This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2017.2724547, IEEE Transactions on Vehicular Technology

Cost-Effective Cache Deployment in Mobile Heterogeneous Networks

Shan Zhang, *Member, IEEE,* Ning Zhang, *Member, IEEE,* Peng Yang, *Student Member, IEEE,* and Xuemin (Sherman) Shen, *Fellow, IEEE*

***Abstract*—This paper investigates one of the fundamental issues in cache-enabled heterogeneous networks (HetNets): how many cache instances should be deployed at different base stations, in order to provide guaranteed service in a cost-effective manner. Specifically, we consider two-tier HetNets with hierarchical caching, where the most popular files are cached at small cell base stations (SBSs) while the less popular ones are cached at macro base stations (MBSs). For a given network cache deployment budget, the cache sizes for MBSs and SBSs are optimized to maximize network capacity while satisfying the file transmission rate requirements. As cache sizes of MBSs and SBSs affect the traffic load distribution, inter-tier traffic steering is also employed for load balancing. Based on stochastic geometry analysis, the optimal cache sizes for MBSs and SBSs are obtained, which are threshold-based with respect to cache budget in the networks constrained by SBS backhauls. Simulation results are provided to evaluate the proposed schemes and demonstrate the applications in cost-effective network deployment.**

***Index Terms*—mobile edge caching, heterogeneous networks, constrained backhaul, stochastic geometry**

I. Introduction

Heterogeneous networks (HetNets), consisting of macro base stations (MBSs) and ultra-densely deployed small cell base stations (SBSs), are envisioned as the dominant theme to meet the 1000× capacity enhancement in 5G networks and beyond [1], [2]. With network further densified, deploying ideal backhaul with unconstrained capacity for each small cell may be impractical, due the unacceptably high costs of deployment and operation [3], [4]. Thus, one of the key problems towards 5G is to reduce the required backhaul capacity while keeping the system capacity. Mobile edge caching provides a promising solution to address the problem, by exploiting the content information [5], [6]. As the requested content of mobile users, e.g., video, may show high similarity, caching popular contents at base stations can effectively alleviate the backhaul pressure and enhance network service capability [7]. Meanwhile, the delay performance can be significantly improved, with service demands accommodated locally.

Since the study on mobile edge caching is still nascent, many research issues need to be addressed, such as architecture design [8], content placement [9], [10] and update [11]. However, the caching deployment is overlooked in the existing literature. Specifically, the fundamental problem of cache deployment is to optimize the cache sizes of different BSs in HetNets, so as to minimize network deployment and operational costs while guaranteeing quality of service (QoS) performance. The basic tradeoff for cache deployment exists between caching efficiency and spectrum efficiency. On one hand, the contents cached at MBSs can serve more users due to the large cell coverage, providing high caching efficiency. On the other hand, the densely deployed SBS tier is more likely to be backhaul-constrained, as extensive spatial spectrum reuse introduces substantial access traffic. As a result, deploying more cache instances at SBSs can narrow the gap between backhaul and radio access capacities, and thus improve spectrum efficiency systematically. In this regard, cache instances should be deployed appropriately, such that network resources can be balanced and fully utilized [12], [13]. However, the cache deployment problem is challenging, as different cache size also influences the traffic load distributions across the network. For example, more traffic needs to be served by MBSs when the MBS cache size increases, changing the loads of both radio access and backhaul of MBS and SBS tiers. Therefore, load balancing should be also considered to avoid problems like service outage and resource under-utilization. To this end, traffic steering can be leveraged to tune load distribution, and jointly optimized with cache deployment [14].

In this paper, the cache deployment problem is investigated in two-tier HetNets, where each SBS caches the most popular files while each MBS caches the less popular ones (i.e., hierarchical caching). If cached at the associated MBSs or SBSs, the requested contents will be directly delivered to mobile users through radio access, i.e., content hit. Otherwise, the requested contents will be delivered through remote file fetching via backhaul connections, i.e., content miss. For a given cache deployment budget, we maximize network capacity while guaranteeing the average file transmission rates, by jointly optimizing the MBS/SBS cache sizes and the inter-tier traffic steering ratio of content miss users. However, the problem is of great challenge due to the transmission rate requirements. Specifically, file

Shan Zhang and Xuemin (Sherman) Shen are with the Department of Electrical and Computer Engineering, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada, N2L 3G1 (email: {s327zhan, sshen}@uwaterloo.ca).

Ning Zhang is with the Department of Computing Science, Texas A&M University-Corpus Christi, 6300 Ocean Dr., Corpus Christi, Texas, USA, 78412 (Email: zhangningbupt@gmail.com).

Peng Yang is with the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China (email: yangpeng@hust.edu.cn).

This work is sponsored by the Natural Sciences and Engineering Research Council of Canada (NSERC).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2017.2724547, IEEE Transactions on Vehicular Technology

transmission rates depend on both radio and backhaul access conditions, which should account for multi-randomness of traffic load, user location, channel fading and network topology. Through stochastic geometry analysis, the lower bound of average file transmission rates are derived in closed form, based on which the cache deployment problem is simplified and numerical results can be obtained. To offer insights into practical network design, we then focus on the scenario when the MBSs have sufficiently large backhaul capacity while the SBS tier is backhaul constrained. The optimal cache deployment is obtained, which is threshold-based with respect to the network cache budget. When the cache budget is smaller than certain threshold, all the cache instances should be deployed at SBSs to maximize network capacity. When the cache budget exceeds the threshold, the cache deployment problem has multiple optimal solutions to achieve maximal network capacity, and we find the one which can simultaneously maximize content hit rate. In fact, cache budget threshold can be interpreted as the deficiency of SBS backhaul, i.e., the minimal cache budget required to match the backhaul and radio resources. Moreover, the threshold characterizes the trading relationship between backhaul and cache capacities, which can be applied to cost-effective network deployment.

