Energy Efficiency of Downlink Networks with Caching at Base Stations

Dong Liu, *Student Member, IEEE*, and Chenyang Yang, *Senior Member, IEEE*

***Abstract*—Caching popular contents at base stations (BSs) can reduce the backhaul cost and improve the network throughput. Yet whether locally caching at the BSs can improve the energy efficiency (EE), a major goal for 5th generation cellular networks, remains unclear. Due to the entangled impact of various factors on EE such as interference level, backhaul capacity, BS density, power consumption parameters, BS sleeping, content popularity and cache capacity, another important question is what are the key factors that contribute more to the EE gain from caching. In this paper, we attempt to explore the potential of EE of the cache-enabled wireless access networks and identify the key factors. By deriving closed-form expression of the approximated EE, we provide the condition when the EE can benefit from caching, find the optimal cache capacity that maximizes the network EE, and analyze the maximal EE gain brought by caching. We show that caching at the BSs can improve the network EE when power efficient cache hardware is used. When local caching has EE gain over not caching, caching more contents at the BSs may not provide higher EE. Numerical and simulation results show that the caching EE gain is large when the backhaul capacity is stringent, interference level is low, content popularity is skewed, and when caching at pico BSs instead of macro BSs.**

***Index Terms*—Energy efficiency, Cache, Wireless Access Networks, Downlink**

I. Introduction

To meet the explosive demands for throughput, support sustainable development and reduce global carbon dioxide emission, energy efficiency (EE) has become a major performance metric for 5th generation (5G) cellular networks. While EE of a network can be improved from various aspects such as introducing new network architecture [2], optimizing network deployment and resource allocation [3, 4], an alternative approach is rethinking the goal of the network. Recently, it has been observed that a large portion of mobile multimedia traffic is generated by many duplicate downloads of a few popular contents [5, 6]. This reflects a shift in major goal of the networks from traditional transmitter-receiver communication to content dissemination. On the other hand, the storage capacity of today's memory devices grows rapidly. As a consequence, equipping caches at base stations (BSs) offers a promising way to unleash the potential of cellular networks except continuing densifying the networks [7, 8].

Manuscript received April 30, 2015; revised September 15, 2015; accepted October 26, 2015. This work was supported in part by National Natural Science Foundation of China (NSFC) under Grant 61120106002 and National Basic Research Program of China (973 Program) under Grant 2012CB316003. The preliminary work of this paper was presented at the 2014 IEEE Global Conferencez on Signal and Information Processing (*GlobalSIP*), Atlanta, December 3-5, 2014 [1].

D. Liu and C. Yang are with the School of Electronics and Information Engineering, Beihang University, Beijing, 100191, P.R. China (e-mail: {dliu, cyyang}@buaa.edu.cn).

Caching is a technique to improve performance well known in many wired network domains, e.g., content-centric networks (CCN) [9–11]. In cellular networks, caching popular contents in the edge can reduce the backhaul cost, access latency and energy consumption as well as boost the throughput. Noticing that backhaul becomes a bottleneck in small cell networks (SCNs) (and therefore in ultra dense networks (UDNs) of 5G) while disk size increases quickly at a relatively low cost, the authors in [12] suggested to replace backhaul links by equipping caches at the BSs. By optimizing the caching policies to serve more users under the constraints of file downloading time, large throughput gain was reported. Considering SCNs with backhaul of very limited capacity and caching files based on their popularity, the authors in [13] observed that the backhaul traffic load can be reduced by caching at the BSs. To minimize the total energy consumed by caching and by data transport between BSs or between BSs and servers, a policy of allocating cache size to BSs and service gateway (SGW) was optimized in [14]. To minimize total service cost, caching policy was optimized in [15] where the impact of multicast transmission was taken into account. In [16], data sharing among backhaul and cooperative beamforming were jointly optimized to minimize the backhaul cost and transmit power of cache-enabled systems. For heterogeneous networks, user access and content caching were jointly optimized to minimize the average access delay in [17], and a coded caching scheme was optimized to achieve information-theoretic bounds in [18].

For highly skewed demands, caches should be pushed to the edge, say SGW or BSs of cellular networks [13]. Compared with caching at the SGW, caching at the BSs creates higher levels of redundancy where more replicas of the same content are stored. Since caches also consume power, whether locally caching at the BSs can improve the EE of wireless access network still remains unknown. Somewhat related problems have been investigated in the context of CCN [9–11], but local caching in cellular networks brings new challenges. In CCN, the energy can be effectively saved by reducing user-content distances and eliminating duplicated transmissions. Yet in wireless access networks, duplicated transmissions over the air cannot be removed due to the asynchronous requests from the users [7] despite that caching at the BSs can reduce the traffic load in core and backhaul networks. Instead, in dense cellular networks the energy can be reduced by turning BSs into sleep mode with no or light traffic load [19] and by controlling interference. Furthermore, many factors have entangled impact on the EE of wireless access networks such as backhaul capacity, interference level, power consumption parameters, BS density, BS sleeping, and user access, not to

mention the content popularity, cache size (i.e., cache capacity) and caching policy.

In this paper, we attempt to explore the potential of EE in cache-enabled wireless access networks and identify the key impacting factors. Specifically, we strive to answer the following fundamental questions.

- Will caching at the BSs bring an EE gain? If yes, what is the condition?
- What is the relation between EE and cache size? Is there a tradeoff or does the cache size should be optimized?
- What is the impact of network density? Where to cache in the access networks is more energy efficient?

To this end, we consider a downlink multicell multiuser multi-antenna network. In order to show the EE gain of caching at the BSs over caching at the SGW (i.e., not caching at the BSs), we assume that the contents have been placed at the caches of the BSs by broadcasting during off-peak times, and hence we consider the energy consumed for content delivery but ignore the energy consumed for cache placement. With the aim of finding critical factors that impact the EE gain, we optimize the configuration in cache placement phase (i.e., where to cache and how much to cache) and in delivery phase (i.e., maximal transmit power of each BS) based on statistics of the user demands, where different levels of interference are considered.

The major contributions of this paper are summarized as follows.

- We derive the closed-form expression of approximated EE for cache-enabled networks, where the consumption of transmit and circuit powers at the BSs, and the power consumption for backhauling and caching at the BSs are taken into account.
- We provide the condition when EE can benefit from caching, find the optimal cache capacity that maximizes the network EE, and analyze the maximal EE gain brought by caching.
- We show that caching at the BSs may not improve the network EE. When caching brings an EE gain, caching more contents at the BSs may not always increase the EE. Both numerical and simulation results show that caching at pico BSs can provide higher EE gain than caching at macro BSs.

The rest of this paper is organized as follows. In Section II, we present the system model. The EE of the cache-enabled access network is derived and analyzed in Section III and Section IV, respectively. The numerical and simulation results are provided in Section V, and the conclusions are drawn in Section VI.

II. System Model

Consider a downlink network consisting of N_b BSs. Each BS is with N_t antennas and serves multiple users each with a single antenna. Each BS is equipped with a cache and is connected to the core network with backhaul. In order to understanding the potential of EE of the cache-enabled wireless networks and identifying the key impacting factors, we make the following assumptions in the analysis, which define a simple scenario but can capture the basic elements.

- We use circle cells each with radius D to approximate hexagonal cells for easy analysis.
- Each content is of equal size F bits as in [10, 12, 20] for mathematical tractability and notational simplicity.[1]
- The content popularity distribution changes with time slowly [12] so that can be regarded as static and the energy consumption for refreshing the cached content can be safely neglected. Specifically, we consider a static content catalog that contains N_f contents, ranking from the most popular (the $1st$ content) to the least popular (the N_fth content) based on the popularity. In practice, Zipf-like distribution is widely applied to characterize many real world phenomena. Assume that each user requests one content from the catalog, and the probability of requesting the fth content is [21],
$$p_f = \frac{f^{-\delta}}{\sum_{j=1}^{N_f} j^{-\delta}} \tag{1}$$
where the typical value of δ is between 0.5 and 1.0, which determines the "peakiness" of the distribution [22]. Since δ reflects different levels of skewness of the distribution, it is called skew parameter.
- The spatial distribution of the users is modeled as homogeneous Poisson point process (PPP) [23, 24] where the average number of users in the whole network is λ.[2] Then, the probability that there are K users in each cell is $\frac{(\lambda/N_b)^K}{K!}e^{-\lambda/N_b}$.
- Each user is associated with the closest BS,[3] which is called its local BS, and each BS caches N_c most popular contents. In fact, with the static content catalog, when each user is associated with its local BS and the users' requests are with identical distribution, caching most popular contents everywhere is the optimal caching strategy in terms of maximizing the cache hit ratio [7].
- Each BS serves the associated users with zero-forcing beamforming (ZFBF), which is a widely-used precoder to eliminate multi-user interference [26], and with equal power allocation among multiple users.[4]

Denote $\mathcal{C}_b = \{1, 2, \cdots, N_c\}$ as the set of the contents cached at the bth BS (denoted by BS$_b$), $b = 1, \cdots, N_b$, then the cache capacity of each BS is N_cF. When a user requests a content that is cached at its local BS, the BS will fetch the content from the cache directly and then transmit to the user.

[1]When the content size is random, we can show that the performance depends on the average content size, and the main results do not change.

[2]When this assumption does not hold, say, if the users are distributed within hotpot areas, the network EE will become lower due to stronger interference. Nonetheless, the main results still hold.

[3]User association based on instantaneous channel gain will cause unnecessary handovers (i.e., the so-called "ping-pong effect") [25]. For mathematical tractability, we do not consider shadowing, which will not change the main trends of the performance.

[4]Optimizing power allocation is rather involved in the considered setting with limited-capacity backhaul. Moreover, the closed-formed expression even for an approximated EE with optimal power allocation is hard to obtain if not impossible. Equal power allocation provides an EE lower bound, which however can reflect the main trends of the EE and becomes near optimal when signal-to-interference-plus-noise ratio (SINR) is high.

Otherwise, the BS will fetch the content from the core network via backhaul link.

To reduce energy consumption and avoid interference, we consider BS idling ranging from very short period (less than 1 ms) to longer period (e.g., 100 ms) [19]. Once a BS has no user to serve, the BS is turned into idle mode. Otherwise, the BS operates in active mode. The probability that BS_b is active is $p_a = 1 - e^{-\lambda/N_b}$ according to the spatial distribution of users. Since we do not restrict the type of caching hardwares where some of them can not be switched off when contents are cached (e.g., Dynamic Random Access Memory (DRAM)), we do not consider cache idling.[5]

The network EE is associated with the throughput, which largely depends on the interference level. To capture the essence of the problem and simplify the analysis, we introduce a parameter to reflect the portion of inter-cell interference (ICI) able to be removed in a network, ranging from the best case to the worst case, as detailed later. When the user density is high such that the number of users in a cell exceeds N_t, we can select several users to serve according to a certain criterion. When round-robin scheduling is used to select N_t users to serve, the probability that BS_b serves K_b users can be derived as

$$p_{K_b} = \begin{cases} \left(\frac{\lambda}{N_b}\right)^{K_b} \frac{1}{K_b!} e^{-\frac{\lambda}{N_b}}, & \text{if } K_b < N_t \\ 1 - \sum_{k=0}^{N_t-1} \left(\frac{\lambda}{N_b}\right)^{k} \frac{1}{k!} e^{-\frac{\lambda}{N_b}}, & \text{if } K_b = N_t \end{cases} \tag{2}$$

The probability for other user scheduling can also be derived, which is not shown for conciseness.

Denote $\mathbf{H}_b = [\sqrt{r_{1b}^{-\alpha}}\mathbf{h}_{1b}, \cdots, \sqrt{r_{K_b b}^{-\alpha}}\mathbf{h}_{K_b b}]$ as the downlink channel matrix from BS_b to the K_b users located in the bth cell, where r_{kb} and $\mathbf{h}_{kb}$ are respectively the distance and the small-scale Rayleigh fading channel vector from BS_b to the kth user (denoted by MS_k), and α is the path-loss exponent. When perfect channel is available at each BS, the ZFBF vector at BS_b can be computed as $\mathbf{W}_b = \frac{1}{\sqrt{K_b}}[\mathbf{w}_{1b}, \cdots, \mathbf{w}_{K_b b}]$, where $\mathbf{w}_{kb} = \bar{\mathbf{w}}_{kb}/\|\bar{\mathbf{w}}_{kb}\|$, $\bar{\mathbf{w}}_{kb}$ denotes the kth column vector of $(\mathbf{H}_b^H)^\dagger$, $(\cdot)^\dagger$, $(\cdot)^H$, and $\|\cdot\|$ stand by the Moore-Penrose inverse, conjugate transpose, and Euclidean norm, respectively.

Then, the instantaneous receive SINR of MS_k served by BS_b when the BS is active is

$$\gamma_{kb} = \frac{P r_{kb}^{-\alpha} |\mathbf{h}_{kb}^H \mathbf{w}_{kb}|^2}{K_b(\beta P I_k + \sigma^2)} \tag{3}$$

where $I_k \triangleq \sum_{j=1, j\neq b}^{N_b} \zeta_j r_{kj}^{-\alpha} \|\mathbf{h}_{kj}\mathbf{W}_j\|^2$ is the power of ICI normalized by the transmit power P at BS, ζ_j is an indicator for the status of BS_j, $\zeta_j = 1$ if BS_j is active, $\zeta_j = 0$ otherwise, σ^2 is the variance of the white Gaussian noise, and $\beta \in [0, 1]$ reflects the percentage of how much ICI can be removed by some sort of interference management techniques. For example, $\beta = 0$ reflects the optimistic scenario, where all ICIs are assumed to be completely eliminated. $\beta = 1$ reflects the pessimistic case, where no interference coordination is assumed among the BSs.

[5]Some cache hardwares such as hard drive disk (HDD) or solid state disk (SSD) can be switched off without losing the cached contents. When a BS is turned into in deep sleep (e.g., with period in hours), these cache hardwares can be switched off to further reduce energy consumption.

