Optimal 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 their theoretical understanding and characterization 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, assuming the user requests for different tasks following a Zipf distribution, we derive the analytical expression for the optimal caching policy. Based on this, we next derive the closed-form expression for the optimum outage probability as a function of delay and other network parameters for the case that the Zipf parameter is smaller than 1. Then, for the case that the Zipf parameter is larger than 1, we derive the closed-form expressions for upper and lower bounds of the optimum outage probability. We provide insights and interpretations based on the derived expressions. Computer simulations validate our analytical results and insights.**

***Index Terms*—Edge caching and edge computing, joint caching, computing and communication, delay-outage analysis, scaling laws, noise-limited networks.**

I. Introduction

Numerous new mobile applications have emerged in the past years, e.g., ultra-high definition video services, augmented reality (AR), and virtual reality (VR). This has lead to 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 [3].

However, these new applications not only require high data transmission rates but also fast access to computation and data storage to reduce latency. To satisfy these requirements, mobile edge-computing and edge-caching have been considered as two of the most promising technologies [4], [5]. Edge-computing improves the performance by providing computational power at the wireless edge, eliminating the need to resort to the cloud servers. 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, e.g., [4], [5] and reference herein. In addition, companies and standard development organizations, e.g., European Telecommunications Standards Institute (ETSI), all intensively work toward the standardization of the mobile edge computing and caching implementation platforms [6]. 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 occurs when a user wants to use machine learning-aided facial 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 jointly determined by the caching, computing, and communication (3C) policies. Consequently, the joint 3C design has recently drawn significant attention [7]–[9].

A. Literature Review

Although wireless edge networks with joint 3C have been a popular topic in recent years, to the best of our knowledge, studies in the literature were focusing on the practical design and implementation aspects. For example, in [10], a novel framework for jointly optimizating 3C was proposed. In [11] and [12], joint 3C optimization solutions were discussed using different convexification techniques. To deal with network dynamics, dynamic 3C optimization approaches were investigated in [13] and [14]. Joint 3C designs for specific applications were also investigated. In [15] and [16], the joint 3C designs for tactile networks were discussed. In [17],

M.-C. Lee is with Institute of Communications Engineering, National Yang Ming 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) project RINGS 2148315; in part by the National Science and Technology Council (NSTC) of Taiwan under grants NSTC 110-2222-E-A49-006-, NSTC 111-2221-E-A49-070-, and NSTC 111-2218-E-A49-024-; and in part by The National Defense Science and Technology Academic Collaborative Research Project in 2022.

Part of this work has been presented in the 2022 IEEE International Conference on Communications [1]. A supplement to this paper, which contains detailed proofs of all the lemmas, corollaries, propositions, and theorems, is online available in [2] and is provided as the supplemental material of this paper.

the quality-aware video delivery was optimized via jointly considering 3C. Refs. [18] and [19] investigated the use of machine learning for vehicular edge caching and computing, while [20] and [21] considered the joint 3C designs for IoT networks. In [16] and [22], the specific designs for AR and VR applications were proposed. Ref. [23] developed a joint 3C design framework for federated learning. These investigations are indeed important, but commonly lead to either complicated solutions without closed-form expressions or even purely numerical solutions that could not be easily interpreted for obtaining insights. We note that although the above literature review cites only a sample of papers on the design and implementation for wireless edge networks, the observation also holds true for other papers dealing with the design and implementation aspects.

There exist papers investigating the theoretical aspects of either edge-caching or edge-computing. Regarding the edge-caching, the optimal deterministic caching approach was investigated and analyzed in [24]. In [25], assuming the locations of base stations (BSs) to be random, the optimal randomized caching policy was presented. To understand the performance and find effective designs in heterogeneous caching networks, analysis and design approaches were proposed in [26]–[30]. Taking into account that the delivered video contents can have different qualities, [31] analyzed and optimized wireless caching networks. Considering the case that the BSs are equipped with multiple antennas and that the wireless edge network can have a hierarchical structure, [32] and [33] analyzed and proposed designs for wireless caching networks. Theoretically-optimal wireless D2D caching was comprehensively investigated in [34]–[37].

Theoretical studies for edge-computing were conducted in [38]–[46]. Considering access points equipped with computing servers, [38] studied the communication and computing latency scaling laws as functions of network parameters. Considering the influences of both the remote cloud server and edge cloudlets, [39] analyzed the outage probability in order to obtain the tradeoff between deployment and operation costs. In [40], again considering both edge and cloud servers, the average latency was analyzed via combining the stochastic geometry and queuing theory. To understand the influences of heterogeneous mobile users and tasks, [41] studied the successful edge computing probability and provided design insights. Ref. [42] considered the massive Internet of Things scenario and analyzed the latency for 5G edge-cloud networks. Considering a network with hierarchical computing structure, [43] and [44] analyzed the latency and successful offloading probability, respectively, and provided optimizations. In [45], the computation offloading probability was analyzed assuming that non-orthogonal multiple access (NOMA) is adopted. Assuming that uplink and downlink transmissions can be provided by different BSs and edge servers, [46] analyzed the latency with results showing that such decoupled uplink and downlink structure can improve the performance. We note that although there exist many papers investigating the theoretical aspects of either edge-caching or edge-computing, it is non-trivial to extend their results to wireless edge networks considering joint 3C. This is because edge-caching and edge-computing were analyzed with different frameworks that cannot be easily merged.

B. Contributions

This paper considers the scaling law analysis for noise-limited wireless networks with joint 3C. To the best of our knowledge, there is no previous work to provide a scaling law analysis for wireless edge networks with joint 3C. Note that the scaling law analysis is important because its result can be used to understand the fundamental limits and benefits of the network and to provide guideline for network design [35], [38], [47]. Our results in this paper show the basic dependence of performance on available link-rate, cache size, and computation resources and provides insights based on this analysis.

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, caching, computing, and communications are all 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 the caching, computing, and communication is insufficient, leading to occurrences of 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, and then derive an analytical expression for the optimal caching policy that minimizes the outage probability. Based on them, we then conduct the delay-outage analysis considering Zipf-distributed request probabilities for tasks, with Zipf distribution factor γ fulfilling $\gamma < 1$ or $\gamma > 1$, corresponding to two regimes that have completely different asymptotic behaviors. Since the analysis provides clear characterization for the relationship between the network parameters and the delay and outage probability, we provide insights and interpretations using the analysis results.

Specifically, when $\gamma < 1$, our analysis indicates that the outage probability can decrease exponentially with respect to the cache size and BS density in the regime of most interest. In addition, we show that the minimum achievable latency can be expressed as the sum of computing delay and effective transmission delay, leading to the fundamental interpretation that the overall latency is the combination of computing and transmission delays. Finally, we show that slightly relaxing the delay requirement can significantly improve the outage probability. This thus implies that the challenges of the wireless network indeed are imposed by the time-sensitive applications. In line with intuition, the analysis also shows that the outage probability for the network with $\gamma > 1$ is better than that of the network with $\gamma < 1$. Finally, based on the main analysis in the paper, we provide analyses for some extended networks and reference networks. We also provide computer simulations to validate our analysis and insights.

C. Paper Organization

The remainder of this paper is organized as follows. Sec. II discusses the models, assumptions, and definitions adopted

in this paper. Sec. III provides the analytical results for the optimal caching policy. Sec. IV provides the delay-outage analysis and the corresponding results considering $\gamma < 1$. The analysis and results considering $\gamma > 1$ are presented in Sec. V. Numerical validations of our analysis are provided in Sec. VI. We conclude this paper in Sec. VII. The detailed proofs are relegated to appendices of [2] and the sketch of the proof strategy is provided in Appendix A of this paper.

II. Edge Network Model

In this paper, we consider an infrastructure-based 3C system where BSs serve users. We assume no data communication is possible between BSs and no cloud server is available for the BSs. 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 processing the data to the necessary content for completing the task; and (iv) final content delivery to the user. Such a task process model is fairly general and can be applied to many practical applications, e.g., AR/VR and facial recognition. We assume that there are M tasks to request, and thus the library has M different auxiliary datasets corresponding to the tasks. We assume for simplicity that different datasets have the same size and that different datasets are used for completing different tasks. Thus, a user requesting task f needs to be associated with the BS having dataset f in its storage.

We assume a BS can cache S datasets. We adopt the Poisson point process (PPP) for the locations of users and BSs, where the density of the BSs is λ and the density of users is λ_{u}. 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 between BSs can be ignored. Note that the assumption that each user is offered a fixed amount of bandwidth and computing power could be fulfilled in a system with the resource allocation strategy that always provides the fixed amount of resource for stable service. Also, the assumption for no interference is justified for systems with excellent interference avoidance/mitigation capability. For example, a mmWave system with massive MIMO arrays may fulfill the assumption, as it can avoid interference by suitable beamforming in the highly directional mmWave channels. We note that the analysis that considers interference is an important future direction.