The contributions of this paper are summarized as follows:

1) The average file transmission rates in large-scale cache-enabled HetNets are analyzed theoretically, considering the constraints of both backhaul capacities and radio resources;
2) The cache deployment is optimized in HetNets, which maximizes QoS-guaranteed network capacity with the given cache budget;
3) The inter-tier traffic steering is jointly optimized to balance the loads of MBS and SBS tiers, considering the influence of cache deployment on traffic distributions;
4) The proposed method can provide the cost-optimal combination of backhaul and radio resource provisioning, which can be applied to practical cache-enabled HetNet deployment.

The remaining of this paper is organized as follows. Firstly, related work on mobile edge caching is reviewed in Section II. Then, the system model is presented in Section III, and the cache deployment problem is formulated in Section IV. In Section V, the QoS-constrained network capacity is obtained, based on which the optimal cache deployment is analyzed in Section VI. The analytical results are validated through extensive simulations in Section VII, followed by the cost-effective network deployment illustrations with numerical results. Finally, Section VIII summarizes the work and discusses future research topics.

II. Literature Review

Content caching at mobile edge networks is considered as a promising solution to cope with the mismatch between explosive mobile video traffic and limited backhaul/wireless capacity, which has drawn increasing attention recently. Cache-enabled 5G network architectures have been designed in [15], [16], which were shown to have a great potential to reduce mobile traffic through trace-driven simulations. The performance of cache-enabled networks has also been analyzed theoretically, which was demonstrated to be more spectrum-efficient compared with the conventional HetNets in backhaul-constrained cases [12]. Meanwhile, effective cache placement schemes have been devised with respect to different optimization objectives, such as maximizing content hit rate [17], [18], [19], reducing file downloading delay [20], [21], [22], [23], enhancing user quality of experience (QoE) [24], improving mobility support [25], [26], and minimizing specific cost functions [27], [28].

Although the existing cache placement schemes were designed based on the predefined cache size for each BS, studies on cache deployment were quite limited. In the very recent work [29], the storage costs of different network entities (like remote servers, gateways, and BSs) have been considered, and a multi-layered cache deployment scheme was proposed to maximize the ratio of content hit rate to storage cost. The performances of BS-caching and gateway-caching have been compared in [30], based on which the cache deployment was optimized to achieve Pareto optimal spectrum efficiency and content hit rate. The BS cache sizes are optimized to maximize the minimal user success probability, under the constraints of backhaul capacity and cache deployment budget [31]. Insightful as it is, the algorithm in [31] mainly focused on small-scale networks. Different from existing work, this paper investigates the cache deployment problem in large-scale HetNets for the first time, aiming at maximizing network capacity while meeting the QoS requirements in terms of file transmission rate. Meanwhile, the cache sizes of different BSs are jointly optimized with inter-tier traffic steering. The analytical results have taken into account the multi-randomness of network topology, traffic distribution and channel fading, which can provide a guideline for practical network design with mobile edge caching.

III. System Model

In this section, we present system model and the hierarchical caching framework, with important notations summarized in Table I.

A. Cache-Enabled Heterogeneous Network Architecture

In 5G HetNets, MBSs are responsible for network coverage with control signaling, whereas SBSs are expected to be densely deployed to boost network capacity in a "plug-and-play" manner. The topology of MBSs are modeled as regular hexagonal cells with density ρ_{m}, while the distribution of SBSs are modeled as Poisson Point Process (PPP) of density ρ_{s}. MBSs and SBSs use orthogonal spectrum bands to avoid inter-tier interference, and the spectrum reuse factor within each tier is set to be 1. Denote by W_{m} and W_{s} the bandwidths available to each MBS and SBS, respectively. Both MBSs and SBSs are connected with

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2017.2724547, IEEE Transactions on Vehicular Technology

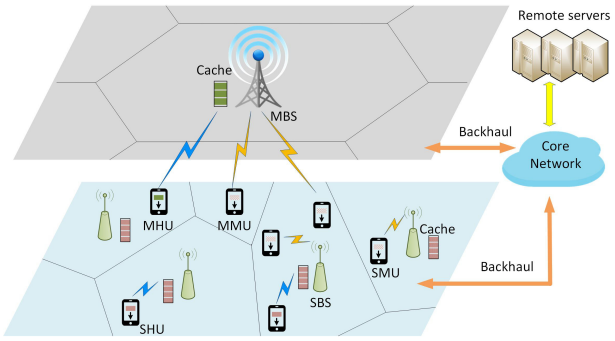


Fig. 1: Cache-enabled heterogeneous networks.

TABLE I: Important notations

$\rho_{\rm m}$	MBS density	$\rho_{\rm s}$	SBS density
$W_{\rm m}$	MBS bandwidth	$W_{\rm s}$	SBS bandwidth
$U_{\rm MBH}$	MBS backhaul capacity	$U_{\rm SBH}$	SBS backhaul capacity
F	number of files	q_f	popularity of file-f
$C_{\rm m}$	MBS cache size	$C_{\rm s}$	SBS cache size
$P_{\rm Hit}^{\rm (m)}$	MBS content hit rate	$P_{\rm Hit}^{\rm (s)}$	SBS content hit rate
C	network cache budget	$P_{\rm Hit}$	aggregated content hit rate
λ	traffic density	σ^2	noise power
φ	traffic steering ratio	ξ_c	cache deployment ratio
$\check{R}_{\rm RAN}$	required rate at RAN	$\check{R}_{\rm BH}$	required rate at backhaul

core network through wired backhauls, with capacities denoted as $U_{\rm MBH}$ and $U_{\rm SBH}$, respectively.

The distribution of active users is modeled as a PPP of density λ, independent of the location of MBSs and SBSs. The service process is illustrated as Fig. 1. Each user keeps dual connectivity with an MBS and an SBS [32], [33], [34], where the MBS (and SBS) which provides the highest average intra-tier signal to interference ratio (SINR) is selected for association. The coverage area of each MBS is a hexagonal cell with side length of $D_{\rm m}$ ($\frac{3\sqrt{3}}{2}{D_{\rm m}}^2=\frac{1}{\rho_{\rm m}}$), and small cells form the Voronoi tessellation, as shown in Fig. 1. As for the service process, mobile users can be directly served through radio access networks if the required files are cached at the MBS or SBS tiers (i.e., content-hit users). Instead, content-miss users will randomly choose the associated SBS or MBS with probability φ and $1-\varphi$, and the chosen SBS/MBS needs to fetch the required file from remote servers via backhaul. Define φ as the inter-tier *traffic steering ratio*, which influences the load of MBS and SBS tiers.

