Throughput-Outage Scaling Behaviors for Wireless Single-Hop D2D Caching Networks with Physical Model

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

***Abstract*—Throughput-Outage scaling laws for single-hop cache-aided device-to-device (D2D) communications have been extensively investigated under the assumption of the protocol model. However, the corresponding performance under physical models has not been explored; in particular it remains unclear whether link-level power control and scheduling can improve the asymptotic performance. This paper thus investigates the throughput-outage scaling laws of cache-aided single-hop D2D networks considering a general physical channel model. By considering the networks with and without the equal-throughput assumption, we analyze the corresponding outer bounds and provide the achievable performance analysis. Results show that when the equal-throughput assumption is considered, using link-level power control and scheduling cannot improve the scaling laws. On the other hand, when the equal-throughput assumption is not considered, we show that the proposed double time-slot framework with appropriate link-level power control and scheduling can significantly improve the throughput-outage scaling laws, where the fundamental concept is to first distinguish links according to their communication distances, and then enhance the throughput for links with small communication distances.**

I. Introduction

In the past few years, the demand of video services has increased rapidly for mobile devices [3], and thus significant efforts have been made to deal with such challenge [4], [5]. Although the improvement from conventional approaches [6], e.g., use of additional spectrum, including massive antenna systems, and adopting network densifications, can partially resolve the challenge, the improvement might still be inadequate and inefficient [7]. In this context, caching at the wireless edge was introduced and investigated to further improve the network performance [7].[1]

The fundamental concept of caching at the wireless edge is to convert memory to bandwidth by pre-fetching and caching the popular video content at the network edge nodes so that the video content can be rapidly delivered to users with low cost when demanded [7]. Thus, this along with the concentration of the popularity can bring huge benefits to the network. Recent investigations have revealed that by introducing the caching technologies at the wireless edge, the network can be improved by orders of magnitude in practical simulations [8], [9] and alter the fundamental scaling in theoretical analysis [10]–[13].

Due to the significance of having caching at the wireless edge, it has been investigated in different scenarios [9], [14]–[16]. These include caching in BSs [7], caching on devices [8], caching in heterogeneous networks [17], caching in vehicles [18], etc. As caching on devices along with high performance device-to-device (D2D) communications can bring huge benefits to the network without installing new infrastructure, cache-aided wireless D2D networks have been widely discussed in the literature [14]–[16]. While many papers investigate the design and implementation aspects, another set of investigations focuses on characterizing the asymptotic behavior as the number of users N goes to infinity. Such investigations are commonly referred to as scaling law analysis, where the analysis results can be used to understand the fundamental limits and benefits for the transmission strategy of the network, to give the quantitative characterization on how much improvement can be obtained from having caching, and to provide guideline for network design [8], [11]–[13]. This paper aims to contribute to this range of investigations for cache-aided wireless D2D networks.

A. Related Literature

The scaling laws for wireless D2D/ad-hoc networks without caching have been studied for many years since the seminal work in [19].[2] Known as one of most representative papers for scaling law analysis, [19] studied the transport capacity of the network under protocol and physical models and characterized the throughput scaling laws, where the derived achievable throughput and upper bound in the case of adopting the multi-hop D2D communication were $\Theta\left(\frac{1}{\sqrt{N\log(N)}}\right)$ and $\Theta\left(\frac{1}{\sqrt{N}}\right)$,

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).

M. Ji is with Department of Electrical and Computer Engineering, University of Utah, Salt Lake City, UT 84112, USA (email: mingyue.ji@utah.edu).

This work was supported in part by the National Science Foundation (NSF) under grants CNS-1816699, CCF-1817154 and SpecEES-1824558, 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.

Part of this work has been presented in the 2021 IEEE International Conference on Communications [1]. An expanded version of this paper, which includes detailed proofs of all the lemmas, corollaries, propositions, and theorems, is online available in [2].

[1] Note that caching at the wireless edge is not a competing technology for the conventional approaches. On the contrary, it can complement the conventional approaches, and thus further improve the network performance.

[2] Scaling law order notations: given two functions f and g, we say that: (1) $f(n) = \mathcal{O}(g(n))$ if there exists a constant c and integer N such that $f(n) \leq cg(n)$ for $n > N$. (2) $f(n) = o(g(n))$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$. (3) $f(n) = \Omega(g(n))$ if $g(n) = \mathcal{O}(f(n))$. (4) $f(n) = \omega(g(n))$ if $g(n) = o(f(n))$. (5) $f(n) = \Theta(g(n))$ if $f(n) = \mathcal{O}(g(n))$ and $g(n) = \mathcal{O}(f(n))$.

respectively. Refs [20] and [21] extended the analysis and showed that the upper bound is $\Theta\left(\frac{1}{\sqrt{N}}\right)$ even under more general conditions. When users are distributed according to a Poisson point process, [22] showed that the optimal achievable throughput for wireless D2D networks is $\Theta\left(\frac{1}{\sqrt{N}}\right)$, matching the upper bound suggested in [20] and [21]. In addition to the above mentioned papers, scaling law analysis has been conducted for networks with more complicated settings and/or channel models. For example, analysis considering fading was dicussed in [21]; analysis considering a specific user mobility model was provided in [23]; analysis considering the multicasting benefit was presented in [24]; and analysis involving a special distributed multiple-input multiple-output (MIMO) structure, known as hierarchical cooperation, was introduced in [25].

Although cache-aided D2D networks have been investigated for years by the computer science community [26]–[29], their fundamental scaling behaviors were not the focus until the early 2010s. Known as one of the earliest scaling law analysis for cache-aided D2D networks, [30] characterized the scaling law of the maximum expected throughput in cache-aided single-hop D2D networks under the protocol model. Since focusing on the maximum expected throughput characterization ignores the outage probability analysis, [31] and [11] remedied this aspect and provided the throughput-outage scaling law analysis again for cache-aided single-hop D2D with protocol model.[3] The results showed that when the outage probability is very small and the popularity distribution is a heavy-tailed Zipf distribution, the throughput scales with $\Theta\left(\frac{S}{M}\right)$, where M is the number of contents in the library and S is the cache space of a device. By considering the more practical MZipf popularity distribution, the results of [11] were then generalized in [12].

Cache-aided multihop D2D scaling laws were firstly analyzed in [32], where the average traffic per node was characterized with users located on a regular grid and with a fairly simplified channel model. Then, the throughput-outage scaling law analysis was provided in [33] under the protocol model. The results showed that when the outage probability is vanishing, the achievable throughput per user is $\Theta\left(\sqrt{\frac{S}{M\log(N)}}\right)$ for heavy-tailed Zipf popularity distributions, while the upper bound is $\Theta\left(\sqrt{\frac{S\log(N)}{M}}\right)$. This upper bound was improved to $\Theta\left(\sqrt{\frac{S}{M}}\right)$ in [34], where the more practical physical model was used and a fully centralized caching policy was considered. Considering the pros and cons in [32]–[34], [13] studied the cache-aided multihop D2D scaling laws adopting the Poisson point process (PPP) for user distribution, physical model for transmissions, decentralized policy for caching, and MZipf distribution for popularity. Results in [13] demonstrated that when the outage probability is very small, the optimal throughput per user is $\Theta\left(\sqrt{\frac{S}{M}}\right)$ for heavy-tailed MZipf distribution and is $\Theta\left(\sqrt{\frac{S}{q}}\right)$ for light-tailed MZipf distribution, where q is the plateau factor of the MZipf distribution.

The above literature considered conventional single-hop and multihop D2D for communications with users more or less uniformly distributed within the network, though there exist papers also considering more complicated settings and communication schemes. For example, in [35], a scaling law analysis was conducted for the cache-aided hierarchical cooperation approach. Besides, the scaling law in cache-aided D2D networks considering nonuniform user distribution was investigated in [36]. Moreover, when involving coding and multicasting schemes, coded cache-aided D2D was proposed and analyzed in [37]–[41].

B. Contributions

In this paper, we focus on the scaling law analysis for cache-aided single-hop D2D networks. By the above literature review, we observe that the scaling law investigations for single-hop cache-aided D2D networks, i.e., [11], [12], [30], were conducted mostly with protocol model. However, the protocol model might be oversimplified, as it cannot incorporate the influence of link-level power control and scheduling into the analysis. A more realistic model to use is the physical model [20], in which the influence of link-level power control and scheduling can be accommodated. Although the suitable scheduling and power control algorithms have been investigated for finite-size networks and performance has been investigated by simulations [42]–[45], to the best of our knowledge, the scaling behaviors for cache-aided single-hop D2D networks with physical model have not been explored, and it is unclear whether and how the link-level power control and scheduling can further improve the scaling laws as compared with those derived under protocol model. This paper thus aims to contribute to this aspect.

Specifically, this paper considers a single-hop cache-aided D2D network with MZipf popularity distribution and with users to be uniformly distributed in the network. We conduct the throughput-outage scaling law analysis with the generalized physical model.[4] We consider two fundamental scenarios for the networks, where scenario 1 assumes that the link-level throughput realizations are equal for all users in the network and scenario 2 relaxes the assumption such that different users can have different link-level throughput realizations, leading to certain unfairness. It should be noted that since the unfairness of the second scenario is only on the realization level, users in either scenario on average have the same throughput, namely, users are treated *statistically fair* in both scenarios.

We conduct both the achievable and outer bound analysis assuming $M, N \to \infty$. We consider the regime that the outage probability is very small or converging to zero for the asymptotic analysis purpose, and such consideration indeed corresponds to the requirement that the desirable network outage probability is small when the parameters are set to be finite. When the MZipf distribution is heavy-tailed, i.e.,

[3] The outage probability notion discussed in this paper, whose definition will be formally provided in Sec. II, is basically the probability that a user experiences a long lack of service.

[4] Although the scaling laws of the bounded physical model could be more realistic, as indicated in [20], they can be viewed as special cases of the scaling laws derived considering the generalized physical channel. Therefore, we in this paper focus on the generalized physical model for brevity. This assumption will discussed in detail in Sec. II.A.

$\gamma < 1$ and $q \to \infty$, where γ is the Zipf factor of the distribution, we show that the upper bound of the throughput per user for scenario 1 is $\Theta\left(\frac{S}{M}\right)$ when the outage probability is very small, which is identical to the throughput-outage scaling laws assuming the protocol model [12]. This indicates that the performance of single-hop cache-aided D2D networks cannot be asymptotically improved by link-level power control and scheduling in scenario 1. Note that the assumption that $q \to \infty$ is to avoid the situation that the MZipf distribution degenerates to a Zipf distribution asymptotically. In contrast, we demonstrate that in scenario 2, by using the proposed throughput-enhancing approach, the throughput per user upper bound can be enhanced to $\Theta\left(\left(\frac{S}{M}\right)^{\frac{1-\gamma}{2-\gamma}}\right)$ while the outage probability retains very small. The fundamental concept of the proposed throughput-enhancing approach is to let transmitter (TX)-receiver (RX) pairs with small communication distances to communicate with a very high speed under appropriate link-level power control and scheduling. In addition to the outer bound analysis, the achievable schemes for both scenarios 1 and 2 adopting $\gamma < 1$ are proposed, and the analysis shows that the proposed schemes can achieve their corresponding outer bounds.

We conduct the analysis also for the light-tailed MZipf distribution, where $\gamma > 1$ and $q \to \infty$ is considered. We show in this case that the user throughput upper bound is $\Theta\left(\frac{S}{q}\right)$ with very small outage probability for scenario 1, again indicating that using link-level power control and scheduling cannot improve the throughput-outage scaling law in scenario 1. On the other hand, we demonstrate that the throughput upper bound can be enhanced to $\Theta\left(\sqrt{\frac{S}{q}}\right)$ with very small outage probability by the proposed throughput-enhancing approach for scenario 2. The schemes that can achieve these outer bounds are proposed and analyzed, respectively. Finally, we analyze the throughput-outage scaling law considering the Zipf distribution under the conditions that $\gamma > 1$ and that the maximum instantaneous power can go to infinity with average user power remaining constant. This is to see whether allowing the maximum instantaneous power going to infinity can allow us to break the $\Theta(1)$ throughput limitation in the physical model when the instantaneous power is upper bounded by some constant. However, the result shows that this is not possible even though we allow the instantaneous power going to infinity with the average user power still being some constant.

