IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received April 11, 2018, accepted May 7, 2018, date of publication May 10, 2018, date of current version June 5, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2834922

Network Throughput Gain of Multicast With User Caching in Heavy Traffic Downlink

JUN-PYO HONG[1], (Member, IEEE), SEONG HO CHAE[2], (Member, IEEE), AND KISONG LEE[3], (Member, IEEE)

[1]Department of Information and Communications Engineering, Pukyoung National University, Busan 48513, South Korea
[2]Department of Electronics Engineering, Korea Polytechnic University, Siheung 15073, South Korea
[3]School of Information and Communication Engineering, Chungbuk National University, Cheongju 28644, South Korea

Corresponding authors: Seong Ho Chae (shchae@kpu.ac.kr) and Kisong Lee (kslee851105@gmail.com)

This work was supported in part by the Institute for Information and Communications Technology Promotion Grant through the Korean Government (a research on a novel communication system using storage as wireless communication resource) under Grant 2015-0-00820 and in part by the National Research Foundation of Korea Grant through the Korean Government under Grant 2016R1C1B2012173.

ABSTRACT In order to cope with the recent dramatic growth in mobile traffic, it is important to understand and exploit its unique features in wireless communications. For example, the fact that a relatively small number of popular contents accounts for most of the traffic is considered one of the most promising factors to explore. This paper proposes a multicast system with user caching in a heavy traffic model. Through the efficient combination of caching and multicasting, the proposed system reduces the redundant communication resource consumptions for the duplicated requests of the same content from different users. The performance of the proposed system is analyzed in closed-form expressions in terms of outage probability and average network throughput by means of the asymptotic analysis. The derived expressions provide insight into the impact of system parameters on performance and can be used to capture the gains in performance obtained from user caching and multicasting in the proposed system. Furthermore, the simulation results confirm the convergence of the asymptotic analysis results in the heavy traffic model and provide additional information on the average network throughput in general cases.

INDEX TERMS Wireless edge caching, multicast, network throughput, content-centric scheduling, and heavy traffic.

I. INTRODUCTION

The recent emergence of new services related to high quality multimedia, such as image, music, and video, has been accompanied by a dramatic increase in mobile data traffic. According to a recent report [1], global mobile data traffic is expected to increase to 48.3 exabytes per month by 2021. It is almost seven times the amount of mobile data traffic seen in 2016. The capacity of the current cellular network is insufficient to cope with rapidly increasing amounts of mobile data traffic because of physical limitations in the available spectrum, and the data rate of current wireless communication systems is already close to the optimum level. The aim of the fifth generation (5G) mobile communication system is to support 1000-fold gains in capacity, connections for at least 100 billion devices, and a 10Gb/s individual user experience with extremely low latency and short response times [2]. A variety of techniques have been developed to meet these requirements, such as small cells, massive MIMO, and mmWave [3].

In addition to these techniques, *wireless edge caching* has recently been the focus of attention as a promising means of addressing the massive increase in traffic. The motivation for wireless edge caching comes from the following two interesting observations: i) There is frequent content reuse in real traffic situations, in which a small number of popular contents accounts for a majority of the mobile traffic [4], [5]. ii) Mobile devices in use today tend to have large storage capacity for caching contents, thanks to the low cost of memory. Motivated by these factors, the wireless edge caching prefetches some popular contents into the storage space of the network edges, such as access point (AP) and mobile devices, during off-peak times. Then, requests on the cached content can be served without downloading the requested content from remote sites at peak times. Wireless edge caching can thus dramatically reduce the traffic load on the network by exploiting cached data to serve the requests on some popular contents.

A variety of caching strategies and caching gains have been studied in cache-enabled device-to-device (D2D) communications and small cell networks [6]–[18]. As a user caching strategy for reducing playback delay in video streaming, user prefix caching was proposed in [7]. Scaling optimal distributed content placement and transmission range for users were discussed in D2D networks [8]–[10]. To reduce the backhaul burden induced by payload exchange in cooperative MIMO, [11] proposed a cooperative MIMO scheme where base stations exploit cached data instead of exchanging payload between them. Distributed content placement for APs whose coverages overlap was considered in [13]. Due to the fact that the search for the optimum content placement is NP-hard problem, several content placement algorithms have been proposed [13], [14]. With a stochastic geometry framework, the tradeoff between cache memory size and base station density was investigated in [15] and [16], and the issue of content placement to maximize channel selection diversity and energy efficiency was investigated in [17] and [18], respectively. However, none of these approaches [6]–[18] have exploited the content reuse feature of mobile traffic in transmissions by means of the wireless multicast.

A. RELATED WORK

In case of delivering the same content to multiple users, wireless multicast is an efficient way to utilize scarce spectrum resource. The idea of using wireless multicast together with caching has been considered in some works. In an initial approach, the basic idea of integrating multicast and caching was discussed in upper network layers [19]. In a heterogeneous cellular network in which a macro BS is capable of multicast, a caching algorithm for small cell base stations was proposed to minimize the total service cost in [21]. With a stochastic geometric framework, a random caching algorithm for helper nodes was proposed to maximize the successful transmission probability in a cellular network [22]. Specifically, an iterative numerical algorithm was proposed to find the optimal solution, and the closed-form expression of the solution was derived for a special case of high SNR, high user density, and small content library size. For cache-enabled cloud RAN, the joint design of multicast beamforming and content-centric BS clustering to minimize network cost was investigated in [23], where the network cost is defined as the weighted sum of transmission power and backhaul cost.

Most previous works on wireless multicast with caching focused on the problem of reducing the traffic burden on the backhaul by exploiting the cached contents at BSs. One of the other most important issues in future wireless networks relates to the scarcity of spectrum resources for wireless communications [24], [25]; however, there has not yet been any investigation of a wireless caching system to mitigate this problem.

B. CONTRIBUTIONS

The content caching at the end-user provides additional benefits over the caching at the BS/AP. First, the user can be served directly from its own memory without the transmission failure and the delay caused by wireless communications. Second, the problem of spectrum scarcity can be significantly reduced because the content caching on the user side reduces the need for wireless communications. Third, since each user caches contents only for itself, the content placement policy is much simpler than the policies for the BS/AP [11]–[18]. In other words, each user caches some popular contents regardless of the deployment and the interaction with other users. In this paper, we consider the multicast with user caching in heavy traffic downlink networks where the number of content requesting users greatly exceeds the maximum number of data streams that an AP can transmit in parallel. In such networks with heavy traffic, the wireless link becomes the bottleneck in the communications. As the main results of this paper, we derive closed-form expressions of outage probability and network throughput of the proposed system. The closed-form expressions explain the gains from caching and multicast in the proposed system by making comparisons with two reference systems, namely multicast without caching and unicast with user caching. These analysis results provide useful information for understanding the role of multicast and user caching in the proposed system and the impact of system parameters on performance. The simulation results also validate the information obtained from the analysis results and show that the proposed system outperforms the system discussed in previous work.

The contributions of this paper can be summarized as follows:

- To the best of our knowledge, this is the first work to discuss user caching and wireless multicasting for mitigating wireless bottlenecks in heavy traffic network.
- We consider that there are infinitely many contents for possible user request contrary to most previous works, which assume a small number of contents for tractable analysis.
- We derive closed-form expressions of network throughput gains from user caching and wireless multicast.