We assume the signal power received by a typical user located at origin $(0,0)$ from 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 average (over the fading) power received at unit distance; $h_{\mathbf{x}}$ is the frequency-flat small-scale fading coefficient such that $|h_{\mathbf{x}}|$ is a unit-variance Nakagami-m_{D} distributed random variable, where $m_{\mathrm{D}} \geq \frac{1}{2}$ and $m_{\mathrm{D}} = 1$ corresponds to the Rayleigh fading; α is the pathloss coefficient. We denote Φ as the set of BSs in the network, and denote Φ_f as the set of BSs that cache dataset f. We assume the association follows the largest received power principle in which the typical user is associated with the BS that has the largest received power among the BSs that cache the required dataset f. Therefore, the received power when requesting task f at the associated BS is:

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

We consider a randomized caching policy [25], where $P_c(f)$ is the probability for a BS to cache dataset f and $\sum_{f=1}^{M} P_c(f) = S$. As a result, the density of Φ_f is $\lambda P_c(f)$. We assume the channel is invariant in a time period with duration D. From the analysis in [27], when the required rate for successfully conducting the transmission between the associated BS in Φ_f and the typical user is ρ_f, we can obtain the probability for successful transmission as:

$$\mathbb{P}\left[R_f \geq \rho_f\right] = 1 - \exp\left(-\kappa P_c(f)\left(\frac{\eta}{2^{\rho_f}-1}\right)^{\delta}\right), \tag{2}$$

where R_f 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, and thus the channel is invariant during the implementation of the task.[1] 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 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 with the unit cycles/bit. Then, the probability to successfully complete task f within a latency requirement D is given as:

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

where d_f is the latency for completing task f; 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 and it would be meaningless to look at the outage probability of a task given that we already know that the task can never be completed. It follows from (2) and (3) that the probability of successfully completing task f is:

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

We denote the probability for the typical user to request task f as $P_r(f)$ and assume that the requesting (popularity)

[1] Note that this assumption implies that the latency requirement for the transmission is shorter than the coherence time, in which case the channel is static for a packet. Comparing typical channel coherence times (tens of ms) to latency requirements for some applications, e.g., URLLC requirements, this is indeed fulfilled in at least some practical cases.

distribution is modeled by a Zipf distribution given as [11], [24]:

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

where γ is the Zipf factor and $H(a,b,\gamma) := \sum_{m=a}^{b}(m)^{-\gamma}$. By using (4) and (5), we obtain the successful probability for completing a task as:

$$P_s = \sum_{f=1}^{M} P_r(f)\mathbb{P}\left[d_f \le D\right] = 1 - \sum_{f=1}^{M} P_r(f)\exp\left(-\kappa P_c(f)\left(\frac{\eta}{2^{\left(\frac{1}{B}\frac{F^{\mathrm{U}}+F^{\mathrm{D}}}{D-\frac{\nu^{\mathrm{U}}F^{\mathrm{U}}+\nu^{\mathrm{D}}F^{\mathrm{D}}}{E_{\mathrm{c}}}}\right)}-1}\right)^{\delta}\right). \tag{6}$$

Hence, the outage probability is:

$$P_o = 1 - P_s = \sum_{f=1}^{M} P_r(f)\exp\left(-\kappa P_c(f)\left(\frac{\eta}{2^{\left(\frac{1}{B}\frac{F^{\mathrm{U}}+F^{\mathrm{D}}}{D-\frac{\nu^{\mathrm{U}}F^{\mathrm{U}}+\nu^{\mathrm{D}}F^{\mathrm{D}}}{E_{\mathrm{c}}}}\right)}-1}\right)^{\delta}\right). \tag{7}$$

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 several sections.[2] As holds generally true in scaling law analysis, the assumptions $M \to \infty$ and $S \to \infty$ are made for the analysis convenience and mathematical tractability, and they are helpful for getting rid of the minor numerical details so that the critical relation between parameters can be revealed. In addition, the analytical results with these assumptions can be regarded as the approximations of the results in finite regime with large M and S, and such approximations become more accurate as M and S become larger. Note that our simulation results in Sec. VI will show good accuracy of our analytical results when considering the practical finite regime of M and S [35], [48]. The frequently used notations in the paper is summarized in Table I on the top of next page.

III. Optimal Caching Policy

In this section, we derive the analytical expression of the optimal caching policy that will be used for the delay-outage performance analysis.[3] To simplify the notation, we define

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

[2]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 $\mathbb{E}_{\mathbf{x}}[|h_{\mathbf{x}}|^2\|\mathbf{x}\|^{-\alpha}] < \infty$.

[3]We note that due to page limitation, we only provide a sketch of the proof strategy in Appendix A of this paper; the detailed proofs are relegated to appendices of [2] which is obtainable also also as supplemental material of this paper.

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

$$P_o = \sum_{f=1}^{M} P_r(f)\exp\left(-\kappa P_c(f)\left(\frac{\eta}{2^{\left(\frac{1}{B}\frac{F^{\mathrm{U}}+F^{\mathrm{D}}}{D-\frac{\nu^{\mathrm{U}}F^{\mathrm{U}}+\nu^{\mathrm{D}}F^{\mathrm{D}}}{E_{\mathrm{c}}}}\right)}-1}\right)^{\delta}\right) = \sum_{f=1}^{M} P_r(f)\exp\left(-\kappa' P_c(f)\right). \tag{9}$$

Using (9), we then derive Proposition 1 which describes the optimal caching policy:

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

$$\begin{aligned} P_c^*(f) &= \min\left(1, \left[\frac{-1}{\kappa'}\log\frac{\zeta}{\kappa' P_r(f)}\right]^+\right) \\ &= \min\left(1, \left[\log\left(\frac{\kappa' P_r(f)}{\zeta}\right)^{\frac{1}{\kappa'}}\right]^+\right), \end{aligned} \tag{10}$$

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

Proof. See Appendix A of [2]. □

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 conditions 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) $0 < m_1^* < M \le m_2^*$. Before providing the theorems, we first note that three frequently used lemmas, i.e., Lemmas 1-3, for proving theorems in this paper are provided in Appendix O in [2]. In the following, we present the theorems that respectively characterize the optimal policy 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 $P_c^*(m_2^*+1) = 0$. We assume that m_2^* is much larger than 1 when $S \to \infty$. The caching distribution $P_c^*(\cdot)$ that minimizes the outage probability P_o is as follows:

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

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

$$m_1^* = c_1 S; \quad m_2^* = c_2 S, \tag{12}$$

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 [2]. □

TABLE I: Summary of Frequently Used Notations

Notations	Descriptions
M; S; D; λ	Number of tasks in the library; cache capability; latency requirement; density of BSs
f; $P_c(f)$; $P_r(f)$; P_o	index of task; probability of caching dataset f; probability of requesting task f; outage probability
α; δ; m_{D}; γ	Pathloss coefficient ; $\delta = \frac{2}{\alpha}$; Nakagami fading coefficient; Zipf factor
B; E_c; P; σ_n^2; η	Bandwidth; computing power; transmit power; noise power; $\eta = \frac{P}{\sigma_n^2}$
$F^{\mathrm{D}}/F^{\mathrm{U}}$; $\nu^{\mathrm{D}}/\nu^{\mathrm{U}}$	Number of bits to download/upload; computational scaling parameter for download/upload
κ; κ'	$\kappa = \pi\lambda\frac{\Gamma(\delta+m_{\mathrm{D}})}{m_{\mathrm{D}}^{\delta}\Gamma(m_{\mathrm{D}})}$; parameter defined in (8)

Theorem 2: Let $M \to \infty$ and $S \to \infty$. Suppose

$$P_c^*(f) = \min\left(1, \left[\left(\log\frac{\kappa' P_r(f)}{\zeta}\right)^{\frac{1}{\kappa'}}\right]^+\right) = \left[\left(\log\frac{\kappa' P_r(f)}{\zeta}\right)^{\frac{1}{\kappa'}}\right]^+ \tag{13}$$

is satisfied, i.e., $P_c^*(f) < 1, \forall f$. 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:

$$P_c^*(f) = \left[\log\left(\frac{z_f}{\nu}\right)\right]^+, \quad f = 1, ..., M, \tag{14}$$

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

$$m^* = \min\left(\frac{S\kappa'}{\gamma}, M\right). \tag{15}$$

Proof. See Appendix C of [2]. □

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 $0 < C_1 \leq 1$ 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:

$$\begin{aligned} P_c^*(f) &= 1, && f = 1, ..., m_1^* \\ P_c^*(f) &= \log\left(\frac{z_f}{\nu}\right), && f = m_1^*+1, ..., M \end{aligned} \tag{16}$$

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

$$m_1^* = C_1 M. \tag{17}$$

Proof. See Appendix D of [2]. □

IV. Delay-Outage Analysis for $\gamma < 1$ Scenarios

In this section, considering $\gamma < 1$, we first conduct the delay-outage analysis based on the optimal caching policy derived in Sec. III. Then, based on the analysis results, insights and some extended results are provided.

