa'}. \tag{21}$$

Furthermore, when κ' is sufficiently large so that $e^{-\kappa'}$ is small, (21) can be approximated as:

$$P_o^* \approx (1-\gamma)e^{\gamma}e^{\frac{-S\kappa'}{M}}. \tag{22}$$

Proof. See Appendix H of [19]. □

[2]We note that the provided theorems slightly abuse the notations as m^*, m_1^*, and m_2^* characterized by them might not be integers.

B. Interpretations and Insights

With results in Sec. III.A, we can obtain fundamental interpretations and insights for the delay-outage performance of the network. First of all, from (9), we see that when all datasets can be cached in a BS, i.e., $S = M$, the outage probability is $e^{-\kappa'}$, serving as the fundamental lower bound for the outage probability when given the network configuration. Then, by using Theorem 4, we see that when the outage probability is high, the reduction of the outage probability follows a power law with respect to the cache size S. By Corollary 6.1, we see that the optimal outage probability of regimes characterized by Theorems 5 and 6 is approximately the same when κ' is large, namely, when the fundamental outage probability lower bound $e^{-\kappa'}$ is small. We then see that when considering the regimes characterized by them, the outage probability decreases exponentially with respect to S. Since we are more interested in Theorems 5 and 6 as regimes characterized by them are regimes that give small outage probability, we further analyze their results in the following.

From (19) and (22), we see that the outage probability decreases exponentially with respect to the critical parameter

$$\kappa' = \kappa\left(\frac{\eta}{2^{\left(\frac{1}{B}\frac{F^{\mathrm{U}}+F^{\mathrm{D}}}{D-\frac{\nu^{\mathrm{U}}F^{\mathrm{U}}+\nu^{\mathrm{D}}F^{\mathrm{D}}}{E_{\mathrm{c}}}}\right)}-1}\right)^{\delta}, \tag{23}$$

which is related to the delay requirement, communication and computing capabilities, and the BS density. Note that $\kappa = \pi\lambda\frac{\Gamma(\delta+m_{\mathrm{D}})}{m_{\mathrm{D}}^{\delta}\Gamma(m_{\mathrm{D}})}$. By using (23), we can see the relations between different parameters. In addition, by further including the caching, we see that the critical parameter is

$$\frac{S\kappa'}{M} = \kappa\left(\frac{\eta}{2^{\left(\frac{1}{B}\frac{F^{\mathrm{U}}+F^{\mathrm{D}}}{D-\frac{\nu^{\mathrm{U}}F^{\mathrm{U}}+\nu^{\mathrm{D}}F^{\mathrm{D}}}{E_{\mathrm{c}}}}\right)}-1}\right)^{\delta}\frac{S}{M}, \tag{24}$$

and the increase of $\frac{S\kappa'}{M}$ can decrease the outage probability exponentially. This critical parameter thus gives the clear characterization of caching, computing, and communications as well as their relations to the outage probability.

Using the results in Theorem 5 and Corollary 6.1, we can reformulate the optimal outage probability expression such that the minimum achievable latency D^* becomes a function of the outage probability and other parameters. We denote

$$\eta_{\text{eff}} = \frac{\eta(\kappa S)^{\frac{1}{\delta}}}{\left(M\log\left(\frac{(1-\gamma)e^{\gamma}}{P_o^*}\right)\right)^{\frac{1}{\delta}}} = \frac{\eta(\kappa S)^{\frac{\alpha}{2}}}{\left(M\log\left(\frac{(1-\gamma)e^{\gamma}}{P_o^*}\right)\right)^{\frac{\alpha}{2}}} \tag{25}$$

as the effective SNR. It follows that we can express the minimum achievable latency as:

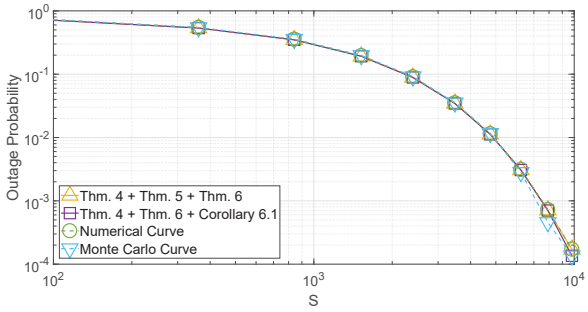
$$D^* = \frac{F^{\mathrm{U}} + F^{\mathrm{D}}}{B \log_2 (1 + \eta_{\mathrm{eff}})} + \frac{\nu^{\mathrm{U}} F^{\mathrm{U}} + \nu^{\mathrm{D}} F^{\mathrm{D}}}{E_{\mathrm{c}}}. \quad (26)$$