C. Paper Organization

The remainder of this paper is organized as follows. Sec. II discusses the models, assumptions, scenarios, and definitions of the throughput and outage adopted in this paper. Sec. III provides the throughput-outage scaling law analysis considering the MZipf distribution with $\gamma < 1$. The scaling law analysis considering the MZipf distribution with $\gamma > 1$ is provided in Sec. IV. The analysis for the Zipf distribution with $\gamma > 1$ and infinite maximum instantaneous power is presented in Sec. V. Conclusions and some discussions of this paper are provided in Sec. VI. The detailed proofs are relegated to appendices of the supplemental file which is online available in [2].

II. Network Model

We consider a random dense network where users are placed according to a binomial point process (BPP) within a unit square-shaped area $[0,1] \times [0,1]$. Accordingly, we assume that the number of users in the network is N, and users are distributed uniformly at random within the network. We assume each device in the network can cache S files and each file has equal size. We consider a library consisting of M files. We assume that users request the files from the library independently according to a request distribution modeled by the MZipf distribution [12]:

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

where γ is the Zipf factor; q is the plateau factor of the distribution; and $H(a,b,\gamma,q) := \sum_{f=a}^{b}(f+q)^{-\gamma}$. Note that the MZipf distribution degenerates to a Zipf distribution when $q = 0$. Thus, it is a more general model. Furthermore, as being reported in [12] based on an extensive real-world dataset, the MZipf distribution is a better model for modeling the on-demand video requests in cellular systems. To simplify the notation, we will in the remainder of this paper use $P_r(f)$ instead of $P_r(f;\gamma,q)$ as the short-handed expression. We consider a decentralized random caching policy for all users [46], in which users cache files independently according to the same caching policy. Denoting $P_c(f)$ as the probability that a user caches file f, the caching policy is fully described by $P_c(1), P_c(2), ..., P_c(M)$, where $0 \leq P_c(f) \leq 1, \forall f$; thus users cache files according to the caching policy $\{P_c(f)\}_{f=1}^{M}$. We consider $\sum_{f=1}^{M} P_c(f) = S$. Then, each user can cache exactly S different files according to the caching mechanism provided in [46]. In this paper, we assume that S and γ are some constants.

We consider the asymptotic analysis in this paper, in which we assume that $N \to \infty$ and $M \to \infty$. We will restrict to $M = o(N)$ and $q = \mathcal{O}(M)$ when $\gamma < 1$; $M = o(N)$ and $q = o(M)$ when $\gamma > 1$. The main reason for restricting to $M = o(N)$ when $\gamma < 1$ is to let users of the network have sufficient ability to cache the whole library. Similarly, the assumption that $q = o(M)$ and $M = o(N)$ when $\gamma > 1$ can give the users of the network a sufficient ability to cache the most popular q files (orderwise); otherwise the outage probability would go to 1.

The plateau factor q can either go to infinity or remain constant. When q goes to infinity, it is sufficient to consider $q = \mathcal{O}(M)$. This is because the MZipf distribution would behave like a uniform distribution asymptotically as $q = \omega(M)$ and such case is less interesting because the concentration property of files in this case is not captured and because this is equivalent to letting γ very close to 0. Consequently, we assume $q = \mathcal{O}(M)$ when $\gamma < 1$. In addition, when $\gamma > 1$, it is more interesting to consider the case that $q = o(M)$ because it gives a clear distinction between the heavy-tailed case ($\gamma < 1$) and the light-tailed case

($\gamma > 1$), where the mathematical definition of a heavy-tailed popularity distribution can be found in Definition 3 of [33]. Furthermore, in practical terms, we see from the measurement results [12] that q is much smaller than M when $\gamma > 1$, which supports the consideration of $q = o(M)$. When q is a constant, i.e., $q = \Theta(1)$, the request distribution generally behaves like a Zipf distribution as $M \to \infty$. Thus, the results for $q = \Theta(1)$ can be representative for the analysis that uses the Zipf distribution for the request distribution. We will consider $q \to \infty$ in Secs. III and IV and consider $q = \Theta(1)$ only in Sec. V.

We consider single-hop D2D communications for file delivery. We assume users can obtain their desired files through only single-hop D2D communications and assume users always have requests to satisfy. Note that we do not eliminate the possibility that a user can find the desired file from its own cache, and such case can be accommodated by letting the distance between the TX and RX be much smaller than the general D2D communication distance. However, we note that since S is some constant, the probability that a user can find the desired file from its own cache goes to zero as q and M go to infinity. Furthermore, as we would assume the link-rate for file delivery is upper bounded by the power of the TX, the self-caching gain is indeed not significant in terms of asymptotic performance. Similar to [11], we assume that different users making the requests on the same file would request different segments of the file. This avoids the gain from naive multicasting.

We define an outage as an occurrence where a user cannot obtain its desired file from the D2D network. Suppose we are given a realization of the placement of the user locations P according to the binomial point process. In addition, we are given a realization of file requests F and a realization of file placement G of users according to the popularity distribution $P_r(\cdot)$ and caching policy $P_c(\cdot)$, respectively. We can define T_u as the throughput of user $u \in \mathcal{U}$ under a feasible single-hop file delivery scheme. Therefore, T_u is defined as:

$$T_u = \frac{1}{T}\sum_{t=1}^{T} C_u(t)A_u(t), \tag{2}$$

where T is the number of time-slots for the transmission, $C_u(t)$ is the link rate for user u in time-slot t, and $A_u(t)$ is the link activation indicator of user u at time-slot t, where $A_u(t) = 1$ if the link of user u is scheduled at time-slot t; otherwise $A_u(t) = 0$. We then define the average throughput of user u as $\overline{T}_u = \mathbb{E}_{\mathsf{P},\mathsf{F},\mathsf{G}}[T_u]$, where the expectation is taken over the placement of user locations P, file requests F of users, the file placement of users G, and the file delivery scheme. Finally, we define the average throughput of a user in the network as

$$T = \min_{u \in \mathcal{U}} \overline{T}_u. \tag{3}$$

When the number of users in the network is N, we define

$$N_o = \sum_{u \in \mathcal{U}} \mathbf{1}\{\mathbb{E}[T_u \mid \mathsf{P},\mathsf{F},\mathsf{G}] = 0\} \tag{4}$$

as the number of users that in outage, where $\mathbf{1}\{\mathbb{E}[T_u \mid \mathsf{P},\mathsf{F},\mathsf{G}] = 0\}$ is the indicator function such that the value is 1 if $\mathbb{E}[T_u \mid \mathsf{P},\mathsf{F},\mathsf{G}] = 0$; otherwise the value is 0. Intuitively, $\mathbf{1}\{\mathbb{E}[T_u \mid \mathsf{P},\mathsf{F},\mathsf{G}] = 0\}$ is equal to zero when the file delivery scheme cannot deliver the desired file to user u. We note that the expectation of $\mathbb{E}[T_u \mid \mathsf{P},\mathsf{F},\mathsf{G}]$ is taken over the file delivery scheme. The outage probability is then defined as

$$p_o = \frac{1}{N}\mathbb{E}_{\mathsf{P},\mathsf{F},\mathsf{G}}[N_o] = \frac{1}{N}\sum_{u \in \mathcal{U}} \mathbb{P}\left(\mathbb{E}[T_u \mid \mathsf{P},\mathsf{F},\mathsf{G}] = 0\right). \tag{5}$$

A. Channel Model

We consider the generalized physical model in this paper. Suppose there is a TX-RX pair u, where user u serves as the TX and user $u^{(\mathrm{r})}$ serves as RX. We denote x_u and $x_{u^{(\mathrm{r})}}$ as the locations of user u and $u^{(\mathrm{r})}$, respectively, and denote Γ_{Co}^u as the set of users transmitting in the same time-frequency resource. Assume that $P_{\max}$ is the maximum power that a user can use for transmission. Then, the generalized physical model defines the link-rate of the TX-RX pair u as [20], [21]:

$$R(u, u^{(\mathrm{r})}) = B_u \log_2\left(1 + \frac{P_u l_{uu^{(\mathrm{r})}}}{B_u N_0 + \sum_{k \neq u, k \in \Gamma_{\text{Co}}^u} P_k l_{ku^{(\mathrm{r})}}}\right), \tag{6}$$

where B_u is the bandwidth used for communication between users u and $u^{(\mathrm{r})}$; $P_u \leq P_{\max}$ is the power of user u; and $l_{uu^{(\mathrm{r})}} = \frac{\chi}{\left(d_{uu^{(\mathrm{r})}}\right)^{\alpha}}$ is the path (power) gain between users u and $u^{(\mathrm{r})}$,[5] where

$$d_{uu^{(\mathrm{r})}} = |x_u - x_{u^{(\mathrm{r})}}| \tag{7}$$

is the distance between users u and $u^{(\mathrm{r})}$, $\chi > 0$ is some calibration factor, and $\alpha > 2$ is the pathloss coefficient. Note that different from our conference version that provides dedicated analysis for both the bounded physical model and the generalized physical model, in this paper, we focus on the analysis of the generalized physical model. This is because the scaling laws derived in consideration of the bounded physical model can be treated as special cases for those derived for the generalized physical model (with the assumption that P_u is finite). Specifically, as indicated in [20], [21], for any configuration of TX-RX pairs, the differences between the link-rates of TX-RX pairs considering these two physical models are simply bounded by some finite constant when $P_u, \forall u$ are finite. Therefore, it is sufficient that we focus on the generalized physical model. Note that to obtain the rigourous derivations for the scaling laws with the bounded physical model, we can repeat the derivations in this paper and apply them to the bounded physical model after some modifications similar to the approach provided in [1].

[5] It should be noted that the adopted path gain model is an approximation of the realistic model, as it could violate the rationale that the transmit power is larger than the receive power in certain regime and that the pathloss coefficient should follow the free-space pathloss model in the near-field regime. Nevertheless, the adopted model can appropriately capture the relative power loss among different TX-RX pairs and provide high tractability. Therefore, this model is effective and useful as we can correctly interpret the derived results in consideration of the potentially unreasonable portions of the results which will be discussed with more details later in this paper.

B. Targeting Scenarios

We in this paper consider two scenarios with different assumptions. Specifically, for *each realization of the network*, we assume (in the order-wise sense) either (i) all TX-RX pairs transmit the same number of bits in T' sec (e.g., users obtain the same amount of segments of files in T' sec) or (ii) different TX-RX pairs can transmit different numbers of bits in T' sec. Since the first assumption forces different TX-RX pairs in a network realization to have equal user throughput, we refer it to as "equal-throughput assumption". On the other hand, when the equal-throughput assumption is relaxed, i.e., when considering the second assumption, different TX-RX pairs of a network realization can have different throughput, indicating receptions of more bits are allowed for some users in a network realization. Note that although the second assumption would lead to $T_u \neq T_v$ for some $u \neq v$, we still have $\overline{T}_u = \overline{T}_v, \forall u, v$ due to symmetry of the network. Hence, the unequal-throughput assumption considered in the second scenario is in the per realization sense, instead of the average sense. This implies that the unfairness happening for the second scenario is only on the realization level, and users are still statistically fair so that the optimal throughput with definition in (3) is non-trivial. We note that the key difference between the equal- and unequal-throughput assumption is that the unequal-throughput assumption allows different TX-RX pairs in an instant to have different throughput; thus in this case, the link variations can be better exploited to improve the throughput at the expense of rate fairness.

In the remainder of this paper, we will analyze the network with and without the equal-throughput assumption, which are denoted as scenario 1 and scenario 2, respectively. Our main results are summarized in Table I.

III. Throughput-Outage Analysis for MZipf Distribution with $\gamma < 1$

In this section, we analyze the throughput-outage performance for the case $\gamma < 1$. We will first provide the outer bound analysis for both scenarios introduced in Sec. II.B. Then, the achievable schemes and the analyses corresponding to each scenario are provided. We start the analysis by providing Lemmas 1 and 2 which characterize the outage probability and by providing Theorem 1 which describes the transport capacity upper bound.