A. Main Results

Theorems 1, 2, and 3 analytically describe the optimal caching policies for different regimes.[4] Based on them, we can have the following theorems, namely, Theorems 4, 5, and 6, which characterize the outage probability as a function of delay requirements and other critical parameters, e.g., S and M, for regimes corresponding to those of Theorems 1, 2, and 3, respectively. Besides, since the expression derived in Theorem 6 might not provide clear insight, we conduct additional approximations to derive a more insightful expression for the outage probability characterized by Theorem 6, leading to Corollary 6.1.

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 is as expressed in (18) on the top of next page, where

$$c_1 = \frac{1}{\frac{\gamma}{\kappa'}\left(e^{\frac{\kappa'}{\gamma}} - 1\right)}; \quad c_2 = \frac{e^{\frac{\kappa'}{\gamma}}}{\frac{\gamma}{\kappa'}\left(e^{\frac{\kappa'}{\gamma}} - 1\right)}. \tag{19}$$

Proof. See Appendix E of [2]. □

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:

$$P_o^* = (1-\gamma)e^{\gamma}e^{\frac{-S\kappa'}{M}}. \tag{20}$$

Proof. See Appendix F of [2]. □

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:

$$P_o^* = \left[(1-\gamma)e^{\gamma}(1-C_1)(C_1)^{\frac{\gamma C_1}{1-C_1}}\right]e^{\frac{-\kappa'(C_2-C_1)}{1-C_1}} + e^{-\kappa'}(C_1)^{1-\gamma}, \tag{21}$$

where C_1 and C_2 are given according to Theorem 3.

Proof. See Appendix G of [2]. □

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:

$$P_o^* \approx (1-\gamma)e^{\gamma}e^{\frac{-S\kappa'}{M}} + (C_1)^{1-\gamma}e^{-\kappa'}. \tag{22}$$

Furthermore, when κ' is sufficiently large so that the outage probability lower bound $e^{-\kappa'}$ is small, the optimal (minimum) achievable outage probability in Theorem 6 can be approximated as:

$$P_o^* \approx (1-\gamma)e^{\gamma}e^{\frac{-S\kappa'}{M}}. \tag{23}$$

Proof. See Appendix H of [2]. □

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

$$P_o^* = 1 - \left[(c_2)^{1-\gamma} - (c_1)^{1-\gamma} e^{-\kappa'} - (1-\gamma) e^{\gamma} (c_2 - c_1)(c_2)^{-\gamma} \left(\frac{c_2}{c_1} \right)^{\frac{-\gamma c_1}{c_2 - c_1}} e^{\frac{-(1-c_1)\kappa'}{c_2 - c_1}} \right] \left(\frac{S}{M} \right)^{1-\gamma} = 1 - \Theta\left(\left(\frac{S}{M} \right)^{1-\gamma} \right). \tag{18}$$

B. Interpretations and Insights

With the results in Sec. IV-A, we can obtain fundamental insights and interpretations 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 outage probability when given the network configuration. In addition, we also see that increasing κ' can lead to an exponential decrease of the outage probability. By using Theorem 4, we see that when S is small, 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 give small outage probability, we further analyze their results in the following.

From (20) and (23), we see that the outage probability decreases exponentially with respect to the critical parameter

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

which depends on the latency requirement, communication and computing capabilities, and the BS density. By using (24), we can see the relations between different parameters. In addition, by further including the caching, we see that the critical parameter is:

$$\kappa_{\mathrm{T}} = \frac{S\kappa'}{M} = \pi \lambda \frac{\Gamma(\delta + m_D)}{m_D^{\delta} \Gamma(m_D)} \left(\frac{\eta}{2^{\left(\frac{1}{B} \frac{F^{\mathrm{U}} + F^{\mathrm{D}}}{D - \frac{\nu^{\mathrm{U}} F^{\mathrm{U}} + \nu^{\mathrm{D}} F^{\mathrm{D}}}{E_{\mathrm{c}}}} \right)} - 1} \right)^{\delta} \frac{S}{M}, \tag{25}$$

and the increase of κ_{T} can decrease the outage probability exponentially. This critical parameter κ_{T} thus gives a 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 expressions such that the minimum achievable latency D^* becomes a function of the outage probability and other parameters. Specifically, from (20) and (23), we can obtain:

$$2^{\left(\frac{1}{B} \frac{F^{\mathrm{U}} + F^{\mathrm{D}}}{D^* - \frac{\nu^{\mathrm{U}} F^{\mathrm{U}} + \nu^{\mathrm{D}} F^{\mathrm{D}}}{E_{\mathrm{c}}}} \right)} - 1 = \frac{\eta (\kappa S)^{\frac{1}{\delta}}}{\left(M \log\left(\frac{(1-\gamma) e^{\gamma}}{P_o^*} \right) \right)^{\frac{1}{\delta}}}. \tag{26}$$

We then denote

$$\eta_{\text{eff}} = \frac{\eta (\kappa S)^{\frac{1}{\delta}}}{\left(M \log\left(\frac{(1-\gamma) e^{\gamma}}{P_o^*} \right) \right)^{\frac{1}{\delta}}} = \frac{\eta (\kappa S)^{\frac{\alpha}{2}}}{\left(M \log\left(\frac{(1-\gamma) e^{\gamma}}{P_o^*} \right) \right)^{\frac{\alpha}{2}}} \tag{27}$$

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

$$D^* = \frac{F^{\mathrm{U}} + F^{\mathrm{D}}}{B \log_2 (1 + \eta_{\text{eff}})} + \frac{\nu^{\mathrm{U}} F^{\mathrm{U}} + \nu^{\mathrm{D}} F^{\mathrm{D}}}{E_{\mathrm{c}}}. \tag{28}$$

From (28), 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 (27), 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 had better 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 a theoretical perspective and quantify them.

It should be noted that the communication delay is inversely proportional to both the bandwidth B and the computing power E_{c}, implying that improving them is a straightforward approach of improving the latency. Therefore, it could be a good idea to trade the computing power against off the bandwidth, as improving computing capability might be easier and cheaper than improving the bandwidth. Furthermore, since increasing S increases η_{eff}, we can also trade off storage against bandwidth. Finally, if we let the computing delay

be negligible as compared to the communication delay and assume $\eta_{\text{eff}} >> 1$, we can have

$$D^* \approx \frac{F^{\mathrm{U}} + F^{\mathrm{D}}}{B \log_2(\eta_{\text{eff}})} = \frac{F^{\mathrm{U}} + F^{\mathrm{D}}}{B \log_2 \left(\frac{\eta (\kappa S)^{\frac{\alpha}{2}}}{\left(M \log\left(\frac{(1-\gamma)e^{\gamma}}{P_o^*} \right) \right)^{\frac{\alpha}{2}}} \right)}. \quad (29)$$

This indicates that the transmission delay-outage tradeoff is on a $\log \log$ scale, implying that increasing D^* slightly can significantly improve the outage probability.

Remark 1: 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.

Remark 2: Our results can generally be applied to the conventional edge-caching and edge-computing scenarios. The results for the conventional edge-caching scenario can be obtain by letting the computing requirement factor be zero, i.e., $\nu^{\mathrm{U}} = \nu^{\mathrm{D}} = 0$; the results for the conventional edge-computing scenario can be obtain by letting $S = M$. However, it should be noted that since we adopt a noise-limited network in this paper, we then have a simple computing model for the conventional edge-computing scenario. Thus, the analysis becomes straightforward for the conventional edge-computing scenario, where the outage probability is simply $P_o = e^{-\kappa'}$.

C. Extended Analysis and Comparisons for Networks with Reference Schemes and Variants

Based on the above analysis and results, in this subsection, we extend our analysis to systems adopting some important reference caching policies and also to some wireless networks with configurations that are variants of our standard network.

