Will Caching at Base Station Improve Energy Efficiency of Downlink Transmission?

Dong Liu and Chenyang Yang
Beihang University, Beijing, China
Email: dliu@buaa.edu.cn, cyyang@buaa.edu.cn

***Abstract*—Caching popular contents at the base station (BS) can reduce the end-to-end delay of service and the cost of cellular networks. Yet it is unknown whether introducing cache to BS can improve the energy efficiency (EE) of downlink systems. With BS caching, energy consumption for backhauling can be reduced, but caches will also consume energy. In this paper, we analyze the EE of two kinds of networks with a macro BS or multiple non-coordinated pico BSs with caches, where the impact of content popularity and request arrival density are considered. Analytical and simulation results show that introducing cache can improve the EE if the file catalog size is not too large and placing the caches at the pico BSs is more energy efficient.**

***Index Terms*—Energy efficiency, Cache, Downlink system**

I. Introduction

To meet the ever-increasing traffic demands in 5th generation (5G) cellular networks with low cost and high energy efficiency (EE), we need to rethink the goal of wireless networks. Recently, it was observed that only a small portion of contents are frequently requested by majority of users although there are massive contents in the network [1]. This suggests that the network should be designed toward content dissemination rather than only for transmitter-receiver communication. Considering such a traffic characteristic and the tremendous amount of storage capacity for today's caches, introducing caching to base station (BS) offers a promising way to unleash the ultimate potential of cellular networks with reduced cost [2].

Local caching of popular files at BSs has been proposed recently in [3–5] to reduce the backhaul cost and access latency as well as to improve the throughput. [2] is the first paper to introduce cache to small cell BS to reduce the backhaul cost, where caching strategies were proposed to serve more users under the constraints of the file downloading time. In [4], a transmission strategy was proposed for cache-enabled wireless network to improve throughput by exploiting multi-user diversity gain. In [5], caching approach was proposed to reduce the traffic loads via multicast transmission. Yet it is unclear until now whether a cache enabled BS can improve the energy efficiency (EE) of a downlink system. This is because if most requested files are stored at the BS cache, the traffic load in backhaul will be reduced and the resulting energy consumption can be saved, but the energy consumed for storage will grow. In fact, the power consumed for backhauling is not negligible for the EE of downlink transmission, especially for small cell networks [6]. On the other hand, when a macro BS serving more users is equipped with cache, the cache will be hit more frequently. By contrast, a small cell BS, say a pico BS, can be turned into sleep mode when there are no users, and is closer to the users despite that the hit rate is lower. Therefore, in order to improve the EE, whether the cache should be placed at macro or pico BS is unknown.

In this paper, we analyze the EE of downlink networks with a macro BS or multiple non-coordinated pico BSs with caches, where the consumption of transmit and circuit powers at the BSs, and the power consumed for backhauling and caching are taken into account but the power consumed in core network is not considered. We derive the EEs for the two kinds of networks, where content popularity and request arrival density are considered. We verify our analysis by simulations, which show that the cache enabled BS can improve the EE and caching at the pico BSs is more energy efficient.

II. System Model

Consider two downlink cellular networks, where either one macro BS or N non-coordinated pico BSs are deployed in a circle area with radius D_c to serve multiple users uniformly distributed in the area, as shown in Fig 1. We use circle cells instead of hexagonal cells for mathematical tractability. Each pico BS is equipped with M antennas and the macro BS is equipped with MN antennas. Both kinds of BSs have local cache and are connected to core network with backhaul.

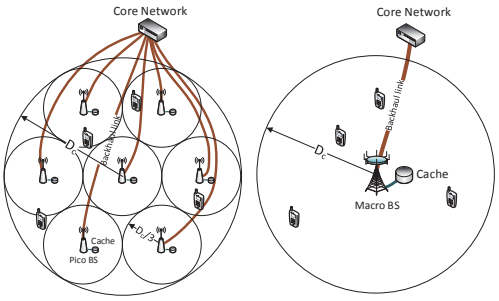

Fig. 1. Example layout of the considered two networks

Consider static content catalog that contains N_f files, ranking from one (the most popular) to N_f (the least popular) based on popularity. Assume that each user requests a file

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61120106002 and National Basic Research Program of China, 973 Program 2012CB316003.

