Flexible Cache-Aided Networks with Backhauling

Italo Atzeni,[1] Marco Maso,[1] Imène Ghamnia,[2] Ejder Baştuğ,[2,3] and Mérouane Debbah[1,2]
[1]Mathematical and Algorithmic Sciences Lab, France Research Center, Huawei Technologies France SASU
[2]Large Networks and Systems Group (LANEAS), CentraleSupélec
[3]Research Laboratory of Electronics, Massachusetts Institute of Technology (MIT)
Email: {italo.atzeni, marco.maso, merouane.debbah}@huawei.com, imene.ghamnia@centralesupelec.fr, ejder@mit.edu

***Abstract*—Caching at the edge is a promising technique to cope with the increasing data demand in wireless networks. This paper analyzes the performance of cellular networks consisting of a tier macro-cell wireless backhaul nodes overlaid with a tier of cache-aided small cells. We consider both *static* and *dynamic* association policies for content delivery to the user terminals and analyze their performance. In particular, we derive closed-form expressions for the area spectral efficiency and the energy efficiency, which are used to optimize relevant design parameters such as the density of cache-aided small cells and the storage size. By means of this approach, we are able to draw useful design insights for the deployment of highly performing cache-aided tiered networks.**

I. Introduction

A tremendous increase in smartphone usage during the last decade resulted in a 4000-fold mobile data traffic explosion in cellular networks [1]. In this setting, mobile operators constantly look for innovative solutions to satisfy performance indicators such as higher spectral and energy efficiency, improved coverage, and lower end-to-end delays: this calls for the development of innovative techniques at both device and network level in the coming years [2]. Caching at the edge is regarded as one of the most promising approaches to cope with the increasing data demand [3]. This is particularly true for heterogeneous and tiered network layouts, for which the extent of the performance gains brought by caching has been extensively assessed in terms of several performance metrics of practical relevance. For instance, [4] and [5] focus on optimal probabilistic caching policies for enhanced content delivery, [6] studies the impact of caching on the coverage and on the delay experienced by the user terminals (UTs), and [7] jointly considers caching, routing, and channel assignment strategies for collaborative small cells (SCs).

Along similar lines, this work analyzes the system performance of cache-aided tiered networks comprised of a tier of macro-cell backhaul (BH) nodes overlaid with a tier of non-cooperative SCs [8]. Each SC in the considered network has (limited) storage capabilities and pre-fetches popular files before the downlink transmission to its associated UTs. In this context, and differently from what has been previously proposed in the literature, our goal is to optimize relevant performance metrics for network planners and operators, and draw design guidelines for the deployment of highly performing cache-aided tiered networks. Using consolidated random spatial models to characterize the location of SCs, BHs and UTs [9], we investigate the area spectral efficiency (ASE) gains brought by caching popular files in the considered network using tools from stochastic geometry. To increase the generality and relevance of our study, we examine both *static* and *dynamic* association policies for content delivery to the UTs. In particular, in the static case, UTs can only be associated with SCs, regardless of the occurrence of cache hit/miss events at the latter; conversely, in the dynamic case, UTs are associated with SCs in case of cache hit and with wireless BHs otherwise.

The first contribution of this paper is the derivation of closed-form expressions and bounds for the ASE and the energy efficiency (EE) with the aforementioned UT association policies. Such expressions are subsequently used to optimize crucial design parameters such as SC density and storage size. Our findings highlight that the impact of caching on the ASE is minor as compared to that of the SC density, i.e., densifying the network is always more convenient than expanding the storage size at the SCs. Conversely, we show that increasing the storage size rather than the SC density is more effective in terms of EE of the network. In practice, the choice of investing on additional SCs or on more storage strongly depends on the goal of the network operator. Finally, it is worth observing that the performance of the considered network heavily hinges on the adopted UT association policy: a dynamic policy is always preferable whenever the network can support the larger complexity necessary for its effective implementation.

This research has been supported in part by the ERC Starting Grant 305123 MORE (Advanced Mathematical Tools for Complex Network Engineering), the U.S. National Science Foundation under Grant CCF-1409228.

II. System Model

A. Network Model

Let us consider a set of mobile UTs in a two-tier ultra-dense network, where the lower tier comprises a set of non-cooperative SCs ensuring network coverage and the upper tier consists in a set of BHs connected to the internet. In our model, each SC is associated with only one UT and one BH. We assume that the spatial distribution of the network nodes follows the stationary, independently marked Poisson point processes (PPP) $\Phi \triangleq \left\{\left(x, u(x), b(x)\right)\right\} \subset \mathbb{R}^2 \times \mathbb{R}^2 \times \mathbb{R}^2$. We use $\Phi_{\text{SC}} \triangleq \{x\}$ to denote the PPP of the SCs with spatial density λ (measured in SCs/m^2), with isotropic marks $\Phi_{\text{UT}} \triangleq u(\Phi_{\text{SC}}) = \{u(x)\}_{x\in\Phi_{\text{SC}}}$ and $\Phi_{\text{BH}} \triangleq b(\Phi_{\text{SC}}) = \{b(x)\}_{x\in\Phi_{\text{SC}}}$ representing the UTs and the BHs, respectively; evidently, Φ_{UT} and Φ_{BH} are dependent on Φ_{SC} and have the same spatial density λ. In this setting, the employment of PPPs allows to capture the randomness of practical ultra-dense network