B. Hierarchical Caching

Denote by $\mathcal{F}=\{1,2,...,f,...F\}$ the set of files that may be requested, and denote by $\mathcal{Q}=\{q_1,q_2,...,q_f,...,q_F\}$ the popularity distribution ($\sum_{f=1}^{F}q_f=1,\ q_f>0$ for $f=1,2,...,F$). Without loss of generality, we assume the files are sorted with descending popularity ($q_f\geq q_{f+1}$ for $f=1,2,...,F-1$) and have the same size of L[1]. A hierarchical content caching framework is adopted, where the SBS tier caches the most popular files while the MBS tier caches the less popular ones to increase content diversity.

[1]If files have different sizes, they can be divided into the same size to conduct analysis.

Denote by $C_{\rm m}$ and $C_{\rm s}$ the cache sizes of each MBS and SBS, respectively. Thus, files $\{1,2,...,C_{\rm s}\}$ are cached at each SBS, and files $\{C_{\rm s+1},C_{\rm s+2},...,C_{\rm s}+C_{\rm m}\}$ are cached at each MBS. Then, the content hit rates of MBS and SBS tiers, $P_{\rm Hit}^{\rm (m)}$ and $P_{\rm Hit}^{\rm (s)}$, are given by

$$P_{\rm Hit}^{\rm (m)}=\sum_{C_{\rm s}+1}^{C_{\rm m}+C_{\rm s}}q_f,\qquad P_{\rm Hit}^{\rm (s)}=\sum_{f=1}^{C_{\rm s}}q_f, \tag{1}$$

and total content hit rate is

$$P_{\rm Hit}=P_{\rm Hit}^{\rm (m)}+P_{\rm Hit}^{\rm (s)}=\sum_{f=1}^{C_{\rm m}+C_{\rm s}}q_f. \tag{2}$$

With the dual connectivity, the equivalent cache size for each mobile user is $C_{\rm m}+C_{\rm s}$ according to Eq. (2)[2]. Thus, caching the most popular $C_{\rm m}+C_{\rm s}$ files can maximize content hit rate, since each mobile user can be served only by the associated MBS or SBS with no intra-tier BS cooperation. In addition, caching more contents at SBSs instead of MBSs can steer more users to the SBS-tier from MBSs. As SBSs are more densely deployed than MBSs in practical networks, steering traffic to the SBS tier can fully utilize rich radio resources with inter-tier load balancing.

Define C the network caching budget, i.e., the number of files cached per unit area:

$$C=\rho_{\rm m}C_{\rm m}+\rho_{\rm s}C_{\rm s}. \tag{3}$$

Cache deployment determines $C_{\rm m}$ and $C_{\rm s}$ to optimize network performance, for the given network caching budget C.

C. File Transmission Rate

With hierarchical caching, users can be classified into four types: (1) MHU (MBS-hit-users), served by the MBS tier with cached contents; (2) SHU (SBS-hit-users), served by the SBS tier with cached contents; (3) MMU (MBS-missed-users), served by the MBS tier through backhaul file fetching; and (4) SMU (SBS-missed-users), served by the SBS tier through backhaul file fetching. Based on the properties of PPP, the four types of users also follow independent PPPs, with densities of $P_{\rm Hit}^{\rm (m)}\lambda$, $P_{\rm Hit}^{\rm (s)}\lambda$, $(1-P_{\rm Hit})(1-\varphi)\lambda$, and $(1-P_{\rm Hit})\varphi\lambda$, respectively. The file transmission rates of MHUs and SHUs only depend on the radio access (i.e., wireless part), whereas the rates of MMUs and SMUs are also constrained by the limited backhaul capacities.

Consider a typical mobile user-u. If user-u is served by the MBS tier, the achievable rate for radio access is given by

$$R_{\rm MR}=\frac{W_{\rm m}}{N_{\rm MR}+1}\log_2\left(1+\gamma_{\rm m}\right), \tag{4}$$

where $N_{\rm MR}$ denotes the number of residual users being served by the associated MBS except user-u (both MHUs and MMUs included), $\gamma_{\rm m}$ is the received SINR given by

$$\gamma_{\rm m}=\min\left(\gamma_{\max},\frac{P_{\rm TM}h_{\rm m}{d_{\rm m}}^{-\alpha_{\rm m}}}{\sigma^2+I_{\rm m}}\right), \tag{5}$$

[2]"Eq." is short for Equation, and "Eqn." is short for inequation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2017.2724547, IEEE Transactions on Vehicular Technology

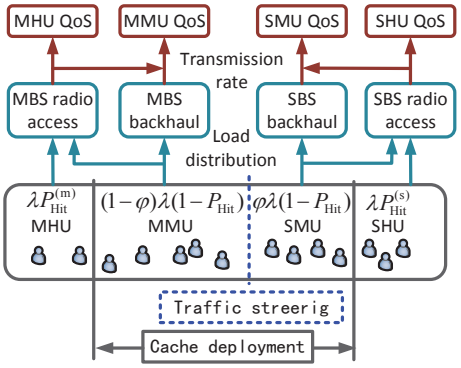


Fig. 2: Influence of cache deployment and traffic steering on QoS performance.

$\gamma_{\max}$ is the maximal received SINR, P_{TM} is the MBS transmit power, h_{m} is an exponential random variable with mean 1 incorporating the effect of Rayleigh fading, α_{m} is the path loss exponent of the MBS-tier, d_{m} denotes the distance from user-u to the associated MBS, σ^2 is the addictive noise power, and I_{m} represents inter-cell interference from other MBSs. In practical systems, N_{MR} varies randomly with the dynamic arrival and departure of file transmission demands, and d_{m} is also uncertain from the network perspective.