The rest of this paper is organized as follows. Section II presents the system model considered in this paper. Section III analyzes the performances of multicast systems with and without user caching and compares them with that of unicast system. Section IV validates the theoretical asymptotic results through numerical simulations. Finally, the conclusion is presented in Section V.

II. SYSTEM MODEL

We consider a downlink network consisting of a single AP and K users as illustrated in figure 1. The AP and all users are assumed to have a single antenna. The AP is assumed to be capable of transmitting at most B different contents over B orthogonal channels for a content transfer interval. Such a limitation of B at the AP could result from limited

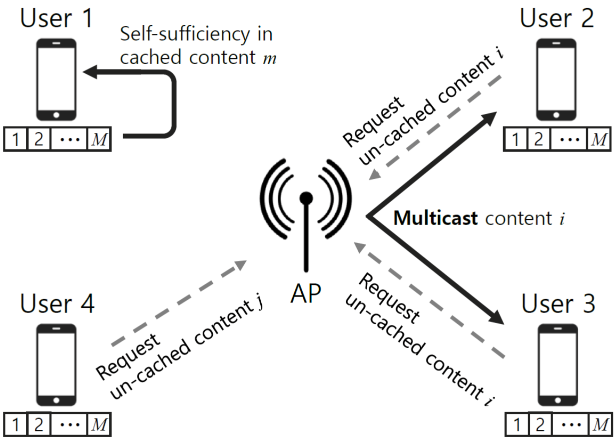

FIGURE 1. Example of system model ($K = 4$, $B = 1$).

communication resources, such as the number of available channels and the number of antennas. The number of users is assumed to be very large such that $K \gg B$, which illustrates scenarios in which a large number of users are densely located in a finite area, such as in a train, a subway station, or an airport during peak times. Each user has a storage space for caching at most M contents and adopts proactive caching strategy that pre-fetches some popular contents during off-peak times. All contents are assumed to be the same size normalized to one for analytical simplicity.[1]

Each user requires content according to a content popularity, and the popularity distribution is assumed to be characterized by Weibull distribution [4], [5]. In other words, the probability that a user requires the i-th most popular content is represented by

$$f_i = e^{-\left(\frac{i-1}{\lambda}\right)^{\kappa}} - e^{-\left(\frac{i}{\lambda}\right)^{\kappa}}, \quad i \in \{1, 2, \ldots, L\}, \tag{1}$$

where L denotes total number of contents, $0 < \kappa \leq 1$ and $\lambda > 0$ are shape and scale parameters, respectively. Note that the probability $f_i > f_j$ for $i < j$. The total number of contents L is assumed to go to infinity. Although the number of contents cannot be infinite in practice, the infinite content assumption can be justified by the fact that there are a tremendous number of content in the Internet. For example, in case of the video service, which is considered as the main application of caching technique, about 300 hours of new video content are uploaded to YouTube every single minute [20]. Since the number of contents are very large, the infinite content assumption may not lead to significant error in the analysis. The larger scaling parameter λ, the more spread out the distribution. The shape parameter affects the shape of the distribution rather than the degree of spread. For example, as the shape parameter κ tends to 0, the tail of the distribution becomes heavier [7]. The cumulative mass function (c.m.f.) of the content popularity is then given by

$$F_i = 1 - e^{-\left(\frac{i}{\lambda}\right)^{\kappa}}. \tag{2}$$

In other words, F_i represents the probability that a user requests content with a popularity ranking among the top i.

In figure 1, each user caches the M most popular contents to maximize cache hit ratio. There are two possible sources from which a user can obtain the required content: the storage space and the AP. The user 1 requires content m which is already cached in its storage. The user 1 can then obtain the required content by directly loading it from the storage instead of requesting it to the AP. However, the users 2, 3, and 4 require content which is not cached in their storages, so each of them requests the required content from the AP. We assume that the requests made by users are synchronized.[2] In order to reduce the resource required to deal with duplicate requests, the AP transmits the B most requested contents through wireless multicast. In other words, if the number of different requested contents is greater than B, the AP transmits only the B most requested contents over B orthogonal channels. Otherwise, the AP transmits all the requested contents. In figure 1, the AP multicasts content i for users 2 and 3 because the content i is one of the $B = 1$ most requested contents. Since users 1 and 4 do not require content i and have no storage space to cache the additional content, they ignore the signal about the content i.

Wireless content transfer from AP to the user suffers from Rayleigh block fading, which is a statistical model for the effect of a propagation environment on a radio signal in wireless communications. In this channel model, the received signal $y_k^{(\beta)}$ of user $k \in \{1, 2, \ldots, K\}$ through channel $\beta \in \{1, 2, \ldots, B\}$ is represented by

$$y_k^{(\beta)} = h_k^{(\beta)} x^{(\beta)} + n_k^{(\beta)}, \tag{3}$$

where $x^{(\beta)}$ denotes a symbol of the content transmitted over a channel β, $n_k^{(\beta)}$ denotes additive white Gaussian noise with zero mean and variance of N_0, $\mathcal{CN}(0, N_0)$, $h_k^{(\beta)}$ denotes the Rayleigh block fading channel gain that follows circularly-symmetric complex normal distribution with zero mean and variance of σ^2, $\mathcal{CN}(0, \sigma^2)$. The probability density function of complex normal distribution with zero mean and variance Γ, $\mathcal{CN}(0, \Gamma)$, is represented by

$$f(z) = \frac{1}{\pi\Gamma} e^{-\frac{|z|^2}{\Gamma}}.$$

The channel gain $h_k^{(\beta)}$ is fixed over a content transfer interval, and the channel state information (CSI) is assumed to be only available to the users.

For a target content transfer rate R, there might be a failure in the process of wireless content transfer from the AP when the wireless channel is in deep fading with a small value

[1] Various content sizes are not addressed in this paper. However, if the contents are divided into the equal-sized small packets and transferred in the unit of packet, the analytical framework of this paper is then applicable to the general case.

[2] In order to mitigate problems resulting from the unsynchronized content requests, the AP accumulates requests from users for a certain length of time before transmission. However, the specific length of time and the impact of this on performance are out of scope of this paper.

of $|h_k^{(\beta)}|$; however, there is no failure in the process of loading it from the storage. Hence, increasing the cache hit ratio is required not only to mitigate the wireless bottleneck but also to reduce the likelihood of failure in the content transfer. Since there is no information sharing and cooperation among users, each user caches the M most popular contents in order to maximize the cache hit ratio.

We focus on delay-limited scenarios where a content should be delivered to the requested user within a content transfer interval. Hence, we define *outage* as an event that a user fails to obtain the required content within a content transfer interval. In our system model, there are two possible cases for outage event.

- A user requests content that is not cached in its storage and is not transmitted from the AP.
- A user receives the signal of the requested content from the AP but cannot decode it successfully if Shannon capacity of Rayleigh fading channel is smaller than the target content transfer rate, $\log_2\left(1+|h_k^{(\beta)}|^2\rho\right) < R$, due to a low channel magnitude $|h_k^{(\beta)}|$,

where $\rho = \frac{\mathbb{E}[|x^{(\beta)}|^2]}{N_0}$ denotes the signal-to-noise ratio (SNR).

III. PERFORMANCE ANALYSIS

In this section, we investigate the network throughput of the proposed multicast system with user caching, and then we determine its throughput improvement compared to conventional systems.