deployments and, at the same time, obtain precise and tractable expressions for the system-level performance metrics [9]–[11]. Let $r_{yz} \triangleq \|y-z\|$ denote the distance between nodes $y, z \in \Phi$: the distances of the UTs and the BHs from their associated SCs are assumed fixed and are denoted by $R_{\text{UT}} \triangleq r_{xu(x)}$ and $R_{\text{BH}} \triangleq r_{b(x)x}$, $\forall x \in \Phi_{\text{SC}}$, respectively, with $R_{\text{BH}} > R_{\text{UT}}$.

B. Caching Model and UT Association

Let $\mathcal{F} \triangleq \{f_i\}_{i=1}^{F}$ be a F-sized subset of all the files available in the internet, entirely accessible by each BH. Without loss of generality, and for simplicity in the notation, we assume that all F files have identical lengths. In the following, we refer to $\mathcal{F}$ as *file catalog*. Each UT requests files from $\mathcal{F}$ with probability denoted by $\mathcal{P} \triangleq \{p_1, p_2, \ldots, p_F\}$, with $\sum_{i=1}^{F} p_i = 1$, which hinges on the files popularity over the whole network.

To offload the overlaying BH infrastructure, each non-cooperative SC x is equipped with a *storage unit* Δ_x of size $S \leq F$ (measured in files/SC). Suppose that, at a given time instant, UT $u(x)$ is interested in downloading file $f_i \in \mathcal{F}$: if f_i is cached at its associated SC x, i.e., $f_i \in \Delta_x$, we have a *cache hit* event; conversely, if $f_i \notin \Delta_x$, we have a *cache miss* event. In this regard, we use $\not\Delta_x$ to denote a cache miss event at x. Accordingly, we introduce the indicator function

$$\mathbb{1}_{\not\Delta_x} \triangleq \begin{cases} 1, & \text{if } \not\Delta_x \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where its logical complement is denoted by $\mathbb{1}_{\Delta_x}$.

The effectiveness of the adopted cache-aided approach depends on the probability that any file requested by a given UT is cached at its serving SC, which is referred to as *cache hit probability* and is denoted by P_{hit}. This quantity typically depends on the adopted caching policy, content placement, and distribution of the UT requests. Further details on these aspects are unnecessary for the scope of our work and we refer to [3] and references therein for further details.

Now, let $R = TB \in \mathbb{R}$ be the time-frequency resource used by the network for delivering one file from the BH tier to the UTs, with $T \in \mathbb{R}$ (resp. $B \in \mathbb{R}$) defined as the amount of necessary time (resp. frequency) resources to perform this operation. Two UT association policies are considered in this work, i.e., *static* and *dynamic*. When the static UT association is adopted, operations in the two tiers occur on orthogonal resources. In this case, the UTs can only be served by their associated SCs, and the BH-to-SC/SC-to-UT links occupy two $\frac{R}{2}$-sized resources: for instance, each link may use the entirety of B for half of the time (i.e., $\frac{T}{2}$) or the entirety of T for half of the bandwidth (i.e., $\frac{B}{2}$). Conversely, under dynamic UT association, operations in the two tiers occur on the same resource. In this case, the UTs are served by either the SCs or the BHs depending on the availability of the requested file at the SCs, and the SC-to-UT/BH-to-UT links occupy a R-sized resource. In practice, the second approach leads to a more efficient use of the time-frequency resource at the cost of more sophisticated inter-tier coordination and possibly signaling. For notational simplicity, in the rest of the paper we differentiate between static and dynamic UT association by means of the super-indices S and D, respectively.

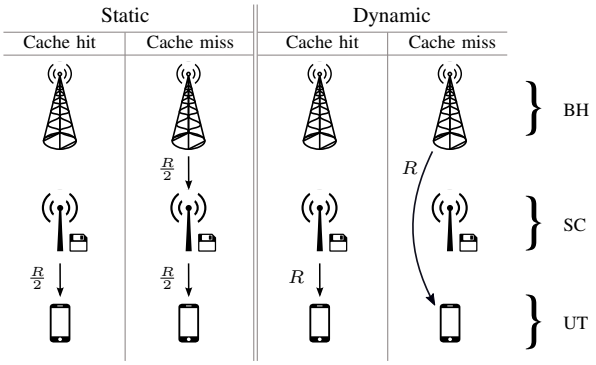

Figure 1. System model for static and dynamic UT association.