978-1-4799-7088-9/14/$31.00 ©2014 IEEE

from the catalog with the same size F, where the file request arrival follows Poisson process with density λ. The probability of requesting file i is $P_{N_f}(i) = i^{-\delta}/\sum_{k=1}^{N_f} k^{-\delta}$ [7], where the parameter δ determines the "peakiness" of the distribution, whose value is between 0.5 and 1.0 [8].

Assume that each pico BS and the macro BS respectively cache the C^{P} and C^{M} most popular files ($C^{\mathrm{P}}, C^{\mathrm{M}} \leq N_f$) [2], then their cache capacities are $C^{\mathrm{P}}F$ and $C^{\mathrm{M}}F$, respectively. When a user requests a file i that is cached at its master BS, i.e., $i \leq C^{\mathrm{P}}$ or $i \leq C^{\mathrm{M}}$, the BS will fetch the file from its cache to serve the user; otherwise the BS will fetch the file from core network via backhaul link.

The overall transmission duration of the system is divided into T time slots each with interval of τ. We consider block fading channels where the channels remain constant during each time slot and are independent among different time slots.

Consider that BS b randomly serves $K_b(t)$ single antenna users at time slot t with zero-forcing beamforming (ZFBF) and equal power allocation, where $K_b(t) = K^{\mathrm{M}}(t) \leq MN$ for macro BS and $K_b(t) = K_b^{\mathrm{P}}(t) \leq M$ for pico BS.[1] Denote $\mathbf{H}_b(t) = [\sqrt{r_{1k}^{-\alpha}(t)}\mathbf{h}_{1b}(t), \cdots, \sqrt{r_{bk}^{-\alpha}(t)}\mathbf{h}_{K_b b}(t)]$ as the downlink channel matrix from BS b to the $K_b(t)$ users, where $r_{kb}(t)$ and $\mathbf{h}_{kb}(t)$ respectively denote the distance and the small-scale fading channel from BS b to user k, and α is the attenuation factor. When perfect channel is available, the precoding vector at BS b can be computed as $\mathbf{W}_b(t) = \sqrt{\frac{P}{K_b(t)}}[\mathbf{w}_{1b}(t), \cdots, \mathbf{w}_{K_b b}(t)]$, where P is the transmit power of each BS, $\mathbf{w}_{kb}(t) = \bar{\mathbf{w}}_{kb}(t)/\|\bar{\mathbf{w}}_{kb}(t)\|$, $\bar{\mathbf{w}}_{kb}(t)$ denotes the kth column vector of $(\mathbf{H}_b^H(t))^\dagger$, $(\cdot)^\dagger$ denotes the Moore-Penrose inverse, $(\cdot)^H$ is the conjugate transpose, and $\|\cdot\|$ denotes the Euclidean norm.

For user k served by BS b in the network with pico BSs, the received signal-to-interference-plus-noise ratio (SINR) is

$$\gamma_k^{\mathrm{P}}(t) = \frac{P}{K_b^{\mathrm{P}}(t) r_{kb}^{\alpha}(I_k + \sigma^2)} |\mathbf{h}_{kb}^H(t)\mathbf{w}_{kb}(t)|^2, \tag{1}$$

where $I_k = \sum_{j=1, j\neq b}^{N} r_{kj}^{-\alpha} \|\mathbf{h}_{kj}^H(t)\mathbf{W}_j(t)\|^2$ is the power of inter-cell interference (ICI) and σ^2 is the variance of the white Gaussian noise. For the network with a macro BS, the SINR of user k is $\gamma_k^{\mathrm{M}}(t) = P|\mathbf{h}_k^H(t)\mathbf{w}_k(t)|^2/(K^{\mathrm{M}}(t) r_k^{\alpha}\sigma^2)$.