1) Analysis for Networks adopting Most Popular and Uniform Random Caching Policies: We first analyze the standard networks adopting two widely used reference caching policies, namely, the most-popular and uniform random caching policies [25], [49]. The most-popular caching policy 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. Clearly, they are two extremes, and thus are good reference schemes. The outage probability results are provided below:

Proposition 2: Suppose $\gamma < 1$, $M \to \infty$, $S \to \infty$, and $S \leq M$. The outage probability of the network adopting the most-popular caching policy is:

$$P_o^{\text{self}} = 1 - \left(1 - e^{-\kappa'}\right) \left(\frac{S}{M}\right)^{1-\gamma}. \quad (30)$$

In addition, the outage probability of the network adopting the uniform random caching policy is:

$$P_o^{\text{Rn}} = e^{\frac{-\kappa' S}{M}}. \quad (31)$$

Proof. See Appendix I of [2]. □

From Proposition 2, we observe that the outage probability for networks adopting the most-popular caching policy has only a power law reduction with respect to S. Although such scaling law is identical to that of the derived optimal scaling law when S is small (see Theorem 4), it cannot have an exponential law when S becomes large, indicating that such policy is not as effective as the optimal policy. On the other hand, the uniform random caching policy can result in an exponential law for the outage probability reduction, and such law is identical to that of the derived optimal scaling law in Theorem 5 and Corollary 6.1. Furthermore, by comparing (31) with (20) and (23), we see that the outage probability given by the uniform random caching policy is different from the optimal caching policy only in the constant factor term $(1-\gamma)e^{\gamma}$, where $(1-\gamma)e^{\gamma} < 1, \forall \gamma < 1$ and $(1-\gamma)e^{\gamma} = 1$ when $\gamma = 0$. This indicates that the uniform random policy is optimal when the popularity distribution is uniform. On the other hand, as γ tends to 1, this factor term would tend to be larger, which differentiates the optimal caching policy from the uniform random policy as the popularity distribution tends to be more concentrated. The above results explain the intuitions that the most-popular caching policy performs poorly while the uniform random caching policy performs effectively when the popularity distribution is not very concentrated, namely, when $\gamma < 1$.

2) Analysis for Networks adopting a Guaranteed Backhaul: Here, we analyze the networks where each BS is equipped with a dedicated backhaul used to provide the desired datasets with probability P_{Ba} and latency d_{B}, and caching is not considered in BSs. Note that this case is equivalent to the case that each dataset is cached with probability P_{Ba}. Therefore, we let users be associated with the BS with the largest received power among all BSs whose backhauls are available. In this case, the received power of the user is then given by

$$\max_{\mathbf{x}_{\text{Ba}} \in \Phi_{\text{Ba}}} P |h_{\mathbf{x}}|^2 \|\mathbf{x}\|^{-\alpha}, \quad (32)$$

where Φ_{Ba} is the set of BSs whose backhauls are available. Then, following the similar derivations in Sec. II and considering the additional $d_{\text{B}} < D$ latency, we can then obtain the following outage probability:

$$P_o^{\text{BaOnly}} = \exp\left(-P_{\text{Ba}} \pi \lambda \underbrace{\frac{\Gamma(\delta + m_D)}{m_D^{\delta} \Gamma(m_D)} \left(\frac{\eta}{2^{\left(\frac{1}{B} \frac{F^{\mathrm{U}} + F^{\mathrm{D}}}{D - d_{\text{B}} - \frac{\nu^{\mathrm{U}} F^{\mathrm{U}} + \nu^{\mathrm{D}} F^{\mathrm{D}}}{E_{\text{c}}}} \right)} - 1} \right)^{\delta}}_{= \kappa_{\text{B}}} \right) = e^{-P_{\text{Ba}} \kappa_{\text{B}}}, \quad (33)$$

where we implicitly assume that the computational delay plus the backhaul latency should be within the requirement latency D. By comparing (33) with our derived optimal outage probability $(1-\gamma)e^{\gamma}e^{\frac{-S\kappa'}{M}}$ in (23), we can understand whether and how replacing the backhaul with caching can still provide the desired network performance.

3) Analysis for Networks adopting both the Optimal Caching Policy and a Guaranteed Backhaul: Now, we consider a network where each BS is equipped with both a storage for caching and a backhaul. We then consider the following association policy. A typical user with a request for task f would first be associated with the BS that can provide the largest received power among all BSs caching dataset f, and therefore the received power in this case is as described by (1). Then, it checks whether the associated BS can complete the task within the latency requirement D. If yes, the request can be satisfied. If not, the user then choose to switch to the BS that can provide the largest received power among all BSs whose backhaul links are available. Considering the above association policy, we then know that the outage probability in this case can be expressed as

$$P_o^{\text{CB}} = \sum_{f=1}^{M} P_r(m)\mathbb{P}\left[d_f^{\text{C}} > D\right] \cdot \mathbb{P}\left[d^{\text{B}} > D\right], \quad (34)$$

indicating that the outage for completing task f happens only if both the latency of using caching d_m and the latency of using backhaul d^{B} are larger than the latency requirement D. Then, since the latency of completing a task using the backhaul is independent of what task to complete, we obtain

$$P_o^{\text{CB}} = \underbrace{\mathbb{P}\left[d^{\text{B}} > D\right]}_{(a)} \cdot \underbrace{\sum_{f=1}^{M} P_r(f)\mathbb{P}\left[d_f^{\text{C}} > D\right]}_{(b)}, \quad (35)$$

where (a) is indeed given by (33) and the optimal value of (b) can be obtained by using the results in Sec. IV-A. Hence, when considering the most interesting regimes, i.e., the regimes characterized by Theorems 5 and 6, we then obtain:

$$P_o^{\text{CB}} = (1-\gamma)e^{\gamma}e^{\frac{-S\kappa'}{M}}e^{-P_{\text{Ba}}\kappa_{\text{B}}}. \quad (36)$$

From (36), we can see that when equipped with a backhaul, the outage probability of the wireless caching network can be improved by the factor $e^{-P_{\text{Ba}}\kappa_{\text{B}}}$, where $e^{-P_{\text{Ba}}\kappa_{\text{B}}}$ is determined by the performance of the backhauls owned by the BSs. This indicates that the caching-based and the conventional backhaul-based 3C approaches are not mutual exclusive; in contrast, they can effectively be combined with each other.

4) Analysis for Networks adopting a Hierarchical Caching Architecture: The analysis in Sec. IV-C.3, where the backhaul is only available with certain probability, can be exploited in the scenario where each BS is independently connected to an external storage through a backhaul with latency d_{B}. Note that this scenario is similar to that the BSs can connect to some inventory in the cloud. We assume that each connected storage can cache S_{B} datasets. We denote the probability of caching the dataset for task f in each external storage as $P_{c,\text{B}}(f)$. Then, notice that the probability that the backhaul is useful is determined by whether the desired dataset is cached in the connected external storage. Thus, by following the similar derivations in Sec. II and by using the result in Sec. IV-C.3, the outage probability in this case can be expressed as:

$$P_o^{\text{HR}} = \sum_{f=1}^{M} P_r(f)\exp\left(-\kappa' P_c(f)\right)\exp\left(-\kappa_{\text{B}} P_{c,\text{B}}(f)\right). \quad (37)$$

To minimize the outage probability in (37), we need to jointly optimize the caching policy of the BSs and the caching policy of the external storages. This leads to the following optimization problem:

$$\begin{aligned}\min_{P_c(f),P_{c,\text{B}}(f),\forall f} \quad & \sum_{f=1}^{M} P_r(f)\exp\left(-\kappa' P_c(f)\right)\exp\left(-\kappa_{\text{B}} P_{c,\text{B}}(f)\right) \\ s.t. \quad & \sum_{f=1}^{M} P_c(f) = S, \quad \sum_{f=1}^{M} P_{c,\text{B}}(f) = S_{\text{B}}, \\ & 0 \le P_c(f) \le 1, \forall f, \quad 0 \le P_{c,\text{B}}(f) \le 1, \forall f.\end{aligned} \quad (38)$$

Although numerically solving (38) is not challenging as it is a convex optimization problem, finding the closed-form expression for the optimal solution of (38) and the corresponding optimal outage probability is very difficult. Thus, instead of obtaining the optimal outage probability, we characterize an upper bound of it. Specifically, we observe that the uniform random caching policy is very effective when $\gamma < 1$. Therefore, we let the caching policy of the external storages to follow the uniform random policy, leading to $P_{c,\text{B}}(m) = \frac{S_{\text{B}}}{M}$. It follows that the outage probability in this case is given by:

$$P_o^{\text{HR,UPP}} = \exp\left(\frac{-\kappa_{\text{B}} S_{\text{B}}}{M}\right)\sum_{f=1}^{M} P_r(f)\exp\left(-\kappa' P_c(f)\right). \quad (39)$$

Note that (39) is an upper bound of the optimal outage probability because the use of uniform random policy for the external storages is indeed suboptimal. Then, with (39), we can optimize the caching policy of BSs by using the same approach as in Sec. IV-A such that the approximate upper bound of the optimal outage probability can be obtained:

$$\begin{aligned}(P_o^{\text{HR}})^* &\lessapprox \exp\left(\frac{-\kappa_{\text{B}} S_{\text{B}}}{M}\right)(1-\gamma)e^{\gamma}\exp\left(\frac{-\kappa' S}{M}\right) \\ &= (1-\gamma)e^{\gamma}\exp\left(\frac{-(\kappa' S + \kappa_{\text{B}} S_{\text{B}})}{M}\right),\end{aligned} \quad (40)$$

where $(P_o^{\text{HR}})^*$ is the optimal outage probability obtained by solving (38). From (40), we observe that the use of external storages can improve the overall outage probability exponentially.