Considering that the requested contents not cached at BS_b need to be fetched via backhaul and the backhaul traffic load is constrained by the backhaul capacity, the instantaneous downlink throughput of the bth cell can be expressed as

$$R_b = \zeta_b \Bigg(\underbrace{B \sum_{f_k \in \mathcal{C}_b} \log_2(1+\gamma_{kb})}_{R_{b,\mathrm{ca}}} + \underbrace{\min\left(B \sum_{f_k \notin \mathcal{C}_b} \log_2(1+\gamma_{kb}), C_{\mathrm{bh}}\right)}_{R_{b,\mathrm{bh}}} \Bigg) \tag{4}$$

where f_k denotes the index of the content requested by MS_k, B is the downlink transmission bandwidth, C_{bh} is the backhaul capacity, and the $\min(x, y)$ function returns the smallest value between x and y.

The first term $R_{b,\mathrm{ca}}$ in (4) is the sum rate of the users in the bth cell whose requested contents are cached at the BS, called *cache-hit users*. The second term $R_{b,\mathrm{bh}}$ is the sum rate of the users whose requested contents are not cached at the BS, called *cache-miss users*.

III. EE of the Cache-Enabled Network

The EE of the downlink network is defined as the ratio of the average number of bits transmitted to the average energy consumed [27–29], which is equivalent to the ratio of the average throughput of the network to the average total power consumption at the BSs

$$EE = \frac{\mathbb{E}\left\{\sum_{b=1}^{N_b} R_b\right\}}{\mathbb{E}\left\{\sum_{b=1}^{N_b} P_{b,\mathrm{BS}}\right\}} \triangleq \frac{\bar{R}}{\bar{P}_{\mathrm{tot}}} \tag{5}$$

where the expectations are taken over small scale fading, user location and the number of users in the network,[6] and $P_{b,\mathrm{BS}}$ is the total power consumed at BS_b, which will be detailed later.

In the following, we first derive the average throughput, and then derive the average total power consumption, from which we can obtain the EE of the network.

A. Average Throughput of the Network

Since the system configuration, caching and transmission strategies of every BS are the same and the users are uniformly located, the average throughput of the network can be obtained as

$$\bar{R} = \mathbb{E}\left\{\sum_{b=1}^{N_b} R_b\right\} = N_b \mathbb{E}\{R_b\} \tag{6}$$

and the average throughput of the bth cell can be expressed as

$$\mathbb{E}\{R_b\} = \sum_{K_b=1}^{N_t} \sum_{K_c=0}^{K_b} p_{(K_b,K_c)} \mathbb{E}\{R_b | (K_b, K_c)\} \tag{7}$$

[6]In this paper, unless otherwise specified, the expectation operator $\mathbb{E}\{\cdot\}$ is taken over all random variables (RVs) inside "$\{\cdot\}$".

where $p_{(K_b,K_c)}$ denotes the probability that K_b users are served by BS_b meanwhile K_c of them are cache-hit users, and $\mathbb{E}\{R_b|(K_b,K_c)\}$ is the average throughput of the bth cell under the condition that K_b users are served by BS_b meanwhile K_c of them are cache-hit users.

Using the conditional probability formula, we have $p_{(K_b,K_c)} = p_{K_b} \cdot p_{K_c|K_b}$, where p_{K_b} is given in (2), and $p_{K_c|K_b}$ denotes the probability of K_c users requesting the contents from local cache under the condition that BS_b serves K_b users, which can be expressed as

$$p_{K_c|K_b} = \binom{K_b}{K_c} p_h^{K_c} (1-p_h)^{K_b-K_c} \tag{8}$$

where p_h is the probability that $f_k \in \mathcal{C}_b$ (i.e., the *cache hit ratio*), which can be obtained from the Zipf-like distribution probability in (1) as

$$p_h = \sum_{f=1}^{N_c} p_f = \frac{\sum_{f=1}^{N_c} f^{-\delta}}{\sum_{j=1}^{N_f} j^{-\delta}} \tag{9}$$

Without loss of generality, we assume that the contents requested by $\mathrm{MS}_1, \cdots, \mathrm{MS}_{K_c}$ are cached at BS_b and the contents requested by $\mathrm{MS}_{K_c+1}, \cdots, \mathrm{MS}_{K_b}$ are not cached at BS_b. Then, from (4), the conditional expectation of the average throughput of the bth cell is given by

$$\mathbb{E}\{R_b|(K_b,K_c)\} = \bar{R}_{\mathrm{ca}}(K_b,K_c) + \bar{R}_{\mathrm{bh}}(K_b,K_c,C_{\mathrm{bh}}) \tag{10}$$

where $\bar{R}_{\mathrm{ca}}(K_b,K_c) \triangleq \mathbb{E}\big\{B\sum_{k=1}^{K_c}\log_2(1+\gamma_{kb})\big\}$ is the average sum rate of the cache-hit users, and $\bar{R}_{\mathrm{bh}}(K_b,K_c,C_{\mathrm{bh}}) \triangleq \mathbb{E}\big\{\min\big(B\sum_{k=K_c+1}^{K_b}\log_2(1+\gamma_{kb}), C_{\mathrm{bh}}\big)\big\}$ is the average sum rate of the cache-miss users.

To obtain a closed-form expression of EE for further analysis, we derive the approximated $\bar{R}_{\mathrm{ca}}(K_b,K_c)$ and $\bar{R}_{\mathrm{bh}}(K_b,K_c,C_{\mathrm{bh}})$ in the following two lemmas.

Lemma 1: The average sum rate of the cache-hit users can be approximated as

$$\begin{aligned}\bar{R}_{\mathrm{ca}}(K_b,K_c) &\approx K_c B\left(\frac{\alpha}{2\ln 2} + \log_2\frac{(N_t-K_b+1)P}{K_b(p_a\beta P 2^{\Phi} + D^{\alpha}\sigma^2)}\right)\\ &\triangleq K_c\left(\frac{\alpha B}{2\ln 2} + \bar{R}_{\mathrm{e}}(K_b)\right)\end{aligned} \tag{11}$$

where Φ is a constant only depending on the path-loss exponent α when $N_b \to \infty$, $\bar{R}_{\mathrm{e}}(K_b) \triangleq B\log_2\frac{(N_t-K_b+1)P}{K_b(p_a\beta P 2^{\Phi}+D^{\alpha}\sigma^2)}$ can be regarded as the average achievable rate of a cell-edge user when BS_b serves K_b users under unlimited-capacity backhaul.

Proof: See Appendix A. ∎

The approximation of $\bar{R}_{\mathrm{ca}}(K_b,K_c)$ is accurate when both SINR and $\frac{\lambda}{N_b}$ are high.

Lemma 2: The average sum rate of the cache-miss users can be approximated as

$$\bar{R}_{\mathrm{bh}}(K_b,K_c,C_{\mathrm{bh}}) \approx \begin{cases}(K_b-K_c)\big(\frac{\alpha B}{2\ln 2}\gamma(K_b-K_c+1,z) + \bar{R}_{\mathrm{e}}(K_b)\gamma(K_b-K_c,z)\big)\\ \quad + C_{\mathrm{bh}}\Gamma(K_b-K_c,z), & \text{if } C_{\mathrm{bh}} > (K_b-K_c)\bar{R}_{\mathrm{e}}(K_b)\\ C_{\mathrm{bh}}, & \text{otherwise}\end{cases} \tag{12}$$

where $z \triangleq \frac{2\ln 2}{\alpha B}\big(C_{\mathrm{bh}} - (K_b-K_c)\bar{R}_{\mathrm{e}}(K_b)\big)$, $\Gamma(k,x) \triangleq e^{-x}\sum_{i=0}^{k-1}\frac{x^i}{i!}$, and $\gamma(k,x) \triangleq 1 - e^{-x}\sum_{i=0}^{k-1}\frac{x^i}{i!}$.

Proof: See appendix C. ∎

The approximation is accurate in high SINR region when $\frac{\lambda}{N_b}$ is high and $N_t, N_b \to \infty$.

Substituting (10) into (7) and then into (6), we obtain the network average throughput as

$$\bar{R} = N_b\sum_{K_b=1}^{N_t}\sum_{K_c=0}^{K_b} p_{K_b}p_{K_c|K_b}\big(\bar{R}_{\mathrm{ca}}(K_b,K_c) + \bar{R}_{\mathrm{bh}}(K_b,K_c,C_{\mathrm{bh}})\big) \tag{13}$$

where p_{K_b} is given in (2), $p_{K_c|K_b}$ is given in (8), and the approximations of $\bar{R}_{\mathrm{ca}}(K_b,K_c)$ and $\bar{R}_{\mathrm{bh}}(K_b,K_c,C_{\mathrm{bh}})$ are given in (11) and (12), respectively.

B. Average Total Power Consumption

To gain useful insight, we consider a basic model for such cache-enabled networks capturing the fundamental challenges and tradeoffs. By extending the typical BS power consumption model in [30] to include caching power consumption, the total power consumed at BS_b can be modeled as follows,

$$P_{b,\mathrm{BS}} = \rho P_{b,\mathrm{tx}} + P_{b,\mathrm{cc}} + P_{b,\mathrm{ca}} + P_{b,\mathrm{bh}} \tag{14}$$

where $P_{b,\mathrm{tx}}$, $P_{b,\mathrm{cc}}$, $P_{b,\mathrm{ca}}$, and $P_{b,\mathrm{bh}}$ respectively denote the power consumed at BS_b for transmitting, operating the base-band and radio frequency circuits, caching, and backhauling, and ρ reflects the impact of power amplifier, cooling and power supply.

The transmit power of BS_b is $P_{b,\mathrm{tx}} = P$ when the BS is in active mode or $P_{b,\mathrm{tx}} = 0$ when the BS is idle. The circuit power is $P_{b,\mathrm{cc}} = P_{\mathrm{cc}_a}$ in active mode or P_{cc_i} in idle mode. Since the active status of the BSs are independent from each other, the total number of active BSs in the network (denoted by N_a) follows Binomial distribution, and hence $\mathbb{E}\{N_a\} = N_b p_a = N_b(1-e^{-\frac{\lambda}{N_b}})$. Therefore, the average total transmit and circuit power consumption at all BSs is

$$\begin{aligned}&\mathbb{E}\left\{\sum_{b=1}^{N_b}\rho P_{b,\mathrm{tx}} + P_{b,\mathrm{cc}}\right\}\\ &= \mathbb{E}\{N_a\}(\rho P + P_{\mathrm{cc}_a}) + (N_b - \mathbb{E}\{N_a\})P_{\mathrm{cc}_i}\\ &= N_b(1-e^{-\frac{\lambda}{N_b}})P_a + N_b e^{-\frac{\lambda}{N_b}}P_i\end{aligned} \tag{15}$$

where $P_a \triangleq \rho P + P_{\mathrm{cc}_a}$ and $P_i \triangleq P_{\mathrm{cc}_i}$ are the total transmit and circuit power consumptions at a BS in the active mode and idle mode, respectively.

Energy-proportional model is widely used in CCN [9–11] as well as radio access network (RAN) [14], which enables the efficient use of caching resources. In this model, the caching power consumption is proportional to the cache capacity, which can be expressed as $P_{b,\mathrm{ca}} = w_{\mathrm{ca}}B_{\mathrm{ca}}$ [9], where B_{ca} is the number of bits cached at BS_b, and w_{ca} is the power coefficient of caching hardware in watt/bit. Since the cached contents of each BS are fixed, when each BS caches N_c contents, the average total caching power consumption of all BSs is

$$\mathbb{E}\left\{\sum_{b=1}^{N_b}P_{b,\mathrm{ca}}\right\} = N_b P_{b,\mathrm{ca}} = N_b w_{\mathrm{ca}} N_c F \tag{16}$$

The backhauling power consumption at BS$_b$ is modeled as [31]

$$P_{\rm b,bh} = \frac{P^0_{\rm bh} R_{b,\rm bh}}{C^0_{\rm bh}} \triangleq w_{\rm bh} R_{b,\rm bh} \tag{17}$$

where $P^0_{\rm bh}$ denotes the power consumed by the backhaul equipment when supporting the maximum data rate $C^0_{\rm bh}$, $w_{\rm bh} \triangleq P^0_{\rm bh}/C^0_{\rm bh}$ is the power coefficient of backhaul equipment, and $R_{b,\rm bh}$ is the backhaul traffic, i.e., the sum rate of cache-miss users as defined in (4). Then, the average backhaul power consumption is

$$\mathbb{E}\left\{\sum_{b=1}^{N_b} P_{b,\rm bh}\right\} = w_{\rm bh}\mathbb{E}\left\{\sum_{b=1}^{N_b} R_{b,\rm bh}\right\} = w_{\rm bh} N_b \mathbb{E}\{R_{b,\rm bh}\} \tag{18}$$

Similar to the derivations for (7) and (10), we can derive that

$$\begin{aligned}\mathbb{E}\{R_{b,\rm bh}\} &= \sum_{K_b=1}^{N_t}\sum_{K_c=0}^{K_b} p_{(K_b,K_c)}\mathbb{E}\{R_{b,\rm bh}|(K_b,K_c)\} \\ &= \sum_{K_b=1}^{N_t}\sum_{K_c=0}^{K_b} p_{K_b}\cdot p_{K_c|K_b}\bar{R}_{\rm bh}(K_b,K_c,C_{\rm bh})\end{aligned} \tag{19}$$

Then, the average total power consumption at all the BSs is

$$\begin{aligned}\bar{P}_{\rm tot} = N_b\Bigg(&\left(1-e^{-\frac{\lambda}{N_b}}\right)P_a + e^{-\frac{\lambda}{N_b}}P_i + w_{\rm ca}N_cF \\ &+ w_{\rm bh}\sum_{K_b=1}^{N_t}\sum_{K_c=0}^{K_b} p_{K_b}p_{K_c|K_b}\bar{R}_{\rm bh}(K_b,K_c,C_{\rm bh})\Bigg)\end{aligned} \tag{20}$$

C. EE of the Network

By substituting (13) and (20) into (5), the EE of the network can be obtained as (21). With the approximated $\bar{R}_{\rm ca}(K_b,K_c)$ and $\bar{R}_{\rm bh}(K_b,K_c,C_{\rm bh})$ introduced in the two lemmas, it is of closed-form and becomes an approximated EE.