Lemma 1 (Lemma 4 in [13]): When $n = \omega(M)$ users are uniformly distributed within a network with unit size, the probability to have N_{D} users within an area of size $A = o\left(\frac{N_{\text{D}}}{n}\right)$ is upper bounded by $o(1)$.

Lemma 2 (Lemma 5 in [13]): Suppose $\gamma < 1$. Then, when a user in the network searches through $n_{\text{s}} = o\left(\frac{M}{S}\right)$ different users, we obtain $p_{\text{miss}}(n_{\text{s}}) \geq 1 - o(1)$, where $p_{\text{miss}}(n_{\text{s}})$ is the probability that the user cannot find the desired file from these n_{s} users. Furthermore, when a user in the network searches through $n_{\text{s}} = \rho' M$ different users for some ρ', we have the following results: (i) $p_{\text{miss}}(n_{\text{s}}) \geq \Theta\left(e^{-\rho'}\right)$ if $\rho' = \Theta(1)$ is large enough; and (ii) $p_{\text{miss}}(n_{\text{s}}) \geq (1-\gamma)e^{-(S\rho' - \gamma)}$ if $\rho' = \omega(1)$.

Theorem 1: We denote the set of TX-RX pairs as Γ and define r_u as the communication distance for the TX-RX pair u. Let $M \to \infty$, $N \to \infty$, and $q \to \infty$. Suppose $\gamma < 1$ and $q = \mathcal{O}(M)$. We let $R_0 = \epsilon_0 \sqrt{\frac{\rho' M}{SN}}$, where ϵ_0 is some small constant. We denote the transport capacity of the network consisting of Γ, defined in terms of meter-bits/s, as $C_\Gamma = \sum_{u \in \Gamma} r_u C_u$, where C_u is the average rate (bits/s) of user u. We denote the set of TX-RX pairs which have the largest powers among the TX-RX pairs in their corresponding time-frequency resources as $\mathcal{W}$. Under the generalized physical model, C_Γ is upper bounded as:

$$\begin{aligned} C_\Gamma \leq\ & B\overline{C}_{\mathcal{W}} + B\overline{C}_{\Gamma_{R_0}} \\ & + B\frac{\log_2(e)}{\epsilon_0}\sqrt{\frac{SN}{\rho' M}}\left(\alpha\left(3\sqrt{2}+1\right) + 2(2(\sqrt{2}+1))^\alpha\right), \end{aligned} \tag{8}$$

where B is the total bandwidth of the network; $\overline{C}_{\mathcal{W}}$ is the average transport capacity efficiency, defined in terms of meter-bits per second per Hz, of the TX-RX pairs in $\mathcal{W}$; $\overline{C}_{\Gamma_{R_0}}$ is the average transport capacity efficiency of TX-RX pairs that are not in $\mathcal{W}$ and have communication distances smaller than R_0. The definitions of $\overline{C}_{\mathcal{W}}$ and $\overline{C}_{\Gamma_{R_0}}$ are formally given below (66) at the end of Appendix A of [2].

Proof. See Appendix A of [2]. □

Remark 1: From Lemmas 1 and 2, we conclude that to have a non-vanishing probability for a user to obtain the desired file (i.e., $p_{\text{miss}}(n)$ does not go to 1), with high probability, the distance between the TX and RX is at least $\Theta\left(\sqrt{\frac{\rho' M}{SN}}\right)$, where $\rho' = \Omega(1)$.

A. Outer Bound Analysis for Scenario 1

In this subsection, we consider scenario 1 and provide the outer bound. Since the equal-throughput assumption is considered, different TX-RX pairs transmit the same number of bits in a time period of T' sec. With Remark 1 and Theorem 1, we can obtain the following theorem:

Theorem 2: Let $M \to \infty$, $N \to \infty$, and $q \to \infty$. Suppose $\gamma < 1$. Assume that the powers of users in the network are upper bounded by P_{max}. When $\rho' = \Omega(1)$ is large enough, the throughput-outage performance of the network is dominated by:

$$\begin{aligned} T(P_o) = \Theta\Bigg(& \frac{B}{N}\log_2\left(1 + \frac{P_{\text{max}}}{N_0 B_{\text{s}}}\frac{\chi}{\left(\frac{\rho' M}{SN}\right)^{\frac{\alpha}{2}}}\right) \\ & + B\log_2(e)\left(\alpha\left(3\sqrt{2}+1\right) + 2(2(\sqrt{2}+1))^\alpha\right)\frac{S}{\rho' M}\Bigg), \\ P_o = \Theta & \left(e^{-\rho'}\right), \end{aligned} \tag{9}$$

where P_o can be arbitrarily small or converging to zero.

Proof. See Appendix B of [2]. □

Remark 2: Note that when $N \to \infty$, $T(P_o)$ can be unbounded because the term $\frac{\chi}{\left(\frac{\rho' M}{SN}\right)^{\frac{\alpha}{2}}}$ in $T(P_o)$ in Theorem

TABLE I: Summary of the Main Results

	γ	q	Descriptions
Lemmas 1–3	$\gamma < 1$ and $\gamma > 1$	$q \to \infty$	Preliminary outage probability analysis results
Theorems 1 and 7	$\gamma < 1$ and $\gamma > 1$	$q \to \infty$	Transport capacities for deriving outer bounds
Theorems 2, 3 and 4	$\gamma < 1$	$q \to \infty$, $q = \mathcal{O}(M)$	Throughput-outage outer bounds for scenarios 1 and 2
Proposition 1	Arbitrary	Arbitrary	Link-rate guarantee for all achievable schemes
Theorems 5 and 6	$\gamma < 1$	$q \to \infty$, $q = \mathcal{O}(M)$	Achievable throughput-outage performance for scenarios 1 and 2
Theorems 8, 9 and 10	$\gamma > 1$	$q \to \infty$, $q = o(M)$	Throughput-outage outer bounds for scenarios 1 and 2
Theorems 11 and 12	$\gamma > 1$	$q \to \infty$, $q = o(M)$	Achievable throughput-outage performance for scenarios 1 and 2
Theorem 13	$\gamma > 1$	$q = \Theta(1)$	Optimal throughput-outage performance

2 can go to infinity. This unreasonable result is brought by having the cases that the signal-to-interference-plus noise ratio (SINR) becomes unbounded due to the unbounded path gain. Thus, to correctly interpret the result in Theorem 2, we should consider that the term $\frac{\chi}{\left(\frac{\rho' M}{SN}\right)^{\frac{\alpha}{2}}}$ in $T(P_o)$ in Theorem 2 is upper bounded by 1, which corresponds to the physical reality of the pathloss laws. Note that such consideration applies to all results in the remainder of this paper.

Remark 3: Theorem 2 shows that when $\gamma < 1$, M is the dominant factor while q does not impact the asymptotic scaling law. In addition, it shows that when the maximum transmit power is some constant, the throughput-outage performance outer bound considering the generalized physical model has the same scaling law as the throughput-outage performance considering the protocol model [11], [12]. Note that in contrast to the physical model here which enables the link-level power allocation and scheduling, the protocol model and the approaches in [11], [12] only consider the simple clustering network and the system-level changing of the cluster size. As a result, this indicates that the link-level power allocation and scheduling cannot improve the throughput-outage scaling law, i.e., the asymptotic growth rate of the throughput-outage performance, when requests of users are served with equal-throughput assumption. That being said, in practice, the constant factor of the throughput-outage performance might still be improved by a good power control and link scheduling approach.

Remark 4: Slightly different from Remark 3, when we allow the maximum instantaneous transmit power to be infinity while the average power is still some constant, Theorem 2 suggests that the asymptotic performance might be improved if $\frac{S}{\rho' M} = o\left(\frac{\log_2(N)}{N}\right)$. Such improvement could be possible if we let a user to exclusively transmit with the power being $\Theta(N)$ once every $\Theta(N)$ time-slots. However, in this case, it indicates that simple time-division multiple access (TDMA) can dominate the performance, and thus the relevant discussion becomes trivial.

B. Outer Bound Analysis for the Proposed Throughput-Enhancing Approach in Scenario 2

From the result in Sec. III.A, we see that having link-level power allocation and scheduling cannot improve the throughput-outage scaling law when the equal-throughput assumption is considered. To break such limitation, we drop the equal-throughput assumption here and propose a throughput-enhancing approach that can improve the scaling law by appropriately using link-level power allocation and scheduling. We conduct the outer bound analysis for the proposed throughput-enhancing approach in this subsection. The achievable performance for the throughput-enhancing approach will then be discussed later in Sec. III.D.

Since the equal-throughput assumption is dropped in this case, we can take advantage of letting the TX-RX pairs with small communication distances transmit at a much higher throughput to enhance the network throughput. Based on this concept, we propose a double time-slot throughput-enhancing approach as follows. We first split the overall transmission period T' into two time-slots; each has the duration $\frac{T'}{2}$. The first time-slot is used for the general file delivery which adopts the same approach as in scenario 1. We then use the second time-slot to enhance the overall throughput. To do this, we let TX-RX pairs with communication distances smaller than $\sqrt{\epsilon'}R_0$, where $\epsilon' = \mathcal{O}(1)$, transmit with high throughput in the second time-slot. Note that users allowed to transmit in the second time-slot are assumed to transmit with equal throughput in that time-slot, but such throughput should be much larger than the throughput transmitted by users in the first time-slot. In addition, the split of the overall transmission period into two equal time-slots does not lose the optimality of the scaling law as compared to other time-slot splits using different fractions. This is because the scaling law is to characterize the order gain, and we only have a constant factor gain even if we can magically allow the transmission durations for both time-slots to be extended to T', i.e., both time-slots occupy a duration of T' (which is not possible in reality as the overall transmission period is only T'). With the above described approach, we can obtain the following theorem that characterizes the outer bound for the proposed double time-slot scheme:

Theorem 3: Let $M \to \infty$, $N \to \infty$, and $q \to \infty$. Assume $\gamma < 1$ and $q = \mathcal{O}\left(\epsilon' \rho' \frac{M}{S}\right)$. Suppose the double time-slot framework discussed in Sec. III.B is used and the $\epsilon' = \mathcal{O}(1)$ is selected. Assume that the powers of users in the network are upper bounded by P_{max}. Then, when $\rho' = \Omega(1)$ is large enough and λ_2 is feasible, the throughput-outage performance of the network is dominated by:

$$T(P_o) = \Theta\left(\frac{B}{2}\frac{\log_2\left(1+\frac{P_{\text{max}}}{N_0 B_s}\frac{\chi}{\left(\frac{\epsilon'\rho' M}{SN}\right)^{\frac{\alpha}{2}}}\right)}{N} + \frac{B}{2}\frac{S}{\epsilon'\rho' M}\right),$$

$$P_o = \Theta\left(e^{-\rho'}\right). \tag{10}$$

Furthermore, when considering a *network instance*, the throughput per user for users with communication distances $d_u > \sqrt{\epsilon'}R_0$ is dominated by:

$$\lambda_1 = \Theta\left(\frac{B}{N}\log_2\left(1+\frac{P_{\text{max}}}{N_0 B_{\text{s}}}\frac{\chi}{\left(\frac{\rho' M}{SN}\right)^{\frac{\alpha}{2}}}\right)+\frac{BS}{\rho' M}\right); \quad (11)$$

the throughput per user for users with communication distance $d_u \leq \sqrt{\epsilon'}R_0$ is dominated by:

$$\lambda_2 = \Theta\left(\frac{B}{2}\frac{\log_2\left(1+\frac{P_{\text{max}}}{N_0 B_{\text{s}}}\frac{\chi}{\left(\frac{\epsilon'\rho' M}{SN}\right)^{\frac{\alpha}{2}}}\right)}{\delta' N}\right)+\Theta\left(\frac{BS}{\delta'\epsilon'\rho' M}\right), \quad (12)$$

where $\delta' N$ is the number of users with communication distance $d_u \leq \sqrt{\epsilon'}R_0$.