III. EE Analysis of the Two Systems with Cache

The EE is defined as the ratio of averaged number of bits totally transmitted to averaged energy consumed in T time slots. For the network with pico BSs, the EE is

$$EE^{\mathrm{P}} = \frac{\mathbb{E}\Big\{\sum_{t=1}^{T}\sum_{b=1}^{N}\sum_{k=1}^{K_b^P(t)} \tau R_k^{\mathrm{P}}(t)\Big\}}{\mathbb{E}\Big\{\sum_{t=1}^{T}\sum_{b=1}^{N} \tau P_{b,\mathrm{BS}}(t)\Big\}} \triangleq \frac{B^{\mathrm{P}}}{E^{\mathrm{P}}}, \tag{2}$$

where $R_k^{\mathrm{P}}(t) = \log_2(1+\gamma_k^{\mathrm{P}}(t))$ is the data rate of user k served by the the pico BS b and $P_{b,\mathrm{BS}}(t)$ is the power consumed at the pico BS in time slot t.

For the network with a macro BS, the EE is

$$EE^{\mathrm{M}} = \frac{\mathbb{E}\Big\{\sum_{t=1}^{T}\sum_{k=1}^{K^M(t)} \tau R_k^{\mathrm{M}}(t)\Big\}}{\mathbb{E}\Big\{\sum_{t=1}^{T} \tau P_{\mathrm{BS}}(t)\Big\}} \triangleq \frac{B^{\mathrm{M}}}{E^{\mathrm{M}}}, \tag{3}$$

[1]If the number of users in the bth cell exceeds M, user b will select M closest users and the other users will be served by nearby BSs.

where $R_k^{\mathrm{M}}(t) = \log_2(1+\gamma_k^{\mathrm{M}}(t))$, and $P_{\mathrm{BS}}(t)$ is the power consumption of the macro BS.

A. Averaged Number of Bits Totally Transmitted

When λ is small such that the required sum rate is less than the achievable sum rate of the network, the files requested by all users can be successfully transmitted. Then, the average number of bits totally transmitted in the network is

$$B_1^{\mathrm{P}} = B_1^{\mathrm{M}} = F\lambda T\tau, \tag{4}$$

where $\lambda T\tau$ is the average number of files requested by all users in macro or pico network in T time slots.

When λ is large such that all the BSs are fully occupied, we have $K_b^{\mathrm{P}}(t) = M$ for each pico BS and $K^{\mathrm{M}}(t) = NM$ for the macro BS in all time slots $0 \leq t \leq T$. For notional simplicity, we omit (t) in $K(t)$ and $R_k(t)$ in the sequel. Then, the average number of bits totally transmitted is

$$B_2 = WT\tau NM\mathbb{E}\{R_k\}, \tag{5}$$

where W is the bandwidth, $T\tau$ is the transmission time, $B_2 = B_2^{\mathrm{P}}$ when $R_k = R_k^{\mathrm{P}}$ and $B_2 = B_2^{\mathrm{M}}$ when $R_k = R_k^{\mathrm{M}}$.

When served by the macro BS, the achievable rate of user k averaged over small scale channel can be derived by approximate the distribution of $1 + \gamma_k^{\mathrm{M}}$ as a Gamma approximation by matching the first and second moments [9].

$$\mathbb{E}_{\mathbf{h}_k}\{R_k^{\mathrm{M}}\} \approx \log_2\big(1 + \tfrac{(NM - K^{\mathrm{M}} + 1)P}{K^{\mathrm{M}} r_k^{\alpha}\sigma^2}\big). \tag{6}$$