If user-u is served by the SBS tier, the achievable rate for radio access can be given by

$$R_{\text{SR}} = \frac{W_{\text{s}}}{N_{\text{SR}}+1}\log_2\left(1+\gamma_{\text{s}}\right), \tag{6}$$

where N_{SR} denotes the number of residual users being served by the associated SBS except user-u (including both SHUs and SMUs),

$$\gamma_{\text{s}} = \min\left(\gamma_{\max}, \frac{P_{\text{TS}}h_{\text{s}}{d_{\text{s}}}^{-\alpha_{\text{s}}}}{\sigma^2+I_{\text{s}}}\right), \tag{7}$$

P_{TS} is the SBS transmit power, h_{s} is an exponential random variable with mean 1 incorporating the effect of Rayleigh fading, α_{s} is the path loss exponent of the SBS-tier, d_{s} denotes the distance from user-u to the associated SBS, and I_{s} represents the inter-cell interference from other SBSs. Similarly, N_{SR} and d_{s} are also random variables. In addition, the probability distribution functions (PDFs) of N_{SR} and d_{s} can be more complex due to the uncertain small cell sizes.

The MBS and SBS backhaul transmission rates only depend on the corresponding traffic loads and capacities:

$$R_{\text{MBH}} = \frac{U_{\text{MBH}}}{N_{\text{MBH}}+1}, \quad R_{\text{SBH}} = \frac{U_{\text{SBH}}}{N_{\text{SBH}}+1}, \tag{8}$$

where N_{MBH} (or N_{SBH}) represents the number of residual MMUs (or SMUs) sharing the MBS (or SBS) backhaul expect the considered user-u.

IV. Capacity-Optimal Caching Formulation

To meet QoS requirements, the transmission rates should to be guaranteed for successful file delivery, which depend on the transmission rates of each network part (i.e., MBS radio access, MBS backhaul, SBS radio access and SBS backhaul), as shown in Fig. 2. In addition, cache deployment (i.e., cache sizes $[C_{\text{m}}, C_{\text{s}}]$) determines the traffic distributions across the network together with traffic steering ratio φ, thus influencing the transmission rates. Therefore, cache deployment should be jointly optimized with traffic steering, which can be formulated as follows:

$$\begin{aligned}
\max_{C_{\text{s}},\varphi}\quad & \mu(C_{\text{s}},\varphi) && \text{(9a)}\\
\text{(P1)}\quad \text{s.t.}\quad & \mathbb{E}\left[R_{\text{MR}}\right] \geq \check{R}_{\text{RAN}}, && \text{(9b)}\\
& \mathbb{E}\left[R_{\text{MBH}}\right] \geq \check{R}_{\text{BH}}, && \text{(9c)}\\
& \mathbb{E}\left[R_{\text{SR}}\right] \geq \check{R}_{\text{RAN}}, && \text{(9d)}\\
& \mathbb{E}\left[R_{\text{SBH}}\right] \geq \check{R}_{\text{BH}}, && \text{(9e)}\\
& 0 \leq C_{\text{s}} \leq C/\rho_{\text{s}}, \quad 0 \leq \varphi \leq 1, && \text{(9f)}
\end{aligned}$$

where the objective function $\mu(C_{\text{s}},\varphi) = \max_{\lambda}\left(\lambda|_{\{C_{\text{s}},\varphi\}}\right)$ is the network capacity for the given SBS cache size C_{s} and traffic steering ratio φ (i.e., the maximal traffic density that can be catered), $\check{R}_{\text{RAN}}$ and $\check{R}_{\text{BH}}$ denote the per user rate requirements for radio access and backhaul transmissions[3], respectively. The cache size of MBSs C_{m} can be determined with C_{s} according to Eq. (3). Constraint (9b) and (9d) guarantee the QoS requirements of content hit users, while (9c) and (9e) further account for the file fetching delay requirements of content miss users. The average transmission rate is adopted for QoS guarantee as the services suitable for pro-active caching are mostly elastic in practical networks, such as popular video streaming. Furthermore, when the average per user transmission rate is guaranteed, the objective function $\mu(C_{\text{s}},\varphi)$ can also reflect the network goodput.

According to the properties of PPP, the traffic distributions of each network part also follow PPP. Denote by λ_{MR}, λ_{MBH}, λ_{SR} and λ_{SBH} the equivalent user density for MBS radio access, MBS backhaul, SBS radio access, and SBS backhaul:

$$\begin{aligned}
\lambda_{\text{MR}} &= \left[P_{\text{Hit}}^{(\text{m})} + (1-P_{\text{Hit}})(1-\varphi)\right]\lambda, && \text{(10a)}\\
\lambda_{\text{MBH}} &= (1-P_{\text{Hit}})(1-\varphi)\lambda, && \text{(10b)}\\
\lambda_{\text{SR}} &= \left[P_{\text{Hit}}^{(\text{s})} + (1-P_{\text{Hit}})\varphi\right]\lambda, && \text{(10c)}\\
\lambda_{\text{SBH}} &= (1-P_{\text{Hit}})\varphi\lambda. && \text{(10d)}
\end{aligned}$$

The constraints (9b)-(9e) provide the maximal value of λ_{MR}, λ_{MBH}, λ_{SR}, and λ_{SBH}, denoted by $\hat{\lambda}_{\text{MR}}$, $\hat{\lambda}_{\text{MBH}}$, $\hat{\lambda}_{\text{SR}}$, and $\hat{\lambda}_{\text{SBH}}$, respectively. In addition, $\hat{\lambda}_{\text{MR}}$, $\hat{\lambda}_{\text{MBH}}$, $\hat{\lambda}_{\text{SR}}$, and $\hat{\lambda}_{\text{SBH}}$ further constrains the traffic arrival rate λ with Eq. (10), for the given C_{s} and φ. Thus, the network capacity depends on the bottleneck:

$$\mu(C_{\text{s}},\varphi) = \min\left(\frac{\hat{\lambda}_{\text{MR}}}{P_{\text{Hit}}^{(\text{m})}+(1-P_{\text{Hit}})(1-\varphi)}, \frac{\hat{\lambda}_{\text{MBH}}}{(1-P_{\text{Hit}})(1-\varphi)}, \frac{\hat{\lambda}_{\text{SR}}}{P_{\text{Hit}}^{(\text{s})}+(1-P_{\text{Hit}})\varphi}, \frac{\hat{\lambda}_{\text{SBH}}}{(1-P_{\text{Hit}})\varphi}\right). \tag{11}$$

[3]In practical systems, the backhaul rate requirement $\check{R}_{\text{BH}}$ is generally much higher than that of the radio access part $\check{R}_{\text{RAN}}$, considering the end-to-end delay and the serial transmission structure.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2017.2724547, IEEE Transactions on Vehicular Technology

The key issue is the transmission rates analysis, which will be addressed in the next section.