Despite that the approximated EE is in closed form, it is still complex for further analysis. To gain useful insight on how caching impacts the network EE, in the sequel we analyze a special scenario where each BS selects one user in each time slot from the associated users [24, 32].

IV. EE Analysis for the Cache-enabled Network

In this section, we analyze the impact of several key factors on the EE and reveal their interactions for a special case when each BS serves at most one user in each time slot.[7]

In this case, the average throughput of the network in (13) degenerates into,

$$\bar{R} = N_b p_a\big(p_h\bar{R}_{\rm ca} + (1-p_h)\bar{R}_{\rm bh}\big) \tag{22}$$

where $\bar{R}_{\rm ca}$ and $\bar{R}_{\rm bh}$ are respectively the approximate average achievable rate of cache-hit user and cache-miss user derived from (11) and (12) as

$$\bar{R}_{\rm ca} \approx \frac{\alpha B}{2\ln 2} + \bar{R}_{\rm e} \tag{23}$$

[7]This can be also regarded as a special case where no more than one user exists in each cell.

$$\bar{R}_{\rm bh} \approx \begin{cases} C_{\rm bh}, & \text{if } C_{\rm bh} \le \bar{R}_{\rm e} \\ \frac{\alpha B}{2\ln 2} + \bar{R}_{\rm e} - \frac{\alpha B}{2\ln 2}2^{-\frac{2(C_{\rm bh}-\bar{R}_{\rm e})}{\alpha B}}, & \text{otherwise}\end{cases} \tag{24}$$

and $\bar{R}_{\rm e} = B\log_2\frac{N_tP}{p_a\beta P 2^{\Phi}+D^{\alpha}\sigma^2}$ is given by (11).

Remark 1: The average throughput of the network increases with the cache hit ratio p_h and the backhaul capacity $C_{\rm bh}$. In other words, we can improve the throughput by caching more contents and increasing backhaul capacity. When $C_{\rm bh}$ is low and the contents are not with uniform popularity (i.e., $\delta > 0$), the throughput increases with the cache size first rapidly then saturates, i.e., there is a *tradeoff between throughput and memory*.

The backhauling power consumption in (18) degenerates into

$$\begin{aligned}&\mathbb{E}\left\{\sum_{b=1}^{N_b} P_{\rm bh}\right\} = w_{\rm bh}N_b p_a(1-p_h)\bar{R}_{\rm bh} \\ &= \begin{cases} w_{\rm bh}N_b p_a(1-p_h)C_{\rm bh}, & \text{if } C_{\rm bh} \le \bar{R}_{\rm e} \\ w_{\rm bh}N_b p_a(1-p_h)\big(\bar{R}_{\rm ca} - \frac{\alpha B}{2\ln 2}2^{-\frac{2(C_{\rm bh}-\bar{R}_{\rm e})}{\alpha B}}\big), & \text{otherwise}\end{cases}\end{aligned} \tag{25}$$

which decreases with p_h but increases with $C_{\rm bh}$.

Substituting (22), (25), (15) and (16) into (5), the EE of the network can be approximated as,

$$EE \approx \frac{p_a\big(p_h\bar{R}_{\rm ca} + (1-p_h)\bar{R}_{\rm bh}\big)}{p_aP_a + (1-p_a)P_i + w_{\rm ca}N_cF + p_a w_{\rm bh}(1-p_h)\bar{R}_{\rm bh}} \tag{26}$$

where $p_a p_h \bar{R}_{\rm ca}$ and $p_a(1-p_h)\bar{R}_{\rm bh}$ are the average sum rates of the cache-hit and cache-miss users of each cell, $p_aP_a + (1-p_a)P_i$, $w_{\rm ca}N_cF$ and $p_a w_{\rm bh}(1-p_h)\bar{R}_{\rm bh}$ are the average powers consumed for transmission and circuits, caching, and backhauling of each BS, respectively.

Given that the caches in the network somewhat play a role of replacing the backhaul links, and the transmit power affects both the throughput and the total power consumption, the cache capacity N_cF, backhaul capacity $C_{\rm bh}$, and the transmit power of each BS P have an interactive impact on the EE. In what follows, we separately analyze the relation between the network EE and cache capacity or transmit power for a given backhaul capacity. To simplify the analysis, we only consider the case where $\delta = 1$ in the following. The impact of other values of δ will be evaluated later by simulations.

A. Relation Between Network EE and Cache Capacity

With given backhaul capacity and transmit power, we first answer the following question: *whether caching at the BSs can always improve the network EE?*

Proposition 1: When the following condition holds,

$$w_{\rm ca}F\sum_{j=1}^{N_f} j^{-1} < \left(\frac{\bar{R}_{\rm ca}}{\bar{R}_{\rm bh}} - 1\right)\left(p_aP_a + (1-p_a)P_i\right) + p_a w_{\rm bh}\bar{R}_{\rm ca} \tag{27}$$

caching can improve the network EE. Otherwise, caching can not improve the EE.

Proof: See Appendix D. ∎

To help understand this condition, we consider two extreme cases in the following corollary.

$$EE = \frac{\sum_{K_b=1}^{N_t}\sum_{K_c=0}^{K_b} p_{K_b}p_{K_c|K_b}\left(\bar{R}_{\rm ca}(K_b,K_c)+\bar{R}_{\rm bh}(K_b,K_c,C_{\rm bh})\right)}{\left(1-e^{-\frac{\lambda}{N_b}}\right)P_a + e^{-\frac{\lambda}{N_b}}P_i + w_{\rm ca}N_cF + w_{\rm bh}\sum_{K_b=1}^{N_t}\sum_{K_c=0}^{K_b} p_{K_b}p_{K_c|K_b}\bar{R}_{\rm bh}(K_b,K_c,C_{\rm bh})} \tag{21}$$

Corollary 1: When $C_{\rm bh}=0$, caching at BSs can always improve the network EE. When $C_{\rm bh}\to\infty$, the condition in (27) becomes,

$$\frac{p_a w_{\rm bh}\bar{R}_{\rm ca}}{w_{\rm ca}F} > \sum_{j=1}^{N_f} j^{-1} \approx \ln N_f \tag{28}$$

Proof: When $C_{\rm bh}=0$, it is easy to see that (27) always holds. When $C_{\rm bh}\to\infty$, it is shown from (23) and (24) that $\lim_{C_{\rm bh}\to\infty}\bar{R}_{\rm bh}=\bar{R}_{\rm ca}$. Then, by substituting $\bar{R}_{\rm bh}=\bar{R}_{\rm ca}$ and using $\sum_{j=1}^{N_f} j^{-1}=\varepsilon+\ln N_f+\mathcal{O}(\frac{1}{N_f})$ with $\varepsilon\approx 0.577$ as the Euler-Mascheroni constant, (27) becomes (28) and the approximation is accurate when $N_f\gg 1$. ■

Remark 2: In (28), $p_a w_{\rm bh}\bar{R}_{\rm ca}$ is the average backhaul power consumption of each BS without caching, and $w_{\rm ca}F$ is the average cache power consumption of each BS when only the most popular content is cached at each BS. This suggests that whether caching benefits EE largely depends on the power consumption parameters for the cache and backhaul hardwares.

In what follows, we consider the scenario where the condition holds, and strive to answer the second question: *what is the relation between maximal EE of the network and the cache size?* To this end, we first provide the cache hit ratio p_h for large values of N_c and N_f. To reflect the impact of the content catalog size N_f, we analyze a normalized cache capacity $\eta\triangleq N_c/N_f$, $\eta\in[0,1]$. Then, from (9) we can derive

$$p_h = \frac{\sum_{f=1}^{N_c} f^{-1}}{\sum_{j=1}^{N_f} j^{-1}} = \frac{\varepsilon+\ln N_c+\mathcal{O}(\frac{1}{N_c})}{\varepsilon+\ln N_f+\mathcal{O}(\frac{1}{N_f})} \approx \frac{\ln N_c}{\ln N_f} = 1+\frac{\ln\eta}{\ln N_f} \tag{29}$$

where the approximation in (29) is accurate when $N_c\gg 1$ and $N_f\gg 1$.

By substituting (29) into (26), we can approximate the network EE as

$$EE \approx \frac{p_a\left(\bar{R}_{\rm bh}+(\bar{R}_{\rm ca}-\bar{R}_{\rm bh})\left(1+\frac{\ln\eta}{\ln N_f}\right)\right)}{p_aP_a+(1-p_a)P_i+w_{\rm ca}\eta N_fF-p_aw_{\rm bh}\bar{R}_{\rm bh}\frac{\ln\eta}{\ln N_f}} \tag{30}$$

Denote $W(x)$ as the Lambert-W function satisfying $W(x)e^{W(x)}=x$. Then, the relation between EE and cache capacity is shown in the following proposition.

Proposition 2: The solution of the equation $\frac{dEE}{d\eta}\big|_{\eta=\eta_0}=0$ is

$$\eta_0 = \frac{\Omega}{N_fW\left(\Omega e^{-1+\frac{\bar{R}_{\rm bh}}{\bar{R}_{\rm ca}-\bar{R}_{\rm bh}}\ln N_f}\right)} \tag{31}$$

where

$$\Omega \triangleq \frac{\frac{\bar{R}_{\rm ca}\bar{R}_{\rm bh}}{\bar{R}_{\rm ca}-\bar{R}_{\rm bh}}w_{\rm bh}p_a+p_aP_a+(1-p_a)P_i}{w_{\rm ca}F} \tag{32}$$

When $\eta_0<1$, the EE-maximal normalized cache capacity is $\eta^*=\eta_0$. When $\eta_0\geq 1$, $\eta^*=1$.

Proof: See Appendix E. ■

Remark 3: If $\eta_0<1$, the EE will first increase and then decrease with the cache capacity. Otherwise, if $\eta_0\geq 1$, the EE will be maximized when all contents in the catalog are cached at each BS, i.e., there is a *tradeoff between the maximal EE and the cache size*.

To understand when the EE-memory tradeoff exists, we rewrite (31) in a form of $\frac{x}{W(x)}$ as

$$\eta_0 = \frac{\Omega e^{-1+\frac{\bar{R}_{\rm bh}}{\bar{R}_{\rm ca}-\bar{R}_{\rm bh}}\ln N_f}}{W\left(\Omega e^{-1+\frac{\bar{R}_{\rm bh}}{\bar{R}_{\rm ca}-\bar{R}_{\rm bh}}\ln N_f}\right)}\cdot\frac{e^{1-\frac{\bar{R}_{\rm bh}}{\bar{R}_{\rm ca}-\bar{R}_{\rm bh}}\ln N_f}}{N_f} \tag{33}$$

As shown in (32), Ω increases with the average power consumed for transmission and circuits at each BS $p_aP_a+(1-p_a)P_i$ and the backhaul power coefficient $w_{\rm bh}$, and decreases with the content size F and cache power coefficient $w_{\rm ca}$. Further considering that $\frac{x}{W(x)}$ is an increasing function of x [33], η_0 increases with $p_aP_a+(1-p_a)P_i$ and $w_{\rm bh}$, and decreases with F and $w_{\rm ca}$. Moreover, it is shown from (31) that η_0 increases when the content catalog size N_f decreases since $W(x)$ an increasing function of x [33].

Remark 4: $\eta_0\geq 1$ for the systems with high transmit power, large circuit and backhauling power consumptions, power-saving caching hardware, small content size F and small catalog size N_f. Otherwise, $\eta_0<1$, where caching more contents is not always energy efficient.

To further identify the key impacting factors on network EE and gain useful insight on network configuration, in what follows we consider the case when backhaul capacity is unlimited.

1) An Extreme Case of $C_{\rm bh}\to\infty$: In this case, $\lim_{C_{\rm bh}\to\infty}\bar{R}_{\rm bh}=\bar{R}_{\rm ca}$. Then, the network EE in (30) can be expressed as

$$\begin{aligned} EE &\approx \frac{p_a\bar{R}_{\rm ca}}{p_aP_a+(1-p_a)P_i+w_{\rm ca}\eta N_fF-p_aw_{\rm bh}\frac{\ln\eta}{\ln N_f}\bar{R}_{\rm ca}} \\ &= \frac{p_a\bar{R}_{\rm ca}}{p_aP_a+(1-p_a)P_i+\bar{P}_{\rm ca}+\bar{P}_{\rm bh}} \end{aligned} \tag{34}$$

Remark 5: In (34), only the powers consumed for caching and backhauling depend on η. Because $\bar{P}_{\rm ca}$ increases with η linearly, while $\bar{P}_{\rm bh}$ decreases with η first rapidly and then slowly, the total power consumption first increases and then decreases with η. Hence, the relation between network EE and cache capacity relies on the trade-off between backhauling and caching powers.

From (34) and considering the expression of $\bar{R}_{\rm ca}$ in (23), we obtain the following corollary.

Corollary 2: When $C_{\rm bh}\to\infty$, the solution of the equation

$\frac{dEE}{d\eta}\big|_{\eta=\eta_0}=0$ is

$$\eta_0 = p_a \cdot \frac{w_{\rm bh}}{w_{\rm ca}} \cdot \frac{B}{F} \cdot \frac{1}{N_f \ln N_f}\left(\frac{\alpha}{2\ln 2}+\log_2 \frac{N_t}{p_a \beta 2^{\Phi}+\left(\frac{P}{D^{\alpha}\sigma^2}\right)^{-1}}\right) \tag{35}$$

where Φ is the constant only depending on α, and $\frac{P}{D^{\alpha}\sigma^2}$ is the average cell-edge signal-to-noise-ratio (SNR).