As a performance metric, we consider the average network throughput which quantifies the information bits successfully delivered to users in the network as follows:

$$\begin{aligned} T &= R \times \mathbb{E}\left[K - K_{\text{out}}\right] \\ &= RK(1 - p_{\text{out}}), \end{aligned} \tag{4}$$

where K_{out} and p_{out} denote the number of users in outage and the probability that a user is in outage, respectively.

Since the network throughput T is a simple linear function of user outage probability p_{out}, it is possible directly to gauge the effect of system parameters on the network throughput if the closed-form expression of p_{out} is given.

In the following subsections, we first focus on deriving the closed-form expression of the outage probability and the throughput of the various systems considered. From these analysis results, we show the performance gain of the proposed system compared with the two conventional systems, namely the multicast system without user caching and the unicast system with user caching.

A. PROPOSED MULTICAST SYSTEM WITH USER CACHING

According to the definition of outage, the outage event of user k can be written as

$$\begin{aligned} \mathcal{E}_{\text{out},k} &= \{k \notin \mathcal{B}, k \notin \mathcal{C}\} \\ &\cup \left\{k \in \mathcal{B}, \log_2\left(1+|h_k^{(\beta_k)}|^2\rho\right) < R\right\} \end{aligned} \tag{5}$$

where β_k denotes the sub-channel that conveys the requested content of user k, $\mathcal{C}$ and $\mathcal{B}$ denote the set of users who require the cached content and the set of users scheduled to receive the requested content from the AP via multicast, respectively.

The probability that a user requires one of M cached contents is $F_M = 1 - e^{-\left(\frac{M}{\lambda}\right)^\kappa}$. Since the probability of the self-sufficiency $\Pr[k \in \mathcal{C}] = F_M$ grows with M, the outage probability $\Pr\left[\mathcal{E}_{\text{out},k}\right]$ converges to zero as the storage space M increases. In other words, almost all users obtain the required contents from their storage without traversing the AP if M is large.

On the one hand, let us consider the case with a general value of M. Based on (5), the outage probability can be rewritten as

$$\begin{aligned} p_{\text{c,out}} &= \frac{1}{K}\sum_{k=1}^{K}\Pr\left[\mathcal{E}_{\text{out},k}\right] = \frac{1}{K}\mathbb{E}\left[K_{\text{out}}\right] \\ &= \frac{1}{K}\sum_{i=0}^{K}\mathbb{E}[K_{\text{out}}|k_c = i]\Pr[k_c = i] \end{aligned} \tag{6}$$

where $k_c = |\mathcal{C}|$ denotes the number of users obtaining the required contents from their storage. The conditional expectation in (6) can be expanded as

$$\begin{aligned} &\mathbb{E}[K_{\text{out}}|k_c = i] \\ &= K - i - \sum_{j=0}^{K-k_c} j\Pr[k_b = j|k_c = i]\Pr\left[\log_2\left(1+|h_k^{(\beta)}|^2\rho\right) \ge R\right] \\ &= K - i - \Pr\left[|h_k^{(\beta)}|^2 \ge \frac{1}{\rho}\left(2^R - 1\right)\right]\sum_{j=0}^{K-k_c} j\Pr[k_b = j|k_c = i] \\ &= K - i - e^{-\frac{2^R-1}{\sigma^2\rho}}\sum_{j=0}^{K-k_c} j\Pr[k_b = j|k_c = i] \\ &= \begin{cases} K - i - e^{-\dfrac{2^R-1}{\sigma^2\rho}}\displaystyle\sum_{j=B}^{K-i} j\Pr[k_b = j|k_c = i] & ,0 \le i < K - B \\ K - i - e^{-\dfrac{2^R-1}{\sigma^2\rho}}(K-i) & ,K - B \le i \le K, \end{cases} \end{aligned} \tag{7}$$

where $k_b = |\mathcal{B}|$ denotes the number of users who are scheduled to receive the required contents via wireless multicast. Based on (2) and (7), the outage probability (6) is represented by

$$\begin{aligned} p_{\text{c,out}} = 1 - F_M - \frac{e^{-\frac{2^R-1}{\sigma^2\rho}}}{K}\sum_{l=K-B}^{K}\Pr[k_c = l](K-l) \\ - \frac{e^{-\frac{2^R-1}{\sigma^2\rho}}}{K}\sum_{i=0}^{K-B-1}\Pr[k_c = i]\sum_{j=B}^{K-i} j\Pr\left[k_b = j|k_c = i\right]. \end{aligned} \tag{8}$$

From (8), it is intractable to calculate the exact closed-form expression of the probability $\Pr\left[k_b = j|k_c = i\right]$. In order to circumvent this difficulty, we provide upper and lower bounds of the outage probability in the following propositions.

Proposition 1: Given a limited number of channels B and a storage space M, the outage probability is bounded above by

$$p_{\mathrm{c,out}} \le e^{-\left(\frac{M}{\lambda}\right)^{\kappa}} - \left(1-(1+\epsilon)\left(1-e^{-\left(\frac{M}{\lambda}\right)^{\kappa}}\right)\right) \times\left(1-e^{\left(\frac{M}{\lambda}\right)^{\kappa}-\left(\frac{M+B}{\lambda}\right)^{\kappa}}\right)e^{-\frac{2^R-1}{\sigma^2\rho}} + O\left(e^{-2\epsilon^2F_M^2K}\right), \tag{9}$$

for any constant $\epsilon > 0$.[3]

Proof: From (8), the outage probability is bounded above by

$$\begin{aligned} p_{\mathrm{c,out}} &\le e^{-\left(\frac{M}{\lambda}\right)^{\kappa}} - \frac{e^{-\frac{2^R-1}{\sigma^2\rho}}}{K}\sum_{i=0}^{K-B-1}\Pr[k_{\mathrm{c}}=i]\sum_{j=B}^{K-i} j\Pr[k_{\mathrm{b}}=j|k_{\mathrm{c}}=i] \\ &\overset{(a)}{\le} e^{-\left(\frac{M}{\lambda}\right)^{\kappa}} - \frac{e^{-\frac{2^R-1}{\sigma^2\rho}}}{K}\sum_{i=(1-\epsilon)\bar{k}_{\mathrm{c}}}^{(1+\epsilon)\bar{k}_{\mathrm{c}}}\Pr[k_{\mathrm{c}}=i] \times\sum_{j=B}^{K-(1+\epsilon)\bar{k}_{\mathrm{c}}} j\Pr\left[k_{\mathrm{b}}=j|k_{\mathrm{c}}=(1+\epsilon)\bar{k}_{\mathrm{c}}\right] \\ &\overset{(b)}{\le} e^{-\left(\frac{M}{\lambda}\right)^{\kappa}} - \frac{e^{-\frac{2^R-1}{\sigma^2\rho}}}{K}\left(1-2e^{-2K\epsilon^2F_M^2}\right) \times\sum_{j=B}^{K-(1+\epsilon)\bar{k}_{\mathrm{c}}} j\Pr\left[k_{\mathrm{b}}=j|k_{\mathrm{c}}=(1+\epsilon)\bar{k}_{\mathrm{c}}\right] \\ &\overset{(c)}{\le} e^{-\left(\frac{M}{\lambda}\right)^{\kappa}} - \frac{e^{-\frac{2^R-1}{\sigma^2\rho}}}{K}\left(1-2e^{-2K\epsilon^2F_M^2}\right) \times\sum_{j=0}^{K-(1+\epsilon)\bar{k}_{\mathrm{c}}} j\binom{K-(1+\epsilon)\bar{k}_{\mathrm{c}}}{j}F_{\mathrm{b|c}}^{j}(1-F_{\mathrm{b|c}})^{K-j} \\ &= e^{-\left(\frac{M}{\lambda}\right)^{\kappa}} - e^{-\frac{2^R-1}{\sigma^2\rho}}\left(1-2e^{-2K\epsilon^2F_M^2}\right)(1-(1+\epsilon)F_M)F_{\mathrm{b|c}} \\ &= e^{-\left(\frac{M}{\lambda}\right)^{\kappa}} - (1-(1+\epsilon)F_M)F_{\mathrm{b|c}}e^{-\frac{2^R-1}{\sigma^2\rho}} + O\left(e^{-2\epsilon^2F_M^2K}\right) \\ &= e^{-\left(\frac{M}{\lambda}\right)^{\kappa}} - \left(1-(1+\epsilon)\left(1-e^{-\left(\frac{M}{\lambda}\right)^{\kappa}}\right)\right) \times\left(1-e^{\left(\frac{M}{\lambda}\right)^{\kappa}-\left(\frac{M+B}{\lambda}\right)^{\kappa}}\right)e^{-\frac{2^R-1}{\sigma^2\rho}} + O\left(e^{-2\epsilon^2F_M^2K}\right), \end{aligned} \tag{10}$$