The adopted UT association policy has a strong impact on the network response to a cache hit/miss event when UT $u(x)$ requests file f_i. In particular:

1) *Static*: If a cache hit occurs, then SC x transmits f_i to $u(x)$. Conversely, if a cache miss occurs, first x retrieves f_i from its associated BH $b(x)$, then transmits it to UT $u(x)$. In both cases, an $\frac{R}{2}$-sized resource is used by the SC-to-UT link due to the static UT association;
2) *Dynamic*: If a cache hit occurs, then x transmits f_i to $u(x)$. Conversely, if a cache miss occurs, $u(x)$ can bypass x and connect directly to $b(x)$ to download f_i. In both cases, the whole resource R is used by either the SC-to-UT or the BH-to-UT link.

A graphical representation of these operations, and their impact on the network resource usage, is provided in Figure 1.

C. Channel Model

In this work, all nodes have a single transmit/receive antenna. In addition, we assume that the SCs and the BHs transmit with powers ρ_{SC} and ρ_{BH} (measured in W), respectively. The propagation through the wireless channel combines pathloss attenuation and small-scale fading. For the former, we adopt the standard power-law pathloss model and define the pathloss function $\ell(y, z) \triangleq r_{yz}^{-\alpha}$ between nodes $y, z \in \Phi$, with pathloss exponent $\alpha > 2$. For the latter, we use h_{yz} to denote the channel power fading gain between nodes $y, z \in \Phi$: all channels are assumed to be subject to Rayleigh fading and, hence, $h_{yz} \sim \exp(1)$.[1] The signal-to-interference-plus-noise ratio (SINR) at SC x reads as

$$\mathsf{SINR}_x \triangleq \frac{\rho_{\text{BH}} R_{\text{BH}}^{-\alpha} h_{b(x)x}}{I_x + \sigma^2} \tag{2}$$

where I_x is the overall interference power at SC x and σ^2 is the additive noise power. Likewise, the SINR at UT $u(x)$ reads as

$$\mathsf{SINR}_{u(x)} \triangleq \frac{\rho_{\text{SC}} R_{\text{UT}}^{-\alpha} h_{xu(x)}}{I_{u(x)} + \sigma^2} \tag{3}$$

[1]For more involved types of small-scale fading, e.g., Nakagami$-m$ fading, we refer to [11].

where $I_{u(x)}$ is the overall interference power at $u(x)$. For static UT association, we have $I_x = I_x^{(\mathrm{S})}$ and $I_{u(x)} = I_{u(x)}^{(\mathrm{S})}$, with

$$I_x^{(\mathrm{S})} \triangleq \sum_{y\in\Phi_{\mathrm{SC}}\setminus\{x\}} \rho_{\mathrm{BH}} r_{b(y)x}^{-\alpha} h_{b(y)x} \mathbb{1}_{\not\Delta_y} \tag{4}$$

$$I_{u(x)}^{(\mathrm{S})} \triangleq \sum_{y\in\Phi_{\mathrm{SC}}\setminus\{x\}} \rho_{\mathrm{SC}} r_{yu(x)}^{-\alpha} h_{yu(x)} \tag{5}$$

respectively. On the other hand, in case of dynamic UT association, we are only interested in the interference at the UT and we thus have $I_{u(x)} = I_{u(x)}^{(\mathrm{D})}$, with

$$I_{u(x)}^{(\mathrm{D})} \triangleq \sum_{y\in\Phi_{\mathrm{SC}}\setminus\{x\}} \big(\rho_{\mathrm{SC}} r_{yu(x)}^{-\alpha} h_{yu(x)} \mathbb{1}_{\Delta_y} + \rho_{\mathrm{BH}} r_{b(y)u(x)}^{-\alpha} h_{b(y)u(x)} \mathbb{1}_{\not\Delta_y}\big). \tag{6}$$

III. Performance Analysis

The system performance is closely related to the probability of the UTs successfully receiving the requested content, which is termed as *success probability*. This depends on two factors, i.e., the cache hit probability (cf. Section II-B) and the SINRs of the file transmissions (cf. Section II-C). Thus, in the following, we derive the success probability using both the considered UT association policies.