5) Comparison between Co-located and Distributed Wireless Caching Networks: Using results in Sec. IV-A, we can compare between having co-located caching and distributed caching. Specifically, according to our results, the outage

probability of the interesting regimes can be (approximately) expressed as:

$$P_o^* = (1-\gamma)e^{\gamma}e^{\frac{-S\kappa'}{M}} = (1-\gamma)e^{\gamma} \cdot \exp\left(\frac{-S}{M}\pi\lambda\frac{\Gamma(\delta+m_D)}{m_D^{\delta}\Gamma(m_D)}\left(\frac{\eta}{2^{\left(\frac{1}{B}\frac{F^{\mathrm{U}}+F^{\mathrm{D}}}{D-\frac{\nu^{\mathrm{U}}F^{\mathrm{U}}+\nu^{\mathrm{D}}F^{\mathrm{D}}}{E_{\mathrm{c}}}}\right)}-1}\right)^{\delta}\right). \tag{41}$$

The difference between the co-located caching and distributed caching is whether the cache space is distributedly located in different BSs or co-located in a small number of BSs. Thus, for the comparison, we consider $\lambda S = S_{\text{tot}}$ to be a constant. Then, the caching is more distributed/co-located when letting λ to be larger/smaller. It follows that since λS is a constant, we see from (41) that the outage probability is also a constant. Therefore, the co-located caching and distributed caching indeed provide the same performance. We note that, however, this conclusion only applies to the scenarios that the BS distribution is a homogeneous PPP and the network is noise-limited. Thus, when considering more complicated scenarios, the result might be different.

V. Delay-Outage Analysis for $\gamma > 1$ Scenarios

In this section, considering $\gamma > 1$, we first conduct the delay-outage analysis based on the optimal caching policy derived in Sec. III. Then, based on the analysis results, insights are provided.

A. Main Results

In the following, we first obtain Theorems 7, 8, 9, and 10 that characterize the upper and lower bounds of the outage probability in different regimes. Then, to obtain insights, we conduct an additional approximation using Theorem 10, leading to Corollary 10.1.

Theorem 7: Let $M \to \infty$ and $S \to \infty$. Consider $\gamma > 1$. Suppose the caching policy is given by Theorem 1 with $m_1^* < 1$. Then, the optimal (minimum) achievable outage probability is lower and upper bounded as:

$$\frac{1}{\gamma}\left[(\gamma-1)e^{\gamma}(c_2)^{1-\gamma}e^{\frac{-\kappa'}{c_2}}+(c_2)^{1-\gamma}\right]\left(\frac{1}{S}\right)^{\gamma-1} - \frac{1}{\gamma}\left(\frac{1}{M}\right)^{\gamma-1} \le P_o^* \le \left[(\gamma-1)e^{\gamma}(c_2)^{1-\gamma}e^{\frac{-\kappa'}{c_2}}+(c_2)^{1-\gamma}\right]\left(\frac{1}{S}\right)^{\gamma-1} - \left(\frac{1}{M}\right)^{\gamma-1}. \tag{42}$$

Proof. See Appendix J of [2]. □

Theorem 8: Let $M \to \infty$ and $S \to \infty$. Consider $\gamma > 1$. Suppose the caching policy is given by Theorem 1 with $m_1^* \ge 1$. Then, the optimal (minimum) achievable outage probability is lower and upper bounded as in (43) on the top of next page.

Proof. See Appendix K of [2]. □

Theorem 9: 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 lower and upper bounded as:

$$\frac{\gamma-1}{\gamma}e^{\gamma}\left(\frac{1}{M}\right)^{\gamma-1}e^{\frac{-S\kappa'}{M}} \le P_o^* \le (\gamma-1)e^{\gamma}\left(\frac{1}{M}\right)^{\gamma-1}e^{\frac{-S\kappa'}{M}}. \tag{44}$$

Proof. See Appendix L of [2]. □

Theorem 10: 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 lower and upper bounded as:

$$\frac{1}{\gamma}e^{-\kappa'}+\frac{\gamma-1}{\gamma}e^{\gamma}(1-C_1)(C_1)^{\frac{\gamma C_1}{1-C_1}}e^{-\kappa'\frac{C_2-C_1}{1-C_1}}\left(\frac{1}{M}\right)^{\gamma-1} \le P_o^* \le \gamma e^{-\kappa'}+(\gamma-1)e^{\gamma}(1-C_1)(C_1)^{\frac{\gamma C_1}{1-C_1}}e^{-\kappa'\frac{C_2-C_1}{1-C_1}}\left(\frac{1}{M}\right)^{\gamma-1}. \tag{45}$$

Proof. See Appendix M of [2]. □

Corollary 10.1: 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 because we are interested in the case that the caching space of a BS is much smaller than the library size. Then, the lower and upper bounds in Theorem 10 for the optimal (minimum) achievable outage probability can be approximated as:

$$\frac{1}{\gamma}e^{-\kappa'}+\frac{\gamma-1}{\gamma}e^{\gamma}\left(\frac{1}{M}\right)^{\gamma-1}e^{\frac{-S\kappa'}{M}} \le P_o^* \le \gamma e^{-\kappa'}+(\gamma-1)e^{\gamma}\left(\frac{1}{M}\right)^{\gamma-1}e^{\frac{-S\kappa'}{M}}. \tag{46}$$

Furthermore, when κ' is sufficiently large so that the outage probability lower bound $e^{-\kappa'}$ is small, the lower and upper bounds in Theorem 10 for the optimal (minimum) achievable outage probability can be approximated as:

$$\frac{\gamma-1}{\gamma}e^{\gamma}\left(\frac{1}{M}\right)^{\gamma-1}e^{\frac{-S\kappa'}{M}} \le P_o^* \le (\gamma-1)e^{\gamma}\left(\frac{1}{M}\right)^{\gamma-1}e^{\frac{-S\kappa'}{M}}. \tag{47}$$

Proof. The proof follows the similar procedure of proving Corollary 6.1. □

B. Interpretations and Insights

Based on the results in Sec. V-A, we can obtain fundamental insights and interpretations for the delay-outage performance of the network. First of all, since (9) is applied also for the case of $\gamma > 1$, we observe that an increase of κ' can again lead to the exponential decrease of the outage probability. In addition, from Theorems 7 and 8, we observe that the outage probability reduction follows a power law with respect to the cache size S when S is small. Then, by Theorem 9 and Corollary 10.1, we see that the outage probability in their corresponding regimes decreases following an exponential law with respect to the

$$\frac{1}{\gamma}e^{-\kappa'}-\frac{1}{\gamma}\left(\frac{1}{M}\right)^{\gamma-1}+$$
$$\frac{1}{\gamma}\left[(\gamma-1)e^{\gamma}(c_2-c_1)(c_2)^{-\gamma}\left(\frac{c_2}{c_1}\right)^{\frac{-\gamma c_1}{c_2-c_1}}e^{\frac{-(1-c_1)\kappa'}{c_2-c_1}}+\left(\frac{1}{c_2}\right)^{\gamma-1}-e^{-\kappa'}\left(\frac{1}{c_1+\frac{1}{S}}\right)^{\gamma-1}\right]\left(\frac{1}{S}\right)^{\gamma-1}\leq P_o^*\leq \tag{43}$$
$$\gamma e^{-\kappa'}-\left(\frac{1}{M}\right)^{\gamma-1}+\left[(\gamma-1)e^{\gamma}(c_2-c_1)(c_2)^{-\gamma}\left(\frac{c_2}{c_1}\right)^{\frac{-\gamma c_1}{c_2-c_1}}e^{\frac{-(1-c_1)\kappa'}{c_2-c_1}}+\left(\frac{1}{c_2}\right)^{\gamma-1}-e^{-\kappa'}\left(\frac{1}{c_1}\right)^{\gamma-1}\right]\left(\frac{1}{S}\right)^{\gamma-1}.$$

cache size S. By comparing Theorem 9 (Corollary 10.1) with Theorem 5 (Corollary 6.1), we observe that although both theorems (corollaries) follow an exponentially decreasing law, the additional $\left(\frac{1}{M}\right)^{\gamma-1}$ term indicates that the outage probability in the case of $\gamma>1$ is much smaller. Furthermore, we see that when the the caching probability is given by Theorems 2 and 3, we have $S=\Theta(M)$. This thus explains the intuition that when the cache space is orderwise comparable to the library size and the popularity distribution is sharp, the caching capability might not be the restrictive factor for the network as both Theorem 9 and corollary 10.1 imply that the outage probability would be small and would decrease very fast with respect to S as long as the fundamental outage probability law bound $e^{-\kappa'}$ is not restrictive. Note that to improve $e^{-\kappa'}$, we shall improve the computing and communication capabilities of the network. Therefore, similar to the case of $\gamma<1$, κ' is again a critical parameter for the network optimization, and it is even more critical when $\gamma>1$ because the caching capability is less likely to be the limiting factor.