From (26), we observe that the minimum achievable latency D^* is the sum of two terms, where the first term represents the delay due to transmissions and the second term represents the delay due to computations. We see that the caching capability affects the first term through the effective SNR η_{eff}. Then, by (25), we observe that the effective SNR is proportional to $\left(\frac{S}{M}\right)^{\frac{\alpha}{2}}$, indicating that the caching is more influential when the pathloss factor α is larger. However, this should not be mis-interpreted as that the minimum latency would be smaller when the pathloss factor α is larger. On the contrary, since η_{eff} is inversely proportional to $\left(\log\left(\frac{(1-\gamma)e^{\gamma}}{P_o^*}\right)\right)^{\frac{\alpha}{2}}$, when having the same required outage probability P_o^*, η_{eff} would be smaller when the pathloss factor α is larger. We see that since the minimum achievable latency is the sum of two terms, it is clear that we need to improve caching, computing, and communications in a balanced manner when improving the network. In other words, when the transmission delay is dominant, we need to think of improving the caching and/or communications. On the other hand, if the computational delay is dominant, we should resort to improving the computation capability for efficient performance improvement. Note that although the above statements are intuitive, our results indeed rigorously validate the intuitions from the theoretical aspect.

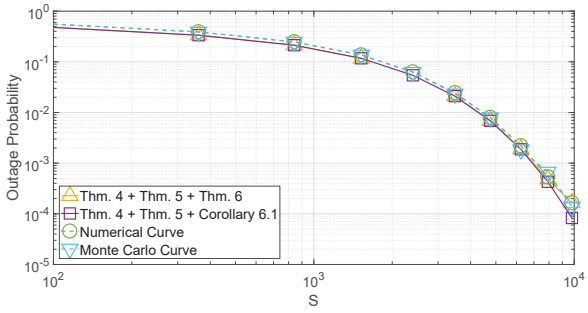
IV. Computer Simulations

In this section, we validate our analysis using simulations. Unless otherwise indicated, we consider $D = 10^{-3}$ sec, $m_{\mathrm{D}} = 1$, $P = 20$ dBm, $\alpha = 3.5$, $B = 20$ MHz, the noise power spectral density $N_0 = -173$ dBm/Hz, $F^{\mathrm{U}} = 10^4$ bis, $F^{\mathrm{D}} = 10^5$ bits, $\nu^{\mathrm{U}} = \nu^{\mathrm{D}} = 1$ cycle/bit, $E_{\mathrm{c}} = 10^9$ cycles/sec, $M = 10^4$, and $\lambda = 5$ per km^2.

In Fig. 1, we validate our theorems proposed in Sec. III.A as a function of the caching capability S, where the numerical curves are results obtained by first numerically solving the outage probability minimization problem and then evaluating the outage probability using (9); the Monte-Carlo curves are results obtained by directly conducting Monte-Carlo simulations of the considered network with the numerically optimized caching policy. Note that since different theorems in Sec. III.A characterize different regimes, the theoretical curves are the combinations of different theorems. Furthermore, the curves with "+ Corollary 6.1" are obtained via replacing Theorem 6 with Corollary 6.1. The results show that our proposed analysis is very accurate, except for the endpoint of the curve with "+ Corollary 6.1" in Fig. 1(b), where the small divergence comes from that the conditions for the good approximation of Corollary 6.1 might not be well-satisfied. However, we note that the accurate results in Fig. 1 even for the constant factors are already not common for the asymptotic analysis of wireless networks. From Fig. 1, we also see that the outage probability is generally exponentially decreasing with respect to S. Finally, we note that the outage probability in Fig. 1 indeed is lower bounded by $e^{-\kappa'}$, which is the fundamental bound due to the network configuration.



(a) $\gamma = 0.6$.