V. QoS-Constrained Capacity Analysis

In this section, the file transmission rates of different networks parts are analyzed respectively, based on which the constraints (9b)-(9e) can be simplified with respect to λ.

A. MBS backhaul

N_{MBH} follows Poisson distribution of mean $\lambda_{\mathrm{MBH}}/\rho_{\mathrm{m}}$, according to Slivnyak-Mecke theorem [35]. Thus, based on Eq. (8), the average file transmission rate of MBS backhaul can be derived:

$$\begin{aligned}\mathbb{E}[R_{\mathrm{MBH}}] &= \sum_{n=0}^{\infty} \frac{U_{\mathrm{MBH}}}{n+1} \Pr\{N_{\mathrm{MBH}} = n\} \\ &= \sum_{n=0}^{\infty} \frac{U_{\mathrm{MBH}}}{n+1} \frac{\left(\frac{\lambda_{\mathrm{MBH}}}{\rho_{\mathrm{m}}}\right)^n}{n!} e^{-\frac{\lambda_{\mathrm{MBH}}}{\rho_{\mathrm{m}}}} = \frac{U_{\mathrm{MBH}}\rho_{\mathrm{m}}}{\lambda_{\mathrm{MBH}}}\left(1 - e^{-\frac{\lambda_{\mathrm{MBH}}}{\rho_{\mathrm{m}}}}\right).\end{aligned} \tag{12}$$

Combining Eq. (12) with Eq. (10b), the SBS backhaul constraint Eqn. (9c) can be simplified with respect to traffic density λ. Denote by $\hat{\lambda}_{\mathrm{MBH}} = \max\left\{\lambda_{\mathrm{MBH}} | \mathbb{E}[R_{\mathrm{MBH}}] \geq \check{R}_{\mathrm{BH}}\right\}$, the maximal traffic load on MBS backhaul. As the average rate $\mathbb{E}[R_{\mathrm{MBH}}]$ decreases with λ_{MBH}, $\hat{\lambda}_{\mathrm{MBH}}$ satisfies

$$\frac{\rho_{\mathrm{m}}}{\hat{\lambda}_{\mathrm{MBH}}}\left(1 - e^{-\frac{\hat{\lambda}_{\mathrm{MBH}}}{\rho_{\mathrm{m}}}}\right) = \frac{\check{R}_{\mathrm{BH}}}{U_{\mathrm{MBH}}}, \tag{13}$$

according to Eq. (12).

B. SBS backhaul

Compared with MBS backhaul, the transmission rate of SBS backhaul is more complex due to the random small cell size. Denote by A_{s} the cell area size, which follows Gamma distribution with shape $\kappa = 3.575$ and scale $1/\kappa\rho_{\mathrm{s}}$ [36]. Thus, the PDF of A_{s} is given by

$$f_{A_{\mathrm{s}}}(A) = A^{\kappa-1} e^{-\kappa\rho_{\mathrm{s}} A} \frac{(\kappa\rho_{\mathrm{s}})^{\kappa}}{\Gamma(\kappa)}, \tag{14}$$

where $\Gamma(\cdot)$ is the gamma function. Furthermore, the number of SMUs served through SBS backhaul follows Poisson distribution of mean $\lambda_{\mathrm{SBH}} A$ given the cell size $A_{\mathrm{s}} = A$. Thus, based on Eq. (8), the average transmission rate can be derived:

$$\begin{aligned}\mathbb{E}[R_{\mathrm{SBH}}] &= \int_{A=0}^{\infty} \left(\sum_{n=0}^{\infty} \frac{U_{\mathrm{SBH}}}{n+1} \Pr\{N_{\mathrm{SBH}} = n | A\}\right) f_{A_{\mathrm{s}}}(A) \mathrm{d}A \\ &= \int_{A=0}^{\infty} \frac{U_{\mathrm{SBH}}}{\lambda_{\mathrm{SBH}} A}\left(1 - e^{-\lambda_{\mathrm{SBH}} A}\right) A^{\kappa-1} e^{-\kappa\rho_{\mathrm{s}} A} \frac{(\kappa\rho_{\mathrm{s}})^{\kappa}}{\Gamma(\kappa)} \mathrm{d}A \\ &= \frac{U_{\mathrm{SBH}}}{\lambda_{\mathrm{SBH}}} \left\{ \int_{A=0}^{\infty} A^{\kappa-2} e^{-\kappa\rho_{\mathrm{s}} A} \frac{(\kappa\rho_{\mathrm{s}})^{\kappa}}{\Gamma(\kappa)} \mathrm{d}A \right. \\ &\quad \left. - \int_{A=0}^{\infty} A^{\kappa-2} e^{-(\kappa\rho_{\mathrm{s}} + \lambda_{\mathrm{SBH}}) A} \frac{(\kappa\rho_{\mathrm{s}})^{\kappa}}{\Gamma(\kappa)} \mathrm{d}A \right\} \\ &= \frac{U_{\mathrm{SBH}}\kappa\rho_{\mathrm{s}}}{\lambda_{\mathrm{SBH}}} \frac{\Gamma(\kappa-1)}{\Gamma(\kappa)} \left(1 - \frac{1}{\left(1 + \frac{\lambda_{\mathrm{SBH}}}{\kappa\rho_{\mathrm{s}}}\right)^{\kappa-1}}\right).\end{aligned} \tag{15}$$