Our analysis focuses on a randomly chosen SC x, and its marks $u(x)$ and $b(x)$, referred to as *typical SC*, *typical UT* and *typical BH*, respectively: due to Slivnyak's theorem and to the stationarity of Φ, these nodes are representative of the whole network [9]. For a given SINR threshold θ, we assume that a file is successfully received by the typical UT if $\mathsf{SINR}_x > \theta \wedge \mathsf{SINR}_{u(x)} > \theta$ and $\mathsf{SINR}_{u(x)} > \theta$ for static and dynamic UT association, respectively. The success probability is formalized in the following theorem, where we use the notation

$$\Upsilon(z) \triangleq \frac{\pi z^{\frac{2}{\alpha}} \csc\left(\frac{2\pi}{\alpha}\right)}{\alpha}. \tag{7}$$

Theorem 1. For static UT association, the success probability is given by

$$\mathsf{P}_{\mathrm{suc}}^{(\mathrm{S})}(\theta) \triangleq \exp\big(-\theta\tfrac{\sigma^2}{\rho_{\mathrm{SC}}} R_{\mathrm{UT}}^{\alpha}\big)\mathcal{L}_{I_{u(x)}}^{(\mathrm{S})}\big(\theta\rho_{\mathrm{SC}}^{-1} R_{\mathrm{UT}}^{\alpha}\big) \times \big(\mathsf{P}_{\mathrm{hit}} + (1-\mathsf{P}_{\mathrm{hit}})\exp\big(-\theta\tfrac{\sigma^2}{\rho_{\mathrm{BH}}} R_{\mathrm{BH}}^{\alpha}\big)\mathcal{L}_{I_x}^{(\mathrm{S})}\big(\theta\rho_{\mathrm{BH}}^{-1} R_{\mathrm{BH}}^{\alpha}\big)\big) \tag{8}$$

where

$$\mathcal{L}_{I_x}^{(\mathrm{S})}(s) \triangleq \exp\big(-2\pi\lambda(1-\mathsf{P}_{\mathrm{hit}})\Upsilon(\rho_{\mathrm{BH}} s)\big) \tag{9}$$

$$\mathcal{L}_{I_{u(x)}}^{(\mathrm{S})}(s) \triangleq \exp\big(-2\pi\lambda\Upsilon(\rho_{\mathrm{SC}} s)\big) \tag{10}$$

are the Laplace transforms of the interference terms $I_x^{(\mathrm{S})}$ and $I_{u(x)}^{(\mathrm{S})}$ in (4)–(5), respectively. For dynamic UT association, the success probability is given by

$$\mathsf{P}_{\mathrm{suc}}^{(\mathrm{D})}(\theta) \triangleq \mathsf{P}_{\mathrm{hit}} \exp\big(-\theta\tfrac{\sigma^2}{\rho_{\mathrm{SC}}} R_{\mathrm{UT}}^{\alpha}\big)\mathcal{L}_{I_{u(x)}}^{(\mathrm{D})}\big(\theta\rho_{\mathrm{SC}}^{-1} R_{\mathrm{UT}}^{\alpha}\big) + (1-\mathsf{P}_{\mathrm{hit}}) \times \exp\big(-\theta\tfrac{\sigma^2}{\rho_{\mathrm{BH}}} R_{\mathrm{BH}}^{\alpha}\big)\int_0^{2\pi} \mathcal{L}_{I_{u(x)}}^{(\mathrm{D})}\big(\theta\rho_{\mathrm{BH}}^{-1}\Omega(R_{\mathrm{UT}}, R_{\mathrm{BH}}, \phi)\big)\frac{\mathrm{d}\phi}{2\pi} \tag{11}$$

with $\Omega(R_{\mathrm{UT}}, R_{\mathrm{BH}}, \phi) \triangleq \big(R_{\mathrm{UT}}^2 + R_{\mathrm{BH}}^2 + 2R_{\mathrm{UT}}R_{\mathrm{BH}}\cos\phi\big)^{\frac{\alpha}{2}}$ and where

$$\mathcal{L}_{I_{u(x)}}^{(\mathrm{D})}(s) \triangleq \exp\big(-2\pi\lambda\mathsf{P}_{\mathrm{hit}}\Upsilon(\rho_{\mathrm{SC}} s)\big) \times \exp\big(-2\pi\lambda(1-\mathsf{P}_{\mathrm{hit}})\Upsilon(\rho_{\mathrm{BH}} s)\big) \tag{12}$$

is the Laplace transform of the interference term $I_{u(x)}^{(\mathrm{D})}$ in (6).

Proof: See Appendix I. ∎

Since a closed-form expression for the success probability $\mathsf{P}_{\mathrm{suc}}^{(\mathrm{D})}(\theta)$ in (11) is not available, we provide a useful lower bound in the following corollary.