By taking the expectation over r_k for the users uniformly distributed in the cell with radius D_c, the average achievable rate of user k when the cell-edge signal-to-noise ratio (SNR) $\frac{P}{\sigma^2 D_c^{\alpha}}$ is high can be approximated as

$$\begin{aligned}\mathbb{E}\{R_k^{\mathrm{M}}\} &\approx \int_0^{D_c} \log_2\big(1 + \tfrac{(NM - K^{\mathrm{M}} + 1)P}{K^{\mathrm{M}} r^{\alpha}\sigma^2}\big)\tfrac{2r}{D_c^2} dr \\ &\approx \tfrac{\alpha}{2\ln 2} + \log_2 \tfrac{(NM - K^{\mathrm{M}} + 1)P}{K^{\mathrm{M}} D_c^{\alpha}\sigma^2}.\end{aligned} \tag{7}$$

By substituting (7) into (5) with $K^{\mathrm{M}} = NM$, the averaged number of bits totally transmitted when λ is large can be approximated as

$$B_2^{\mathrm{M}} \approx WT\tau NM\big(\tfrac{\alpha}{2\ln 2} + \log_2 \tfrac{P}{NMD_c^{\alpha}\sigma^2}\big). \tag{8}$$

When served by the pico BSs, deriving the average rate is very challenging owing to the ICI. To simplify the derivation, we replace I_k in (1) by an approximated average ICI I by only accounting for the interference from the nearest BSs. For a user in the shaded area in Fig. 2, the average ICI can be approximated by those caused by two nearest cells as

$$I = \tfrac{2}{A}\int_{\frac{D_c}{3}}^{\frac{2D_c}{3}} \tfrac{P}{r^{\alpha}} \cdot \tfrac{\pi}{6} r dr = \tfrac{4P}{3^{1-\alpha}(\alpha-2)}(D_c^{-\alpha} - 4(2D_c)^{-\alpha}), \tag{9}$$

where $A = \pi D_c^2/36$ is the area of the shaded area.

Analogously, by averaging over the small scale fading channel and the distance r_{kb} and introducing approximations as in (6) and (8), the averaged number of bits transmitted by all N pico BSs when λ is large can be approximated as

$$B_2^{\mathrm{P}} \approx WT\tau NM \log_2\Big(\tfrac{\alpha}{2\ln 2} + \log_2 \tfrac{P}{MD_c^{\alpha}(I+\sigma^2)}\Big). \tag{10}$$

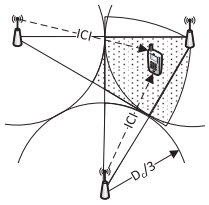

Fig. 2. Approximated average interference caused by the nearest cells

By combining (4), (8) and (10), the averaged number of bits totally transmitted in the network can be approximated as

$$B^{\mathrm{P}} \approx \min\{B_1^{\mathrm{P}}, B_2^{\mathrm{P}}\}, \quad B^{\mathrm{M}} \approx \min\{B_1^{\mathrm{M}}, B_2^{\mathrm{M}}\}. \tag{11}$$

B. Energy Consumption

We extend the typical BS power consumption model in [10] to include caching power consumption as follows,

$$P_{\mathrm{BS}}(t) = \rho P_{\mathrm{TX}}(t) + P_{\mathrm{CC}}(t) + P_{\mathrm{CA}}(t) + P_{\mathrm{BH}}(t), \tag{12}$$

where $P_{\mathrm{TX}}(t)$, $P_{\mathrm{CC}}(t)$, $P_{\mathrm{CA}}(t)$ and $P_{\mathrm{BH}}(t)$ respectively denote the power consumed at each BS for transmitting, operating circuit, caching and backhauling, and ρ reflects the impact of power amplifier, cooling and power supply.