By using results in Theorem 9 and corollary 10.1, we see that the outage probability expressions with $\gamma>1$ can have the same structure as those with $\gamma<1$. Therefore, we can follow a similar procedure as in Sec. IV-B to reformulate the outage probability expression such that the minimum achievable latency expression as well as the similar insights described in Sec. IV-B can be obtained. We stress that such reformulation can be insightful when κ' and S are large, namely, the outage probability is very small and $e^{-\kappa'}$ is not a restrictive factor. However, in certain situations, we observe that the outage probability would already be very low when S is moderate. Thus, in those cases, it would be more straightforward to directly use numerical analysis with the derived outage probability expressions to gain insights.

Similar to the analysis in Sec. IV-C, we can compare the optimal outage probability with the outage probabilities of the most-popular caching policy and the uniform random caching policy to gain insights. To do this, we first provide Proposition 3 which describes the outage probabilities of the most-popular caching policy and the uniform random caching policy:

Proposition 3: Suppose $\gamma>1$, $M\to\infty$, and S is a sufficiently large number with $S\leq M$. The outage probability of the network adopting the most popular caching policy is lower and upper bounded as:

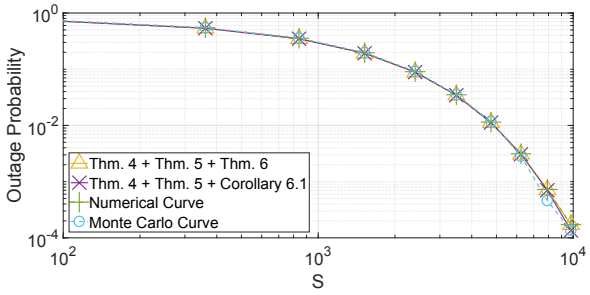
$$\frac{1}{\gamma}\left(1-\left(\frac{1}{S+1}\right)^{\gamma-1}\right)e^{-\kappa'}+\frac{1}{\gamma}\left(\left(\frac{1}{S}\right)^{\gamma-1}-\left(\frac{1}{M+1}\right)^{\gamma-1}\right)$$
$$\leq P_o^{\text{self}}\leq$$
$$\left(\gamma-\left(\frac{1}{S+1}\right)^{\gamma-1}\right)e^{-\kappa'}+\left(\frac{1}{S}\right)^{\gamma-1}-\left(\frac{1}{M+1}\right)^{\gamma-1}. \tag{48}$$

In addition, the outage probability of the network adopting the uniform random caching policy is:

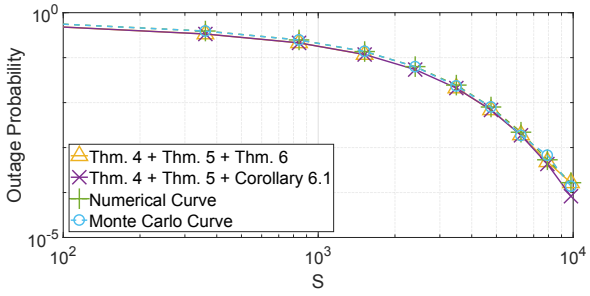
$$P_o^{\text{Rn}}=e^{\frac{-\kappa' S}{M}}. \tag{49}$$

Proof. See Appendix N of [2]. □

By comparing the outage probability lower and upper bounds of most-popular caching in Proposition 3 with those in Theorems 7 and 8, we see that they share a similar structure. Therefore, it can be expected that the most-popular caching policy has the similar behavior to the optimal caching policy in the regimes characterized by Theorems 7 and 8, i.e., the regimes that S is much smaller than M. On the other hand, when S is much smaller than M, the outage probability performance of the uniform random caching policy could be very poor as $P_o^{\text{Rn}}=e^{\frac{-\kappa' S}{M}}$ would be close to 1 when $\frac{S}{M}$ is small. This is different from the case that when $\gamma<1$, the uniform random caching policy has a good performance. However, we also see that the uniform random caching policy has a better outage probability decreasing law as compared to that of the most-popular caching policy. This implies that when we keep increasing S to the order comparable to M, the uniform random caching policy might start to perform better than the most-popular caching policy, though the crossing points depend on the specific network parameters. By comparing the proposed optimal caching policy with these reference policies, we observe that the optimal caching policy can have the merits of both reference policies, in which when S is small, the optimal policy would focus on caching datasets of popular tasks, and then when S is large, it would act more like the uniform random caching policy where datasets are cached in a more cooperative manner. Finally, we note that extensions similar to those discussed in Sec. IV-C can also be conducted via using results in this section. However, since the procedure and insights would be very similar, we omit the relevant discussion here for brevity.

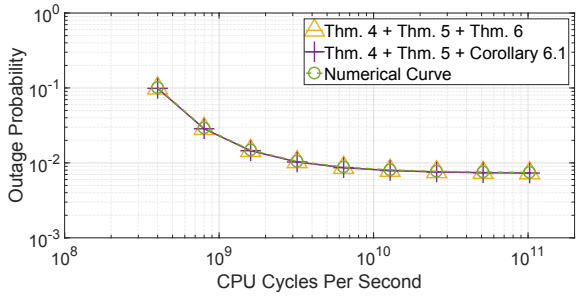


(a) $\gamma = 0.6$.

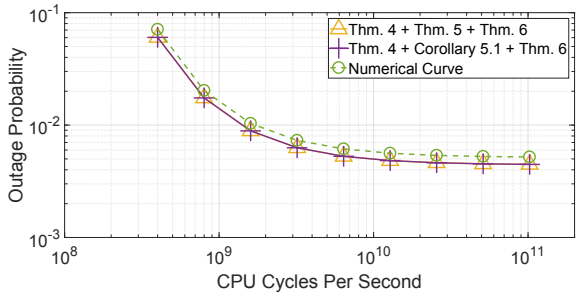


(b) $\gamma = 0.8$.

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



(a) $\gamma = 0.6$.



(b) $\gamma = 0.8$.

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

VI. Computer Simulations

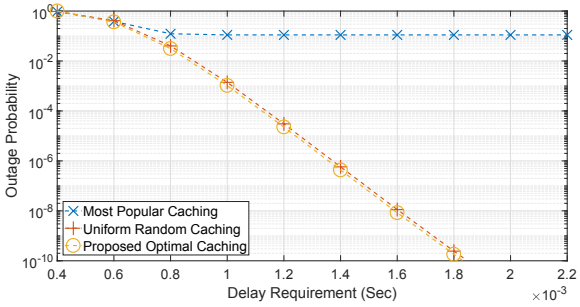
In this section, we validate our analysis using simulations. Unless otherwise indicated, in the simulations, we consider $D = 10^{-3}$ sec, $m_{\rm D} = 1$, $P = 20$ dBm, $\alpha = 3.5$, $B = 20$ MHz, the noise power spectral density $N_0 = -173$ dBm/Hz, $F^{\rm U} = 10^4$ bits, $F^{\rm D} = 10^5$ bits,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

A. Simulation Results for networks with $\gamma < 1$

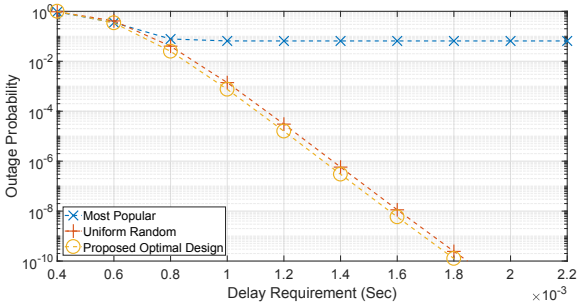
In Fig. 1, we validate our theorems proposed in Sec. IV-A for the outage probability as a function of the caching capability S, where the "numerical" curves are results obtained by first numerically solving the outage probability minimization problem and then evaluating the outage probability using (9); 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. IV-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 accurate, except for the endpoint of the curve with "+ Corollary 6.1" in Fig. 1(b), where the small divergence comes from the fact that the conditions for the good approximation of Corollary 6.1 might not be well-satisfied, namely, C_2 is not small. However, we note that accurate results for not only the functional form, but also the constant factor, 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.

In Fig. 2, we consider $S = 4000$ and evaluate the outage probability as a function of the computing power $E_{\rm c}$. Since we already saw 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 accurate, and only some small constant factor differences can be observed. 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 an approach of improving the performance by only increasing the computing capability could be limited.