Corollary 1. For dynamic UT association, the success probability can be lower-bounded as

$$\mathsf{P}_{\mathrm{suc}}^{(\mathrm{D})}(\theta) \geq \mathsf{P}_{\mathrm{hit}} \exp\big(-\theta\tfrac{\sigma^2}{\rho_{\mathrm{SC}}} R_{\mathrm{UT}}^{\alpha}\big)\mathcal{L}_{I_{u(x)}}^{(\mathrm{D})}\big(\theta\rho_{\mathrm{SC}}^{-1} R_{\mathrm{UT}}^{\alpha}\big) + (1-\mathsf{P}_{\mathrm{hit}})\exp\big(-\theta\tfrac{\sigma^2}{\rho_{\mathrm{BH}}} R_{\mathrm{BH}}^{\alpha}\big)\mathcal{L}_{I_x}^{(\mathrm{D})}\big(\theta\rho_{\mathrm{BH}}^{-1}\big(R_{\mathrm{BH}} + R_{\mathrm{UT}}\big)^{\alpha}\big). \tag{13}$$

Proof: The lower bound is obtained by considering the worst-case distance between BH and UT, i.e., $R_{\mathrm{BH}} + R_{\mathrm{UT}}$ (cf. [9]). ∎

The expressions above can be used to obtain other useful performance metrics. In this paper, we focus on the achievable ASE (measured in bps/Hz/m^2), which is readily obtained as

$$\mathsf{ASE}^{(\mathrm{S})}(\theta,\lambda) \triangleq \frac{1}{2}\lambda\mathsf{P}_{\mathrm{suc}}^{(\mathrm{S})}(\theta)\log_2(1+\theta) \tag{14}$$

$$\mathsf{ASE}^{(\mathrm{D})}(\theta,\lambda) \triangleq \lambda\mathsf{P}_{\mathrm{suc}}^{(\mathrm{D})}(\theta)\log_2(1+\theta) \tag{15}$$

for static and dynamic UT association, respectively. At this stage, it is worth observing that, in practice, the distances R_{UT} and R_{BH} depend on the SC density λ. Thus, in the following we set $R_{\mathrm{UT}} = \frac{\beta_{\mathrm{UT}}}{2\sqrt{\lambda}}$ and $R_{\mathrm{BH}} = \frac{\beta_{\mathrm{BH}}}{2\sqrt{\lambda}}$, with $\beta_{\mathrm{BH}} > \beta_{\mathrm{UT}}$; observe that $\frac{1}{2\sqrt{\lambda}}$ represents the average distance between nodes in a PPP with spatial density λ. Figure 2 illustrates the achievable ASEs as defined in (14)–(15) against the cache hit probability $\mathsf{P}_{\mathrm{hit}}$, with $\lambda = 10^{-2}$ SCs/m^2, $\alpha = 4$, $\theta = 1$, $\sigma^2 = 0$ (i.e., interference limited case), $\rho_{\mathrm{SC}} = 0.5$ W, $\rho_{\mathrm{BH}} = 1$ W, $\beta_{\mathrm{SC}} = 0.5$, and $\beta_{\mathrm{BH}} = 1$ (the same parameters will be used in Section IV-A). The beneficial impact of caching is evident: as compared with the cache-free setup (which corresponds to $\mathsf{P}_{\mathrm{hit}} = 0$), the achievable ASE with $\mathsf{P}_{\mathrm{hit}} = 0.25$ (resp. $\mathsf{P}_{\mathrm{hit}} = 0.5$) is increased by 188% (resp. 264%) for dynamic UT association and by 158% (resp. 218%) for static UT association. Furthermore, we observe that the dynamic UT association always outperforms the static UT association. Lastly, the lower bound on (15), derived using Corollary 1, is increasingly tight as $\mathsf{P}_{\mathrm{hit}}$ grows large.

IV. Cache-Aided SC Network Deployment

In this section, we look at possible design problems of interest to network operators and we aim at providing simple design guidelines for the deployment of cache-aided SC networks. As design parameters, we focus on the SC density $\lambda \in [\lambda^{(\min)}, \lambda^{(\max)}]$ and the storage size $S \in [0, S^{(\max)}]$, with $S^{(\max)} \leq F$: therefore, we make the dependence on these parameters explicit in our performance metrics. For

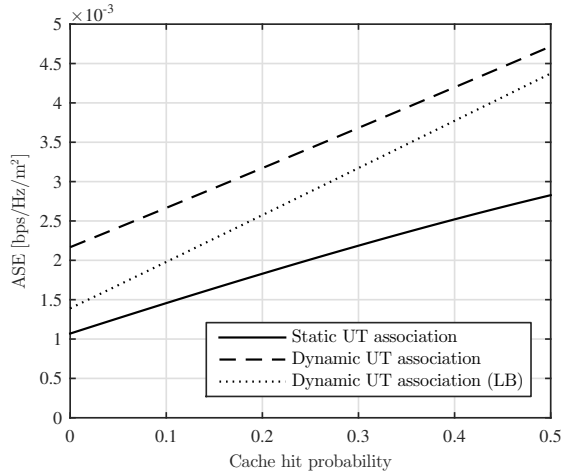

Figure 2. Achievable ASE as a function of the cache hit probability.