Proof. See Appendix C of [2]. □

Remark 5: By Theorem 3, we see that the concept of leveraging high-throughput transmissions of the TX-RX pairs with small distances can effectively enhance the overall throughput outer bound as our proposed double time-slot approach is realized with $\epsilon' = o(1)$. We stress that this is an outer bound for the proposed double time-slot scheme, while we make no claims about outer bounds for all possible transmission schemes. Therefore, it is likely that introducing the multiple (more than two) time-slot approach might further enhance the throughput. However, as the performance enhancement ability is dependent on ϵ' and δ', and the characteristics of such performance enhancement ability are unclear, we thus in this paper focus on studying the performance of the double time-slot approach, and the investigations of the multiple time-slot approach are considered as possible future works.

We see from Theorem 3 that the network throughput T can be increased via decreasing ϵ'. However, due to the physical limitation, namely the TX-RX link feasibility, there is a lower bound on ϵ', leading to an upper bound of the throughput. To find this upper bound, we in the following analyze ϵ'. Note that the first term of (10) is asymptotically irrelevant to ϵ' because we interpret $\frac{\chi}{\left(\frac{\epsilon'\rho' M}{SN}\right)^{\frac{\alpha}{2}}}$ is upper bounded by some constant according to Remark 2. Thus, the benefit of the first term of (10) comes only from letting the transmission power go to infinity. It follows that we can without loss of generality focus on the $\frac{B}{2}\frac{S}{\epsilon'\rho' M}$ term when characterizing the lower bound of ϵ'. We thus in the following assume that P_{max} is some constant for simplicity, and then derive the following Corollary:

Corollary 1: Following Theorem 3, when maximizing the number of users that can obtain the desired files within the distance $\sqrt{\epsilon'}R_0$, namely when maximizing δ', the throughput for users with communication distances being $d_u \leq \sqrt{\epsilon'}R_0$ is dominated by:

$$\lambda_2 = \Theta\left(\frac{B}{2}\frac{\log_2\left(1+\frac{P_{\text{max}}}{N_0 B_{\text{s}}}\frac{\chi}{\left(\frac{\epsilon'\rho' M}{SN}\right)^{\frac{\alpha}{2}}}\right)}{(\epsilon'\rho')^{1-\gamma}N}\right)+\Theta\left(\frac{BS}{(\epsilon'\rho')^{2-\gamma}M}\right). \quad (13)$$

Proof. See Appendix D of [2]. □

With Corollary 1, we can then derive the outer bound for the proposed throughput-enhancing approach. This is elaborated as follows. By using the same analysis as that for scenario 1, we first know that the throughput per user in the second time-slot leads to the following upper bound (see Appendix B of [2] for details):

$$\lambda_2\delta' N\Theta\left(\sqrt{\frac{\epsilon'\rho' M}{SN}}\right) = \mathcal{O}\left(B\log_2\left(1+\frac{P_{\text{max}}}{N_0 B_{\text{s}}}\frac{\chi}{\left(\frac{\epsilon'\rho' M}{SN}\right)^{\frac{\alpha}{2}}}\right)\sqrt{\frac{\epsilon'\rho' M}{SN}}+B\sqrt{\frac{SN}{\epsilon'\rho' M}}\right). \quad (14)$$

With (14) and that P_{max} is some constant, we obtain

$$\lambda_2\delta' = \mathcal{O}\left(\frac{S}{\epsilon'\rho' M}\right). \quad (15)$$

Then, observe that if we want λ_2 to be less likely to hit its upper bound for a given ϵ', we shall maximize δ'. Recall that the maximum is $\delta' = \Theta\left((\epsilon'\rho')^{1-\gamma}\right)$ as indicated in Corollary 1. This along with that the TX-RX feasibility condition to satisfy is $\lambda_2 \leq \eta$, where η is some constant indicating that the link-rate cannot be infinitely large, leads to that if we want to minimize ϵ' while maintaining the tightness of the upper bound, we should have

$$\eta(\epsilon'\rho')^{1-\gamma} = \Theta\left(\frac{S}{\epsilon'\rho' M}\right) \quad (16)$$

Since η is some constant, this then lead to

$$\epsilon'\rho' = \Theta\left(\left(\frac{S}{M}\right)^{\frac{1}{2-\gamma}}\right). \quad (17)$$

Finally, by using Theorem 3, Corollary 1, and (17), we obtain Theorem 4 as following:

Theorem 4: Let $M \to \infty$, $N \to \infty$, and $q \to \infty$. Assume $\gamma < 1$ and $q = \mathcal{O}\left(\left(\frac{M}{S}\right)^{\frac{1-\gamma}{2-\gamma}}\right)$. Suppose the double time-slot framework discussed in Sec. III.B is used and the $\epsilon' = \mathcal{O}(1)$ is selected. Assume that the powers of users in the network are upper bounded by P_{max}. When $\rho' = \Omega(1)$ is large enough, the throughput-outage performance of the network is dominated by:

$$
\begin{aligned}
T &= \\
&\Theta\left(\frac{B}{2}\frac{\log_2\left(1+\frac{P_{\text{max}}}{N_0 B_{\text{s}}}\frac{\chi}{\left(\left(\frac{M}{S}\right)^{\frac{1-\gamma}{2-\gamma}}\frac{1}{N}\right)^{\frac{\alpha}{2}}}\right)}{N}+\frac{B}{2}\left(\frac{S}{M}\right)^{\frac{1-\gamma}{2-\gamma}}\right), \\
P_o &= \Theta\left(e^{-\rho'}\right).
\end{aligned}
\tag{18}
$$

Furthermore, when considering a *network instance*, the throughput per user for users with communication distances $d_u > \sqrt{\epsilon'} R_0$ is dominated by:

$$
\begin{aligned}
\lambda_1 = \Theta\Bigg(&\frac{B}{N}\log_2\left(1+\frac{P_{\text{max}}}{N_0 B_{\text{s}}}\frac{\chi}{\left(\frac{\rho' M}{SN}\right)^{\frac{\alpha}{2}}}\right) \\
&+ B\log_2(e)\left(\alpha\left(3\sqrt{2}+1\right)+2(2(\sqrt{2}+1))^{\alpha}\right)\frac{S}{\rho' M}\Bigg);
\end{aligned}
\tag{19}
$$

the throughput per user for users with communication distances $d_u \leq \sqrt{\epsilon'} R_0$ is dominated by:

$$
\lambda_2 = \Theta\left(\frac{B}{2}\frac{\log_2\left(1+\frac{P_{\text{max}}}{N_0 B_{\text{s}}}\frac{\chi}{\left(\left(\frac{M}{S}\right)^{\frac{1-\gamma}{2-\gamma}}\frac{1}{N}\right)^{\frac{\alpha}{2}}}\right)}{\left(\left(\frac{S}{M}\right)^{\frac{1-\gamma}{2-\gamma}}\right)N}\right)+\Theta(B).
\tag{20}
$$

Proof. This is directly obtained by using Theorem 3, Corollary 1, and (17). □

Remark 6: By comparing between Theorem 2 and Theorem 4, we observe that the double time-slot approach can significantly improve the throughput performance without sacrificing the outage probability. The benefit is from that we judiciously let TX-RX pairs with very small communication distances transmit at a much higher throughput in the second time-slot. Note that here we implicitly assume that users have sufficient demands so that the only limitation for a user to increase its throughput is the link-rate. In addition, we see that $T_2 = \Theta(B)$ is some constant if the maximum transmit power is some constant. This indicates that the TX-RX pairs with enhanced throughput satisfy the link-rate feasibility condition. Finally, we note that the fundamental reason for improving the scaling laws in scenario 2 is not that the generalized physical model allows the flexible instantaneous link-rates for different TX-RX pairs. Instead, the improvement comes from that the generalized physical model enables the flexible link-level scheduling and power control such that the network can allow more TX-RX pairs to transmit at the same time in the second time-slot, leading to a much better spectrum reuse of the network. We also stress that the throughput-enhancing result here is not because we let TX-RX pairs with high throughput to transmit the same amount of information bits as those TX-RX pairs with low throughput, and then let them finish their transmissions fast so that the remaining amount of resource for other TX-RX pairs to transmit can be increased. On the other hand, the overall throughput is enhanced because the TX-RX pairs with high throughput indeed successfully transmit and receive much more information bits than those TX-RX pairs with low throughput in a given time period.

Remark 7: From Theorem 4, we observe that using the double time-slot approach indeed gives rise to some degree of unfairness, as for each network realization, the users allowed to transmit in the second time-slot can enjoy a much higher instantaneous throughput, though different users would have the same average throughput. Furthermore, as indicated by Theorem 4 that the network throughput T is independent of P_o, the double time-slot approach can ultimately decouple the tradeoff between throughput and outage, and thus the throughput-outage tradeoff no longer exists in terms of the average user throughput. However, this is because the TX-RX pairs with small communication distances can maintain the overall network throughput when increasing ρ' at the cost of introducing further unfairness, and thus the overall throughput-outage tradeoff has been converted to fairness-outage tradeoff. Finally, it should be noted that although we can also enhance the overall network throughput by letting TX-RX pairs with very small communication distances transmit all the time, this indeed would lead to that most of the users cannot transmit, and thus significantly increase the outage probability. This explains why we split the transmission duration into two time-slots and enhance the throughput only using the second time-slot.

C. Achievable Scheme and Analysis for Scenario 1

In this subsection, we provide the achievable scheme and its corresponding analysis for scenario 1. We consider the following achievable scheme. Suppose the communications are in T' sec. We first split this T'-second period into two time-slots; each has $\frac{T'}{2}$ sec. Then, in the first time-slot, all N users in the network are served in a round-robin manner using time division multiple access (TDMA) approach, in which each of them can transmit with the maximum power P_{max} in a period of $\frac{T'}{2N}$ sec. In the second time-slot, we adopt the clustering network with the frequency reuse scheme and cluster-wise round-robin scheduling, i.e., users in the same cluster are served in the round-robin manner and different clusters can be activated simultaneously. The clustering approach used in the second time-slot is as follows. We split the network into equally-sized square clusters whose side length is $d = \Theta\left(\sqrt{\frac{\rho M}{SN}}\right)$. We assume users in a cluster can obtain the desired file only from users in the same cluster. In each cluster, a user is served at a time and users in the same cluster are served in a round-robin manner. Interference between different clusters is avoided via the frequency reuse approach with reuse factor $(2(K+1))^2$, where $K \in \mathcal{N} > 0$ is a finite positive integer. Then, by symmetry of the clusters and by the facts that $N \to \infty$ and $d = o(1)$, when computing the outage probability, we can assume that clusters are independent to one another and that the number of users N_{cluster} in a cluster follows the Poisson distribution, given as $N_{\text{cluster}} \backsim F_{\text{Poi}}\left(\frac{\rho M}{S}\right)$, where $\frac{\rho M}{S}$ is the mean value of the Poisson distribution. Note that the latter assumption is because the binomial distribution converges to a Poisson distribution when the number of trials goes to infinity while the probability of success goes to zero.

We use the randomized caching policy described in Sec. II. It follows from Lemma 1 in [13] that the outage probability of the described network with clustering is

$$p_o = \sum_{f=1}^{M} P_r(f) e^{-\frac{\rho M}{S} P_c(f)}, \tag{21}$$

where $P_c(f), \forall f$ are determined by the caching policy minimizing (21), as provided in Theorem 1 of [13]. With the aforementioned transmission and caching policies, users in the network are served by the combination of two types of delivery approaches, i.e., the simple TDMA in the first time-slot and the clustering in the second time-slot. Note that the split here is not to enable the throughput-enhancing approach introduced in Sec. III.B as scenario 1 is considered here. On the contrary, it is simply used for achieving the outer bound in Sec. III.A. To derive the achievable throughput-outage performance, we start with the following proposition:

Proposition 1: Suppose the clustering network is considered with the frequency reuse approach adopting the reuse factor $(2(K+1))^2$, where $K \in \mathcal{N} > 0$ is some finite positive integer. Assume that the powers of users in the network are upper bounded by some constant $\nu_{\text{upp}} = \Theta(1)$ and lower bounded by some constant $\nu_{\text{low}} = \Theta(1)$, i.e., $\nu_{\text{low}} \leq P_u \leq \nu_{\text{upp}}, \forall u$. Then, when each cluster at most activates a single user in the cluster for transmission at a time, there must exist some constant ϑ such that for any activated TX-RX pair u, we obtain

$$\begin{aligned} R(u, u^{(\text{r})}) &= B_u \log_2\left(1 + \frac{P_u l_{uu^{(\text{r})}}}{B_u N_0 + \sum_{k \neq u, k \in \Gamma_{\text{Co}}^u} P_k l_{ku^{(\text{r})}}}\right) \\ &\geq B_u \log_2(1 + \vartheta), \end{aligned} \tag{22}$$

where $B_u = \frac{B}{(2(K+1))^2}$ and ϑ is monotonically increasing with respect to the reuse factor K.