1) Transmit and Circuit Energy Consumption: When there is no request in a cell, the BS will be turned into idle mode. The transmit power of BS in time slot t is $P_{\mathrm{TX}}(t) = P$ in active mode or 0 in idle mode. The circuit power in slot t is $P_{\mathrm{CC}}(t) = P_{\mathrm{CC,a}}$ in active mode or $P_{\mathrm{CC,i}}$ in idle mode. Then, the average transmit and circuit energy consumption of the network with pico BSs and the network with a macro BS are respectively

$$\begin{aligned} E_{\mathrm{TC}}^{\mathrm{P}} &= \mathbb{E}\left\{\sum\nolimits_{t=1}^{T}\sum\nolimits_{b=1}^{N}\tau\big(P_{b,\mathrm{TX}}(t)+P_{b,\mathrm{CC}}(t)\big)\right\} \\ &= \sum\nolimits_{b=1}^{N}\big(\bar{T}_b^{\mathrm{P}}\tau(\rho P+P_{\mathrm{CC,a}}^{\mathrm{P}})+(T-\bar{T}_b^{\mathrm{P}})\tau P_{\mathrm{CC,i}}^{\mathrm{P}}\big), \quad (13) \\ E_{\mathrm{TC}}^{\mathrm{M}} &= \bar{T}^{\mathrm{M}}\tau(\rho P+P_{\mathrm{CC,a}}^{\mathrm{M}})+(T-\bar{T}^{\mathrm{M}})\tau P_{\mathrm{CC,i}}^{\mathrm{M}}, \quad (14) \end{aligned}$$

where $\bar{T}_b^{\mathrm{P}}$ and $\bar{T}^{\mathrm{M}}$ are respectively the number of active time slots of pico BS b and that of the macro BS averaged over small-scale fading, user location and file request arrival.

Proposition 1: When $T \to \infty$, we have

$$\bar{T}_b^{\mathrm{P}} \le \bar{T}_{\max}^{\mathrm{P}} = \min\left\{T, \tfrac{\lambda TF}{NW\bar{R}_{\min}^{\mathrm{P}}}\right\}, \tag{15}$$

$$\bar{T}^{\mathrm{M}} \le \bar{T}_{\max}^{\mathrm{M}} = \min\left\{T, \tfrac{\lambda TF}{W\bar{R}_{\min}^{\mathrm{M}}}\right\}, \tag{16}$$

where $\bar{R}_{\min}^{\mathrm{P}} \approx \frac{\alpha}{2\ln 2} + \log_2\frac{MP}{D_c^{\alpha}(I+\sigma^2)}$, and $\bar{R}_{\min}^{\mathrm{M}} \approx \frac{\alpha}{2\ln 2} + \log_2\frac{NMP}{D_c^{\alpha}\sigma^2}$.

Proof: See Appendix A ∎

By substituting (15) and (16) into (13) and (14) respectively, we can obtain the upper bounds of $E_{\mathrm{TC}}^{\mathrm{P}}$ and $E_{\mathrm{TC}}^{\mathrm{M}}$.

2) Caching Energy Consumption: The caching power consumption is modeled as $P_{\mathrm{CA}}(t) = B_{\mathrm{CA}}w_{\mathrm{CA}}$ [11], where B_{CA} is the number of cached bits, and w_{CA} is the power efficiency of caching hardware in watt/bit. Then, the energy consumption for caching can be expressed as

$$E_{\mathrm{CA}}^{\mathrm{P}} = T\tau C^{\mathrm{P}}Fw_{\mathrm{CA}}, \quad E_{\mathrm{CA}}^{\mathrm{M}} = T\tau C^{\mathrm{M}}Fw_{\mathrm{CA}}, \tag{17}$$

where $T\tau$ is the total duration of caching, $C^{\mathrm{P}}F$ and $C^{\mathrm{M}}F$ are the number of bits cached at each pico and at the macro BS, respectively.

3) Backhauling Energy Consumption: The backhauling power consumption is modeled as [12]

$$P_{\mathrm{BH}}(t) = \frac{p_{\mathrm{BH}}R_{\mathrm{BH}}(t)}{C_{\mathrm{BH}}} \triangleq w_{\mathrm{BH}}R_{\mathrm{BH}}(t), \tag{18}$$

where p_{BH} is the power consumption under the backhaul capacity C_{BH}, $w_{\mathrm{BH}} \triangleq p_{\mathrm{BH}}/C_{\mathrm{BH}}$, and $R_{\mathrm{BH}}(t)$ is the backhaul traffic in time slot t.