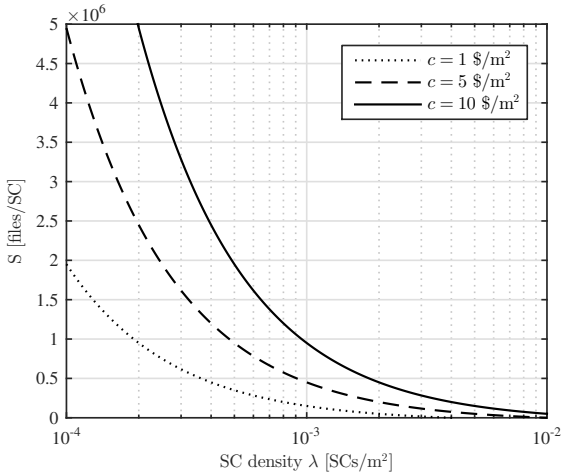

Figure 3. Feasible solution set of problems (P1)–(P2) for different values of the monetary budget c.

instance, the minimum SC density $\lambda^{(\min)}$ can be chosen to ensure an interference-limited system performance (i.e., the background noise becomes negligible), whereas the maximum SC density $\lambda^{(\max)}$ can be due to spatial limitations; similarly, the maximum storage size $S^{(\max)}$ may arise from practical constraints regarding space or power consumption at the SCs.

In our analysis, we consider the simplest case of uniform files popularity, i.e., $p_i = \frac{1}{F}$, $i = 1, \ldots, F$, with $\mathsf{P}_{\rm hit}(S) = \frac{S}{F}$, where we have made the dependence of the cache hit probability on S explicit. Note that, for the more realistic case of decreasing files popularity, i.e., $p_1 > p_2 > \ldots > p_F$, expanding the storage will have an increasingly less significant impact on the system performance.

A. Deployment Cost

Suppose that a network operator has a fixed monetary budget per m^2 c (measured in \$/m^2) to deploy a cache-aided SC network. Let p_λ denote the SC price (measured in \$/SC) and let p_S denote the storage price (measured in \$/file). The following simple question naturally arises in this context: *is it better to invest in deploying SCs or storage?* The answer hinges on the choice of the performance metric and can be found by means of an appropriate optimization. For instance, to maximize the achievable ASE in (14)–(15), the answer is given by the solution of the following optimization problem:

$$\begin{aligned} \max_{\lambda,S} \quad & \mathsf{ASE}(\lambda, S) \\ \text{s.t.} \quad & p_\lambda \lambda + p_S \lambda S \leq c \\ & \lambda \in [\lambda^{(\min)}, \lambda^{(\max)}] \\ & S \in [0, S^{(\max)}]. \end{aligned} \tag{P1}$$

Alternatively, the system performance can be measured in terms of its EE. Before formalizing the corresponding problem, some definitions are in order. Let

$$E_{\rm tot}(S) \triangleq \mathsf{P}_{\rm hit}(S) E_{\rm hit} + \big(1 - \mathsf{P}_{\rm hit}(S)\big) E_{\rm miss} \tag{16}$$

be the total energy consumed to provide a content to a UT (measured in J), where $E_{\rm hit}$ is the energy needed by the SC to retrieve a cached file from its storage unit and transmit it to the UT in case of cache hit, and $E_{\rm miss}$ is the energy needed by the BH to retrieve a non-cached file from the internet and transmit it to the UT in case of cache miss (via the SC for static UT association and directly for dynamic UT association). Moreover, we define the area energy consumption (measured in J/m^2) as $\mathsf{AEC}(\lambda, S) \triangleq \lambda E_{\rm tot}(S)$ and the EE (measured in bit/J) as

$$\mathsf{EE}(\lambda, S) \triangleq \frac{\mathsf{ASE}(\lambda, S)}{\mathsf{AEC}(\lambda, S)}. \tag{17}$$

Hence, the corresponding optimization problem is given by:

$$\begin{aligned} \max_{\lambda,S} \quad & \mathsf{EE}(\lambda, S) \\ \text{s.t.} \quad & p_\lambda \lambda + p_S \lambda S \leq c \\ & \lambda \in [\lambda^{(\min)}, \lambda^{(\max)}] \\ & S \in [0, S^{(\max)}]. \end{aligned} \tag{P2}$$

Using the same parameters as in Section III, we additionally set $F = 10^7$ files, $\lambda^{(\min)} = 10^{-4}$ SCs/m^2, $\lambda^{(\max)} = 10^{-2}$ SCs/m^2, $S^{(\max)} = 5 \times 10^6$ files/SC, $p_\lambda = 250$ \$/SC, $p_\lambda = 0.005$ \$/file, $E_{\rm hit} = 1$ J, and $E_{\rm miss} = 10$ J. Assuming $p_\lambda \lambda^\star + p_S \lambda^\star S^\star = c$, which excludes the trivial solution $\{\lambda^\star, S^\star\} = \{\lambda^{(\max)}, S^{(\max)}\}$, Figure 3 illustrates the feasible solution set of the considered problems with $c = \{1, 2.5, 5\}$ \$/m^2.