(b) $\gamma = 0.8$.

Fig. 1: Outage probability evaluation of the proposed theorems as a function of S.

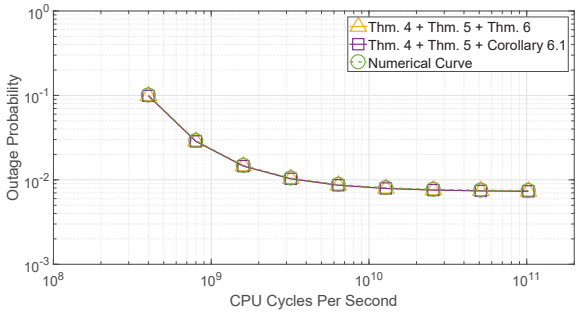


Fig. 2: Outage probability evaluation of the proposed theorems as a function of E_{c}.

In Fig. 2, we consider $S = 4000$ and $\gamma = 0.6$ and evaluate the outage probability as a function of the computing power E_{c}. Since we already see that the Monte-Carlo results are identical to the numerical results in Fig. 1, we only provide the numerical validation here. Results again show that our theorems are very accurate. In addition, we observe that the outage probability saturates when the computing power increases to a large number. This is because when the computing power is sufficient, the performance is then limited by the caching and communication capabilities, showing that the approach improving only the computing capability could be limited.

Finally, in Fig. 3, we evaluate the delay-outage performance of the considered network with $S = 7500$ and $\gamma = 0.6$.

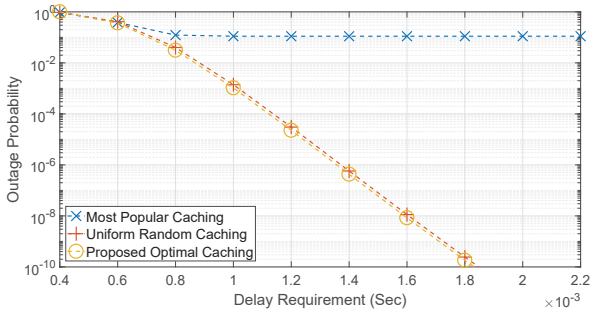


Fig. 3: Delay-Outage performance evaluation.

In addition to evaluating the network adopting our proposed caching policy, we also evaluate two reference caching policies, namely, the most-popular caching and uniform random caching, where the most-popular caching let BSs only cache datasets relevant to the most popular tasks until the storage is full; the uniform random caching on the other hand let BSs cache datasets uniformly at random. Results show that the proposed optimal caching can provide the best performance. Besides, we observe that the most-popular caching performs poorly because BSs with this caching policy do not collaboratively cache the datasets so that there are numerous tasks that do not have their datasets cached in the network and thus can never be completed even if the delay requirement can be relaxed. On the other hand, the uniform random caching is indeed near-optimal as its outage probability is $e^{-\frac{S}{M}\kappa'}$, which only differs from the optimal outage probability in the constant factor (see (19) and (22) to compare). However, we note that the good performance of the uniform random caching only exists when $\gamma < 1$, i.e., the requests are not very concentrated on the popular tasks. In our journal version [18], we will show that when $\gamma > 1$, i.e., the requests are more concentrated on popular tasks, the uniform random caching could perform poorly. Finally, we see that changing the delay requirement significantly affects the outage probability, showing that by slightly relaxing the latency requirement, the outage probability performance can be significantly improved. Thus, the challenges of the wireless network indeed are imposed by the time-sensitive applications, which matches our intuition well.

V. Conclusions

In this paper, we provide the asymptotic delay-outage analysis for obtaining the fundamental understanding of the behaviors of wireless edge networks considering collaborative 3C. Our analysis clearly reveals the relations between the delay-outage performance and 3C parameters. Specifically, we see that increasing S and κ' can bring an exponential-law improvement to the outage probability. In addition, we see that the delay is composed of the effective transmission and computing delays, where improving only either of them can lead to the situation that the delay is dominated by the other. Hence, to efficiently improve the network, an approach having a balanced view on 3C is necessary. Finally, we observe that slightly relaxing the delay requirement can significantly improve the outage probability. This implies that the challenges of the wireless network indeed are imposed by the time-sensitive applications.