Then, the energy consumption for backhauling averaged over small scale fading, user location and file request arrival can be expressed as

$$E_{\mathrm{BH}} = \mathbb{E}\Big\{\sum\nolimits_{t=1}^{T}\tau P_{\mathrm{BH}}(t)\Big\} = w_{\mathrm{BH}}B_{\mathrm{BH}}, \tag{19}$$

where B_{BH} is the average total number of bits fetched from the core network via backhaul links.

Denote $p_{\mathrm{CA}}^{\mathrm{P}}$ and $p_{\mathrm{CA}}^{\mathrm{M}}$ as the probability of the requested files being cached at a pico BS and at the macro BS (i.e., the hit ratio of cache), respectively. Then, we have

$$B_{\mathrm{BH}}^{\mathrm{P}} = B^{\mathrm{P}}(1-p_{\mathrm{CA}}^{\mathrm{P}}), \quad B_{\mathrm{BH}}^{\mathrm{M}} = B^{\mathrm{M}}(1-p_{\mathrm{CA}}^{\mathrm{M}}), \tag{20}$$

where B^{P} and B^{M} are given in (11), and

$$p_{\mathrm{CA}}^{\mathrm{P}} = \sum\nolimits_{i=1}^{C^{\mathrm{P}}}P_{N_f}(i) = \frac{\sum_{i=1}^{C^{\mathrm{P}}}i^{-\delta}}{\sum_{k=1}^{N_f}k^{-\delta}}, \quad p_{\mathrm{CA}}^{\mathrm{M}} = \frac{\sum_{i=1}^{C^{\mathrm{M}}}i^{-\delta}}{\sum_{k=1}^{N_f}k^{-\delta}},$$

from which the scaling law of E_{BH} can be derived as

$$\lim_{C^{\mathrm{P}}\to\infty} E_{\mathrm{BH}}^{\mathrm{P}} = w_{\mathrm{BH}}B^{\mathrm{P}}\big(1-\tfrac{(C^{\mathrm{P}})^{1-\delta}}{N_f^{1-\delta}}\big), \ 0 \le \delta < 1, \tag{21}$$

$$\lim_{C^{\mathrm{M}}\to\infty} E_{\mathrm{BH}}^{\mathrm{M}} = w_{\mathrm{BH}}B^{\mathrm{M}}\big(1-\tfrac{(C^{\mathrm{M}})^{1-\delta}}{N_f^{1-\delta}}\big), \ 0 \le \delta < 1. \tag{22}$$

This implies that the energy consumption for backhauling will reduce sharply with a little increase of the cache size but decrease more and more slowly with the growth of cache size.

Finally, the average total energy consumption in the networks are respectively $E^{\mathrm{P}} = E_{\mathrm{TC}}^{\mathrm{P}} + E_{\mathrm{CA}}^{\mathrm{P}} + E_{\mathrm{BH}}^{\mathrm{P}}$ and $E^{\mathrm{M}} = E_{\mathrm{TC}}^{\mathrm{M}} + E_{\mathrm{CA}}^{\mathrm{M}} + E_{\mathrm{BH}}^{\mathrm{M}}$, and the EEs can be respectively computed with (2) and (3).

IV. Analytical and Simulation Results

In this section, we evaluate the impact of deploying cache at the macro BS or pico BS on the EE of the network.

We consider the network either with $N=7$ non-coordinated pico BSs each with $M=4$ antennas or one macro BS with $NM=28$ antennas as shown in Fig 1. $W=10$ MHz, $D_c=250$ m, $\sigma^2=-95$ dBm, $\tau=5$ ms, and $T=10^4$. The path-loss models for pico and macro cell are $30.6+36.7\log_{10}(r_{kb})$ in dB and $35.3+37.6\log_{10}(r_k)$ in dB, respectively [13]. Unless otherwise specified, the file catalog contains $N_f=1000$ files each with size of $F=30$ MB (MegaByte) [2]; the file request density within the area is $\lambda=2$ files/second and the popularity distribution of the files follows Zipf-like distribution with parameter $\delta=0.8$ [14]. The power consumption parameters of the system are listed in Table I.