Figures 4 and 5 plot the achievable ASE and the EE, respectively, against the SC density λ, where each value of λ yields a value of S (and thus of $\mathsf{P}_{\rm hit}$) according to the constraint on the monetary budget (cf. Figure 3). The achievable ASE in Figure 4 is monotonically increasing in λ: this means that, to maximize the network ASE, one should invest in deploying more SCs (even if that means $S = 0$), whereas S should be increased only if there is remaining budget. Therefore, the solution of problem (P1) is given by

$$\lambda^\star = \min\left\{\frac{c}{p_\lambda}, \lambda^{(\max)}\right\}, \quad S^\star = \frac{1}{p_S}\left(\frac{c}{\lambda^\star} - p_\lambda\right). \tag{18}$$

On the contrary, the EE in Figure 5 is monotonically decreasing in λ: this means that, to maximize the network EE, one should invest in expanding the storage as much as

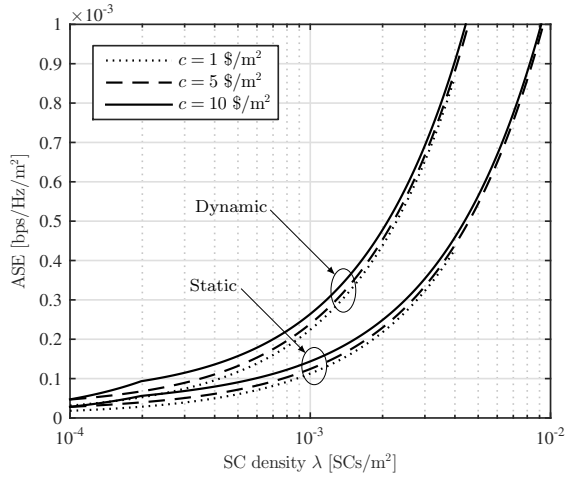

Figure 4. Achievable ASE as a function of the SC density.

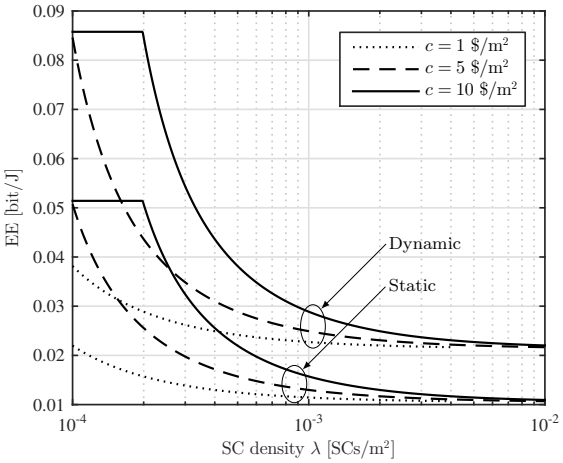

Figure 5. EE as a function of the SC density.

possible (even if that means $\lambda = \lambda^{(\min)}$), whereas λ should be increased only if there is remaining budget. Hence, the solution of problem (P2) reads as

$$S^\star = \min\left\{\frac{1}{p_S}\left(\frac{c}{\lambda^{(\min)}} - p_\lambda\right), S^{(\max)}\right\}, \quad \lambda^\star = \frac{c}{p_\lambda + p_S S^\star}. \tag{19}$$

In practice, increasing the SCs caching capabilities seems to be very useful for increasing the EE. Conversely, the impact of additional storage on the ASE is minor if the operator disposes of additional resources to increase the SC density. These results suggest that the answer to the question initially posed in this section, i.e., the choice of investing on additional SCs or on more storage, is not unique but strongly hinges on the ultimate goal of the network operator. Finally, we observe that a dynamic UT association policy is always preferable in terms of network performance whenever the larger complexity necessary to implement it can be afforded.

V. Conclusions

In this work, we study the performance of a tiered network where a tier of macro-cell wireless backhaul nodes is overlaid with a tier of cache-aided SCs. To frame a more realistic scenario, we consider both *static* and *dynamic* UT association policies. Building on random spatial models and using tools from stochastic geometry, we derive closed-form expressions and bounds for the achievable ASE for both UT association policies. Then, we study suitable optimization problems to identify deployment strategies for maximizing the ASE and EE of the network. Our findings show that the ASE is maximized by first increasing the SC density and then the storage size at the SCs, whereas an opposite approach should be adopted for maximizing the EE.