where $\bar{k}_c = \mathbb{E}[k_c]$, $F_{\mathrm{b|c}} = \frac{F_{M+B}-F_M}{1-F_M}$, and ϵ denotes some positive constant. The inequality (a) in (10) is obtained by confining the summations to some values around $\bar{k}_c$. The inequality (b) in (10) comes from Hoeffding's inequality (11) given blow.

Lemma 1 (Hoeffding's Inequality): Let $X_1, \ldots, X_K$ be i.i.d. random variables, where $\Pr[X_i \in [a_i, b_i]] = 1$ for $1 \le i \le K$. The probability that the empirical mean of the random variables $\tilde{X} = \frac{1}{K}\sum_{i=1}^{K} X_i$ deviates from its expected value is bounded above by

$$\Pr\left[\left|1-\frac{\tilde{X}}{\mathbb{E}[\tilde{X}]}\right| \ge \epsilon\right] \le 2\exp\left[-\frac{2K^2\epsilon^2\left(\mathbb{E}[\tilde{X}]\right)^2}{\sum_{i=1}^{K}(b_i-a_i)^2}\right], \tag{11}$$

$$\Pr\left[1-\frac{\tilde{X}}{\mathbb{E}[\tilde{X}]} \ge \epsilon\right] \le \exp\left[-\frac{2K^2\epsilon^2\left(\mathbb{E}[\tilde{X}]\right)^2}{\sum_{i=1}^{K}(b_i-a_i)^2}\right], \tag{12}$$

for some positive ϵ.

Specifically, in the inequality (b),

$$\begin{aligned} \sum_{(1-\epsilon)\bar{k}_{\mathrm{c}}}^{(1+\epsilon)\bar{k}_{\mathrm{c}}}\Pr[k_{\mathrm{c}}=i] &= \Pr\left[(1-\epsilon)\bar{k}_{\mathrm{c}} \le k_{\mathrm{c}} \le (1+\epsilon)\bar{k}_{\mathrm{c}}\right] \\ &= \Pr\left[\left|1-\frac{k_{\mathrm{c}}}{\bar{k}_{\mathrm{c}}}\right| \le \epsilon\right] \\ &= \Pr\left[\left|1-\frac{\sum_{k=1}^{K}\mathbf{1}(k\in\mathcal{C})}{\mathbb{E}\left[\sum_{k=1}^{K}\mathbf{1}(k\in\mathcal{C})\right]}\right| \le \epsilon\right] \\ &\ge 1-2e^{-2K\epsilon^2F_M^2}, \end{aligned}$$

where $\mathbf{1}(k \in \mathcal{C})$ is an indicator function which returns 1 if user k requests one of cached content and returns 0 otherwise. Since users are assumed to require content independently according to the popularity distribution, the returns $\mathbf{1}(k \in \mathcal{C})$ for all users $k \in \mathcal{K}$ are independent Bernoulli random variables with probability F_M.

The inequality (c) in (10) comes from the fact that k_{b} is bounded below by the number of users requesting one of the contents which ranked from $M+1$ to $M+B$. ■

The lower bound of the outage probability $p_{\mathrm{c,out}}$ is given in the following proposition.

Proposition 2: Given a limited number of channels B and a storage space M, the outage probability is bounded below by

$$p_{\mathrm{c,out}} \ge 1-(1+\epsilon)\left(1-e^{-\left(\frac{M}{\lambda}\right)^{\kappa}}\right) -(1+\epsilon')\left(e^{-\left(\frac{M}{\lambda}\right)^{\kappa}}-e^{-\left(\frac{M+B}{\lambda}\right)^{\kappa}}\right)e^{-\frac{2^R-1}{\sigma^2\rho}} + O\left(e^{-\delta_1K}\right), \tag{13}$$

for any constants $\epsilon > 0$, $\epsilon' > 0$, and δ_1 as explained in (18).

Proof: Let us define the following two events

$$\mathcal{E}_{\mathrm{c}} \triangleq \left\{\left|1-\frac{\sum_{i=1}^{M}K_i}{\sum_{i=1}^{M}\bar{K}_i}\right| \le \epsilon\right\}, \tag{14}$$

[3] The notation $f(x) = O(g(x))$ denotes that there exists a constant c such that $f(x) \le cg(x)$ as x goes to infinity.

$$\mathcal{E}_{\mathrm{b|c}} \triangleq \left\{\{M+1, M+2, \dots, M+B\} = \arg\max_{\mathcal{S},\, |\mathcal{S}|=B} \sum_{i\in\mathcal{S}, i>M} K_i\right\} \cap \left\{\bigcap_{i=M+1}^{M+B} \left|1 - \frac{K_i}{\bar{K}_i}\right| \leq \epsilon'\right\}, \quad (15)$$

where K_i denotes the number of users who want the ith most popular content, and $\bar{K}_i$ denotes the mean of K_i.

According to Hoeffding's inequality (11), the probability of the complementary event of (14) is bounded above by

$$\Pr\left[\mathcal{E}_c^c\right] \leq 2e^{-2K\epsilon^2 F_M^2}. \quad (16)$$

Based on (16), the outage probability (8) is bounded below by, for some constant $\epsilon > 0$ and $\epsilon' > 0$,