Remark 6: As shown in (35), η_0 increases with N_t and P. This suggests that BS with larger number of antennas and transmit power should cache more to achieve the maximal EE.

According to Proposition 2, when $\eta_0 \geq 1$, there exists a trade-off between EE and η. Considering that $y = x \ln x$ can be rewritten as $x = e^{W(y)}$, from $\eta_0 \geq 1$ and (35) we can obtain the following corollary.

Corollary 3: When $C_{\rm bh} \to \infty$, there exists a trade-off between EE and η if $N_f \leq N_{th}$, where

$$\begin{aligned} N_{th} &= e^{W\left(p_a \cdot \frac{w_{\rm bh}}{w_{\rm ca}} \cdot \frac{B}{F}\left(\frac{\alpha}{2\ln 2}+\log_2 \frac{N_t}{p_a\beta 2^{\Phi}+(P/D^{\alpha}\sigma^2)^{-1}}\right)\right)} \\ &= e^{W\left(p_a \cdot \frac{w_{\rm bh}}{w_{\rm ca}} \cdot \frac{\bar{R}_{\rm ca}}{F}\right)} \end{aligned} \tag{36}$$

Remark 7: As shown in (36), when the average cell-edge SNR is high, the interference level β dominates the value of $\bar{R}_{\rm ca}$. If the interference can be reduced to a low level, $\bar{R}_{\rm ca}$ will increase and the value of N_{th} will be large, and then the EE-memory trade-off will exist even for a large content catalog size.

Again according to Proposition 2, when $\eta_0 < 1$, the EE optimal normalized cache capacity is $\eta^* = \eta_0$. From (35), we can further analyze the impact of network density.

Corollary 4: When $C_{\rm bh} \to \infty$, for a given total coverage area of the cells $N_b \pi D^2$, $\eta^* = \eta_0$ decreases with N_b, and $N_b \eta$ increases with N_b for $\frac{\lambda}{N_b} \to 0$.

Proof: See Appendix F. ∎

Remark 8: Corollary 4 indicates that when the network becomes denser, each BS should cache less contents but the total cache capacity of the network should increase in order to maximize the network EE. Further considering that η_0 decreases with N_t and P as mentioned in Remark 6, this implies that a pico BS should cache less contents than a macro BS to achieve the maximal EE.

Since (35) gives the optimal cache capacity maixmizing the network EE when $\eta_0 < 1$, we can further analyze the impacts of different factors on the maximal EE gain brought by caching.

Corollary 5: When $C_{\rm bh} \to \infty$ and $\eta_0 < 1$, the gain of maximal EE with caching over that without caching is

$$EE_{\rm gain} = \frac{1}{1-G} \tag{37}$$

where

$$G = \frac{\frac{1}{\ln N_f}\left(\ln \frac{p_a w_{\rm bh} \bar{R}_{\rm ca}}{w_{\rm ca} F \ln N_f} - 1\right)}{\frac{p_a P_a + (1-p_a)P_i}{p_a \bar{R}_{\rm ca} w_{\rm bh}} + 1} \tag{38}$$

Proof: By substituting (35) into (34), we can obtain the maximal EE denoted as $EE_{\max}$. Denoting the network EE without caching (i.e., $N_c = 0$) as $EE_{\rm no}$, we can obtain the maximal EE gain with caching over that without caching as $EE_{\rm gain} \triangleq \frac{EE_{\max}}{EE_{\rm no}}$, which can be written as (37). ∎

Remark 9: As shown in (38), G increases with $\bar{R}_{\rm ca}$ since the numerator increases with $\bar{R}_{\rm ca}$ while the denominator decreases with $\bar{R}_{\rm ca}$. This implies that the EE gain of caching at the BSs can be improved significantly by mitigating ICI because the value of $\bar{R}_{\rm ca}$ largely depends on the interference level β as we mentioned before and $EE_{\rm gain}$ increases with G. We can also see from (38) that G increases when the ratio of total transmit and circuit power to the backhauling power without caching (i.e., $\frac{p_a P_a + (1-p_a)P_i}{p_a \bar{R}_{\rm ca} w_{\rm bh}}$) decreases. This implies that caching at the pico BSs may provide higher EE gain than caching at the macro BSs since backhaul power consumption usually takes a larger portion of the energy in the pico cells [34].

When $\eta_0 \geq 1$, the results are similar and the conclusion is the same.

B. Relation Between Network EE and Transmit Power

When the backaul capacity is unlimited, by substituting $\bar{R}_{\rm ca}$ in (23), and P_a and P_i in (15) into (34), the network EE can be expressed as a function of transmit power P as (39).

Corollary 6: When $C_{\rm bh} \to \infty$ and the network is interference limited, i.e., $p_a \beta P 2^{\Phi} \gg D^{\alpha}\sigma^2$,[8] the EE decreases with the transmit power P.

Proof: Since $p_a \beta P 2^{\Phi} \gg D^{\alpha}\sigma^2$, by omitting the term $D^{\alpha}\sigma^2$ in (39), we can see that EE decreases with the transmit power P. ∎

Corollary 7: When $C_{\rm bh} \to \infty$ and the network is noise limited, i.e., $p_a \beta P 2^{\Phi} \ll D^{\alpha}\sigma^2$, the EE first increases and then decreases with the transmit power, and the optimal transmit power maximizing the EE is

$$P_0 = \frac{(\bar{P}_{\rm cc} + \bar{P}_{\rm ca})}{p_a \rho W\left(\frac{p_a N_t (\bar{P}_{\rm cc}+\bar{P}_{\rm ca})}{\rho D^{\alpha}\sigma^2} e^{\frac{\alpha}{2}-1}\right)} \tag{40}$$

where $\bar{P}_{\rm cc} = p_a P_{{\rm cc}_a} + (1-p_a) P_{{\rm cc}_i}$ is the average circuit power consumption of each BS, and $\bar{P}_{\rm ca} = w_{\rm ca} \eta N_f F$ is the average cache power consumption of each BS.

Proof: See Appendix G. ∎

Remark 10: As shown in (40), P_0 increases with $\bar{P}_{\rm ca}$ since $\frac{x}{W(x)}$ increases with x. This means that the transmit power should increase with the cache capacity to maximize the EE.

We can show that the EE is not joint concave in η and P, despite that the EE is an unimodal function respectively of η and P when the network is noise limited. Therefore, the point (P_0, η_0) satisfying $\frac{dEE}{dP} = 0$ in (40) and $\frac{dEE}{d\eta} = 0$ in (35) may not be joint optimal. In the next section, we provide numerical results to show that (P_0, η_0) is joint optimal in the considered system setup.

When the backaul capacity is very low, i.e., $C_{\rm bh} \to 0$, almost the same results and conclusion can be obtained, which are not shown for conciseness.

From previous analysis in this section, we can draw the following conclusions.

- If the backhaul capacity is unlimited, then the average throughput of the network will not change no matter

[8]This condition can be rewritten as $\beta \gg \frac{1}{p_a 2^{\Phi}} \cdot \frac{D^{\alpha}\sigma^2}{P}$, which is $\beta \gg 0.015$ for $p_a = 0.8$ and 20 dB cell-edge SNR.

$$EE \approx \frac{p_a B \left(\frac{\alpha}{2\ln 2} + \log_2 \frac{N_t P}{p_a \beta P 2^{\Phi} + D^{\alpha}\sigma^2}\right)}{p_a(\rho P + P_{\mathrm{cc}_a}) + (1-p_a)P_{\mathrm{cc}_i} + w_{\mathrm{ca}} N_c F + p_a w_{\mathrm{bh}} B(1-p_h)\left(\frac{\alpha}{2\ln 2} + \log_2 \frac{N_t P}{p_a \beta P 2^{\Phi} + D^{\alpha}\sigma^2}\right)} \tag{39}$$

whether each BS is equipped with cache. If the backhaul is with limited capacity, there will exist a tradeoff between throughput and memory.

- Whether caching at the BSs brings an EE gain depends on the backhaul capacity, and the power consumption parameters of the cache and backhaul hardware.
- If the backhaul capacity is unlimited, the EE gain of caching will come from trading off the backhaul power consumption with the cache power consumption. If the backhaul capacity is limited, the caching gain will come from both the increase of network throughput and the decrease of backhaul power consumption.
- When the content catalog size is small, there is a tradeoff between EE and memory. Otherwise, the cache size of each BS should be optimized to maximize the EE of the network.

V. Numerical and Simulation Results

In this section, we validate the analysis and evaluate the EE of the cache-enabled networks. We show when caching at BSs has EE gain and how much gain we can expect in real systems.

While in the derivation we have assumed circle cells, in the simulation we consider a hexagonal region with radius 250 m. To demonstrate the impact of interference, we deploy three tiers of hexagonal pico cells in the region. Then, $N_b = 37$, and the radius of each pico cell is $D = \frac{250}{\sqrt{N_b}} \approx 40$ m. In order to remove the boundary effect, we deploy three more tiers of cells to ensure that every cell is surrounded by no less than three tiers of cells. Each pico BS is equipped with four antennas, and the transmission bandwidth is set as 20 MHz. The noise power is set as $\sigma^2 = -95$ dBm and the path-loss model is $30.6 + 36.7\log_{10}(r_{kb})$ in dB [35].[9] The catalog contains $N_f = 10^4$ contents each with size of $F = 30$ MB (MegaByte) [7]. Recall that the EE analysis in Section IV is obtained for a special scenario where each BS serves at most one user. To show that the analytical results are also true for more general scenarios, in the following, each BS can schedule at most N_t users with ZFBF. The user distribution in the whole network follows PPP and the average number of users in the network is $\lambda = 30$. Then, the ratio of user density to BS density is $\frac{\lambda}{N_b} \approx 0.8$.[10] The popularity of the contents follows Zipf-like distribution with typical parameter $\delta = 0.8$ [38]. The power consumption parameters of the system are $\rho = 15.13$, $P = 21$ dBm, P_{cc_i} is 3.85 W, P_{cc_a} is 10.16 W for typical pico BS [27], $w_{\mathrm{bh}} = 5 \times 10^{-7}$ J/bit for microwave backhaul link [31], and $w_{\mathrm{ca}} = 6.25 \times 10^{-12}$ W/bit for high-speed SSD [9]. Unless otherwise specified, the above setups will be used for all simulations and numerical results.

A. Validation of the Analysis

To validate the assumption that the energy consumption for content update is negligible when content popularity changes slowly, we estimate the energy consumption for updating contents. Suppose that u percent of the cached contents are updated at interval T. Then, the percentage of energy consumption for content update to the total energy consumption during T is

$$E_{\mathrm{ud}} = \frac{u N_b N_c F w_{\mathrm{bh}}}{T \bar{P}_{\mathrm{tot}}} \tag{41}$$

where uN_bN_cF is the total number of bits conveyed through backhaul links and thus $uN_bN_cFw_{\mathrm{bh}}$ is the energy consumed for updating contents. Considering that the popularity of many contents changes slowly,[11] we set $u = 10\%$ and $T = 12$ hours. Then, when $N_c = 10^3$, $E_{\mathrm{ud}} < 3\%$.

To validate the approximation made for $\mathbb{E}\{\log_2(\beta I_k + \frac{\sigma^2}{P})\}$ in Appendix A, we compare the simulation results of this term with the numerical results of its approximation given in (A.5) in Fig. 1. Since the term depends on $p_a = 1 - e^{-\frac{\lambda}{N_b}}$ and β, the results for different values of $\frac{\lambda}{N_b}$ and β are provided. We can see that the simulation and numerical results almost overlap for all values of $\beta \in [0, 1]$ especially when $\frac{\lambda}{N_b}$ is high, i.e., the approximation is accurate.

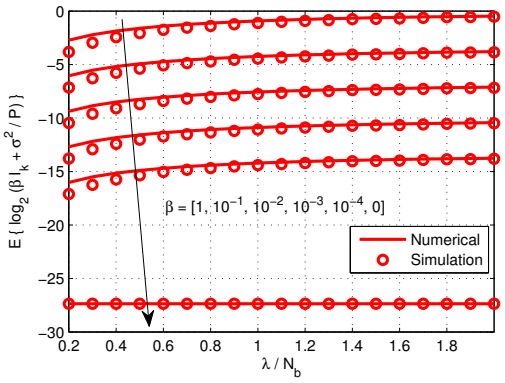


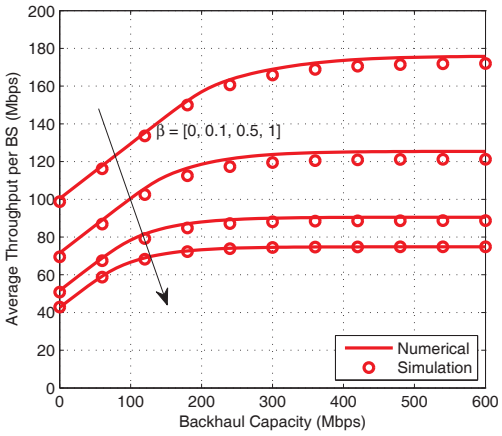
Fig. 1. The accuracy of the approximation of $\mathbb{E}\{\log_2(\beta I_k + \frac{\sigma^2}{P})\}$.

To validate the approximation introduced in (C.1), we compare the simulation results of the average throughput per cell with the numerical results obtained from (13) versus C_{bh} in Fig. 2(a). We can see that the simulation and numerical results almost overlap, i.e., the approximation is accurate, although $N_t = 4$ and $N_b = 37$ that are far from infinity. To show

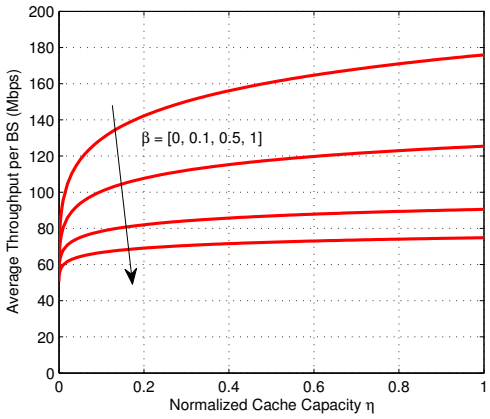
[9] In practice, the propagation environment may change and the line of sight (LoS) paths may exist between BS and user with a certain probability. In this scenario, the EE will reduce due to stronger ICI but the EE-cache size relation will not change.