References

[1] H. Tataria, M. Shafi, A. F. Molisch, M. Dohler, H. Sjöland, and F. Tufvesson, "6G wireless systems: Vision, requirements, challenges, insights, and opportunities," *Proceedings of the IEEE*, vol. 109, no. 7, pp. 1166–1199, July 2021.

[2] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2322–2358, Fourthquarter 2017.

[3] L. Li, G. Zhao, and R. S. Blum, "A survey of caching techniques in cellular networks: Research issues and challenges in content placement and delivery strategies," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 1710–1732, 2018.

[4] C. Wang, Y. He, F. R. Yu, Q. Chen, and L. Tang, "Integration of networking, caching, and computing in wireless systems: A survey, some research issues, and challenges," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 1, pp. 7–38, Firstquarter 2018.

[5] S. Wang, X. Zhang, Y. Zhang, L. Wang, J. Yang, and W. Wang, "A survey on mobile edge networks: Convergence of computing, caching and communications," *IEEE Access*, vol. 5, pp. 6757–6779, 2017.

[6] M. Chen, Y. Hao, L. Hu, M. S. Hossain, and A. Ghoneim, "Edge-cocaco: Toward joint optimization of computation, caching, and communication on edge cloud," *IEEE Wireless Commun. Mag.*, vol. 25, no. 3, pp. 21–27, June 2018.

[7] M. Tang, L. Gao, and J. Huang, "Enabling edge cooperation in tactile internet via 3c resource sharing," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 11, pp. 2444–2454, November 2018.

[8] Y. Sun, Z. Chen, M. Tao, and H. Liu, "Communications, caching, and computing for mobile virtual reality: Modeling and tradeoff," *IEEE Trans. Commun.*, vol. 67, no. 11, pp. 7573–7586, November 2019.

[9] A. Ndikumana, N. H. Tran, T. M. Ho, Z. Han, W. Saad, D. Niyato, and C. S. Hong, "Joint communication, computation, caching, and control in big data multi-access edge computing," *IEEE Trans. Mobile Comput.*, vol. 19, no. 6, pp. 1359–1374, June 2019.

[10] W. Wen, Y. Cui, T. Q. Quek, F.-C. Zheng, and S. Jin, "Joint optimal software caching, computation offloading and communications resource allocation for mobile edge computing," *IEEE Trans. Veh. Technol.*, vol. 69, no. 7, pp. 7879–7894, July 2020.

[11] S. Luo, X. Chen, Z. Zhou, and S. Yu, "Fog-enabled joint computation, communication and caching resource sharing for energy-efficient iot data stream processing," *IEEE Trans. Veh. Technol.*, vol. 70, no. 4, pp. 3715–3730, April 2021.

[12] K. Shanmugam, N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, December 2013.

[13] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in *2015 IEEE international conference on communications (ICC)*, 2015, pp. 3358–3363.

[14] S. H. Chae and W. Choi, "Caching placement in stochastic wireless caching helper networks: Channel selection diversity via caching," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 6626–6637, October 2016.

[15] S.-W. Ko, K. Han, and K. Huang, "Wireless networks for mobile edge computing: Spatial modeling and latency analysis," *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 5225–5240, August 2018.

[16] M.-C. Lee, M. Ji, A. F. Molisch, and N. Sastry, "Throughput-outage analysis and evaluation of cache-aided d2d networks with measured popularity distributions," *IEEE Trans. on Wireless Commun.*, vol. 18, no. 11, pp. 5316–5332, November 2019.

[17] Y. Gu, Y. Yao, C. Li, B. Xia, D. Xu, and C. Zhang, "Modeling and analysis of stochastic mobile edge computing wireless networks," *IEEE Internet of Things J.*, vol. 8, no. 18, pp. 14 051–14 065, September 2021.

[18] M.-C. Lee and A. F. Molisch, "Optimal delay-outage analysis for noise-limited wireless networks with caching, computing, and communications," *IEEE Trans. on Wireless Commun.*, (submitted to).

[19] ——, "Optimal delay-outage analysis for noise-limited wireless networks with caching, computing, and communications – Derivations and proofs," online available at `https://arxiv.org/abs/2111.05535`.