$$\begin{aligned}
&p_{\mathrm{c,out}} \\
&= 1 - \frac{1}{K}\Pr[\mathcal{E}_c]\mathbb{E}\left[k_c + e^{-\frac{2^R-1}{\sigma^2\rho}} k_{\mathrm{b}} \middle| \mathcal{E}_c\right] - \frac{1}{K}\Pr[\mathcal{E}_c^c]\mathbb{E}\left[k_c + e^{-\frac{2^R-1}{\sigma^2\rho}} k_{\mathrm{b}} \middle| \mathcal{E}_c^c\right] \\
&\geq 1 - \frac{1}{K}\mathbb{E}\left[k_c + e^{-\frac{2^R-1}{\sigma^2\rho}} k_{\mathrm{b}} \middle| \mathcal{E}_c\right] + O\left(e^{-2\epsilon^2 F_M^2 K}\right) \\
&\geq 1 - \frac{1}{K}\mathbb{E}\left[k_c + e^{-\frac{2^R-1}{\sigma^2\rho}} k_{\mathrm{b}} \middle| k_c = (1+\epsilon)\bar{k}_c\right] + O\left(e^{-2\epsilon^2 F_M^2 K}\right) \\
&\geq 1 - (1+\epsilon)F_M - \frac{e^{-\frac{2^R-1}{\sigma^2\rho}}}{K}\mathbb{E}\left[k_{\mathrm{b}} \middle| k_c = \bar{k}_c\right] + O\left(e^{-2\epsilon^2 F_M^2 K}\right) \\
&= 1 - (1+\epsilon)F_M - \frac{e^{-\frac{2^R-1}{\sigma^2\rho}}}{K}\Big(\mathbb{E}\left[k_{\mathrm{b}} \middle| k_c = \bar{k}_c, \mathcal{E}_{b|c}\right] \times \Pr\left[\mathcal{E}_{b|c} \middle| k_c = \bar{k}_c\right] + \mathbb{E}\left[k_{\mathrm{b}} \middle| k_c = \bar{k}_c, \mathcal{E}_{b|c}^c\right] \\
&\qquad \times \Pr\left[\mathcal{E}_{b|c}^c \middle| k_c = \bar{k}_c\right]\Big) + O\left(e^{-2\epsilon^2 F_M^2 K}\right) \\
&\geq 1 - (1+\epsilon)F_M - \frac{e^{-\frac{2^R-1}{\sigma^2\rho}}}{K}\Bigg(\left\{\sum_{i=M+1}^{M+B}(1+\epsilon')\bar{K}_i\right\} + (K - \bar{k}_c)\Pr\left[\mathcal{E}_{b|c}^c \middle| k_c = \bar{k}_c\right]\Bigg) + O\left(e^{-2\epsilon^2 F_M^2 K}\right)
\end{aligned} \quad (17)$$

The expression (17) is further bounded below by

$$\begin{aligned}
p_{\mathrm{c,out}} &\geq 1 - (1+\epsilon)F_M - \frac{e^{-\frac{2^R-1}{\sigma^2\rho}}}{K}\left((1+\epsilon')(K - \bar{k}_c)F_{\mathrm{b|c}}\right) + O\left(e^{-\delta_1 K}\right) \\
&= 1 - (1+\epsilon)\left(1 - e^{-\left(\frac{M}{\lambda}\right)^\kappa}\right) - (1+\epsilon') \times \left(e^{-\left(\frac{M}{\lambda}\right)^\kappa} - e^{-\left(\frac{M+B}{\lambda}\right)^\kappa}\right) e^{-\frac{2^R-1}{\sigma^2\rho}} + O\left(e^{-\delta_1 K}\right),
\end{aligned} \quad (18)$$

where $\delta_1 = \min\{2\epsilon^2 F_M^2, \delta_2\}$, the inequality comes from the fact that $\Pr\left[\mathcal{E}_{\mathrm{b|c}}^c \middle| k_{\mathrm{c}} = \bar{k}_{\mathrm{c}}\right] = O\left(e^{-\delta_2 K}\right)$, for $\delta_2 = \min\left\{2\left(f_{M+B|\mathrm{c}} - f_{M+B+1|\mathrm{c}}\right)^2(1-F_M), 2(1-f_M)\epsilon'^2 f_{M+B|\mathrm{c}}^2\right\}$ as derived below.

$$\begin{aligned}
&\Pr\left[\mathcal{E}_{\mathrm{b|c}}^c \middle| k_{\mathrm{c}} = \bar{k}_{\mathrm{c}}\right] = \Pr\Bigg[\bigcup_{i=M+1}^{M+B} \bigcup_{j=M+B+1}^{L} \{K_i < K_j\} \cup \left\{\bigcup_{l=M+1}^{M+B}\left|1 - \frac{K_l}{\bar{K}_l}\right| \geq \epsilon'\right\} \Bigg| k_{\mathrm{c}} = \bar{k}_{\mathrm{c}}\Bigg] \\
&= \Pr\Bigg[\Bigg\{\bigcup_{i=M+1}^{M+B} \bigcup_{j=M+B+1}^{\eta_c} \{K_i < K_j\} \cup \left\{K_i < \sum_{u=\eta_c+1}^{L} K_u\right\}\Bigg\} \cup \left\{\bigcup_{l=M+1}^{M+B}\left|1 - \frac{K_l}{\bar{K}_l}\right| \geq \epsilon'\right\} \Bigg| k_{\mathrm{c}} = \bar{k}_{\mathrm{c}}\Bigg] \\
&\leq B(\eta_{\mathrm{c}} - M - B)\Pr\left[K_{M+B} < K_{M+B+1} \middle| k_{\mathrm{c}} = \bar{k}_{\mathrm{c}}\right] + B\Pr\left[K_{M+B} < \sum_{u=\eta_c+1}^{L} K_u \middle| k_{\mathrm{c}} = \bar{k}_{\mathrm{c}}\right] \\
&\quad + B\Pr\left[\left|1 - \frac{K_{M+B}}{\bar{K}_{M+B}}\right| \geq \epsilon' \middle| k_{\mathrm{c}} = \bar{k}_{\mathrm{c}}\right] \\
&\leq B(\eta_{\mathrm{c}} - M - B)e^{-2\left(f_{M+B|\mathrm{c}} - f_{M+B+1|\mathrm{c}}\right)^2(K - \bar{k}_{\mathrm{c}})} + Be^{-2\left(f_{M+B|\mathrm{c}} + F_{\eta_{\mathrm{c}}|\mathrm{c}} - 1\right)^2(K - \bar{k}_{\mathrm{c}})} + 2Be^{-2(K - \bar{k}_{\mathrm{c}})\epsilon'^2 f_{M+B|\mathrm{c}}^2} \\
&= B(\eta_{\mathrm{c}} - M - B)e^{-2\left(f_{M+B|\mathrm{c}} - f_{M+B+1|\mathrm{c}}\right)^2(1 - F_M)K} + Be^{-2\left(f_{M+B|\mathrm{c}} + F_{\eta_{\mathrm{c}}|\mathrm{c}} - 1\right)^2(1 - F_M)K} + 2Be^{-2K(1 - f_M)\epsilon'^2 f_{M+B|\mathrm{c}}^2},
\end{aligned} \quad (19)$$

where $\eta_{\mathrm{c}} = \lceil \lambda \left(-\log\left(e^{-\left(\frac{M+B-1}{\lambda}\right)^\kappa} - e^{-\left(\frac{M+B}{\lambda}\right)^\kappa}\right)\right)^{1/\kappa} \rceil + 1$ indicates the smallest content index i that satisfy $f_{M+B} \geq 1 - F_i$, and $f_{i|\mathrm{c}} = \frac{f_i}{1-F_M}$. The last inequality in (19) comes from the Hoeffding's inequality. Specifically, K_{M+B} and K_{M+B+1} can be considered as the binomial distribution with probabilities $f_{M+B|c}$ and $f_{M+B+1|c}$, respectively. Then, the probability $\Pr\left[K_{M+B} < K_{M+B+1} \middle| k_{\mathrm{c}} = \bar{k}_{\mathrm{c}}\right]$ can be rewritten as

$$\begin{aligned}
&\Pr\left[K_{M+B} < K_{M+B+1} \middle| k_{\mathrm{c}} = \bar{k}_{\mathrm{c}}\right] \\
&= \Pr\left[K_{M+B+1} - K_{M+B} > 0 \middle| k_{\mathrm{c}} = \bar{k}_{\mathrm{c}}\right] \\
&= \Pr\Big[K_{M+B+1} - K_{M+B} + K(f_{M+B|c} - f_{M+B+1|c}) > K(f_{M+B|c} - f_{M+B+1|c}) \Big| k_{\mathrm{c}} = \bar{k}_{\mathrm{c}}\Big] \\
&= \Pr\left[1 - \frac{K_{M+B}/K - K_{M+B+1}/K}{f_{M+B|c} - f_{M+B+1|c}} > 1 \middle| k_{\mathrm{c}} = \bar{k}_{\mathrm{c}}\right].
\end{aligned} \quad (20)$$