[10] The ratio of user density to BS density is typically around one for SCNs [23, 36] and is much smaller than one for future UDNs in 5G [37].

[11] For example, new movies are posted (or change popularity) every week, and new music videos are posted about every month [12].



(a) Average throughput versus $C_{\rm bh}$, $\eta = 0.1$.



(b) Average throughput versus η, $C_{\rm bh} = 100$Mbps.

Fig. 2. Average throughput versus backhaul capacity and cache capacity.

the impact of caching on the throughput of the network, we also provide the numerical results obtained from (13) versus η in Fig. 2(b). We can see from Fig. 2(a) and Fig. 2(b) that the throughput increases with both the backhaul capacity and cache capacity, which agrees with the result in (22) derived in the special scenario. Moreover, the throughput increases with η more sharply when β is small. This suggests that the throughput can be boosted more efficiently by caching at the BSs if the ICI level can be reduced.

B. When EE Benefits from Caching?

In Table I, we use numerical results to show when the condition in (27) holds for different content catalog size N_f, backhaul hardware and cache hardware.

A typical pico BS in LTE system is considered, where the transmission and power consumption parameters have been defined in the beginning of this section. The interference level is set as $\beta = 1$. In such a worst case, the condition is more prone to be invalid. While there are various kinds of memory technologies, we consider the two kinds that are most likely employed due to their higher power efficiencies and larger cache sizes. Except for the high speed SSD cache hardware with $w_{\rm ca} = 6.25 \times 10^{-12}$ W/bit and microwave backhaul link with $w_{\rm bh} = 5 \times 10^{-7}$ J/bit, we also consider DRAM as cache hardware and optical fiber as backhaul link (with capacity 1 Gbps), whose power coefficients are respectively $w_{\rm ca} = 2.5 \times 10^{-9}$ W/bit [9] and $w_{\rm bh} = 4 \times 10^{-8}$ J/bit [9, 14]. Considering that N_f has a wide range in literatures, e.g., $N_f = 10^2 \sim 10^3$ with a large content size $F = 10^2 \sim 10^3$ MB [39, 40] and $N_f = 10^4 \sim 10^5$ with a small content size $F = 1 \sim 10$ MB [12, 41], we also investigate the impact of N_f and F on the condition.

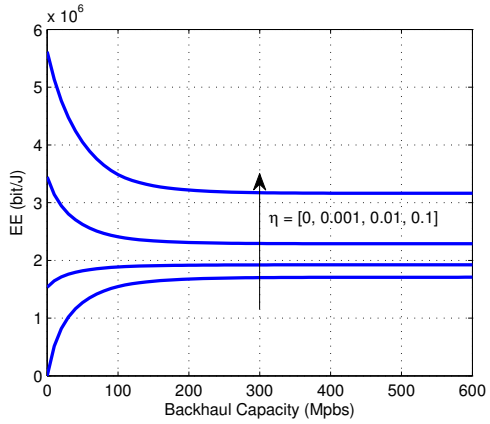
TABLE I
NUMERICAL EXAMPLE, $\delta = 1$

Condition	(27) LHS	(27) RHS	$w_{\rm ca}$	$w_{\rm bh}$	N_f	F
Hold	0.006	34.4	SSD	microwave	10^5	10 MB
Hold	0.006	2.31	SSD	optical fiber	10^5	10 MB
Hold	0.37	2.31	SSD	optical fiber	10^3	10^3 MB
Hold	2.41	34.4	DRAM	microwave	10^5	10 MB
Not hold	2.41	2.31	DRAM	optical fiber	10^5	10 MB
Not hold	149.7	34.4	DRAM	microwave	10^3	10^3 MB

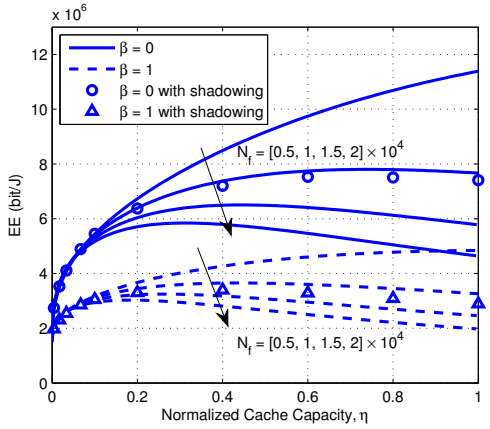
As expected, when the values of $w_{\rm ca}$ is large and $w_{\rm bh}$ is small, the EE does not benefit from caching at the BSs. Moveover, with the same value of $N_f F$, the condition is more prone to be invalid when the content size F is large.

C. Impact of Key Parameters on EE

In Fig. 3, we show the numerical results of EE obtained from (21) respectively versus backhaul capacity and normalized cache capacity. We can see from Fig. 3(a) that when no content or a little portion of the contents are cached at each BS (i.e., $\eta = 0$ and 0.001), EE increases with the backhaul capacity, and when the portion is large (i.e., $\eta = 0.01, 0.1$), EE decreases with $C_{\rm bh}$. This is because although the throughput increases with $C_{\rm bh}$, the backhaul power consumption also increases with more backhaul traffic. Moreover, the EE gain of caching over not caching is high when the backhaul capacity is very limited, and the gain approaches a constant when $C_{\rm bh}$ is large, say 200 Mbps. Fig. 3(b) shows that when the catalog size N_f is relatively small (i.e., $N_f < N_{th}$), say $N_f = 5000$, EE increases with η until all contents are cached, and the maximal EE gain of caching over not caching is about 575% when $\beta = 0$ and 250% when $\beta = 1$. When N_f is large (i.e., $N_f > N_{th}$), EE first increases and then decreases with η. In fact, we can compute the values of N_{th} from (36) for unlimited-capacity backhaul or numerically from (31) for limited-capacity backhaul. In the considered setting, the values of N_{th} range from 3000 to 20000 contents. Note that these results are obtained when each BS can schedule at most N_t users. Nonetheless, the results are consistent with the analysis in Section IV-A and Proposition 2, which are obtained in the special case where each BS only serves at most one user. By comparing Fig. 3(b) with Fig. 2(b), we can see that the EE gain from caching is much higher than the throughput gain from caching if ICI can be perfectly controlled (i.e., $\beta = 0$). This is because when backhaul capacity is limited, the throughput gain of caching only comes from reducing ICI, but the EE gain also comes from reducing the proportion of power consumed

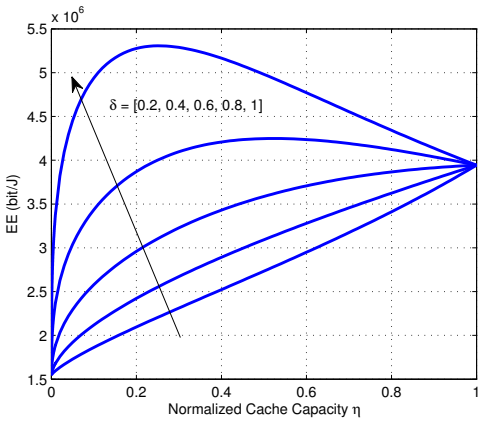


(a) EE versus backhaul capacity, $\beta = 0.5$.

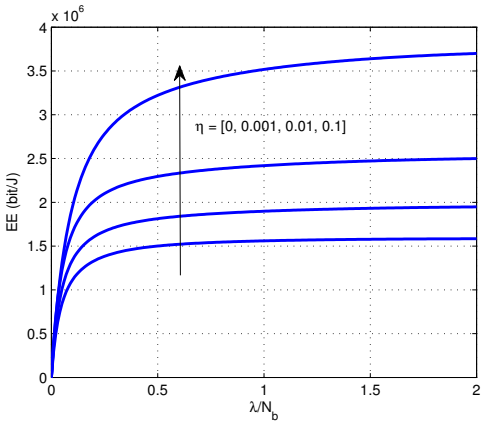


(b) EE versus the cache capacity, $C_{\rm bh} = 100$ Mbps.

Fig. 3. EE versus backhaul capacity and cache capacity.



(a) EE versus η under different skew parameter δ.



(b) EE versus $\frac{\lambda}{N_b}$ under different cache capacity η.

Fig. 4. EE versus cache capacity and user density, $\beta = 0.5$, $C_{\rm bh} = 100$ Mbps.

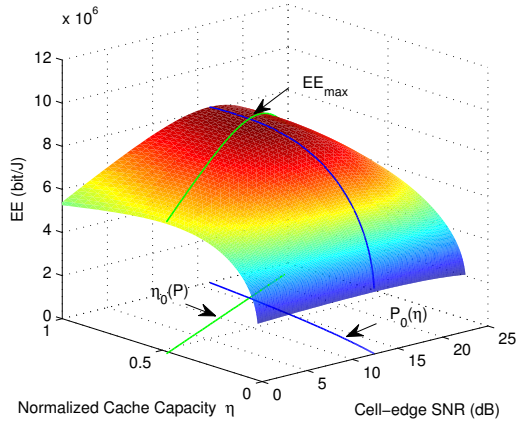
for backhauling. To show the impact of shadowing, we also provide the simulation result of EE in Fig. 3(b), where the shadowing is subject to log-normal distribution with 8 dB deviation. We can see that the network EE is slightly lower when shadowing is considered but the main trend of EE-cache relationship does not change.

In Fig. 4(a), we show the numerical results of EE obtained from (21) versus the normalized cache capacity with different skew parameter δ. We can see that the optimal cache capacity decreases with δ. With the same cache capacity, EE increases with δ. This is because the cache hit ratio increases with δ as shown in (9). When $\delta = 1$, the EE gain of caching with optimized η over not caching is about 350%. In Fig. 4(b), we show the numerical results of EE obtained from (21) versus the ratio of user density to BS density. We can see that EE first increases with $\frac{\lambda}{N_b}$ quickly and then saturates gradually because the throughput is finally limited by ICI. Moreover, the EE increases more sharply when cache is enabled. This is because the throughput is increased and the backhaul power consumption is reduced by caching. When $\frac{\lambda}{N_b}$ is around one, which is typical for SCNs, the EE gain is about 230%.

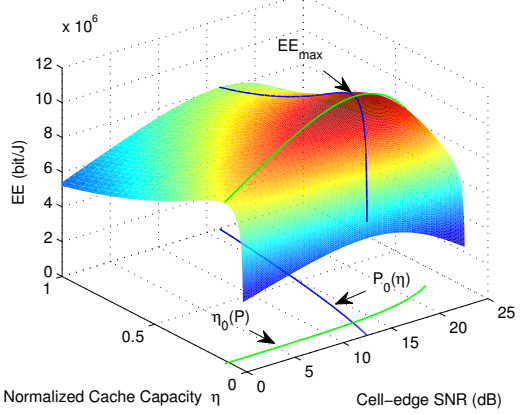
In Fig. 5, we show the numerical results of EE obtained from (26) versus the cell-edge SNR (which is controlled by changing the transmit power and hence reflects the impact of transmit power) and normalized cache capacity under unlimited-capacity backhaul and very stringent-capacity backhaul. As we analyzed in section IV-B, with a given cache capacity, the EE first increases with P and then decreases with P. We also plot the optimal transmit power P_0 as a function of η denoted as $P_0(\eta)$, as well as the optimal normalized cache capacity η_0 as a function of P denoted as $\eta_0(P)$. We can see that $P_0(\eta)$ increases with η slowly as we analyzed in Section IV-B, and $\eta_0(P)$ increases with P slowly with very stringent-capacity backhaul. This implies that in a cache-enabled network with stringent-capacity backhaul, the value of transmit power has minor impact on the EE-optimal cache capacity and the value of cache capacity has little impact on the optimal transmit power. Besides, it is easy to find that the joint optimal values of η and P maximizing the network EE is the crossing point of $\eta_0(P)$ and $P_0(\eta)$. This means that (P_0, η_0) satisfying both $\frac{dEE}{dP} = 0$ in (40) and $\frac{dEE}{d\eta} = 0$ in (35) are the joint optimal transmit power and cache capacity with the considered system setting, although the EE is not joint concave in P and η as we analyzed in Section IV-B.

D. Where to Cache Can Provide Higher EE?

To illustrate where to deploy the caches can provide higher EE, we compare the throughput and EE achieved by caching

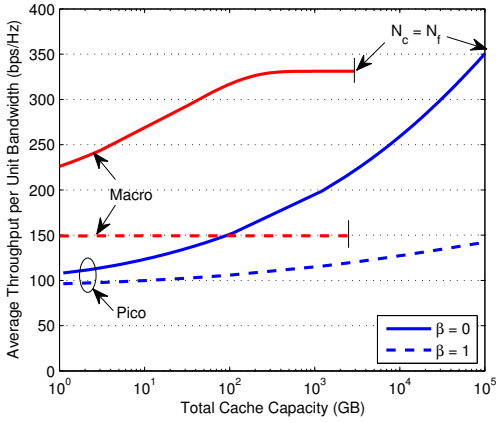


(a) $C_{\mathrm{bh}} \to \infty$

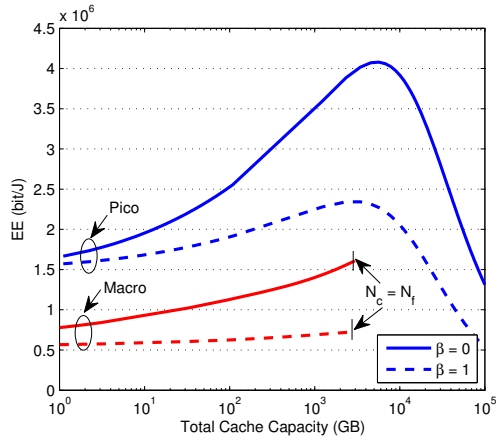


(b) $C_{\mathrm{bh}} \to 0$

Fig. 5. EE versus cell-edge SNR and normalized cache capacity, $\beta = 0$, $\delta = 1$.