In Fig. 3, we provide the analytical results of the energy consumption and EE of the network with pico BSs versus the

TABLE I
POWER CONSUMPTION PARAMETERS

	Macro BS	Pico BS
ρ	3.22	15.13
P	46 dBm	21 dBm
P_{CC}	$P^{\mathrm{M}}_{\mathrm{CC,i}} = 381.63$ W	$P^{\mathrm{P}}_{\mathrm{CC,i}} = 3.87$ W
	$P^{\mathrm{M}}_{\mathrm{CC,a}} = 724.64$W	$P^{\mathrm{P}}_{\mathrm{CC,a}} = 10.16$ W [15]
w_{BH}	5×10^{-7} J/bit (Microwave link) [12]	
w_{CA}	6.25×10^{-12} W/bit (High-speed solid state disk) [11]	

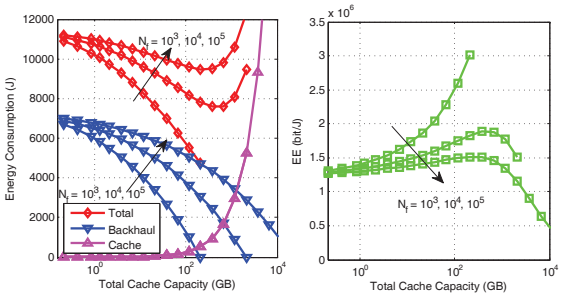

Fig. 3. Impact of file catalog size N_f, the network with pico BSs

total cache capacity $NC^{\mathrm{P}}F$ in the network with different file catalog sizes. Because simulation results are very close to the analytical results and the two kinds of networks behave similarly, we only show the analytical results of the network with pico BSs due to space limitation. It is shown that the energy consumption for backhauling decreases with the growing of cache capacity and increases with N_f. When N_f is small, the total energy consumption continues to drop until C^{P} reaches N_f, otherwise it first decreases and then increases. This is because the energy consumed for backhauling decreases with cache capacity as shown in (21) while that for cache increases linearly with the cache size as in (17). Since the cache capacity and N_f do not affect the total number of bits transmitted in the network, the EE behaves oppositely to the total energy consumption.

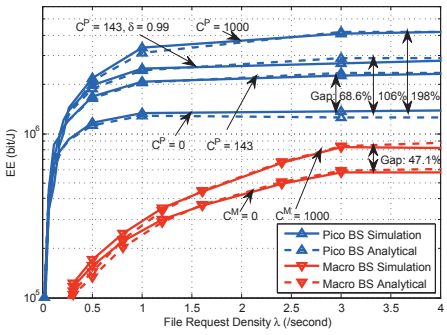

Fig. 4. EE of the macro or pico network vs. λ, $N_f = 1000$

In Fig. 4, we compare the analytical results of the EE for both networks with the simulation results. It is shown that the analytical results are close to the simulation results. When λ increases, the EE of both systems first increase and then remain constant, because each BS will become fully occupied and the total number of bits will not increase anymore. When the total cache capacities in the two kinds of networks are equal, i.e., $C^{\mathrm{P}} = 143$ and $C^{\mathrm{M}} = 1000$, the EE gain from using caches in pico BSs is larger than that from using cache in macro BS, because the pico BSs have more opportunities to idle and have low transmit and circuit power. If the cache capacity at each pico BS is the same as that at the macro BS or δ gets larger, the EE gain in the pico network will be even higher. This suggests that the network with pico BSs benefits more from caching by saving the energy consumption for backhauling.