Combining Eq. (15) with Eq. (10d), the SBS backhaul constraint Eqn. (9e) can be simplified. According to Eq. (15), the maximal traffic load on SBS backhaul $\hat{\lambda}_{\mathrm{SBH}}$ can be given by

$$\frac{\kappa\rho_{\mathrm{s}}}{\hat{\lambda}_{\mathrm{SBH}}} \frac{\Gamma(\kappa-1)}{\Gamma(\kappa)} \left(1 - \frac{1}{\left(1 + \frac{\hat{\lambda}_{\mathrm{SBH}}}{\kappa\rho_{\mathrm{s}}}\right)^{\kappa-1}}\right) = \frac{\check{R}_{\mathrm{BH}}}{U_{\mathrm{SBH}}}. \tag{16}$$

C. MBS Radio Access

According to Eq. (4), the transmission rate of MBS radio access can be given by

$$\mathbb{E}[R_{\mathrm{MR}}] = \mathop{\mathbb{E}}_{\{N_{\mathrm{MR}}, d_{\mathrm{m}}\}} \left[\frac{W_{\mathrm{m}}}{1 + N_{\mathrm{MR}}} \log_2(1 + \gamma_{\mathrm{m}})\right] \tag{17}$$

where the user number N_{MR} follows Poisson distribution of mean $\lambda_{\mathrm{MR}}/\rho_{\mathrm{m}}$:

$$p_{N_{\mathrm{MR}}}(n) = \frac{(\frac{\lambda_{\mathrm{MR}}}{\rho_{\mathrm{m}}})^n}{n!} e^{-\frac{\lambda_{\mathrm{MR}}}{\rho_{\mathrm{m}}}}, \tag{18}$$

and the communication distance d_{m} can be considered to follow:

$$f_{d_{\mathrm{m}}}(d) = \frac{2d}{{D_{\mathrm{m}}}^2}, \tag{19}$$

by approximating MBS coverage as a circle of radius D_{m}. Then, the lower bound of average transmission rate for MBS radio access can be obtained by approximating the random inter-cell interference with the average value, which can be quite accurate under the condition of high signal-to-noise ratio (SNR) [37].

Lemma 1. The lower bound of average transmission rate of MBS radio access is given by:

$$\mathbb{E}[R_{\mathrm{MR}}] \geq \frac{\tau_{\mathrm{m}} W_{\mathrm{m}}}{\bar{N}_{\mathrm{MR}}}, \tag{20}$$

where

$$\begin{aligned}\tau_{\mathrm{m}} &= \log_2 \frac{P_{\mathrm{TM}} D_{\mathrm{m}}^{-\alpha_{\mathrm{m}}}}{(1 + \theta_{\mathrm{m}})\sigma^2} + \frac{\alpha_{\mathrm{m}}}{2\ln 2}\left(1 - \frac{D_{\min}^2}{D_{\mathrm{m}}^2}\right), \\ \bar{N}_{\mathrm{MR}} &= \frac{\lambda_{\mathrm{MR}}}{\rho_{\mathrm{m}}}\left(1 - e^{-\frac{\lambda_{\mathrm{MR}}}{\rho_{\mathrm{m}}}}\right)^{-1},\end{aligned} \tag{21}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2017.2724547, IEEE Transactions on Vehicular Technology

$D_{\min}$ is the transmission distance corresponding to the maximal received SINR (i.e., $\frac{P_{\mathrm{TM}} D_{\min}^{-\alpha_{\mathrm{m}}}}{(1+\theta_{\mathrm{m}})\sigma^2} = \gamma_{\max}$), and θ_{m} denotes the ratio of average inter-cell interference to noise (i.e., $\theta_{\mathrm{m}}\sigma^2 = \mathbb{E}[I_{\mathrm{m}}]$). The equality of Eqn. (20) holds when $\frac{\sigma^2}{P_{\mathrm{TM}}} \to 0$.

Proof: Please refer to Appendix A. ∎

Remark: The physical meaning of τ_{m} is the average spectrum efficiency of MBS tier, and $\bar{N}_{\mathrm{MR}}$ reflects the average number of users accessing each MBS. In practical cellular networks, the received SINR is usually guaranteed to be high enough for reliable communications, through methods like inter-cell interference control. Therefore, Lemma 1 can be applied to approximate average data rate, and constraint Eqn. (9b) can be simplified with respect to traffic load λ based on Eq. (10a). In addition, the maximal traffic load on MBS radio access $\hat{\lambda}_{\mathrm{MR}}$ can be given by:

$$\frac{\rho_{\mathrm{m}}}{\hat{\lambda}_{\mathrm{MR}}}\left(1 - e^{-\frac{\hat{\lambda}_{\mathrm{MR}}}{\rho_{\mathrm{m}}}}\right) = \frac{\check{R}_{\mathrm{RAN}}}{\tau_{\mathrm{m}} W_{\mathrm{m}}}. \tag{22}$$

D. SBS Radio Access

According to Eq. (6), the average transmission rate for SBS radio access is given by

$$\mathbb{E}[R_{\mathrm{SR}}] = \mathop{\mathbb{E}}_{\{A_{\mathrm{s}}, N_{\mathrm{SR}}, d_{\mathrm{s}}\}}\left[\frac{W_{\mathrm{s}}}{1+N_{\mathrm{SR}}}\log_2(1+\gamma_{\mathrm{s}})\right]. \tag{23}$$

The accurate average transmission rate cannot be derived in closed form, due to the random SBS topology and user location. Similarly, the lower bound of average transmission rate can be obtained by approximating the random inter-cell interference by the average value, given by Lemma 2.