(a) Throughput

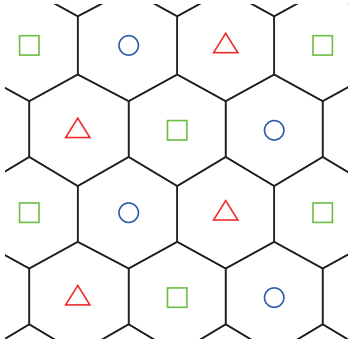


(b) EE

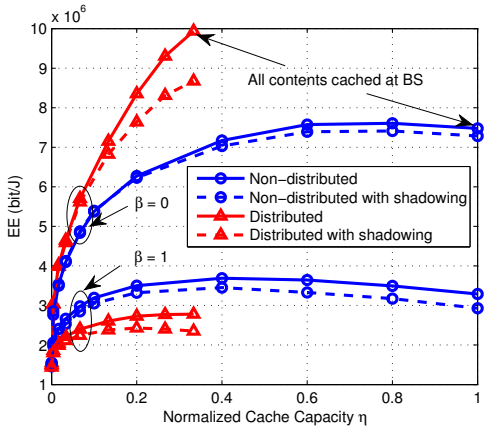
Fig. 6. Throughput and EE comparison between macro and pico networks, $N_f = 10^5$. The throughput is evaluated within a region of 250 m radius including one macro cell and 37 pico cells. The curves stop when $N_c = N_f$, i.e, all contents are cached at each BS. The curves of pico network stop earlier because each pico BS caches less contents than each macro BS because the two networks are set with identical total cache capacity.

at the macro and pico BSs. For a fair comparison, we deploy three tiers of macro BSs similar to the pico network. The radius of each macro cell is 250 m, i.e., the coverage area of each macro cell is the same as that of $N_b = 37$ pico cells. To ensure that the pico network and the macro network have the same total number of antennas and the same sum backhaul capacity within the same coverage area, each macro BSs is equipped with 4×37 antennas and the backhaul capacity for each pico BS and macro BS is 100 Mbps and 100×37 Mbps. The power consumption parameters of the macro BS are $\rho = 3.22$, $P = 46$ dBm, $P_{\mathrm{cc}_i} = 2.01 \times 10^3$ W (13.6 W per antenna), $P_{\mathrm{cc}_a} = 3.81 \times 10^3$ W (25.8 W per antenna) [27]. If each BS caches N_c contents, the total cache capacities of the macro and pico networks will be $N_c F$ and $N_b N_c F$, respectively. In this simulation, we set the two networks with the same total cache capacity, hence each pico BS caches less contents.

We can see from Fig. 6(a) that when the total cache capacity of the network is low, the throughput of the macro network is higher than the pico network due to higher backhaul capacity for each BS. When $\beta = 1$, the throughput of the macro network does not change with cache capacity, but the throughput of the pico network increases with cache capacity. This is because the backhaul capacity of each macro BS is large such that interference is the limiting factor of throughput, while the backhaul capacity of each pico BS network is low so that increasing cache capacity can relieve the backhaul congestion and hence increase the throughput. When there is no interference and $\beta = 0$, backhaul capacity becomes the bottleneck of both networks and thus their throughputs increase with cache capacity. We can see from Fig. 6(b) that the EE of the pico network is higher than the macro network since the pico BSs have more opportunities to idle and have low transmit and circuit powers although the cache capacity of each pico BS is smaller than each macro BS. The EE of the pico networks benefits more from caching, despite that more replicas of the same content are cached. This is because the backhaul capacity limits the throughput of each pico BS meanwhile the backhaul power consumption takes a large portion of the energy consumed in the pico network.



(a) An illustration of distributed caching.



(b) EE comparison, $C_{\rm bh} = 100$ Mbps.

Fig. 7. Impact of user association with distributed caching and shadowing.

E. Impact of User Association

In the system model, we have assumed that each user is associated with the closest BS, and hence caching most popular contents in each BS is optimal. Now we relax this assumption and consider a user association based on both location and content. As shown in (26), EE increases with the cache hit ratio p_h. To increase p_h, we consider a distributed caching strategy where every three adjecent BSs cache different contents and each user associates with the nearest BS that caches the user's requested contents. As illustrated in Fig. 7(a), the BS marked with "Δ" caches the 1st, 4th, 7th, $\cdots$, $(3N_c-2)$th popular contents, the BS marked with "$\square$" caches the 2nd, 5th, 8th, $\cdots$, $(3N_c-1)$th popular contents, and the BS marked with "$\bigcirc$" caches the 3rd, 6th, 9th, $\cdots$, $3N_c$th popular contents. This way of caching can reduce content redundancy by storing different contents in different BSs. Then, when each BS caches N_c contents with the distributed caching, each user can access to $3N_c$ cached contents, i.e, the equivalent cache capacity seen from each user can be regarded as three times over that with non-distributed caching.

In Fig. 7(b), we show the simulation results of EE with distributed caching and non-distributed caching. We can see that when $\beta = 0$, i.e., no interference, distributed caching can achieve higher EE due to higher cache hit ratio. When $\beta = 1$, i.e., in the worst case of interference, distributed caching achieves lower EE than non-distributed caching. This is because each user may not always associate with the nearest BS with distributed caching and hence the nearest BS may generate strong interference to the user, which results in the EE reduction. When shadowing is considered and each user is associated to the BS with highest average channel gain, the network EE is slightly lower but the main trend of EE-cache relationship doed not change for both non-distributed and distributed caching.

VI. Conclusion

In this paper, we investigated whether and how caching at BSs can improve EE of wireless access networks. By analyzing the EE for the cache-enabled network, we found the condition of whether EE can benefit from caching, the EE-memory relation, and the maximal EE gain from caching. Analytical results showed that EE can be improved by caching at the BSs when power efficient cache hardware is used. A key observation is that the EE gain of caching comes from boosting the throughput, reducing the backhaul consumption and exploiting the content popularity when the backhaul is limited. The EE gain is large when the interference level is low, the backhaul capacity is stringent, and the content popularity distribution is skewed. Another key observation is that EE-memory relation is not a simple tradeoff. When the content catalog size is not very large, there is a tradeoff between EE and cache capacity. Otherwise, optimizing cache capacity of each BS can maximize the EE of the network. The EE-optimal cache capacity depends on the system setting, and decreases when the network becomes denser. Numerical and simulation results validated the analysis and showed that caching at pico BS can provide higher EE gain than caching at macro BS. Finally, we provided simulation results to illustrate that distributed caching will achieve much higher EE gain than simply caching popular contents everywhere if inter-cell interference can be successfully eliminated, but will be inferior to the simple caching policy if the interference can not be coordinated.

Appendix A
Proof of Lemma 1

Considering that the SINRs for the users shown in (3) are identically distributed, $\bar{R}_{\rm ca}(K_b, K_c)$ can be derived as

$$\bar{R}_{\rm ca}(K_b,K_c) = K_c B \mathbb{E}\left\{\log_2\left(1+\frac{r_{kb}^{-\alpha}|\mathbf{h}_{kb}^H\mathbf{w}_{kb}|^2}{K_b(\beta I_k+\frac{\sigma^2}{P})}\right)\right\}$$

$$\overset{(a)}{\approx} K_c B\Bigg(\mathbb{E}\left\{\log_2|\mathbf{h}_{kb}^H\mathbf{w}_{kb}|^2\right\} - \log_2 K_b + \mathbb{E}\{\log_2 r_{kb}^{-\alpha}\} - \mathbb{E}\left\{\log_2\left(\beta I_k+\frac{\sigma^2}{P}\right)\right\}\Bigg)$$

$$\overset{(b)}{=} K_c B\Bigg(\frac{1}{\ln 2}\psi(N_t-K_b+1) - \log_2 K_b + \int_0^D \log_2\left(r_{kb}^{-\alpha}\right)\frac{2r_{kb}}{D^2}dr_{kb} - \mathbb{E}\left\{\log_2\left(\beta I_k+\frac{\sigma^2}{P}\right)\right\}\Bigg)$$

$$\overset{(c)}{\approx} K_c B\Bigg(\log_2\frac{N_t-K_b+1}{K_b} + \frac{\alpha}{2\ln 2} + \log_2 D^{-\alpha}$$

$$- \mathbb{E}\left\{\log_2\left(\beta I_k + \frac{\sigma^2}{P}\right)\right\}\Bigg) \tag{A.1}$$

where the approximation in step (a) is from omitting the term "1" inside the log function, which is accurate in high SINR region, step (b) comes from the facts that $|\mathbf{h}_{kb}^H \mathbf{w}_{kb}|^2$ follows Gamma distribution $\mathbb{G}(N_t - K_b + 1, 1)$ [42] and $\frac{2r_{kb}}{D^2}$ is the probability density function (PDF) of r_{kb} when the user is uniformly distributed in the circle cell, and step (c) is obtained by applying the asymptotic approximation of the Digamma function $\psi(n)$, i.e., $\psi(n) = \ln(n) + \mathcal{O}(\frac{1}{n}) \approx \ln n$ [43] and the approximation is accurate when $N_t - K_b + 1 > 1$.

When the network is interference-limited, i.e., the interference power $\beta P I_k \gg \sigma^2$,

$$\mathbb{E}\left\{\log_2\left(\beta I_k + \frac{\sigma^2}{P}\right)\right\} \approx \mathbb{E}\{\log_2(\beta I_k)\} \tag{A.2}$$

Considering the expression of I_k defined in (3) and $\mathbb{E}\{\zeta_j\} = 1 \cdot p_a + 0 \cdot (1 - p_a) = p_a$, we have

$$\mathbb{E}\{\log_2(\beta I_k)\} = \mathbb{E}\{\log_2(I_k D^\alpha)\} + \log_2(\beta D^{-\alpha}) \tag{A.3}$$

where $\mathbb{E}\{\log_2(I_k D^\alpha)\}$ can be derived as

$$\begin{aligned}
&\mathbb{E}\{\log_2(I_k D^\alpha)\} \\
&= \mathbb{E}_{r_{kj}, \mathbf{h}_{kj}, \zeta_j}\left\{\log_2\left(\sum_{j=1, j\neq b}^{N_b} \zeta_j \left(\frac{D}{r_{kj}}\right)^\alpha \|\mathbf{h}_{kj}\mathbf{W}_j\|^2\right)\right\} \\
&\overset{(a)}{\leq} \mathbb{E}_{r_{kj}, \mathbf{h}_{kj}}\left\{\log_2\left(\sum_{j=1, j\neq b}^{N_b} \mathbb{E}\{\zeta_j\} \left(\frac{D}{r_{kj}}\right)^\alpha \|\mathbf{h}_{kj}\mathbf{W}_j\|^2\right)\right\} \\
&= \mathbb{E}_{r_{kj}, \mathbf{h}_{kj}}\left\{\log_2\left(\sum_{j=1, j\neq b}^{N_b} \left(\frac{D}{r_{kj}}\right)^\alpha \|\mathbf{h}_{kj}\mathbf{W}_j\|^2\right)\right\} + \log_2 p_a \\
&\triangleq \Phi + \log_2 p_a = \log_2 p_a 2^\Phi
\end{aligned} \tag{A.4}$$

where the upper bound in step (a) is from using the Jensen's inequality and the bound is tight when $\frac{\lambda}{N_b}$ is high (then $p_a \to 1$ and hence $\zeta_j \to \mathbb{E}\{\zeta_j\}$), and Φ is a constant only depending on the path-loss exponent α when $N_b \to \infty$ (to be proved in Appendix B). By substituting (A.4) into (A.3) and then into (A.2), we obtain

$$\begin{aligned}
\mathbb{E}\left\{\log_2\left(\beta I_k + \frac{\sigma^2}{P}\right)\right\} &\leq \log_2(p_a \beta 2^\Phi D^{-\alpha}) \\
&\approx \log_2\left(p_a \beta 2^\Phi D^{-\alpha} + \frac{\sigma^2}{P}\right)
\end{aligned} \tag{A.5}$$

where the approximation comes from the fact that when $\beta P I_k \gg \sigma^2$, we have $\log_2(p_a \beta 2^\Phi D^{-\alpha}) \geq \mathbb{E}\{\log_2(\beta I_k)\} \gg \log_2(\frac{\sigma^2}{P})$ which means $p_a \beta 2^\Phi D^{-\alpha} \gg \frac{\sigma^2}{P}$.

When the network is noise-limited, i.e., $\beta P I_k \ll \sigma^2$, we also have $\mathbb{E}\{\log_2(\beta I_k + \frac{\sigma^2}{P})\} \approx \log_2 \frac{\sigma^2}{P} \approx \log_2(p_a \beta 2^\Phi D^{-\alpha} + \frac{\sigma^2}{P})$, which is the same as the result in (A.5).[12]

By substituting (A.5) into (A.1), $\bar{R}_{\mathrm{ca}}(K_b, K_c)$ can be approximated as

$$\bar{R}_{\mathrm{ca}}(K_b, K_c) \approx K_c B\left(\frac{\alpha}{2\ln 2} + \log_2 \frac{(N_t - K_b + 1)P}{K_b(p_a \beta P 2^\Phi + D^\alpha \sigma^2)}\right)$$

$$\triangleq K_c\left(\frac{\alpha B}{2\ln 2} + \bar{R}_{\mathrm{e}}(K_b)\right) \tag{A.6}$$

where $\bar{R}_{\mathrm{e}}(K_b) \triangleq B\log_2 \frac{(N_t - K_b + 1)P}{K_b(p_a \beta P 2^\Phi + D^\alpha \sigma^2)}$ can also be derived from $\mathbb{E}\left\{B\log_2 \frac{P D^{-\alpha} |\mathbf{h}_{kb}^H \mathbf{w}_{kb}|^2}{K_b(\beta P I_k + \sigma^2)}\right\}$. Hence, $\bar{R}_{\mathrm{e}}(K_b)$ can be regarded as the average achievable rate of a cell-edge user when the backhaul capacity is unlimited and BS_b serves K_b users.