Proof. See Appendix E of [2]. □

Proposition 1 indicates that by using the proposed clustering with frequency reuse scheme, users in different clusters are guaranteed to have some constant link-rate. Then, by combining the transmissions in the first and second time-slots and leveraging Proposition 1, we obtain the following theorem:

Theorem 5: Let $M \to \infty$, $N \to \infty$, and $q \to \infty$. Suppose $\gamma < 1$ and $q = \mathcal{O}(M)$. Consider the caching policy in Theorem 1 of [13] and that the side length of a cluster is $\sqrt{\frac{\rho M}{SN}}$. Assume equal-throughput transmissions of users. Assume that the powers of users in the network are upper bounded by $P_{\max}$. When $\rho = \Omega(1)$ is large enough, the following throughput-outage performance of the network is achievable:

$$\begin{aligned} T(P_o) &= \Theta\left(\frac{B}{N}\log_2\left(1 + \frac{P_{\max}}{N_0 B}\frac{\chi}{\left(\frac{\rho M}{SN}\right)^{\frac{\alpha}{2}}}\right)\right) + \Theta\left(\frac{BS}{\rho M}\right), \\ P_o &= \Theta\left(e^{-\rho}\right), \end{aligned} \tag{23}$$

where P_o can be very small or converging to zero.

Proof. See Appendix F of [2]. □

Remark 8: By comparing between Theorems 2 and 5, we see that the proposed outer bound is achievable. This indicates that when the equal-throughput assumption is considered, the simple clustering scheme is asymptotically optimal even though we are allowed to use the link-level power allocation and scheduling.

D. Achievable Scheme and Analysis for Scenario 2 with the Proposed Throughput-Enhancing Approach

In this subsection, an achievable scheme for scenario 2 is presented and analyzed. The achievable scheme for scenario 2 is a combination of the achievable scheme for scenario 1 and the double time-slot throughput-enhancing approach introduced in Sec. III.B. Thus, for the achievable scheme here, by following the throughput-enhancing approach, we first split the transmission duration T' sec into two time-slots; each has $\frac{T'}{2}$ sec. We assume without loss of generality that S is an even number. Then, in the first time-slot, the achievable scheme proposed for scenario 1 in Sec. III.C is directly used, where cluster size in this case is set to $d_1 = \sqrt{\frac{\rho M}{SN}}$. In the second time-slot, we also consider the achievable scheme proposed for scenario 1 in Sec. III.C. However, in this case, the side length of the cluster is changed to $d_2 = \sqrt{\frac{\epsilon \rho M}{SN}}$, where $\epsilon = \mathcal{O}(1)$. By the above descriptions, we observe that the achievable scheme in Sec. III.C is used twice, and each has a dedicated cluster size for realizing the clustering approach used in each time-slot.

For the caching scheme, we also split the whole cache space into two subspaces, and each has size $\frac{S}{2}$. For the first caching subspace, we consider the caching policy proposed in Theorem 1 of [13] with $g_{c,1}(M) = \frac{2\rho M}{S}$. For the second caching subspace, we consider again the same caching policy and let $g_{c,2}(M) = \frac{2\epsilon\rho M}{S}$. By the above described transmission and caching policies, we can then obtain the following theorem which characterizes the achievable performance:

Theorem 6: Let $M \to \infty$, $N \to \infty$, and $q \to \infty$. Suppose $\gamma < 1$ and $q = \mathcal{O}\left(\left(\frac{M}{S}\right)^{\frac{1-\gamma}{2-\gamma}}\right)$. Suppose the proposed achievable scheme in Sec. III.D is used and ϵ is selected such that $\epsilon\rho = \Theta\left(\left(\frac{S}{M}\right)^{\frac{1}{2-\gamma}}\right)$. Assume that the powers of users in the network are upper bounded by $P_{\max}$. Then, when $\rho = \Omega(1)$

is large enough, the following throughput-outage performance of the network is achievable:

$$\begin{aligned}&T(P_o) = \\ &\Theta\left(\frac{B}{2}\frac{\log_2\left(1+\frac{P_{\max}}{N_0 B_{\mathrm{s}}}\frac{\chi}{\left(\left(\frac{M}{S}\right)^{\frac{1-\gamma}{2-\gamma}}\frac{1}{N}\right)^{\frac{\alpha}{2}}}\right)}{N}+\frac{B}{2}\left(\frac{S}{M}\right)^{\frac{1-\gamma}{2-\gamma}}\right), \\ &P_o = \Theta\left(e^{-\rho}\right).\end{aligned} \tag{24}$$

Furthermore, when considering a *network instance*, the achievable throughput per user for users with communication distances $d_u \geq \sqrt{\frac{\epsilon\rho M}{SN}}$ is:

$$T_1 = \Theta\left(\frac{B}{N}\log_2\left(1+\frac{P_{\max}}{N_0 B_{\mathrm{s}}}\frac{\chi}{\left(\frac{\rho M}{SN}\right)^{\frac{\alpha}{2}}}\right)+\frac{BS}{\rho M}\right); \tag{25}$$

the achievable throughput per user for users with communication distances $d_u < \sqrt{\frac{\epsilon\rho M}{SN}}$ is:

$$T_2 = \Theta\left(\frac{B}{2}\frac{\log_2\left(1+\frac{P_{\max}}{N_0 B_{\mathrm{s}}}\frac{\chi}{\left(\left(\frac{M}{S}\right)^{\frac{1-\gamma}{2-\gamma}}\frac{1}{N}\right)^{\frac{\alpha}{2}}}\right)}{\left(\left(\frac{S}{M}\right)^{\frac{1-\gamma}{2-\gamma}}\right)N}\right)+\Theta(B). \tag{26}$$

Proof. See Appendix G of [2]. □

Remark 9: By comparing between Theorems 4 and 6, we see that the proposed outer bound is achievable. However, different from scenario 1 where the optimality can be achieved without resorting to link-level power control and scheduling, the achievable scheme of scenario 2 exploits the link-level power control and scheduling to enhance the throughput of users in the second time-slot so that the overall throughput is increased. Note that the use of different cluster sizes for different time-slots implies that power control and scheduling are used in link-level such that TX-RX pairs scheduled in different time-slots can follow the required cluster size and scheduling. This indicates that the link-level power control and scheduling can significantly improve the network throughput at the cost of some degree of fairness.

IV. Throughput-Outage Analysis for MZipf Distribution with $\gamma > 1$

In this section, we analyze the throughput-outage performance considering $\gamma > 1$ and $q = o(M)$. Similar to Sec. III, we will in this section first derive the outer bounds for scenario 1 and scenario 2 with the throughput-enhancing approach, and then provide the achievable schemes along with the corresponding throughput-outage performance analysis. We start the analysis by providing Lemma 3 and Theorem 7 which characterize the outage probability lower bound and the transport capacity upper bound, respectively.

Lemma 3 (Lemma 8 in [13]): Suppose $\gamma > 1$. Considering $q = o(M)$, we have the following results: (i) when a user searches through $n_{\mathrm{s}} = o\left(\frac{q}{S}\right)$ different users in the network, we obtain $p_{\text{miss}}(n) \geq 1 - o(1)$; and (ii) when a user searches through $n_{\mathrm{s}} = \frac{\alpha_1' q}{S} < \frac{M}{S}$ different users, where $\alpha_1' = \Omega\left(1\right)$ but $\alpha_1' = \mathcal{O}\left(q^{\frac{1}{\gamma-1}}\right)$, we obtain $p_{\text{miss}}(n) \geq \Theta\left(\frac{1}{(\alpha_1')^{\gamma-1}}\right)$.

Theorem 7: We denote the set of TX-RX pairs as Γ and define r_u as the communication distance for the TX-RX pair u. Let $M \to \infty$ and $N \to \infty$. Suppose $\gamma > 1$ and $\alpha_1' q = o(M)$. We let $R_0' = \epsilon_0\sqrt{\frac{\alpha_1' q}{SN}}$, where ϵ_0 is some small constant. We denote the transport capacity of the network consisting of Γ, defined in terms of meter-bits/s, as $C_\Gamma = \sum_{u\in\Gamma} r_u C_u$, where C_u is the average throughput (bits/s) of user u. We denote the set of TX-RX pairs which have the largest powers among the TX-RX pairs in their corresponding time-frequency resources as $\mathcal{W}$. Under the generalized physical model, C_Γ is upper bounded as:

$$\begin{aligned}C_\Gamma \leq\ & B\overline{C}_{\mathcal{W}} + B\overline{C}_{\Gamma_{R_0'}} \\ & + B\frac{\log_2(e)}{\epsilon_0}\sqrt{\frac{SN}{\alpha_1' q}}\left(\alpha\left(3\sqrt{2}+1\right)+2(2(\sqrt{2}+1))^{\alpha}\right),\end{aligned} \tag{27}$$

where B is the total bandwidth of the network; $\overline{C}_{\mathcal{W}}$ is the average transport capacity efficiency, defined in terms of meter-bits per second per Hz, of the TX-RX pairs in $\mathcal{W}$; $\overline{C}_{\Gamma_{R_0'}}$ is the average transport capacity efficiency of TX-RX pairs that are not in $\mathcal{W}$ and have communication distances smaller than R_0'. The definitions of $\overline{C}_{\mathcal{W}}$ and $\overline{C}_{\Gamma_{R_0'}}$ are formally given below (107) at the end of Appendix H of [2].

Proof. This can be proved by following the same procedure as in Appendix A of [2]. We thus only illustrate the proof in Appendix H of [2], while some details are omitted for brevity. □

Remark 10: By using Lemma 3, we conclude that to have a non-vanishing probability for a user to obtain the desired file (i.e., $p_{\text{miss}}(n)$ does not go to 1), with high probability, the distance between the TX and RX is at least $\Theta\left(\sqrt{\frac{\alpha_1' q}{SN}}\right)$, where $\alpha_1' = \Omega(1)$ and $\alpha_1' = \mathcal{O}\left(q^{\frac{1}{\gamma-1}}\right)$.

A. Outer Bound Analysis for Scenario 1

In scenario 1, we assume equal-throughput for users. Then, by following the same procedure as in Sec. III.A and considering Theorem 7 and Remark 10, we can obtain the following theorem:

Theorem 8: Let $M \to \infty$, $N \to \infty$, and $q \to \infty$. Suppose $\gamma > 1$ and $\alpha_1' q = o(M)$. Assume that the powers of users in the network are upper bounded by $P_{\max}$. When $\alpha_1' = \Omega(1)$ is large enough and $\alpha_1' = \mathcal{O}\left(q^{\frac{1}{\gamma-1}}\right)$, the throughput-outage performance of the network is dominated by (28) on the top of next page, where P_o can be arbitrarily small or converging to zero.

$$\begin{aligned} T(P_o) &= \Theta\left(\frac{B}{N}\log_2\left(1+\frac{P_{\max}}{N_0B_{\mathrm{s}}}\frac{\chi}{\left(\frac{\alpha_1'q}{SN}\right)^{\frac{\alpha}{2}}}\right)+B\log_2(e)\left(\alpha\left(3\sqrt{2}+1\right)+2(2(\sqrt{2}+1))^{\alpha}\right)\frac{S}{\alpha_1'q}\right), \\ P_o &= \Theta\left(\frac{1}{(\alpha_1')^{\gamma-1}}\right). \end{aligned} \tag{28}$$

Proof. The proof can be done by directly using Lemma 3 and Theorem 7 and following the similar proof of proving Theorem 2. We thus omit the proof for brevity. □

Remark 11: Theorem 8 shows that when $\gamma > 1$, q is the dominant factor. In addition, similar to Remark 3, it shows that when the maximum transmit power is some constant, the throughput-outage performance bound considering the generalized physical model has the same scaling law as that considering the protocol model [12]. As a result, this again indicates that the link-level power allocation and scheduling cannot improve the asymptotic throughput-outage performance when users are served under equal-throughput assumption.