Based on Hoeffding's inequality (12), the probability (20) is bounded above by

$$\begin{aligned}
&\Pr\left[K_{M+B} < K_{M+B+1} \middle| k_{\mathrm{c}} = \bar{k}_{\mathrm{c}}\right] \\
&\leq \exp\left[-2(f_{M+B|c} - f_{M+B+1|c})^2(K - \bar{k}_c)\right].
\end{aligned} \quad (21)$$

In a similar way, from (11), we can obtain the upper bound of the probability $\Pr\left[\left|1-\frac{K_{M+B}}{\bar{K}_{M+B}}\right| \geq \epsilon' \middle| k_c = \bar{k}_c\right]$ as $2e^{-2(K-\bar{k}_c)\epsilon'^2 f^2_{M+B|c}}$. ∎

The upper bound (9) and the lower bound (13) converge as K increases. Eventually, for an infinitely large number of users K, the probability $p_{c,out}$ is asymptotically approximated by

$$\lim_{K\to\infty} p_{c,out} = e^{-\left(\frac{M}{\lambda}\right)^{\kappa}} - \left(e^{-\left(\frac{M}{\lambda}\right)^{\kappa}} - e^{-\left(\frac{M+B}{\lambda}\right)^{\kappa}}\right) e^{-\frac{2^R-1}{\sigma^2\rho}}. \tag{22}$$

Correspondingly, the asymptotic network throughput can be represented by

$$T_c = KR\left(1 - e^{-\left(\frac{M}{\lambda}\right)^{\kappa}} + \left(e^{-\left(\frac{M}{\lambda}\right)^{\kappa}} - e^{-\left(\frac{M+B}{\lambda}\right)^{\kappa}}\right) e^{-\frac{2^R-1}{\sigma^2\rho}}\right). \tag{23}$$

Remarks:

- As the storage space M grows, the asymptotic outage probability $p_{c,out}$ strictly decreases.
- The network throughput T_c linearly increases with the number of users K, and the slope of that linear relationship is determined by the size of storage space M. As the storage space M grows, the increasing rate of T_c with respect to K reduces.

B. PERFORMANCE COMPARISONS WITH CONVENTIONAL SYSTEMS

In order to see the throughput gains resulting from user caching and multicasting, we additionally consider two reference systems: multicast without user caching and unicast with user caching.

1) COMPARISON WITH MULTICAST WITHOUT USER CACHING

When there is no caching at the users, only users who request one of the B most requested contents are served by the AP. By substituting $M = 0$ into (22) and (23), the asymptotic outage probability and network throughput of the conventional multicast system are represented, respectively, by

$$\lim_{K\to\infty} p_{b,out} = 1 - \left(1 - e^{-\left(\frac{B}{\lambda}\right)^{\kappa}}\right) e^{-\frac{2^R-1}{\sigma^2\rho}}, \tag{24}$$

$$T_b = KR\left(1 - e^{-\left(\frac{B}{\lambda}\right)^{\kappa}}\right) e^{-\frac{2^R-1}{\sigma^2\rho}}. \tag{25}$$

Remarks:

- Both network throughputs T_c and T_b are increasing functions of B; however, the amount of the network throughput originated from the AP multicast is reduced from (25) to $KR\left(e^{-\left(\frac{M}{\lambda}\right)^{\kappa}} - e^{-\left(\frac{M+B}{\lambda}\right)^{\kappa}}\right) e^{-\frac{2^R-1}{\sigma^2\rho}}$ through user caching. Consequently, as the storage space M grows, the throughput originating from the AP multicast decreases even though the network throughput increases.
- The caching gain in the multicast system is given by

$$G_c = \frac{T_c}{T_b} = \frac{1 - e^{-\left(\frac{M}{\lambda}\right)^{\kappa}} + \left(e^{-\left(\frac{M}{\lambda}\right)^{\kappa}} - e^{-\left(\frac{M+B}{\lambda}\right)^{\kappa}}\right) e^{-\frac{2^R-1}{\sigma^2\rho}}}{\left(1 - e^{-\left(\frac{B}{\lambda}\right)^{\kappa}}\right) e^{-\frac{2^R-1}{\sigma^2\rho}}}. \tag{26}$$

The caching gain G_c is an increasing function of the storage space M and a decreasing function of the number of channels B.

2) COMPARISON WITH UNICAST WITH USER CACHING

We discuss the performance gain of multicast against unicast where the AP transmits the requested contents only to B users. In the unicast system, there is no need for the AP to transmit the most requested contents because each transmission over a channel is dedicated to a single user. The network throughput of the unicast system with caching is then represented by

$$\begin{aligned} T_c^{(u)} &= KR \Pr\left[\{k \in \mathcal{C}\} \cup \left\{k \in \mathcal{B}^{(u)}, \log_2\left(1 + |h_k^{(\beta_k)}|^2\rho\right) > R\right\}\right] \\ &= KR \Pr[k \in \mathcal{C}] + BR \Pr\left[\log_2\left(1 + |h_k^{(\beta_k)}|^2\rho\right) \geq R\right] \\ &= KR\left(1 - e^{-\left(\frac{M}{\lambda}\right)^{\kappa}}\right) + BRe^{-\frac{2^R-1}{\sigma^2\rho}}, \end{aligned} \tag{27}$$

where $\mathcal{B}^{(u)}$ denotes the set of B users who are scheduled for the unicast, $|\mathcal{B}^{(u)}| = B$. With user caching, the multicast gain compared with unicast is given by

$$G_m = \frac{T_c}{T_c^{(u)}} = \frac{1 - e^{-\left(\frac{M}{\lambda}\right)^{\kappa}} + \left(e^{-\left(\frac{M}{\lambda}\right)^{\kappa}} - e^{-\left(\frac{M+B}{\lambda}\right)^{\kappa}}\right) e^{-\frac{2^R-1}{\sigma^2\rho}}}{1 - e^{-\left(\frac{M}{\lambda}\right)^{\kappa}} + \frac{B}{K} e^{-\frac{2^R-1}{\sigma^2\rho}}}. \tag{28}$$

If each user has no storage space for content caching $M = 0$, the multicast gain can be reduced to

$$G_m = \frac{K}{B}\left(1 - e^{-\left(\frac{B}{\lambda}\right)^{\kappa}}\right). \tag{29}$$

On the other hand, for general value of M, the multicast gain can be approximated by

$$G_m \approx 1 + \frac{\left(e^{-\left(\frac{M}{\lambda}\right)^{\kappa}} - e^{-\left(\frac{M+B}{\lambda}\right)^{\kappa}}\right) e^{-\frac{2^R-1}{\sigma^2\rho}}}{1 - e^{-\left(\frac{M}{\lambda}\right)^{\kappa}}}, \tag{30}$$

where the approximation comes from the fact that $\frac{B}{K}$ is too small to be ignored.