V. Conclusion

In this paper, we investigated whether introducing cache to BS will improve the EE of downlink networks and where the cache should be placed. Analysis and simulation results showed that the EE will be improved by introducing cache if the file catalog size is not too large. Compared with caching at a macro BS, caching at multiple pico BSs is more energy efficient by saving the energy consumed for backhauling.

Appendix A
Proof of Proposition 1

When the macro BS schedules only one user at each time slot, the BS operates at its minimal service capability. Then, the active time will reach the maximal value $T^{\mathrm{M}}_{\max}$ ($T^{\mathrm{M}}_{\max} \leq T$). Denote $R^{\mathrm{M}}_{\min}(t_i)$ as the sum rate of the system when serving one user at the ith active time slot t_i, $0 \leq t_i \leq T^{\mathrm{M}}_{\max}$. When $T^{\mathrm{M}}_{\max} < T$, the total number of bits of requested files during T time slots should be equal to the total number of bits transmitted, which is

$$B_{\mathrm{RQ}} = W \sum\nolimits_{i=1}^{T^{\mathrm{M}}_{\max}} \tau R^{\mathrm{M}}_{\min}(t_i). \tag{23}$$

By taking the inverse of both sides of (23) and multiplying both sides by $T^{M}_{\max}/T$, we can obtain

$$\frac{T^{\mathrm{M}}_{\max}}{T} = \frac{\frac{B_{\mathrm{RQ}}}{T}}{\frac{W\tau}{T^{\mathrm{M}}_{\max}} \sum_{i=1}^{T^{\mathrm{M}}_{\max}} R^{\mathrm{M}}_{\min}(t_i)}. \tag{24}$$

If $T \to \infty$, then $T^{M}_{\max} \to \infty$, and we have

$$\begin{aligned} \bar{R}^{\mathrm{M}}_{\min} &\triangleq \lim_{T\to\infty} \frac{1}{T^{\mathrm{M}}_{\max}} \sum\nolimits_{i=1}^{T^{\mathrm{M}}_{\max}} R^{\mathrm{M}}_{\min}(t_i) = \mathbb{E}\{R^{\mathrm{M}}_{\min}(t_i)\} \\ &= \mathbb{E}\{R^{\mathrm{M}}_{k}\} \overset{(a)}{\approx} \frac{\alpha}{2\ln 2} + \log_2 \frac{NMP}{D_c^{\alpha}\sigma^2}, \end{aligned} \tag{25}$$

where in step (a) we use the approximation in (7) by letting $K^{\mathrm{M}} = 1$. Substituting (25) into (24) for $T \to \infty$ and further taking the expectation of both sides of (24) over small scale fading, user location and file request arrival, we have

$$\lim_{T\to\infty} \frac{\bar{T}^{\mathrm{M}}_{\max}}{T} = \frac{\lambda F}{W \bar{R}^{\mathrm{M}}_{\min}}, \tag{26}$$

where we use $\mathbb{E}\{B_{\mathrm{RQ}}\} = F\lambda T\tau$ and $\bar{T}^{\mathrm{M}}_{\max} = \mathbb{E}\{T^{\mathrm{M}}_{\max}\}$. Combining (25) and (26), the result for the macro BS is proved. The result for the pico BSs can be similarly proved.

REFERENCES

[1] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," in *Proc. ACM SIGCOMM Conference on Internet measurement*, 2007.

[2] 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.

[3] 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.

[4] S. P. Shariatpanahi, H. Shah-Mansouri, and B. H. Khalaj, "Caching gain in wireless networks with fading: A multi-user diversity perspective," in *Proc. IEEE WCNC*, 2014.

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

[6] 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.

[7] 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.

[8] M. Cha, P. Rodriguez, J. Crowcroft, S. Moon, and X. Amatriain, "Watching television over an ip network," in *Proc. ACM SIGCOMM Conference on Internet Measurement*, 2008.

[9] N. I. Akhiezer and N. Kemmer, *The classical moment problem: and some related questions in analysis*. Oliver & Boyd Edinburgh, 1965, vol. 5.

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.