B. Outer Bound Analysis for the Proposed Throughput-Enhancing Approach in Scenario 2

In this subsection, we consider scenario 2, where the network is not confined to the equal-throughput assumption. Similar to the case with $\gamma < 1$, we can benefit from letting TX-RX pairs with small communication distances to transmit with a much higher throughput. Hence, we again adopt the throughput-enhancing approach proposed in Sec. III.B, where the first time-slot is used for ordinary file delivery and the second time-slot is used to enhance the overall throughput by letting TX-RX pairs with communication distances smaller than $\sqrt{\epsilon'}R_0'$, where $\epsilon' = \mathcal{O}(1)$, transmit with high speed. By adopting the proposed throughput-enhancing approach, we can obtain the following theorem:

Theorem 9: Let $M \to \infty$, $N \to \infty$, and $q \to \infty$. Suppose $\gamma > 1$ and $\alpha_1'q = o(M)$. Suppose the double time-slot approach discussed in Sec. III.B is used and the $\epsilon' = \mathcal{O}(1)$ is selected. Assume that the powers of users in the network are upper bounded by $P_{\max}$. When $\alpha_1' = \Omega(1)$ is large enough, $\alpha_1' = \mathcal{O}\left(q^{\frac{1}{\gamma-1}}\right)$, and λ_2 is feasible, the throughput-outage performance of the network is dominated by:

$$\begin{aligned} T(P_o) &= \mathcal{O}\left(\frac{B}{2}\frac{\log_2\left(1+\frac{P_{\max}}{N_0B_{\mathrm{s}}}\frac{\chi}{\left(\frac{\epsilon'\alpha_1'q}{SN}\right)^{\frac{\alpha}{2}}}\right)}{N}+\frac{B}{2}\frac{S}{\epsilon'\alpha_1'q}\right), \\ P_o &= \Theta\left(\frac{1}{(\alpha_1')^{\gamma-1}}\right). \end{aligned} \tag{29}$$

Furthermore, when considering a *network instance*, the throughput per user for users with communication distances $d_u > \sqrt{\epsilon'}R_0'$ is dominated by:

$$\lambda_1 = \mathcal{O}\left(\frac{B}{N}\log_2\left(1+\frac{P_{\max}}{N_0B_{\mathrm{s}}}\frac{\chi}{\left(\frac{\alpha_1'q}{SN}\right)^{\frac{\alpha}{2}}}\right)+\frac{BS}{\alpha_1'q}\right); \tag{30}$$

the throughput per user for users with communication distances $d_u \leq \sqrt{\epsilon'}R_0'$ is dominated by:

$$\lambda_2 = \Theta\left(\frac{B}{2}\frac{\log_2\left(1+\frac{P_{\max}}{N_0B_{\mathrm{s}}}\frac{\chi}{\left(\frac{\epsilon'\alpha_1'q}{SN}\right)^{\frac{\alpha}{2}}}\right)}{\epsilon'\alpha_1'N}\right)+\Theta\left(\frac{BS}{(\epsilon'\alpha_1')^2q}\right). \tag{31}$$

Proof. See Appendix I of [2]. □

By Theorem 9, we understand that minimizing ϵ' can maximize the overall network throughput. However, there is also a lower bound on ϵ' due to the physical limitation of the link-rate. The feasibility condition we need to satisfy is again given as:

$$\lambda_2 \leq \eta, \tag{32}$$

where η is some constant. Then, notice that by using Theorem 7, the throughput per user in the second time-slot satisfies

$$\lambda_2\delta'N\Theta\left(\sqrt{\frac{\epsilon'\alpha_1'q}{SN}}\right) = \mathcal{O}\left(B\log_2\left(1+\frac{P_{\max}}{N_0B_{\mathrm{s}}}\frac{\chi}{\left(\frac{\epsilon'\alpha_1'q}{SN}\right)^{\frac{\alpha}{2}}}\right)\sqrt{\frac{\epsilon'\alpha_1'q}{SN}}+B\sqrt{\frac{SN}{\epsilon'\alpha_1'q}}\right), \tag{33}$$

and that $P_{\max}$ is some constant. It follows that we have

$$\lambda_2\delta' = \mathcal{O}\left(\frac{BS}{\epsilon'\alpha_1'q}\right). \tag{34}$$

Then, by the derivations in Appendix I of [2] and by using the same arguments as in Sec. III.B, we know from (115) in Appendix I of [2] that

$$\delta' = \Theta\left(\epsilon'\alpha_1'\right) \tag{35}$$

should be adopted. It follows that we must have

$$\eta(\epsilon'\alpha_1') = \Theta\left(\frac{BS}{\epsilon'\alpha_1'q}\right) \tag{36}$$

if we want to minimize ϵ' subject to the feasibility condition and the tightness of the upper bound. Since η is some constant, this then leads to

$$\epsilon'\alpha_1' = \Theta\left(\left(\frac{S}{q}\right)^{\frac{1}{2}}\right). \tag{37}$$

Finally, by using Theorem 9 and (37), we obtain the following theorem:

Theorem 10: Let $M \to \infty$, $N \to \infty$, and $q \to \infty$. Suppose $\gamma > 1$ and $\alpha_1' q = o(M)$. Suppose the double time-slot framework discussed in Sec. III.B is used and the $\epsilon' = \mathcal{O}(1)$ is selected. Assume that the powers of users in the network are upper bounded by P_{max}. When $\alpha_1' = \Omega(1)$ is large enough and $\alpha_1' = \mathcal{O}\left(q^{\frac{1}{\gamma-1}}\right)$, the throughput-outage performance of the network is dominated by:

$$T(P_o) = \Theta\left(\frac{B}{2}\frac{\log_2\left(1+\frac{P_{\text{max}}}{N_0 B_{\text{s}}}\frac{\chi}{\left(\left(\frac{q}{S}\right)^{\frac{1}{2}}\frac{1}{N}\right)^{\frac{\alpha}{2}}}\right)}{N} + \frac{B}{2}\left(\frac{S}{q}\right)^{\frac{1}{2}}\right), P_o = \Theta\left(\frac{1}{(\alpha_1')^{\gamma-1}}\right). \tag{38}$$

Furthermore, when given a *network instance*, the throughput per user for users with communication distances $d_u > \sqrt{\epsilon'}R_0'$ is dominated by:

$$\lambda_1 = \Theta\left(\frac{B}{N}\log_2\left(1+\frac{P_{\text{max}}}{N_0 B_{\text{s}}}\frac{\chi}{\left(\frac{\alpha_1' q}{SN}\right)^{\frac{\alpha}{2}}}\right) + \frac{BS}{\alpha_1' q}\right); \tag{39}$$

the throughput per user for users with communication distances $d_u \leq \sqrt{\epsilon'}R_0'$ is dominated by:

$$\lambda_2 = \Theta\left(\frac{B}{2}\frac{\log_2\left(1+\frac{P_{\text{max}}}{N_0 B_{\text{s}}}\frac{\chi}{\left(\left(\frac{q}{S}\right)^{\frac{1}{2}}\frac{1}{N}\right)^{\frac{\alpha}{2}}}\right)}{\left(\left(\frac{S}{q}\right)^{\frac{1}{2}}\right)N}\right) + \Theta(B). \tag{40}$$

Proof. This is directly obtained by applying (37) to Theorem 9. □

Remark 12: Similar to Remarks 6 and 7, we observe that the double time-slot approach can significantly improve the throughput performance without sacrificing the outage probability. The benefit is again from that we judiciously let TX-RX pairs with very small communication distances transmit at a much higher throughput in the second time-slot. In addition, as indicated by Theorem 10, the double time-slot approach again decouples the tradeoff between throughput and outage and converts the throughput-outage tradeoff to the fairness-outage tradeoff.

C. Achievable Scheme and Analysis for Scenario 1

In this subsection, we provide the achievable scheme and its corresponding analysis for scenario 1, in which the equal-throughput assumption is considered. We consider the achievable scheme similar to that in Sec. III.C, which is as follows. We first equally split the transmission period T' into two equally-sized time-slots. Then, we let the transmissions in first time-slot follow the TDMA approach and let the transmissions in second time-slot follow the clustering approach. Specifically, for the clustering, we split the network into equally-sized square clusters in which the side length of each cluster is $d = \Theta\left(\sqrt{\frac{\alpha_1 q}{SN}}\right)$. Users in a cluster can only obtain its desired file through users in the same cluster. In each cluster, a user is served at a time and users in the same cluster are served in a round-robin manner. Interference between different clusters are avoided via the frequency reuse approach with reuse factor $(2(K+1))^2$, where $K \in \mathcal{N} > 0$ is a finite positive integer. Then, by symmetry of the clusters and by the fact that $N \to \infty$ and that $d = o(1)$, we can follow the same arguments in Sec. III.C such that the number of users in a cluster is given as $N_{\text{cluster}} \backsim F_{\text{Poi}}\left(\frac{\alpha_1 q}{S}\right)$, where $\frac{\alpha_1 q}{S}$ is the mean value of the Poisson distribution. It follows again by Lemma 1 in [13], the outage probability is

$$p_o = \sum_{f=1}^{M} P_r(f) e^{-\frac{\alpha_1 q}{S}P_c(f)}, \tag{41}$$

where $P_c(f), \forall f$ are determined by the caching policy provided in Theorem 1 of [13]. With the above clustering and caching policy, the same scheduling approach as in Sec. III.C is adopted, and we thus can obtain the following theorem:

Theorem 11: Let $M \to \infty$, $N \to \infty$, and $q \to \infty$. Suppose that $\gamma > 1$ and that $q = o(M)$. Assume that the powers of users in the network are upper bounded by P_{max}. Consider the caching policy in Theorem 1 of [13] and that the side length of a cluster is $\sqrt{\frac{\alpha_1 q}{SN}}$, where $\Theta\left(\frac{\alpha_1 q}{S}\right) = o(M)$. Assume equal-throughput transmissions of users. When $\alpha_1 = \Omega(1)$ is large enough and $\alpha_1' = \mathcal{O}\left(q^{\frac{1}{\gamma-1}}\right)$, the following throughput-outage performance of the network is achievable

$$\begin{aligned} T(P_o) &= \Theta\left(\frac{1}{N}\log_2\left(1+\frac{P_{\text{max}}}{N_0 B}\frac{\chi}{\left(\frac{\alpha_1 q}{SN}\right)^{\frac{\alpha}{2}}}\right) + \frac{BS}{\alpha_1 q}\right), \\ P_o &= \Theta\left(\frac{1}{(\alpha_1)^{\gamma-1}}\right), \end{aligned} \tag{42}$$

where P_o can be very small or converging to zero.

Proof. See Appendix K of [2]. □

Remark 13: By comparing between Theorems 8 and 11, we see that the proposed outer bound is achievable. This indicates that when the equal-throughput assumption is considered, the simple clustering scheme is asymptotically optimal even though we are allowed to use the link-level power allocation and scheduling.