Remarks:

- Based on (29) and (30), the multicast gain G_m linearly increases with the number of users K if there is no caching $M = 0$ but is not significantly affected by K if the user is capable of caching $M \geq 1$.

- The multicast gain G_m is inversely proportional to the number of channels B with $M = 0$; however, the effect of B on the gain G_m is negligible with a large storage space $M \gg B$.
- The multicast gain G_m diminishes as the storage space M grows. As an extreme example, if $M \to \infty$, there is no multicast gain $G_m = 1$.

IV. SIMULATION RESULTS

In this section, we provide simulation results to validate the analysis contained in the previous section. For all simulation results in this section, we consider a content transfer rate $R = 1.5$ [bits/symbol/Hz], a SNR $\rho = 10$ [dB], and a popularity distribution with parameters $\lambda = 100$ and $\kappa = 0.5$. All the following simulation results are obtained from an average of 10000 trials. In order to compare the performance with that described in relevant previous work, we adopt the multicasting technique proposed in [22] to the network with user caching as a reference system. One key difference of the reference system compared with the proposed one is that the AP multicasts all the requested contents with equal bandwidth allocation, instead of multicasting the B most requested contents.

Figure 2 shows the outage probabilities of the multicast systems with and without user caching versus the number of users K. The storage space for caching at users and the number of channels are set to $M = 100$ [contents] and $B = 10, 80$ [streams], respectively. The figure confirms that the simulation results for the cases with and without caching converge to their asymptotic analysis results (22) and (24), respectively, as K increases. It is shown that exploiting cached data instead of downloading data from the AP significantly reduces the outage probability. From propositions 1 and 2, it is shown that the errors in the asymptotic analysis results (22) could be at most $O(e^{-\delta_1 K})$. Since δ_1 is a decreasing function of B, the error in the asymptotic analyses increases as B grows. From the figure, we can also confirm the effect of B on the convergence rate of the asymptotic analyses to the simulation results. Furthermore, the reference system shows the comparable outage probabilities with the proposed system for a relatively small number of users K; however, its performance becomes significantly degraded for large K. This is because no one can successfully obtain requested content with the reference system when there is a large variety of different content requests from a great number of users.

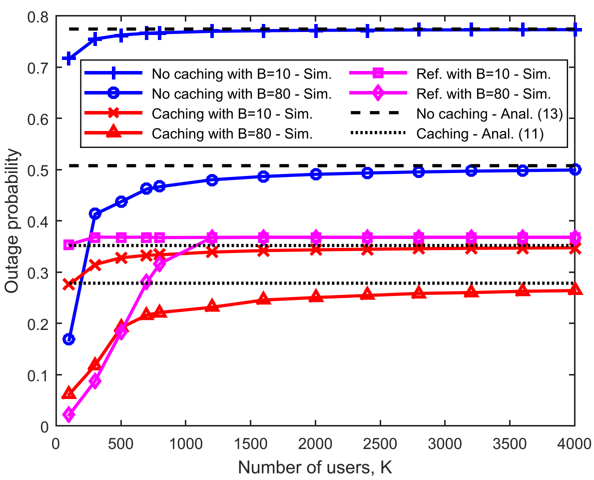

FIGURE 2. Outage probability versus the number of users ($M = 100$ [contents], $B = 10, 80$ [streams]).

Figure 3 shows the average network throughputs of multicast systems with respect to the number of users K. The simulation environment is the same as that of figure 2. As shown in figure 2, the outage probability p_{out} is not significantly affected by the value of K if K is large. Accordingly, the network throughputs of the systems with and without caching increase almost linearly with K as derived in (23) and (25). It is confirmed that the simulation results are in good agreement with the corresponding analysis results regardless of the value of K. It is also shown that the proposed system outperforms the reference system in terms of throughput, and the performance gap between them increases as K grows.

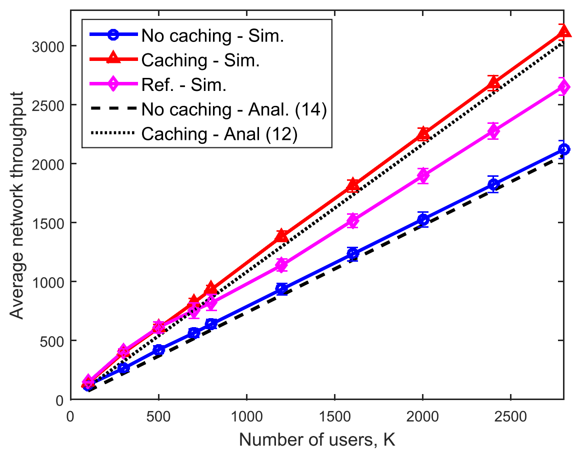

FIGURE 3. Average network throughput versus the number of users ($M = 100$ [contents], $B = 80$ [streams]).

Figure 4 shows the average network throughputs of the multicast systems with respect to the user storage space M. The number of channels and users are set to $B = 50$ [streams] and $K = 2000$ [users], respectively. As M increases, so does the amount of mobile traffic handled as cached contents. Accordingly, the outage caused by the limited number of channels and the failure of wireless transmission can be reduced by increasing M. As derived in (26), the figure shows that the increasing rate of throughput with respect to M reduces as M grows. From propositions 1 and 2, it is shown that the error in the asymptotic analysis (22) could be at most $O(e^{-\delta_1 K})$. Since δ_1 is a decreasing function of M, the error in the asymptotic analysis can increase with M. From the figure, we confirm that the gap between the simulation and the asymptotic analysis also increases with M. It is shown that the throughput gap between the proposed and reference systems reduces as the storage space of users M increases. This is because a large proportion of content

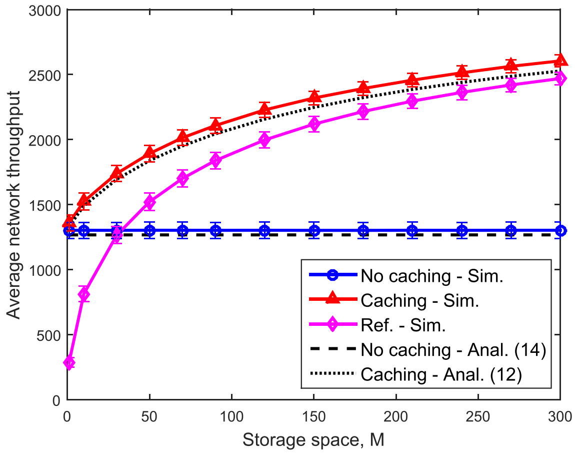

FIGURE 4. Average network throughput versus storage space for caching ($K = 2000$ [users], $B = 50$ [streams]).

requests are handled by cached data in both systems if M is large.

Figure 5 shows the gain of multicast over unicast with respect to the number of orthogonal channels B (user capacity of AP). The number of users is set to $K = 2000$ [users]. It is confirmed that if there is no user caching, the multicast gain G_{m} reduces as B grows as derived in (29). On the other hand, if the users have enough storage space, the multicast gain compared with the unicast is negligible for a large storage space M as derived in (30).