Appendix B
Proof of the constant Φ when $N_b \to \infty$

In the following, we first prove Φ only depends on α and N_b, and then prove Φ converges when $N_b \to \infty$. Without loss of generality, we assume the coordinate of BS_b as $(0, 0)$. Denoting (x_k, y_k) and (u_j, v_j) as the coordinate of MS_k and BS_j, respectively, then $r_{kb} = \sqrt{x_k^2 + y_k^2}$ and $r_{kj} = \sqrt{(x_k - u_j)^2 + (y_k - v_j)^2}$. Denoting $I_{kj} \triangleq \|\mathbf{h}_{kj}\mathbf{W}_j\|^2$ and taking the expectation over user location in (A.4), we obtain

$$\begin{aligned}
\Phi = \frac{1}{\pi D^2} \iint_{x_k^2 + y_k^2 \leq D^2} \mathbb{E}_{I_{kj}}\Bigg\{\log_2\Bigg(\sum_{j=1, j\neq b}^{N_b} \\
\left(\frac{D}{\sqrt{(x_k - u_j)^2 + (y_k - v_j)^2}}\right)^\alpha I_{kj}\Bigg)\Bigg\} dx_k dy_k
\end{aligned} \tag{B.1}$$

We normalize the coordinates of MS_k and BS_j with the cell radius D as $(\bar{x}_k, \bar{y}_k) = \left(\frac{x_k}{D}, \frac{y_k}{D}\right)$ and $(\bar{u}_j, \bar{v}_j) = \left(\frac{u_j}{D}, \frac{v_j}{D}\right)$, respectively. After changing the integration variables as $\bar{x}_k$ and $\bar{y}_k$, (B.1) can be rewritten as

$$\begin{aligned}
\Phi = \frac{1}{\pi} \iint_{\bar{x}_k^2 + \bar{y}_k^2 \leq 1} \mathbb{E}_{I_{kj}}\Bigg\{\log_2\Bigg(\sum_{j=1, j\neq b}^{N_b} \\
\left((\bar{x}_k - \bar{u}_j)^2 + (\bar{y}_k - \bar{v}_j)^2\right)^{-\frac{\alpha}{2}} I_{kj}\Bigg)\Bigg\} d\bar{x}_k d\bar{y}_k
\end{aligned} \tag{B.2}$$

Since the normalized coordinates $(\bar{x}_k, \bar{y}_k)$ and $(\bar{u}_j, \bar{v}_j)$ do not depend on D, and I_{kj} is averaged over small-scale fading channel in (B.2), Φ only depend on α and N_b.

By using the Jensen's inequality in (B.2) to move the expectation into the log function and considering $\mathbb{E}\{I_{kj}\} = 1$, we obtain

$$\begin{aligned}
\Phi \leq \frac{1}{\pi} \iint_{\bar{x}_k^2 + \bar{y}_k^2 \leq 1} \log_2\Bigg(\sum_{j=1, j\neq b}^{N_b} \\
\left((\bar{x}_k - \bar{u}_j)^2 + (\bar{y}_k - \bar{v}_j)^2\right)^{-\frac{\alpha}{2}}\Bigg) d\bar{x}_k d\bar{y}_k
\end{aligned} \tag{B.3}$$

Considering $\alpha > 2$ in practice and after some manipulations, we can show that $\sum_{j=1, j\neq b}^{N_b} \left((\bar{x}_k - \bar{u}_j)^2 + (\bar{y}_k - \bar{v}_j)^2\right)^{-\frac{\alpha}{2}}$ converges when $N_b \to \infty$. Therefore, Φ has an upper bound. Further considering Φ increases with N_b, Φ converges when $N_b \to \infty$.

[12] In section V-A, we use simulations to show that (A.5) is accurate for all values of $\beta \in [0, 1]$.

APPENDIX C
PROOF OF LEMMA 2

Consider that when $N_t \to \infty$, $\mathbb{E}_{\mathbf{h}_{kb}}\{\frac{|\mathbf{h}_{kb}\mathbf{w}_{kb}|^2}{N_t}\} \to 1$ and the variance of $\frac{|\mathbf{h}_{kb}\mathbf{w}_{kb}|^2}{N_t}$ approaches to zero resulting from channel hardening [44]. Besides, when the interference power from each BS is independent and identically distributed (i.i.d.),[13] the interference power per BS $\frac{\beta P I_k}{N_b} = \frac{\beta P}{N_b}\sum_{j=1,j\neq b}^{N_b}\zeta_j r_{kj}^{-\alpha}\|\mathbf{h}_{kj}\mathbf{W}_j\|^2$ approaches to its expectation $\frac{\beta P}{N_b}\mathbb{E}\{I_k\}$ when $N_b \to \infty$ according to the law of large numbers. This suggests that the distance between each user and its local BS r_{kb} dominates the comparison between $\sum_{k=K_c+1}^{K_b} B\log_2(1+\gamma_k)$ and $C_{\rm bh}$ when N_b is large, and therefore the second term in (10) can be approximated as

$$\bar{R}_{\rm bh}(K_b,K_c,C_{\rm bh}) = \mathbb{E}\left\{\min\left(B\sum_{k=K_c+1}^{K_b}\log_2(1+\gamma_k), C_{\rm bh}\right)\right\} \approx \mathbb{E}_{r_{kb}}\left\{\min\left(B\sum_{k=K_c+1}^{K_b}\mathbb{E}_{\mathbf{h},r_{kj},\zeta_j}\{\log_2(1+\gamma_k)\}, C_{\rm bh}\right)\right\} \tag{C.1}$$

which is accurate as shown via simulations in Section V-A.

By omitting the term "1" inside the log function, approximating $\psi(n)$ by $\ln(n)$ similar to the derivation for (A.1), and further considering (A.5) and the definition of $\bar{R}_{\rm e}(K_b)$, we have

$$\mathbb{E}_{\mathbf{h},r_{kj},\zeta_j}\{\log_2(1+\gamma_k)\} \approx \log_2\frac{(N_t-K_b+1)P}{K_b(p_a\beta P 2^{\Phi}D^{-\alpha}+\sigma^2)} + \log_2 r_{kb}^{-\alpha} = \frac{\bar{R}_{\rm e}(K_b)}{B} + \alpha\log_2\frac{D}{r_{kb}} \tag{C.2}$$

By substituting (C.2) into (C.1), we obtain

$$\bar{R}_{\rm bh}(K_b,K_c,C_{\rm bh}) \approx \mathbb{E}_{r_{kb}}\left\{\min\left((K_b-K_c)\bar{R}_{\rm e}(K_b) + \frac{\alpha B}{2\ln 2}\sum_{k=K_c+1}^{K_b}2\ln\frac{D}{r_{kb}}, C_{\rm bh}\right)\right\} \triangleq \mathbb{E}_{r_{kb}}\{\tilde{R}_{\rm bh}\} \tag{C.3}$$

where we define $\tilde{R}_{\rm bh}$ to denote the term inside $\mathbb{E}_{r_{kb}}\{\cdot\}$ in (C.3) for notation simplicity.

With the PDF of r_{kb}, i.e., $\frac{2r_{kb}}{D^2}$, we can prove that $\{2\ln\frac{D}{r_{kb}}, k=1,\cdots,K_b, b=1,\cdots,N_b\}$ are independent exponential distributed RVs with unit mean. Hence, the term $y \triangleq \sum_{k=K_c+1}^{K_b}2\ln\frac{D}{r_{kb}}$ in (C.3) is a Gamma distributed RV following $\mathbb{G}(K_b-K_c,1)$, i.e., it is positive, and the PDF of this term is $f(y) = \frac{y^{K_b-K_c-1}e^{-y}}{(K_b-K_c-1)!}$, $y>0$. This gives rise to the following results.

When $C_{\rm bh} \le (K_b-K_c)\bar{R}_{\rm e}(K_b)$, i.e., the backhaul capacity is less than the average achievable sum-rate of all the cache-miss users under unlimited-capacity backhaul when they are located at the cell edge, the right hand side (RHS) of (C.3) becomes

$$\mathbb{E}_{r_{kb}}\{\tilde{R}_{\rm bh}\} = C_{\rm bh} \tag{C.4}$$

[13]When the spatial distribution of the BSs also follows PPP, the interference power from each BS is indeed i.i.d. [45].

When $C_{\rm bh} > (K_b-K_c)\bar{R}_{\rm e}(K_b)$, considering

$$\tilde{R}_{\rm bh} = \begin{cases}(K_b-K_c)\bar{R}_{\rm e}(K_b)+\frac{\alpha B}{2\ln 2}y, & \text{if } y<z\\ C_{\rm bh}, & \text{if } y\ge z\end{cases} \tag{C.5}$$

where $z \triangleq \frac{2\ln 2}{\alpha B}\big(C_{\rm bh}-(K_b-K_c)\bar{R}_{\rm e}(K_b)\big)$, the RHS of (C.3) can be derived as

$$\begin{aligned}&\mathbb{E}_{r_{kb}}\{\tilde{R}_{\rm bh}\}\\ &= \int_0^\infty \min\left((K_b-K_c)\bar{R}_{\rm e}(K_b)+\frac{\alpha B}{2\ln 2}y, C_{\rm bh}\right)f(y)dy\\ &= \int_0^z\left((K_b-K_c)\bar{R}_{\rm e}(K_b)+\frac{\alpha B}{2\ln 2}y\right)f(y)dy\\ &\quad+\int_z^\infty C_{\rm bh}f(y)dy\\ &= (K_b-K_c)\bigg(\frac{\alpha B}{2\ln 2}\gamma(K_b-K_c+1,z)\\ &\quad+\bar{R}_{\rm e}(K_b)\gamma(K_b-K_c,z)\bigg)+C_{\rm bh}\Gamma(K_b-K_c,z)\end{aligned} \tag{C.6}$$

Combine (C.4) and (C.6), Lemma 2 is proved.

APPENDIX D
PROOF OF PROPOSITION 1

With $N_c = 0$ and $p_h = 0$, from (26) the EE without caching can be obtained as $EE_{\rm no} = \frac{p_a\bar{R}_{\rm bh}}{p_aP_a+(1-p_a)P_i+p_aw_{\rm bh}\bar{R}_{\rm bh}}$. If $EE_{\rm no}$ exceeds the EE with caching in (26), then with (9) we have

$$w_{\rm ca}N_cF\sum_{j=1}^{N_f}j^{-1} > \big((p_aP_a+(1-p_a)P_i)\bar{R}_{\rm ca}+p_aw_{\rm bh}\bar{R}_{\rm ca}\bar{R}_{\rm bh}\big)\sum_{f=1}^{N_c}f^{-1} \tag{D.1}$$

If (D.1) holds for $N_c = 1$, then

$$w_{\rm ca}F\sum_{j=1}^{N_f}j^{-1} > \big((p_aP_a+(1-p_a)P_i)\bar{R}_{\rm ca}+p_aw_{\rm bh}\bar{R}_{\rm ca}\bar{R}_{\rm bh}\big) \tag{D.2}$$

Multiplying both side of (D.2) by N_c, we obtain

$$w_{\rm ca}N_cF\sum_{j=1}^{N_f}j^{-1} > \big((p_aP_a+(N_b-p_a)P_i)\bar{R}_{\rm ca}+p_aw_{\rm bh}\bar{R}_{\rm ca}\bar{R}_{\rm bh}\big)N_c \tag{D.3}$$

Furthering considering that $N_c > \sum_{f=1}^{N_c}f^{-1}$ for $N_c > 1$, (D.3) turns into

$$w_{\rm ca}N_cF\sum_{j=1}^{N_f}j^{-1} > \big((p_aP_a+(1-p_a)P_i)\bar{R}_{\rm ca}+p_aw_{\rm bh}\bar{R}_{\rm ca}\bar{R}_{\rm bh}\big)\sum_{f=1}^{N_c}f^{-1} \tag{D.4}$$

which is the same as (D.1). This suggests that if caching one content can not improve EE, then for any $N_c > 1$ caching can not improve EE. Therefore, (D.2) is the condition of whether caching can increase EE. (D.2) can be rewritten as (27), and Proposition 1 is proved.

APPENDIX E
PROOF OF PROPOSITION 2

From $\frac{dEE}{d\eta}\big|_{\eta=\eta_0} = 0$, we can obtain $\frac{\Omega}{\eta_0 N_f} + \ln\frac{1}{\eta_0 N_f} = \frac{\bar{R}_{\rm bh}}{\bar{R}_{\rm ca}-\bar{R}_{\rm bh}}\ln N_f - 1$. Adding $\ln\Omega$ on both sides of the equation, we obtain

$$\frac{\Omega}{\eta_0 N_f} + \ln\frac{\Omega}{\eta_0 N_f} = \ln\Omega + \frac{\bar{R}_{\rm bh}}{\bar{R}_{\rm ca}-\bar{R}_{\rm bh}}\ln N_f - 1 \quad \text{(E.1)}$$

Taking the exponential of both sides of (E.1), we have $\frac{\Omega}{\eta_0 N_f} e^{\frac{\Omega}{\eta_0 N_f}} = \Omega e^{\frac{\bar{R}_{\rm bh}}{\bar{R}_{\rm ca}-\bar{R}_{\rm bh}}\ln N_f - 1}$. Since $W(x)$ satisfies $W(x)e^{W(x)} = x$, we obtain

$$\frac{\Omega}{\eta_0 N_f} = W\left(\Omega e^{\frac{\bar{R}_{\rm bh}}{\bar{R}_{\rm ca}-\bar{R}_{\rm bh}}\ln N_f - 1}\right) \quad \text{(E.2)}$$

Since $\frac{\Omega}{\eta N_f} + \ln\frac{\Omega}{\eta N_f}$ decreases with η, $\frac{dEE}{d\eta} > 0$ when $\eta < \eta_0$ and $\frac{dEE}{d\eta} < 0$ when $\eta > \eta_0$. Rewriting (E.2) as (31) and further considering $\eta \le 1$, Proposition 2 can be proved.