D. Achievable Scheme and Analysis for Scenario 2 with Throughput-Enhancing Approach

In this subsection, the achievable scheme and corresponding analysis for scenario 2 with the throughput-enhancing approach are provided. We consider the achievable scheme having the same framework as that in Sec. III.D. Therefore, we again split the transmission duration T' into two time-slots; each has the duration of $\frac{T'}{2}$. We assume without loss of generality that S is an even number. Then, in the first timeslot, we adopt the clustering with side length $d_1 = \sqrt{\frac{\alpha_1 q}{SN}}$; in the second timeslot, we adopt the clustering with side length $d_2 = \sqrt{\frac{\epsilon\alpha_1 q}{SN}}$, where $\epsilon = \mathcal{O}(1)$. For the caching scheme, we split the whole cache space into two subspace, in which each subspace has size $\frac{S}{2}$. For the first caching subspace, we consider the caching policy proposed in Theorem 1 of [13] with $g_{c,1}(M) = \frac{2\alpha_1 q}{S}$; for the second caching subspace, we adopt the same caching policy and let $g_{c,2}(M) = \frac{2\epsilon\alpha_1 q}{S}$. By the above described scheme, we can then obtain the following theorem:

Theorem 12: Let $M \to \infty$, $N \to \infty$, and $q \to \infty$. Suppose that $\gamma > 1$ and that $q = o(M)$. Assume that the powers of users in the network are upper bounded by P_{max}. Suppose the proposed achievable scheme in Sec. IV.D is used and that the ϵ is selected such that $\epsilon\alpha_1 = \Theta\left(\left(\frac{S}{q}\right)^{\frac{1}{2}}\right)$. Then, when $\alpha_1 = \Omega(1)$ is large enough, the following throughput-outage performance of the network is achievable:

$$T(P_o) = \Theta\left(\frac{B}{2}\frac{\log_2\left(1 + \frac{P_{\text{max}}}{N_0 B_{\text{s}}}\frac{\chi}{\left(\left(\frac{q}{S}\right)^{\frac{1}{2}}\frac{1}{N}\right)^{\frac{\alpha}{2}}}\right)}{N} + \frac{B}{2}\left(\frac{S}{q}\right)^{\frac{1}{2}}\right)$$
$$P_o = \Theta\left(\frac{1}{(\alpha_1')^{\gamma-1}}\right). \tag{43}$$

Furthermore, when given a *network instance*, the achievable throughput per user for users with communication distances $d_u > \sqrt{\frac{\epsilon'\alpha_1 q}{SN}}$ is:

$$T_1 = \Theta\left(\frac{B}{N}\log_2\left(1 + \frac{P_{\text{max}}}{N_0 B}\frac{\chi}{\left(\frac{\alpha_1 q}{SN}\right)^{\frac{\alpha}{2}}}\right) + \frac{BS}{\alpha_1 q}\right); \tag{44}$$

the achievable throughput per user for users with communication distances $d_u \leq \sqrt{\frac{\epsilon'\alpha_1 q}{SN}}$ is:

$$T_2 = \Theta\left(\frac{B}{2}\frac{\log_2\left(1 + \frac{P_{\text{max}}}{N_0 B_{\text{s}}}\frac{\chi}{\left(\left(\frac{q}{S}\right)^{\frac{1}{2}}\frac{1}{N}\right)^{\frac{\alpha}{2}}}\right)}{\left(\left(\frac{S}{q}\right)^{\frac{1}{2}}\right)N}\right) + \Theta(B). \tag{45}$$

Proof. See Appendix L of [2]. □

Remark 14: By comparing between Theorems 10 and 12, we see that the proposed outer bound is achievable. Similar to Remark 9, the achievable scheme of scenario 2 requires the link-level power control and scheduling to enhance the overall throughput. In addition, we stress that although the throughput scaling law $\Theta\left(\sqrt{\frac{S}{q}}\right)$ of scenario 2 gives the same scaling law as the scaling law of the multi-hop cache-aided D2D networks derived in [13], we should not misinterpret that the single-hop cache-aided D2D with link-level power control and scheduling can have the same performance as the multi-hop cache-aided D2D. This is because comparing our result here with the result derived in [13] is slightly unfair as [13] indeed provides the scaling law considering the equal-throughput assumption. In other word, the single-hop cache-aided D2D with link-level power control and scheduling can have the same performance as its multi-hop counterpart only if the unfairness (from the realization level) is introduced, while the fairness is retained in the multi-hop case.

V. Throughput-Outage Analysis for Zipf Distribution with $\gamma > 1$

In this section, we consider the analysis for networks considering the Zipf distribution, i.e., $q = 0$ or equivalently $q = \Theta(1)$. We note that since M is dominant in the case of $\gamma < 1$, we are not interested in the case that $\gamma < 1$ for the Zipf distribution as the case of Zipf distribution with $\gamma < 1$ should have the same throughput-outage scaling law as that of the MZipf distribution with $\gamma < 1$. On the other hand, we are interested in the case that $\gamma > 1$ for the Zipf distribution only when $P_{\text{max}} \to \infty$ is allowed with $\frac{P_{\text{max}}}{N} = \mathcal{O}(1)$. This is because we already know that when P_{max} is some constant, the throughput per user is also upper bounded by some constant, and such scaling law has already been achieved without resorting to the link-level power control and scheduling in the literature [12]. Hence, we in this section only focus on whether allowing instantaneous power to be infinity while confining the average power to be constant, i.e., $P_{\text{max}} \to \infty$ with $\frac{P_{\text{max}}}{N} = \mathcal{O}(1)$, can bring additional performance gain. Since we can consider $\gamma > 1$ and $q = \Theta(1)$ for deriving the scaling law for the Zipf case, we indeed can use the same procedure of proving Theorems 8 and 11 while considering $q = \Theta(1)$. This then leads to the following theorem:

Theorem 13: Let $M \to \infty$ and $N \to \infty$. Suppose that $\gamma > 1$ and $q = \Theta(1)$. Consider the equal-throughput assumption. When $\alpha_2' = \Theta(1)$ is large enough, the optimal throughput-outage performance of the network is:

$$T(P_o) = \Theta\left(\frac{B}{2}\frac{\log_2\left(1 + \frac{P_{\text{max}}}{N_0 B_{\text{s}}}\frac{\chi}{\left(\left(\frac{1}{S}\right)^{\frac{\alpha}{2}}\frac{1}{N}\right)^{\frac{\alpha}{2}}}\right)}{N}\right) + \Theta\left(\frac{BS}{\alpha_2'}\right),$$
$$P_o = \epsilon_{\text{zip}}(\alpha_2'), \tag{46}$$

where $\epsilon_{\text{zip}}(\alpha_2')$ can be arbitrarily small.

Proof. The proof simply follows the procedure for proving Theorems 8 and 11. We thus omit the proof for brevity. □

By Remark 2, we understand that $\frac{\chi}{\left(\left(\frac{1}{S}\right)^{\frac{\alpha}{2}}\frac{1}{N}\right)^{\frac{\alpha}{2}}}$ should be upper bounded by 1. Therefore, from Theorem 13, we observe that the throughput per user is:

$$T(P_o) = \Theta\left(\frac{B}{2}\frac{\log_2\left(1 + \frac{P_{\text{max}}}{N_0 B_{\text{s}}}\right)}{N}\right) + \Theta\left(\frac{BS}{\alpha_2'}\right). \tag{47}$$

Then, since we know $\frac{1}{N}\log_2(P_{\text{max}}) \to 0$ when $\frac{P_{\text{max}}}{N} = \mathcal{O}(1)$, we obtain the following corollary:

Corollary 2: Let $M \to \infty$ and $N \to \infty$. Suppose that $\gamma > 1$ and $q = \Theta(1)$. Consider the equal-throughput assumption and $\frac{P_{\max}}{N} = \mathcal{O}(1)$. When $\alpha_2' = \Theta(1)$ is large enough, the throughput-outage performance of the network is dominated by:

$$T(P_o) = \Theta\left(\frac{BS}{\alpha_2'}\right), P_o = \epsilon_{\text{zip}}(\alpha_2'), \tag{48}$$

where $P_o = \epsilon_{\text{zip}}(\alpha_2')$ can be arbitrarily small.

From Corollary 2, we understand that even if we allow the instantaneous power to be infinity while still confining the average power to be some constant, the best average throughput per user is some constant with the outage probability being arbitrarily small. However, since such throughput-outage performance can be achieved without letting the instantaneous power go to infinity, we conclude the following remark:

Remark 15: When considering Zipf distribution with $\gamma > 1$, even if we allow the instantaneous power to be infinity while still confining to that the constant average power, the throughput-outage performance cannot be improved by the capability of having infinite instantaneous power. Therefore, there is no need to let the instantaneous power to be infinity. In practice, this implies that the network is an interference-limited network, which is in line with our expectation.

Remark 16: Although the conclusion in Remark 15 is derived assuming the scenario 1, such conclusion applies to scenario 2. This is because it is not possible to obtain an order gain by letting TX-RX pairs with (orderwise) smaller distances to transmit with much higher throughput as the number of TX-RX pairs with (orderwise) smaller distances goes to zero when considering the Zipf distribution with $\gamma > 1$.

VI. Finite-Dimensional Simulations

In this section, we conduct simulations to better illustrate our results. Specifically, in Fig. 1, we obtain numerical values for the throughput-outage performance provided in our theorems with $S = 10$, $M = 10,000$ and $q = 20$, where the normalized throughput per user indicates that the constant terms of the theoretical results are omitted so that we can focus on the scaling behaviors. From the figure, we observe that the throughput-outage performance of the network considering scenario 2 can significantly outperform the corresponding performance considering scenario 1. This validates that the proposed throughput-enhancing approach is very effective as indicated in our theoretical analysis. In addition, we also see that the throughput of the networks in scenario 2 can be maintained to be constant when decreasing the outage probability. This again corresponds to our results that the throughput-enhancing approach can convert the throughput-outage tradeoff to the fairness-outage tradeoff, as decreasing the outage probability in this case indeed leads to higher unfairness instead of lower throughput.

We then conduct Monte-Carlo simulations to further demonstrate that the insights and implications derived from our theoretical results can be observed and exploited in practice. Specifically, we consider evaluating the throughput-outage performance of $N = 100,000$ users uniformly distributed in a square area with size 1 km^2. We consider $\chi = 1$, $\alpha = 3.5$, and $N_0 = -174$ dBm for the pathloss model and consider the path power gain is upper bounded by 1. We consider the transmissions to be successful if the SINR is larger than 0 dB, and set the link-rate to be 1 bit/s/Hz as long as the transmissions can be successful. We assume the bandwidth is $B = 1$ Hz for simplicity and consider $S = 10$ and $M = 10,000$. We consider using the proposed achievable schemes in our paper for both scenarios 1 and 2, where the clustering approach is adopted. In addition, to obtain different throughput-outage performance in the figures, we adjust the cluster sizes so that the throughput and outage probability can trade off against each other. However, since the achievable schemes for scenario 2 adopt the throughput-enhancing approaches for the small-distance TX-RX pairs, the cluster sizes used by the throughput-enhancing approaches should be different from the cluster sizes used for the general

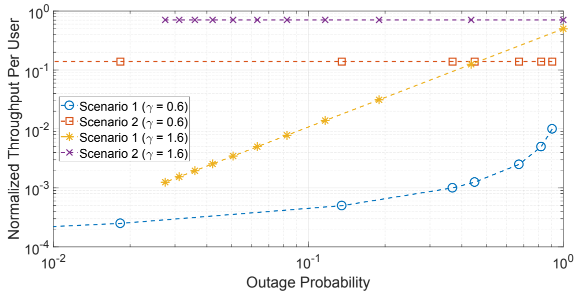
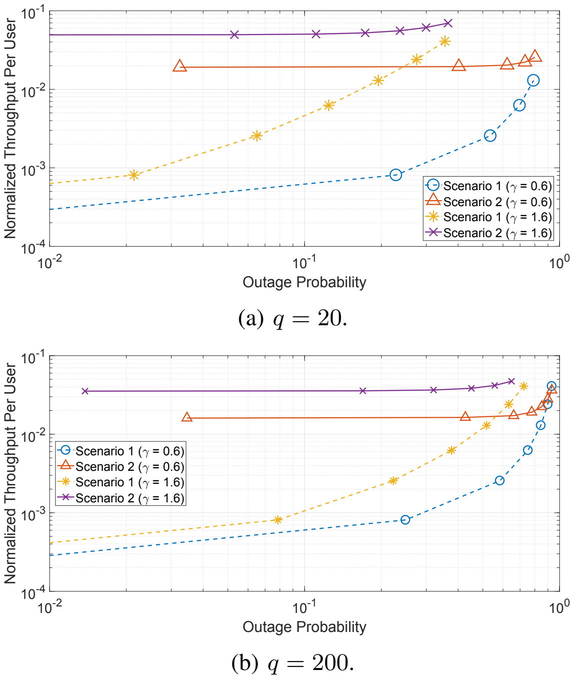


Fig. 1: Numerical evaluations for the throughput-outage performance of scenarios 1 and 2 with $S = 10$, $M = 10,000$, and $q = 20$.