Finally, in Fig. 3, we evaluate the delay-outage performance of the considered network with $S = 7500$. 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. Results show that the proposed optimal caching provides 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 by a constant factor (see Theorem 5, Corollary 6.1, and Proposition 2 to compare). However, we note that the good performance of the uniform random caching only exists when $\gamma < 1$, i.e., the

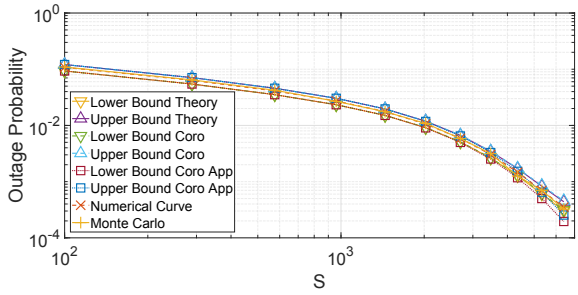


(a) $\gamma = 0.6$.

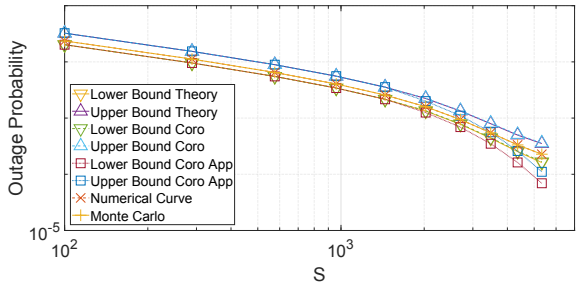


(b) $\gamma = 0.8$.

Fig. 3: Delay-Outage performance evaluation with $S = 7500$.



(a) $\gamma = 1.3$.



(b) $\gamma = 1.6$.

Fig. 4: Outage probability evaluation of the proposed theorems as a function of S.

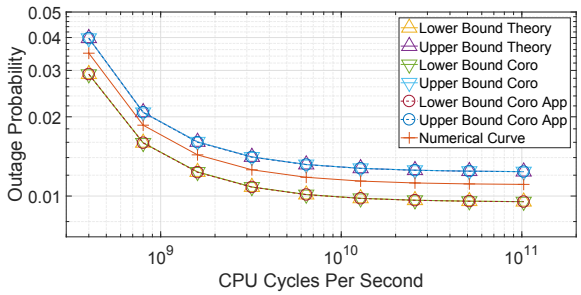
requests are not very concentrated on the popular tasks. We will show below that when $\gamma > 1$, i.e., the requests are more concentrated on popular tasks, the uniform random caching could perform poorly. Finally, we see that changing the delay requirement significantly affects the outage probability, 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.

B. Simulation Results for networks with $\gamma > 1$

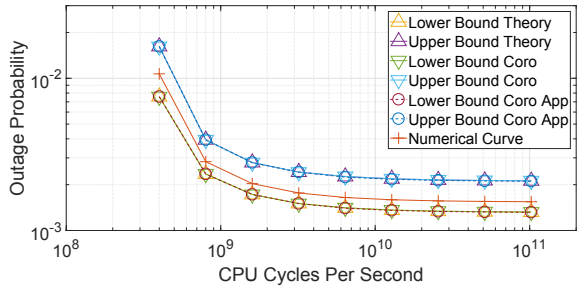
In Fig. 4, we validate our theorems proposed in Sec. V-A as a function of the caching capability S. Again, since different theorems in Sec. V-A characterize different regimes, the theoretical curves are the combinations of different theorems. Specifically, the curves with "Theory" are obtained via collecting results of Theorems 7, 8, 9, and 10; the curves with "Coro" are obtain via replacing Theorem 10 with (46) in Corollary 10.1; and the curves with "Coro App" are obtain via replacing Theorem 10 with (47) in Corollary 10.1. The results show that our proposed analysis can in general effectively characterize the performance via using the derived upper and lower bounds. The only exception is again at the endpoints of the curves with "Coro App" in Fig. 4, where the small divergence comes from that the conditions for the good approximation of Corollary 10.1 might not be well-satisfied. Finally, we see that the outage probability of Fig. 4 for a given S is much lower than the corresponding outage probability in Fig. 1. This validates that the optimal outage probability performance is better when γ is larger.

In Fig. 5, we consider $S = 1500$ and evaluate the outage probability as a function of the computing power E_c. Since we already saw in Fig. 4 that the Monte-Carlo results are essentially identical to the numerical results, we only provide the numerical validation here. Results again show that our theorems are accurate. In addition, we observe that the outage probability saturates when the computing power increases to a large number, validating that when the computing power is sufficient, the performance is then limited by the caching and communication capabilities.

Finally, in Fig. 6, we evaluate the delay-outage performance of the considered network with $S = 1500$. We again compare the optimal caching policy with two reference caching policies, namely, the most-popular caching and uniform random caching. Results show that the proposed optimal caching can provide the best performance. Besides, different from the results in Fig. 3, here we observe that the uniform random caching can no longer provide the near-optimal performance as the uniform caching cannot focus on caching of datasets for very popular tasks. On the other hand, we see that the most-popular caching can be very good in the situations that the delay requirement is stringent, implying that when the requirement is stringent, we should really focus on caching the datasets of most popular tasks. We observe that when keeping relaxing the delay requirement, the uniform random caching ultimately would perform better than the most-popular caching. this is because when the delay requirement is loose, reaching the most-popular tasks in the nearby BSs of users become less important. On the other hand, having more tasks that can be completed at the wireless edge network becomes more important. Note that by comparing Fig. 6(a) with Fig. 6(b), we observe that the crossing point for the most-popular and

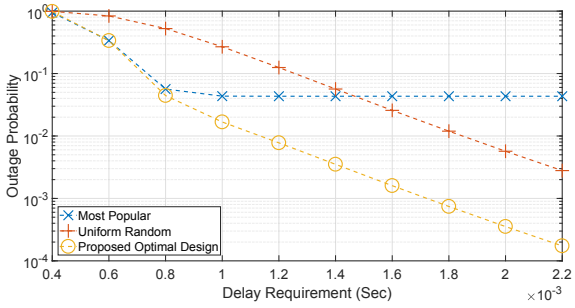


(a) $\gamma = 1.3$.

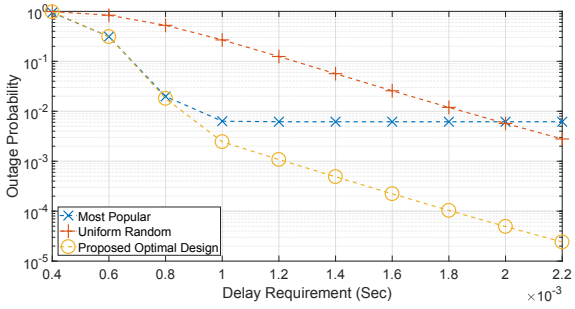


(b) $\gamma = 1.6$.

Fig. 5: Outage probability evaluation of the proposed theorems as a function of E_c.



(a) $\gamma = 1.3$.



(b) $\gamma = 1.6$.

Fig. 6: Delay-Outage performance evaluation with $S = 1500$.

uniform random caching policies shifts right as γ increases. This is because when γ is larger, the popularity distribution is more concentrated. Finally, we again see that changing the delay requirement significantly affects the outage probability, validating our previous observation that via relaxing the latency requirement, the outage probability performance can be significantly improved. Thus, the challenges of the wireless network are imposed by the time-sensitive applications.

VII. Conclusions

In this paper, we provided the asymptotic delay-outage analysis for obtaining the fundamental understanding of the behaviors of wireless edge networks considering joint 3C. Our analysis clearly revealed the relations between the delay-outage performance and 3C parameters and the provided simulations validate our analysis. To be more specific, our analysis showed that increasing κ' can bring an exponential-law improvement to the outage probability. In addition, the analysis showed that the increase of S can first lead to the power-law improvement, and then the exponential-law improvement as S is increased a large number. Furthermore, by some reformulation, we demonstrated 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 observed that slightly relaxing the delay requirement can significantly improve the outage probability. Therefore, the biggest challenges of the wireless network are indeed imposed by the most time-sensitive applications.

Appendix A
Sketch of the Proof Strategy

To obtain the final outage probability as a function of the delay constraint and other network parameters, we conduct the following steps: (i) deriving the general expression for the optimal caching policy that minimizes the outage probability (Proposition 1); (ii) deriving the analytical expressions for the optimal caching policy under different regimes (Theorems 1, 2, and 3); and (iii) deriving the analytical expressions of the optimal outage probability under different regimes (Theorems 4, 5, 6, 8, 9, and 10). Since the derivations for different theorems in the same step follow similar procedure with the differences mainly on the computational details, we in the following sketch the proof procedure of each step described above. Note that since the proof of proposition 1 in step 1 directly applies the Lagrange multiplier and KarushKuhnTucker (KKT) condition that are commonly used in the literature, we skip the sketch of it for brevity and start the discussion from step 2. All details of the proofs can be found in the supplementary material [2].