Appendix I
Proof of Theorem 1

Due to space limitations, we provide only a sketch of the proof. Considering static UT association, we have $\mathsf{P}_{\mathrm{suc}}^{(\mathrm{S})}(\theta) = \mathsf{P}_{\mathrm{hit}}\mathbb{P}[\mathsf{SINR}_{u(x)} > \theta] + (1 - \mathsf{P}_{\mathrm{hit}})\mathbb{P}[\mathsf{SINR}_{u(x)} > \theta]\mathbb{P}[\mathsf{SINR}_x > \theta]$ and, building on [8], [10], we obtain

$$\mathbb{P}[\mathsf{SINR}_{u(x)} > \theta] = \exp\big(-\theta\tfrac{\sigma^2}{\rho_{\mathrm{SC}}}R_{\mathrm{UT}}^{\alpha}\big)\mathcal{L}_{I_x}^{(\mathrm{S})}\big(\theta\rho_{\mathrm{SC}}^{-1}R_{\mathrm{UT}}^{\alpha}\big) \tag{20}$$

$$\mathbb{P}[\mathsf{SINR}_x > \theta] = \exp\big(-\theta\tfrac{\sigma^2}{\rho_{\mathrm{BH}}}R_{\mathrm{BH}}^{\alpha}\big)\mathcal{L}_{I_x}^{(\mathrm{S})}\big(\theta\rho_{\mathrm{BH}}^{-1}R_{\mathrm{BH}}^{\alpha}\big) \tag{21}$$

with $\mathcal{L}_{I_{u(x)}}^{(\mathrm{S})}(s)$ and $\mathcal{L}_{I_x}^{(\mathrm{S})}(s)$ defined in (9)–(10). Likewise, for dynamic UT association, we have $\mathsf{P}_{\mathrm{suc}}^{(\mathrm{D})}(\theta) = \mathbb{P}[\mathsf{SINR}_{u(x)} > \theta]$ where (cf. [8], [10])

$$\begin{aligned}\mathbb{P}[\mathsf{SINR}_{u(x)} > \theta] = {} & \exp\big(-\theta\tfrac{\sigma^2}{\rho_{\mathrm{SC}}}R_{\mathrm{UT}}^{\alpha}\big)\mathcal{L}_{I_{u(x)}}^{(\mathrm{S})}\big(\theta\rho_{\mathrm{SC}}^{-1}R_{\mathrm{UT}}^{\alpha}\big)\mathbb{1}_{\Delta_x} \\ & + \exp\big(-\theta\tfrac{\sigma^2}{\rho_{\mathrm{BH}}}R_{\mathrm{BH}}^{\alpha}\big)\mathcal{L}_{I_{u(x)}}^{(\mathrm{S})}\big(\theta\rho_{\mathrm{BH}}^{-1}R_{\mathrm{BH}}^{\alpha}\big)\mathbb{1}_{\bar{\Delta}_x}\end{aligned} \tag{22}$$

with $\mathcal{L}_{I_{u(x)}}^{(\mathrm{D})}(s)$ defined in (12). ■

References

[1] Cisco, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast, 2016–2021," *White Paper*, 2017.

[2] J. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. Soong, and J. Zhang, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[3] E. Baştuğ, M. Bennis, M. Kountouris, and M. Debbah, "Cache-enabled small cell networks: modeling and tradeoffs," *EURASIP Journal on Wireless Communic. and Netw.*, no. 1, pp. 41–51, 2015.

[4] K. Li, C. Yang, Z. Chen, and M. Tao, "Optimization and analysis of probabilistic caching in n-tier heterogeneous networks," Dec. 2016. [Online]. Available: https://arxiv.org/pdf/1612.04030.pdf

[5] D. Liu and C. Yang, "Caching policy toward maximal success probability and area spectral efficiency of cache-enabled HetNets," 2016. [Online]. Available: http://arxiv.org/pdf/1608.03749v1.pdf

[6] M. A. Abd-Elmagid, O. Ercetin, and T. ElBatt, "Cache-aided heterogeneous networks: Coverage and delay analysis," Jan. 2017. [Online]. Available: https://arxiv.org/pdf/1701.06735.pdf

[7] A. Khreishah, J. Chakareski, and A. Gharaibeh, "Joint caching, routing, and channel assignment for collaborative small-cell cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 8, pp. 2275–2284, Aug. 2016.

[8] M. Maso, I. Atzeni, I. Ghamnia, E. Baştuğ, and M. Debbah, "Cache-aided full-duplex small cells," in *Proc. Int. Symp. Modeling and Optimiz. in Mobile, Ad Hoc, and Wireless Netw. (WiOpt)*, Paris, France, May 2017, pp. 1–6. [Online]. Available: https://arxiv.org/pdf/1702.05064.pdf

[9] I. Atzeni and M. Kountouris, "Full-duplex MIMO small-cell networks with interference cancellation," Dec. 2016. [Online]. Available: https://arxiv.org/pdf/1612.07289v1.pdf

[10] 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.

[11] I. Atzeni, J. Arnau, and M. Kountouris, "Downlink cellular network analysis with LOS/NLOS propagation and elevated base stations," Mar. 2017. [Online]. Available: https://arxiv.org/pdf/1703.01279.pdf