Lemma 2. The lower bound of average transmission rate of SBS radio access is given by:

$$\mathbb{E}[R_{\mathrm{SR}}] \geq \frac{\tau_{\mathrm{s}} W_{\mathrm{s}}}{\bar{N}_{\mathrm{SR}}}, \tag{24}$$

where

$$\begin{aligned}
\tau_{\mathrm{s}} &= \log_2 \frac{P_{\mathrm{TS}}(\pi\rho_{\mathrm{s}})^{\frac{\alpha_{\mathrm{s}}}{2}}}{(1+\theta_{\mathrm{s}})\sigma^2} + \frac{\alpha_{\mathrm{s}}}{2\ln 2}\gamma, \\
\bar{N}_{\mathrm{SR}} &= \frac{\lambda_{\mathrm{SR}}}{\kappa\rho_{\mathrm{s}}}\frac{\Gamma(\kappa)}{\Gamma(\kappa-1)}\left[1 - \frac{1}{\left(1+\frac{\lambda_{\mathrm{SR}}}{\kappa\rho_{\mathrm{s}}}\right)^{\kappa-1}}\right]^{-1},
\end{aligned} \tag{25}$$

$\gamma \approx 0.577$ is Euler-Mascheroni constant, and θ_{s} denotes the ratio of average inter-cell interference to noise at SBS tier. The equality of Eqn. (24) holds when $\frac{\sigma^2}{P_{\mathrm{TS}}} \to 0$.

Proof: Please refer to Appendix B. ∎

Remark: τ_{s} can be interpreted as the average spectrum efficiency of SBS tier, and $\bar{N}_{\mathrm{SR}}$ reflects the average number of users accessing each SBS. The constraint Eqn. (9d) can be simplified by combing Lemma 2 with Eq. (10c). In addition, the maximal traffic load on SBS radio access $\hat{\lambda}_{\mathrm{SR}}$ can be given by

$$\frac{\kappa\rho_{\mathrm{s}}}{\hat{\lambda}_{\mathrm{SR}}}\frac{\Gamma(\kappa-1)}{\Gamma(\kappa)}\left[1 - \frac{1}{\left(1+\frac{\hat{\lambda}_{\mathrm{SR}}}{\kappa\rho_{\mathrm{s}}}\right)^{\kappa-1}}\right] = \frac{\check{R}_{\mathrm{RAN}}}{\tau_{\mathrm{s}} W_{\mathrm{s}}}, \tag{26}$$

VI. Capacity-Optimal Hierarchical Caching

Based on the transmission rates analysis, problem (P1) can be simplified as follows:

$$\begin{aligned}
\max_{C_{\mathrm{s}},\varphi} \quad & \mu(C_{\mathrm{s}},\varphi) && \text{(27a)}\\
\text{(P2)} \quad \text{s.t.} \quad & \left[P_{\mathrm{Hit}}^{(\mathrm{m})} + (1-P_{\mathrm{Hit}})(1-\varphi)\right]\lambda \leq \hat{\lambda}_{\mathrm{MR}}, && \text{(27b)}\\
& (1-P_{\mathrm{Hit}})(1-\varphi)\lambda \leq \hat{\lambda}_{\mathrm{MBH}}, && \text{(27c)}\\
& \left[P_{\mathrm{Hit}}^{(\mathrm{s})} + (1-P_{\mathrm{Hit}})\varphi\right]\lambda \leq \hat{\lambda}_{\mathrm{SR}}, && \text{(27d)}\\
& (1-P_{\mathrm{Hit}})\varphi\lambda \leq \hat{\lambda}_{\mathrm{SBH}}, && \text{(27e)}\\
& 0 \leq C_{\mathrm{s}} \leq C/\rho_{\mathrm{s}}, \quad 0 \leq \varphi \leq 1, && \text{(27f)}
\end{aligned}$$

where $\hat{\lambda}_{\mathrm{MR}}$, $\hat{\lambda}_{\mathrm{MBH}}$, $\hat{\lambda}_{\mathrm{SR}}$, and $\hat{\lambda}_{\mathrm{SBH}}$ are given by Eqs. (22, 13, 26, and 16), while the content hit rate $P_{\mathrm{Hit}}^{(\mathrm{m})}$, $P_{\mathrm{Hit}}^{(\mathrm{s})}$ and P_{Hit} can be derived by Eqs. (1) and (2) with respect to different caching deployment $[C_{\mathrm{s}}, C_{\mathrm{m}}]$. Although different cache deployments can influence the traffic load distribution, problem (P2) differs significantly from the conventional load balancing problems. The total traffic load remains constant in load balancing problems, where the traffic load is shifted from one part to another. Instead, different cache deployments may change backhaul loads, as the content hit rate varies with cache sizes.

A. Problem Analysis and Solutions

Denote by

$$\mu_{\mathrm{MR}} = \frac{\hat{\lambda}_{\mathrm{MR}}}{P_{\mathrm{Hit}}^{(\mathrm{m})} + (1-P_{\mathrm{Hit}})(1-\varphi)}, \tag{28a}$$

$$\mu_{\mathrm{MBH}} = \frac{\hat{\lambda}_{\mathrm{MBH}}}{(1-P_{\mathrm{Hit}})(1-\varphi)}, \tag{28b}$$

$$\mu_{\mathrm{SR}} = \frac{\hat{\lambda}_{\mathrm{SR}}}{P_{\mathrm{Hit}}^{(\mathrm{s})} + (1-P_{\mathrm{Hit}})\varphi}, \tag{28c}$$

$$\mu_{\mathrm{SBH}} = \frac{\hat{\lambda}_{\mathrm{SBH}}}{(1-P_{\mathrm{Hit}})\varphi}, \tag{28d}$$

the maximal network traffic density constrained by the corresponding network part, and the network capacity is given by

$$\mu(C_{\mathrm{s}},\varphi) = \min\left\{\mu_{\mathrm{MR}}, \mu_{\mathrm{MBH}}, \mu_{\mathrm{SR}}, \mu_{\mathrm{SBH}}\right\}. \tag{29}$$

Numerical results of problem (P2) can be obtained by exhaustive search based on Eqs. (13, 16, 22, 26, 28 and 29). Furthermore, low-complexity heuristic algorithms can be designed. To maximize network capacity, the traffic loads of each network part should be balanced according to the corresponding service capabilities. Notice that the traffic load distribution can be manipulated by adjusting cache deployment strategy or traffic steering ratio. Specifically, Table II gives the variations of traffic load distribution with respect to cache size and traffic steering ratio, with proof provided in Appendix C. Table II can provide a guideline to enhance network capacity in practical networks. For example, when the MBS radio access is the performance bottleneck (i.e., $\mu_{\mathrm{MR}} < \mu_{\mathrm{MBH}}$, $\mu_{\mathrm{MR}} < \mu_{\mathrm{SR}}$, $\mu_{\mathrm{MR}} < \mu_{\mathrm{SBH}}$),

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2017.2724547, IEEE Transactions on Vehicular Technology

TABLE II: Capacity variations by increasing cache size or steering ratio

	μ_{MR}	μ_{MBH}	μ_{SR}	μ_{SBH}
C_{s}	Increase	Decrease	Decrease	Decrease
φ	Increase	Increase	Decrease	Decrease

we can either reduce cache size at MBSs, or increase the traffic steering ratio. Instead, when the SBS backhaul is limited, we can either reduce the SBS cache size or lower the traffic steering ratio.