A. Sketch of the Proof Strategy for Step 2

To derive the analytical expression of the optimal caching policy, we first derive the bounding conditions for the critical indices, namely, m_1^*, m_2^*, and m^*, if they exist. Then, based on the bounding conditions, we can obtain the upper and lower bounds of m_1^*, m_2^*, and m^* via substituting the bound conditions into the general expression of the optimal caching bound in Proposition 1. Then, after some algebraical manipulations, we can show that the upper and lower bounds of m_1^*, m_2^*, and m^* are tight, leading to the analytical expressions of m_1^*, m_2^*,

and m^*. Finally, by combining the analytical expressions of m_1^*, m_2^*, and m^* with the expression in Proposition 1, we can derive the final theorem.

B. Sketch of the Proof Strategy for Step 3

To derive the analytical expression of the optimal outage probability, we first substitute the analytical expression derived in step 2 into the outage probability expression given by (9). It follows that by using the concept of Riemann sum calculus, we can obtain the upper and lower bounds of the outage probability via transforming the summation term with respect to different contents in (9) into its corresponding upper and lower bounds with integration forms. Finally, after some algebraical manipulations, we can obtain analytical expressions of the upper and lower bounds that have similar/identical expressions that construct the final theorem. Note that if we can show that the expressions of the upper and lower bounds have the identical expression, we can conclude the exact expression of the optimal outage probability. On the other hand, if the upper and lower bounds are not identical but having many similar terms, we provide both the upper and lower bounds in the theorem and analyze them subsequently.

REFERENCES

[1] M.-C. Lee and A. F. Molisch, "Asymptotic delay–outage analysis for noise-limited wireless networks with caching, computing, and communications," *2022 IEEE International Conference on Communications (ICC)*, (Accepted by).

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

[3] 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.

[4] 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.

[5] 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.

[6] F. Spinelli and V. Mancuso, "Toward enabled industrial verticals in 5g: A survey on mec-based approaches to provisioning and flexibility," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 1, pp. 596–630, 2020.

[7] 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.

[8] 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.

[9] Y. Cai, J. Llorca, A. M. Tulino, and A. F. Molisch, "Compute- and data-intensive networks: The key to the metaverse," in *2022 1st International Conference on 6G Networking (6GNet)*, 2022, pp. 1–8.

[10] 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.

[11] 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.

[12] 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.

[13] Y. Wei, F. R. Yu, M. Song, and Z. Han, "Joint optimization of caching, computing, and radio resources for fog-enabled iot using natural actor–critic deep reinforcement learning," *IEEE Internet of Things J.*, vol. 6, no. 2, pp. 2061–2073, 2018.

[14] K. Kamran, E. Yeh, and Q. Ma, "Deco: Joint computation, caching and forwarding in data-centric computing networks," in *Proceedings of the Twentieth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, 2019, pp. 111–120.

[15] 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.

[16] S. Sukhmani, M. Sadeghi, M. Erol-Kantarci, and A. El Saddik, "Edge caching and computing in 5g for mobile ar/vr and tactile internet," *IEEE MultiMedia*, vol. 26, no. 1, pp. 21–30, 2018.

[17] T.-Y. Kuo, M.-C. Lee, and T.-S. Lee, "Quality-aware caching, computing and communication design for video delivery in vehicular networks," in *ICC 2022 IEEE International Conference on Communications (ICC)*, 2022, pp. 261–266.

[18] L. T. Tan and R. Q. Hu, "Mobility-aware edge caching and computing in vehicle networks: A deep reinforcement learning," *IEEE Trans. on Veh. Technol.*, vol. 67, no. 11, pp. 10 190–10 203, 2018.

[19] G. Qiao, S. Leng, S. Maharjan, Y. Zhang, and N. Ansari, "Deep reinforcement learning for cooperative content caching in vehicular edge computing and networks," *IEEE Internet of Things J.*, vol. 7, no. 1, pp. 247–257, 2019.

[20] Y. Liu, M. Peng, G. Shou, Y. Chen, and S. Chen, "Toward edge intelligence: Multiaccess edge computing for 5g and internet of things," *IEEE Internet of Things J.*, vol. 7, no. 8, pp. 6722–6747, 2020.

[21] 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.

[22] 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.

[23] X. Wang, Y. Han, C. Wang, Q. Zhao, X. Chen, and M. Chen, "In-edge ai: Intelligentizing mobile edge computing, caching and communication by federated learning," *IEEE Network*, vol. 33, no. 5, pp. 156–165, 2019.

[24] 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.

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

[26] C. Yang, Y. Yao, Z. Chen, and B. Xia, "Analysis on cache-enabled wireless heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 131–145, January 2016.

[27] 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.

[28] Y. Cui, D. Jiang, and Y. Wu, "Analysis and optimization of caching and multicasting in large-scale cache-enabled wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 5101–5112, July 2016.

[29] Y. Chen, M. Ding, J. Li, Z. Lin, G. Mao, and L. Hanzo, "Probabilistic small-cell caching: Performance analysis and optimization," *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 4341–4354, May 2017.

[30] K. Li, C. Yang, Z. Chen, and M. Tao, "Optimization and analysis of probabilistic caching in n-tier heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 1283–1297, February 2018.

[31] D. Jiang and Y. Cui, "Analysis and optimization of caching and multicasting for multi-quality videos in large-scale wireless networks," *IEEE Trans. Commun*, vol. 67, no. 7, pp. 4913–4927, July 2019.

[32] X. Xu and M. Tao, "Modeling, analysis, and optimization of caching in multi-antenna small-cell networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5454–5469, November 2019.

[33] X. Li, X. Wang, P.-J. Wan, Z. Han, and V. C. M. Leung, "Hierarchical edge caching in device-to-device aided mobile networks: Modeling, optimization, and design," *IEEE J. Sele. Areas Commun.*, vol. 36, no. 8, pp. 1768–1785, 2018.

[34] M. Ji, G. Gaire, and A. F. Molisch, "The throughput-outage tradeoff of wireless one-hop caching networks," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6833–6859, December 2015.

[35] 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.

[36] M.-C. Lee, M. Ji, and A. F. Molisch, "Optimal throughput-outage analysis of cache-aided wireless multi-hop D2D networks," *IEEE Trans. Commun.*, vol. 69, no. 4, pp. 2489–2504, April 2021.

[37] M.-C. Lee, A. F. Molisch, and M. Ji, "Throughput-outage scaling behaviors for wireless single-hop d2d caching networks with physical model," *IEEE Trans. Wireless Commun.*, 2022.

[38] 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.

[39] H.-S. Lee and J.-W. Lee, "Task offloading in heterogeneous mobile cloud computing: Modeling, analysis, and cloudlet deployment," *IEEE Access*, vol. 6, pp. 14 908–14 925, 2018.

[40] 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.

[41] C. Park and J. Lee, "Mobile edge computing-enabled heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1038–1051, February 2021.

[42] A. Chilwan and Y. Jiang, "Modeling and delay analysis for sdn-based 5g edge clouds," in *2020 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2020, pp. 1–7.

[43] J. Ren, G. Yu, Y. He, and G. Y. Li, "Collaborative cloud and edge computing for latency minimization," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 5031–5044, May 2019.

[44] Y. Wang, M. Tang, S. Zhou, G. Tan, Z. Zhang, and J. Zhan, "Performance analysis of heterogeneous mobile edge computing networks with multi-core server," in *2020 IEEE 20th International Conference on Communication Technology (ICCT)*. IEEE, 2020, pp. 1540–1545.

[45] L. Lin, W. Zhou, and Z. Zhao, "Analytical modeling of noma-based mobile edge computing systems with randomly located users," *IEEE Commun. Lett.*, vol. 24, no. 12, pp. 2965–2968, December 2020.

[46] A. Al-Shuwaili and A. Lawey, "Latency reduction for mobile edge computing in hetnets by uplink and downlink decoupled access," *IEEE Wireless Commun. Lett.*, vol. 10, no. 10, pp. 2205–2209, October 2021.

[47] N. Lu and X. S. Shen, "Scaling laws for throughput capacity and delay in wireless networks – a survey," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 2, pp. 642–657, 2013.

[48] M.-C. Lee, A. F. Molisch, N. Sastry, and A. Raman, "Individual preference probability modeling and parameterization for video content in wireless caching networks," *IEEE/ACM Trans. Netw.*, vol. 27, no. 2, pp. 676–690, April 2019.

[49] N. Golrezaei, P. Mansourifard, A. F. Molisch, and A. G. Dimakis, "Base-station assisted device-to-device communications for high-throughput wireless video networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 3665–3676, July 2014.