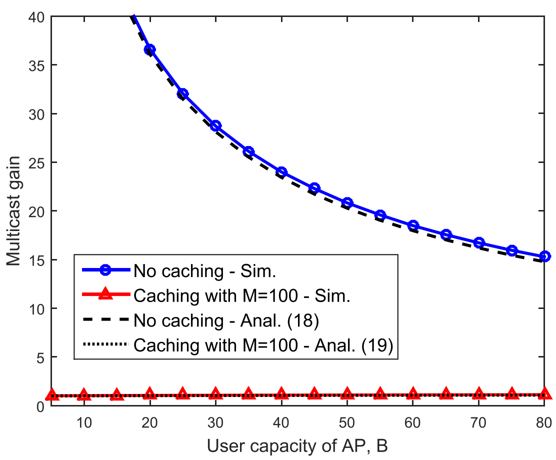

FIGURE 5. Multicast gain versus user capacity of AP ($K = 2000$ [users], $M = 100$ [contents]).

V. CONCLUSION

In this paper, we have proposed a multicast system with user caching to resolve problems induced by massive amounts of mobile traffic. The network throughput gains from user caching and multicasting have been derived in closed-form expressions through asymptotic analysis. For heavy traffic scenarios with a large number of users, asymptotic analysis results have successfully characterized the effect of the system parameters on the performance gains. It has been shown that network throughput is a monotonic increasing function of storage space M, but its increasing rate reduces as M grows. It has been also shown that the gain of multicast compared with unicast is a decreasing function of B in a network without user caching; however, it is a monotonically increasing function of B in a network with user caching. The results of this paper can be used as a guideline for designing a wireless transmission strategy and memory allocation algorithm for caching in mobile devices.

REFERENCES

[1] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update 2016–2021," Cisco, San Jose, CA, USA, White Paper 1454457600805266, Jun. 2017.
[2] T. Wen and P. Zhu, *5G: A Technology Vision*. Guangdong, China: Huawei, 2013.
[3] J. G. Andrews *et al.*, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.
[4] X. Cheng, C. Dale, and J. Liu, "Statistics and social network of YouTube videos," in *Proc. IEEE IWQoS*, Jun. 2008, pp. 229–238.
[5] L. Guo, E. Tan, S. Chen, Z. Xiao, and X. Zhang, "The stretched exponential distribution of Internet media access patterns," in *Proc. ACM Symp. PODC*, Aug. 2008, pp. 283–294.
[6] J. Song, H. Song, and W. Choi, "Optimal content placement for wireless femto-caching network," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4433–4444, Jul. 2017.
[7] J.-P. Hong and W. Choi, "User prefix caching for average playback delay reduction in wireless video streaming," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 377–388, Jan. 2016.
[8] 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, May 2017.
[9] N. Golrezaei, A. G. 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, Jul. 2014.
[10] N. Golrezaei, A. F. Molisch, and A. G. Dimakis, "Base-station assisted device-to-device communications for high-throughput wireless video networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 3665–3676, Jul. 2014.
[11] A. Liu and V. K. N. Lau, "Exploiting base station caching in MIMO cellular networks: Opportunistic cooperation for video streaming," *IEEE Trans. Signal Process.*, vol. 63, no. 1, pp. 57–69, Jan. 2015.
[12] 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 Commun. Mag.*, vol. 51, no. 4, pp. 142–149, Apr. 2013.
[13] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless video content delivery through distributed caching helpers," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 1107–1115.
[14] Z. Chang, Y. Gu, Z. Han, X. Chen, and T. Ristaniemi, "Context-aware data caching for 5G heterogeneous small cells networks," in *Proc. IEEE ICC*, May 2016, pp. 1–6.
[15] E. Baştuğ, M. Bennis, and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," *EURASIP J. Wireless Commun. Netw.*, vol. 2015, no. 41, pp. 1–11, Feb. 2015.
[16] S. Tamoor-ul-Hassan, M. Bennis, P. H. J. Nardelli, and M. Latva-Aho, "Caching in wireless small cell networks: A storage-bandwidth tradeoff," *IEEE Commun. Lett.*, vol. 20, no. 6, pp. 1175–1178, Jun. 2016.
[17] S. H. Chae and W. Choi, "Caching placement in stochastic wireless caching helper networks: Channel selection diversity via caching," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 6626–6637, Oct. 2016.
[18] B. Perabathini, E. G. Baştuğ, M. Kountouris, M. Debbah, and A. Conte, "Caching at the edge: A green perspective for 5G networks," in *Proc. IEEE ICCW*, Jun. 2015, pp. 2830–2835.
[19] J. Y. Kim, G. M. Lee, and J. K. Choi, "Efficient multicast schemes using in-network caching for optimal content delivery," *IEEE Commun. Lett.*, vol. 17, no. 5, pp. 1048–1051, May 2013.
[20] K. Smith. (Dec. 12, 2017). *39 Fascinating and Incredible YouTube Statistics*. Accessed: Apr. 6, 2018. [Online]. Available: https://www.brandwatch.com/blog/39-youtube-stats/

[21] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas, "Exploiting caching and multicast for 5G wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2995–3007, Apr. 2016.
[22] Y. Cui, D. Jiang, and Y. Wu, "Analysis and optimization of caching and multicasting in large-scale cache-enabled wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 5101–5112, Jul. 2016.
[23] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6118–6131, Sep. 2016.
[24] J.-P. Hong, B. Hong, T. Ban, and W. Choi, "On the cooperative diversity gain in underlay cognitive radio systems," *IEEE Trans. Commun.*, vol. 60, no. 1, pp. 209–219, Jan. 2012.
[25] J. P. Hong and W. Choi, "Throughput characteristics by multiuser diversity in a cognitive radio system," *IEEE Trans. Signal Process.*, vol. 59, no. 8, pp. 3749–3763, Aug. 2011.

JUN-PYO HONG (S'08–M'14) received the B.Sc. degree from Information and Communications University, Daejeon, South Korea, in 2008, and the M.S. and Ph.D. degrees from the Korea Advance Institute of Science and Technology, Daejeon, South Korea, in 2010 and 2014, respectively, all in electrical engineering. In 2015, he was a Researcher with the Electronics and Telecommunications Research Institute, where he was involved in RAN architecture for mobile communication systems. He is currently an Assistant Professor with the Department of Information and Communications Engineering, Pukyoung National University, Busan, South Korea.

SEONG HO CHAE (S'12–M'16) received the B.S. degree from the School of Electronic and Electrical Engineering, Sungkyunkwan University, Suwon, South Korea, in 2010, and the M.S. and Ph.D. degrees from the School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2012 and 2016, respectively. He was a Post-Doctoral Research Fellow with KAIST in 2016 and a Senior Researcher with Agency for Defense Development from 2016 to 2018, where he was involved in research and development for military communication system and frequency management software. He is currently an Assistant Professor with Korea Polytechnic University. His research interests include stochastic geometry, wireless edge aching, 5G, UAV, and tactical communication system and radar.

KISONG LEE (S'10–M'14) received the B.S. degree from the Department of Electrical Engineering, Information and Communications University, South Korea, in 2007, and the M.S. and Ph.D. degrees from the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, South Korea, in 2009 and 2013, respectively. He was a Researcher with the Electronics and Telecommunications Research Institute from 2013 to 2015. From 2015 to 2017, he was an Assistant Professor with the Department of Information and Telecommunication Engineering, Kunsan National University. He is currently an Assistant Professor with the School of Information and Communication Engineering, Chungbuk National University, Cheongju, South Korea. His research interests include self-organizing networks, radio resource management, magnetic induction communication, wireless power transfer, and energy harvesting networks.