(a) $q = 20$.

(b) $q = 200$.

Fig. 2: Monte-Carlo evaluations for the throughput-outage performance of clustering networks considering scenarios 1 and 2 with $S = 10$, $M = 10,000$, and $N = 100,000$.

TX-RX pairs. Thus, we let the cluster size used by small-distance TX-RX pairs to be 14.92 meters and 10 meters for cases that $\gamma = 0.6$ and $\gamma = 1.6$, respectively. We assume the powers and frequency reuse factors used for the simulations are manually selected without using advanced optimization techniques as their dedicated optimization approaches are not our focus.

Results are shown in Fig. 2. We observe that the networks considering scenario 2 can significantly outperform the networks considering scenario 1 in all cases. Besides, we can see that the throughput-outage tradeoff for curves with scenario 2 is not obvious, validating our remark that the throughput-outage tradeoff has been converted to the fairness-outage tradeoff when the throughput-enhancing approach is used in scenario 2. Finally, we see that the change of q is more influential in the case that $\gamma > 1$ and the throughput-outage performance in the case of $\gamma > 1$ is better than that of $\gamma < 1$. These also validate our theorems that q is dominant for the throughput-outage scaling laws considering $\gamma > 1$, while M is dominant when considering $\gamma < 1$.

VII. Conclusions and Discussions

This paper investigated the throughput-outage scaling laws of cache-aided single-hop D2D networks considering the generalized physical channel. The main purpose of this paper was to understand whether and how including link-level power control and scheduling can improve the scaling laws. Results showed that when the equal-throughput assumption is considered, i.e., users are served fairly on the realization level, having link-level power control and scheduling cannot improve the scaling laws. On the other hand, if the equal-throughput assumption is dropped, i.e., we allow users to have different throughput on the realization level, the scaling laws indeed can be significantly improved by first appropriately using link-level power control and scheduling, and then letting TX-RX pairs with small communication distances to transmit at high speed. These results implied that if we want the users to always be fairly served, the link-level power control and scheduling might not have a significant influence even if a very clever link-level power control and scheduling is used. On the other hand, when we only schedule the users such that they are on average fairly served, while for each instant, link variations are exploited, the network performance can be significantly improved by some advanced link-level power control and scheduling.

Our results are expected to motivate the following practical applications: (i) since we know that the scaling law considering the link-level power control and scheduling can only be improved when the users are served with different throughput in each realization, we should focus on exploiting such property when designing the power allocation and scheduling policy; (ii) to improve the scaling law of single-hop D2D caching networks, we should first distinguish between TX-RX pairs with small and large communication distances, and then let the TX-RX pairs with small communication distances to transmit with high speed. Therefore, when designing scheduling and user selection/pairing policy, this property should be exploited; and (iii) since our study characterizes the fundamental scaling law with respect to the increase of the cache space, when evaluating a single-hop D2D caching network, our scaling law can be used as the reference regarding whether the network to be evaluated is already performing well or there is still room for improvement.

Although this paper has conducted the analysis to certain extent, there still remain some possible future directions. First of all, we in this paper focus only on the double time-slot approach for throughput enhancement in scenario 2, i.e., we only distinguish the TX-RX pairs into two types by their communication distances. Clearly, such approach might be extended to considering more than two types of TX-RX pairs, and whether such extension can bring further performance gain is unclear. Besides, our results are derived under the assumption that users are uniformly distributed within the network. Thus, our analysis can only be considered as the *statistically worst-case* analysis, and different scaling laws might be derived when different user distribution assumptions are considered.

References

[1] M.-C. Lee, A. F. Molisch, and M. Ji, “Throughput–outage scaling laws for wireless single-hop D2D caching networks with physical models,” in *2021 IEEE International Conference on Communications (ICC)*, June 2021.

[2] ——, “Throughput-outage scaling behaviors for wireless single-hop D2D caching networks with physical model – analysis and derivations,” Jul. 2021, online available at https://arxiv.org/abs/2106.00300.

[3] “Cisco virtual networking index: Global mobile data traffic forecast update, 2017-2022,” San Jose, CA, USA, Tech. Rep.

[4] 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.

[5] Z. Zhang, Y. Xiao, Z. Ma, M. Xiao, Z. Ding, X. Lei, G. K. Karagiannidis, and P. Fan, “6g wireless networks: Vision, requirements, architecture, and key technologies,” *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 28–41, 2019.

[6] J. G. Andrews, S. Buzzi, W. Choi, and et. al., “What will 5g be?” *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, January 2014.

[7] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, “Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution,” *IEEE Commmun. Mag.*, vol. 51, no. 4, pp. 142–149, April 2013.

[8] M. Ji, G. Caire, and A. F. Molisch, “Wireless device-to-device caching networks: Basic principles and system performance,” *IEEE J. Sel. Area Commun.*, vol. 34, no. 1, pp. 176–189, January 2016.

[9] 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.

[10] M. A. Maddah-Ali and U. Niessen, “Fundamental limits of caching,” *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.

[11] 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.

[12] 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.

[13] 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.

[14] I. Ahmed, M. H. Ismail, and M. S. Hassan, “Video transmission using device-to-device communications: A survey,” *IEEE Access*, vol. 7, pp. 131 019–131 038, 2019.

[15] M. Mehrabi, D. You, V. Latzko, H. Salah, M. Reisslein, and F. H. P. Fitzek, “Device-enhanced mec: Multi-access edge computing (mec) aided by end device computation and caching: A survey,” *IEEE Access*, vol. 7, pp. 166 079–166 108, 2019.

[16] D. Prerna, R. Tekchandani, and N. Kumar, “Device-to-device content caching techniques in 5g: A taxonomy, solutions, and challenges,” *Computer Communications*, 2020.

[17] X. Li, X. Wang, K. Li, Z. Han, and V. C. M. Leung, “Collaborative multi-tier caching in heterogeneous networks: Modeling, analysis, and design,” *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6926–6939, 2017.

[18] R. Q. Hu *et al.*, “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] P. Gupta and P. R. Kumar, “The capacity of wireless networks,” *IEEE Trans. on inf. theory*, vol. 46, no. 2, pp. 388–404, 2000.

[20] A. Agarwal and P. R. Kumar, “Capacity bounds for ad hoc and hybrid wireless networks,” *ACM SIGCOMM Computer Communication Review*, vol. 34, no. 3, pp. 71–81, 2004.

[21] F. Xue and P. R. Kumar, “Scaling laws for ad hoc wireless networks: an information theoretic approach,” *Foundations and Trends® in Networking*, vol. 1, no. 2, pp. 145–270, 2006.

[22] M. Franceschetti, O. Dousse, D. N. Tse, and P. Thiran, “Closing the gap in the capacity of wireless networks via percolation theory,” *IEEE Trans. Inf. Theory*, vol. 53, no. 3, pp. 1009–1018, 2007.

[23] A. El Gamal, J. Mammen, B. Prabhakar, and D. Shah, “Optimal throughput-delay scaling in wireless networks-part i: The fluid model,” *IEEE Trans. on Inf. Theory*, vol. 52, no. 6, pp. 2568–2592, 2006.

[24] S. Shakkottai, X. Liu, and R. Srikant, “The multicast capacity of large multihop wireless networks,” *IEEE/ACM Trans Networking*, vol. 18, no. 6, pp. 1691–1700, 2010.

[25] A. Ozgur, O. Lévêque, and D. N. Tse, “Hierarchical cooperation achieves optimal capacity scaling in ad hoc networks,” *IEEE Trans. inf. theory*, vol. 53, no. 10, pp. 3549–3572, 2007.

[26] T. Hara, “Effective replica allocation in ad hoc networks for improving data accessibility,” in *IEEE INFOCOM 2001*, vol. 3. IEEE, 2001, pp. 1568–1576.

[27] E. Cohen and S. Shenker, “Replication strategies in unstructured peer-to-peer networks,” *ACM SIGCOMM Computer Communication Review*, vol. 32, no. 4, pp. 177–190, 2002.

[28] T. Hara and S. K. Madria, “Data replication for improving data accessibility in ad hoc networks,” *IEEE transactions on mobile computing*, vol. 5, no. 11, pp. 1515–1532, 2006.

[29] J. Zhao, P. Zhang, G. Cao, and C. R. Das, “Cooperative caching in wireless p2p networks: Design, implementation, and evaluation,” *IEEE Trans. Parallel and Distributed Systems*, vol. 21, no. 2, pp. 229–241, 2010.

[30] N. Golrezaei, A. D. Dimakis, and A. F. Molisch, “Scaling behavior for device-to-device communications with distributed caching,” *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4286–4298, July 2014.

[31] M. Ji, G. Caire, and A. F. Molisch, “Fundamental limits of distributed caching in d2d wireless networks,” in *2013 IEEE Information Theory Workshop (ITW)*. IEEE, 2013, pp. 1–5.

[32] S. Gitzenis, G. S. Paschos, and L. Tassiulas, “Asymptotic laws for joint content replication and delivery in wireless networks,” *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 2760–2776, 2012.

[33] S.-W. Jeon, S.-N. Hong, M. Ji, G. Caire, and A. F. Molisch, “Wireless multihop device-to-device caching networks,” *IEEE Trans. Inf. Theory*, vol. 63, no. 3, pp. 1662–1676, 2017.

[34] L. Qiu and G. Cao, “Popularity-aware caching increases the capacity of wireless networks,” *IEEE Trans. Mobile Comput.*, vol. 19, no. 1, pp. 173–187, 2019.

[35] J. Guo, J. Yuan, and J. Zhang, “An achievable throughput scaling law of wireless device-to-device caching networks with distributed mimo and hierarchical cooperations,” *IEEE Trans. on Wireless Commun.*, vol. 17, no. 1, pp. 492–505, 2017.

[36] J. Ren, D. Li, L. Zhang, and G. Zhang, “Scaling performance analysis and optimization based on the node spatial distribution in mobile content-centric networks,” *Wireless Communications and Mobile Computing*, vol. 2021, Jan. 2021.

[37] M. Ji, R.-R. Chen, G. Caire, and A. F. Molisch, “Fundamental limits of distributed caching in multihop d2d wireless networks,” in *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017, pp. 2950–2954.

[38] M. Ji, G. Caire, and A. F. Molisch, “Fundamental limits of caching in wireless d2d networks,” *IEEE Trans. Inf. Theory*, vol. 62, no. 2, pp. 849–869, February 2016.

[39] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr, “Fundamental limits of cache-aided interference management,” *IEEE Transactions on Information Theory*, vol. 63, no. 5, pp. 3092–3107, 2017.

[40] Ç. Yapar, K. Wan, R. F. Schaefer, and G. Caire, “On the optimality of d2d coded caching with uncoded cache placement and one-shot delivery,” *IEEE Trans. Commun.*, vol. 67, no. 12, pp. 8179–8192, 2019.

[41] X. Zhang, N. Woolsey, and M. Ji, “Cache-aided interference management using hypercube combinatorial design with reduced subpacketizations and order optimal sum-degrees of freedom,” *IEEE Trans. Wireless Commun.*, 2021.

[42] L. Zhang, M. Xiao, G. Wu, and S. Li, “Efficient scheduling and power allocation for d2d-assisted wireless caching networks,” *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2438–2452, June 2016.

[43] B. Chen, C. Yang, and Z. Xiong, “Optimal caching and scheduling for cache-enabled d2d communications,” *IEEE Commun. Lett.*, vol. 21, no. 5, pp. 1155–1158, 2017.

[44] M.-C. Lee and A. F. Molisch, “Caching policy and cooperation distance design for base station assisted wireless d2d caching networks: Throughput and energy efficiency optimization and trade-off,” *IEEE Trans. on Wireless Commun.*, vol. 17, no. 11, pp. 7500–7514, November 2018.

[45] M. Choi, A. F. Molisch, and J. Kim, “Joint distributed link scheduling and power allocation for content delivery in wireless caching networks,” *IEEE Trans. on Wireless Commun.*, vol. 19, no. 12, pp. 7810–7824, 2020.

[46] B. Blaszczyszyn and A. Giovanidis, “Optimal geographic caching in cellular networks,” June 2015.