APPENDIX F
PROOF OF COROLLARY 4

Denote $N_b\pi D^2 \triangleq c$, where c is a constant. Substituting $D = (\frac{c}{\pi N_b})^{\frac{1}{2}}$ and $p_a = 1 - e^{-\frac{\lambda}{N_b}}$ into (35) and then taking the derivation of η_0 in (35) with respect to N_b, we obtain

$$\frac{d\eta_0}{dN_b} = \frac{-w_{\rm bh}B}{2N_b w_{\rm ca} F N_f \ln N_f \ln 2}\left(\frac{\frac{2\lambda}{N_b}e^{-\frac{\lambda}{N_b}} + \alpha\left(1-e^{-\frac{\lambda}{N_b}}\right)}{1+\beta 2^{\Phi} D^{\alpha}\left(1-e^{-\frac{\lambda}{N_b}}\right)}\right.$$
$$\left. + \frac{\lambda}{N_b}e^{-\frac{\lambda}{N_b}}\left(\alpha - 2 + \log_2\frac{N_t}{p_a\beta 2^{\Phi} + \left(\frac{P}{D^{\alpha}\sigma^2}\right)^{-1}}\right)\right) \quad \text{(F.1)}$$

Since the path-loss exponent $\alpha > 2$, we have $\frac{d\eta_0}{dN_b} < 0$, i.e., η_0 decreases with N_b.

When $\frac{\lambda}{N_b} \to 0$, we have $p_a = 1 - e^{-\frac{\lambda}{N_b}} \to \frac{\lambda}{N_b}$. Then from (35), $\eta_0 N_b$ can be expressed as

$$\eta_0 N_b \to \frac{\lambda w_{\rm bh} B}{w_{\rm ca} F N_f \ln N_f}\log_2\frac{N_t}{\frac{\lambda}{N_b}\beta 2^{\Phi} + \left(\frac{P}{D^{\alpha}\sigma^2}\right)^{-1}} \quad \text{(F.2)}$$

from which we can see that $\eta_0 N_b$ increases with N_b.

APPENDIX G
PROOF OF COROLLARY 7

By substituting $p_a\beta P 2^{\Phi} \ll D^{\alpha}\sigma^2$ into (39) and letting $\frac{dEE}{dP}\big|_{P=P_0} = 0$, we obtain

$$\bar{P}_{\rm cc} + \bar{P}_{\rm ca} - p_a\rho P_0\left(\ln\frac{N_t P_0}{D^{\alpha}\sigma^2} + \frac{\alpha}{2} - 1\right) = 0 \quad \text{(G.1)}$$

where $\bar{P}_{\rm cc} = p_a P_{{\rm cc}_a} + (1-p_a)P_{{\rm cc}_i}$ is the average circuit power consumption of each BS, and $\bar{P}_{\rm ca} = w_{\rm ca}\eta N_f F$ is the average cache power consumption of each BS. From this equation we can derive (40). Since in practice the path-loss exponent $\alpha > 2$, $\ln\frac{N_t P_0}{D^{\alpha}\sigma^2} + \frac{\alpha}{2} - 1 > 0$ and the left hand side (LHS) of (G.1) decreases with P_0. Therefore, $\frac{dEE}{dP} > 0$ when $P < P_0$ and $\frac{dEE}{dP} < 0$ when $P > P_0$, which indicates that P_0 is the optimal transmit power maximizing the network EE.

REFERENCES

[1] D. Liu and C. Yang, "Will caching at base station improve energy efficiency of downlink transmission?" in *Proc. IEEE GlobalSIP*, 2014.

[2] C.-L. I, C. Rowell, S. Han, Z. Xu, G. Li, and Z. Pan, "Toward green and soft: a 5G perspective," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 66–73, Feb. 2014.

[3] S. Yunas, M. Valkama, and J. Niemela, "Spectral and energy efficiency of ultra-dense networks under different deployment strategies," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 90–100, Jan. 2015.

[4] R. Q. Hu and Y. Qian, "An energy efficient and spectrum efficient wireless heterogeneous network framework for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 94–101, May 2014.

[5] S. Woo, E. Jeong, S. Park, J. Lee, S. Ihm, and K. Park, "Comparison of caching strategies in modern cellular backhaul networks," in *Proc. ACM MobiSys*, 2013.

[6] B. A. Ramanan, L. M. Drabeck, M. Haner, N. Nithi, T. E. Klein, and C. Sawkar, "Cacheability analysis of HTTP traffic in an operational LTE network," in *Proc. IEEE WTS*, 2013.

[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 Commun. Mag.*, vol. 51, no. 4, pp. 142–149, Apr. 2013.

[8] M. Chen and A. Ksentini, "Cache in the air: exploiting content caching and delivery techniques for 5G systems," *IEEE Commun. Mag.*, p. 132, Feb. 2014.

[9] N. Choi, K. Guan, D. C. Kilper, and G. Atkinson, "In-network caching effect on optimal energy consumption in content-centric networking," in *Proc. IEEE ICC*, 2012.

[10] J. Llorca, A. M. Tulino, K. Guan, J. Esteban, M. Varvello, N. Choi, and D. C. Kilper, "Dynamic in-network caching for energy efficient content delivery," in *Proc. IEEE INFOCOM*, 2013.

[11] J. Li, B. Liu, and H. Wu, "Energy-efficient in-network caching for content-centric networking," *IEEE Commun. Lett.*, vol. 17, no. 4, pp. 797–800, Apr. 2013.

[12] 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*, 2012.

[13] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82–89, Aug. 2014.

[14] Y. Xu, Y. Li, Z. Wang, T. Lin, G. Zhang, and S. Ci, "Coordinated caching model for minimizing energy consumption in radio access network," in *Proc. IEEE ICC*, 2014.

[15] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas, "Multicast-aware caching for small cell networks," in *Proc. IEEE WCNC*, 2014.

[16] P. Xi, S. Juei-Chin, Z. Jun, and B. L. Khaled, "Joint data assignment and beamforming for backhaul limited caching networks," in *Proc. IEEE PIMRC*, 2014.

[17] M. Dehghan, A. Seetharam, B. Jiang, T. He, T. Salonidis, J. Kurose, D. Towsley, and R. K. Sitaraman, "On the complexity of optimal routing and content caching in heterogeneous networks," in *Proc. IEEE INFOCOM*, 2015.

[18] J. Hachem, N. Karamchandani, and S. N. Diggavi, "Coded caching for heterogeneous wireless networks with multi-level access," in *Proc. IEEE INFOCOM*, 2014.

[19] T. Levanen, J. Pirskanen, T. Koskela, J. Talvitie, M. Valkama *et al.*, "Radio interface evolution towards 5G and enhanced local area communications," *IEEE Access*, vol. 2, pp. 1005–1029, 2014.

[20] Y. Wang, Z. Li, G. Tyson, S. Uhlig, and G. Xie, "Optimal cache allocation for content-centric networking," in *Proc. IEEE ICNP*. IEEE, 2013.

[21] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *Proc. IEEE INFOCOM*, 1999.

[22] M. Cha, P. Rodriguez, J. Crowcroft, S. Moon, and X. Amatriain, "Watching television over an IP network," in *Proc. ACM SIGCOMM IMC*, 2008.

[23] E. Mugume and D. So, "Spectral and energy efficiency analysis of dense small cell networks," in *Proc. IEEE VTC Spring*, 2015.

[24] C. Li, J. Zhang, and K. Letaief, "Throughput and energy efficiency analysis of small cell networks with multi-antenna base stations," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2505–2517, May 2014.

[25] H.-S. Jo, Y. J. Sang, P. Xia, and J. G. Andrews, "Heterogeneous cellular networks with flexible cell association: A comprehensive downlink SINR analysis," *IEEE Trans. Wireless Commun.*, vol. 11, no. 10, pp. 3484–3495, Oct. 2012.

[26] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE J. on Select. Areas Commun.*, vol. 24, no. 3, pp. 528–541, Mar. 2006.

[27] G. Auer, O. Blume, V. Giannini, I. Godor, M. Imran, Y. Jading, E. Katranaras, M. Olsson, D. Sabella, P. Skillermark *et al.*, "D2. 3: Energy efficiency analysis of the reference systems, areas of improvements and target breakdown," *EARTH*, 2010.

[28] Z. Chong and E. Jorswieck, "Energy-efficient power control for MIMO time-varying channels," in *Proc. IEEE GreenCom*, 2011.

[29] G. Y. Li, Z. Xu, C. Xiong, C. Yang, S. Zhang, Y. Chen, and S. Xu, "Energy-efficient wireless communications: Tutorial, survey, and open issues," *IEEE Wireless Commun. Mag.*, vol. 18, no. 6, pp. 28–35, Dec. 2011.

[30] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. A. Imran, D. Sabella, M. J. Gonzalez, O. Blume *et al.*, "How much energy is needed to run a wireless network?" *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 40–49, Oct. 2011.

[31] A. J. Fehske, P. Marsch, and G. P. Fettweis, "Bit per joule efficiency of cooperating base stations in cellular networks," in *Proc. IEEE GLOBECOM Workshops*, 2010.

[32] A. K. Gupta, H. S. Dhillon, S. Vishwanath, and J. G. Andrews, "Downlink multi-antenna heterogeneous cellular network with load balancing," *IEEE Trans. Commun.*, vol. 62, no. 11, pp. 4052–4067, Nov. 2014.

[33] R. M. Corless, G. H. Gonnet, D. E. Hare, D. J. Jeffrey, and D. E. Knuth, "On the Lambert W function," *Advances in Computational mathematics*, vol. 5, no. 1, pp. 329–359, 1996.

[34] S. Tombaz, P. Monti, K. Wang, A. Vastberg, M. Forzati, and J. Zander, "Impact of backhauling power consumption on the deployment of heterogeneous mobile networks," in *Proc. IEEE GLOBECOM*, 2011.

[35] TR 36.814 V1.2.0, "Further advancements for E-UTRA physical layer aspects (release 9)," *3GPP*, Jun. 2009.

[36] J. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. Reed, "Femtocells: Past, present, and future," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 497–508, April 2012.

[37] J. Park, S.-L. Kim, and J. Zander, "Asymptotic behavior of ultra-dense cellular networks and its economic impact," in *Proc. IEEE GLOBECOM*, 2014.

[38] M. Hefeeda and O. Saleh, "Traffic modeling and proportional partial caching for peer-to-peer systems," *IEEE/ACM Trans. Netw.*, vol. 16, no. 6, pp. 1447–1460, Dec. 2008.

[39] E. Bastug, J.-L. Guénégo, and M. Debbah, "Proactive small cell networks," in *Proc. IEEE ICT*, 2013.

[40] C. Yang, Z. Chen, Y. Yao, and B. Xia, "Performance analysis of wireless heterogeneous networks with pushing and caching," in *Proc. IEEE ICC*, 2015.

[41] F. Pantisano, M. Bennis, W. Saad, and M. Debbah, "Match to cache: Joint user association and backhaul allocation in cache-aware small cell networks," in *Proc. IEEE ICC*, 2015.

[42] J. Zhang, M. Kountouris, J. G. Andrews, and R. W. Heath, "Multi-mode transmission for the MIMO broadcast channel with imperfect channel state information," *IEEE Trans. Commun.*, vol. 59, no. 3, pp. 803–814, Mar. 2011.

[43] R. W. Heath, M. Kountouris, and T. Bai, "Modeling heterogeneous network interference using poisson point processes," *IEEE Trans. on Signal Process.*, vol. 61, no. 16, pp. 4114–4126, Aug. 2013.

[44] Q. Zhang, C. Yang, and A. F. Molisch, "Downlink base station cooperative transmission under limited-capacity backhaul," *IEEE Trans. Wireless Commun.*, vol. 12, no. 8, pp. 3746–3759, Sept. 2013.

[45] J. G. Andrews, F. Baccelli, and R. K. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE Trans. Commun.*, vol. 59, no. 11, pp. 3122–3134, Nov. 2011.

Dong Liu (S'13) received the B.S. degree in electronics engineering from Beihang University (formerly Beijing University of Aeronautics and Astronautics), Beijing, China in 2013. He is currently pursuing Ph.D degree in signal and information processing with the School of Electronics and Information Engineering, Beihang University.

His research interests lie in the area of caching in wireless network and cooperative transmission.

Chenyang Yang (SM'08) received the Ph.D. degree in electrical engineering from Beihang University (formerly Beijing University of Aeronautics and Astronautics), Beijing, China, in 1997.

Since 1999, she has been a Full Professor with the School of Electronic and Information Engineering, Beihang University. She has published more than 200 international journal and conference papers and filed more than 60 patents in the fields of energy-efficient transmission, coordinated multi-point, interference management, cognitive radio, relay, etc. Her recent research interests include green radio, local caching, and other emerging techniques for next generation wireless networks.

Prof. Yang was the Chair of the Beijing chapter of the IEEE Communications Society during 2008-2012 and the Membership Development Committee Chair of the Asia Pacific Board, IEEE Communications Society, during 2011-2013. She has served as a Technical Program Committee Member for numerous IEEE conferences and was the Publication Chair of the IEEE International Conference on Communications in China 2012 and a Special Session Chair of the IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP) in 2013. She served as an Associate Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS during 2009-2014 and a Guest Editor for the IEEE JOURNAL ON SELECTED TOPICS IN SIGNAL PROCESSING published in February 2015. She is currently an Associate Editor-in-Chief of the *Chinese Journal of Communications and the Chinese Journal of Signal Processing*. She was nominated as an Outstanding Young Professor of Beijing in 1995 and was supported by the First Teaching and Research Award Program for Outstanding Young Teachers of Higher Education Institutions by the Ministry of Education during 1999-2004.