B. SBS-Backhaul-Constrained HetNets

In practical systems, MBSs are expected to be equipped with optical fiber backhauls which can provide sufficiently large bandwidth, whereas the capacity for radio access can be much smaller due to spectrum resource scarcity. In this case, problem (P2) can be further simplified by removing constraint (27c). The condition of ideal MBS backhaul is $\hat{\lambda}_{\mathrm{MR}} \leq \hat{\lambda}_{\mathrm{MBH}}$, whereby constraint (27c) always holds if (27b) is satisfied.

In what follows, we focus on HetNets with ideal MBS backhaul, and find the analytical solutions to problem (P2) with different network settings. To begin with, we consider a simple case when both the MBS and SBS tiers have unconstrained backhaul capacity, i.e., $\hat{\lambda}_{\mathrm{MR}} \leq \hat{\lambda}_{\mathrm{MBH}}$ and $\hat{\lambda}_{\mathrm{SR}} \leq \hat{\lambda}_{\mathrm{SBH}}$. In this case, constraints (27c) and (27e) can be both neglected, and the network capacity cannot be improved by deploying cache. This case corresponds to conventional network deployment, where the backhaul capacity is sufficiently reserved while radio resources are limited. Problem (P2) degenerates to the conventional inter-tier load balancing problem which can be easily solved. By adding constraints (27b) and (27d), we have $\lambda \leq \hat{\lambda}_{\mathrm{MR}} + \hat{\lambda}_{\mathrm{SR}}$. Thus, the maximal network capacity is $\mu(C_{\mathrm{s}}, \varphi)^* = \hat{\lambda}_{\mathrm{MR}} + \hat{\lambda}_{\mathrm{SR}}$, and the optimal traffic steering is given by $\varphi^* = \frac{\hat{\lambda}_{\mathrm{SR}}}{\hat{\lambda}_{\mathrm{MR}}+\hat{\lambda}_{\mathrm{SR}}}$ without mobile edge caching. As SBSs further densifies, the network capacity for radio access can scale almost linearly with SBS density, whereas densely deploying high speed fiber backhaul for each SBS is not practical considering the high cost. In addition, the traffic of multiple SBSs can be geographically aggregated and transmitted through a shared backhaul (e.g., the cloud-RAN architecture), which further limits the backhaul capacity of each SBS [38]. Thus, the SBS tier can be backhaul-constrained, i.e., $\hat{\lambda}_{\mathrm{SR}} > \hat{\lambda}_{\mathrm{SBH}}$. In this case, deploying caching at SBSs can improve network capacity by reducing backhaul traffic load, which is equivalent to increasing backhaul capacity. Furthermore, the optimal solutions to (P2) is threshold-based. Denote by $C_{\min}$ and $C_{\max}$ the two thresholds of cache budgets, which are given by

$$\begin{aligned} \sum_{f=1}^{C_{\min}/\rho_{\mathrm{s}}} q_f &= \frac{\hat{\lambda}_{\mathrm{SR}} - \hat{\lambda}_{\mathrm{SBH}}}{\hat{\lambda}_{\mathrm{MR}} + \hat{\lambda}_{\mathrm{SR}}}, \\ \sum_{f=C_{\min}/\rho_{\mathrm{s}}+1}^{C_{\min}/\rho_{\mathrm{s}}+(C_{\max}-C_{\min})/\rho_{\mathrm{m}}} q_f &= \frac{\hat{\lambda}_{\mathrm{MR}}}{\hat{\lambda}_{\mathrm{MR}} + \hat{\lambda}_{\mathrm{SR}}}. \end{aligned} \tag{30}$$

The optimal solution to (P2) is summarized in Propositions 2-4, under different cache budgets.

Proposition 2. If $C < C_{\min}$, the optimal solution to (P2) is given by $C_{\mathrm{s}} = C/\rho_{\mathrm{s}}$, $\varphi = \hat{\lambda}_{\mathrm{SBH}}/(\hat{\lambda}_{\mathrm{MR}} + \hat{\lambda}_{\mathrm{SBH}})$.

Proof: Please refer to Appendix D. ∎

Remark: As deploying caching is equivalent to increasing backhaul capacity, cache instances need to be deployed at the backhaul-constrained SBSs for compensation. $C_{\min}$ can be interpreted as the deficiency of SBS backhaul, and $C_{\min}/\rho_{\min}$ is the minimal SBS cache size needed to match with radio access resources. When the cache budget is smaller than $C_{\min}$, the SBS tier is still backhaul-constrained even when all the cache instances are deployed at SBSs, and the network capacity increases with the cache budget. Furthermore, the SBS radio resources are always redundant compared with SBS backhaul, and thus the performance bottleneck exists at either the SBS backhaul or the MBS radio access. Accordingly, the load of SBS backhaul and MBS radio access should be balanced, by steering the content miss users to the two tiers appropriately.

When the cache budget increases to $C_{\min}$, the SBS backhaul deficiency can be completely compensated, and the network capacity achieves $\hat{\lambda}_{\mathrm{MR}} + \hat{\lambda}_{\mathrm{SR}}$. As cache budget further increases, the network performance will be constrained by radio access instead of SBS backhaul, and the network capacity no longer increases. If $C > C_{\min}$, there exist solutions to achieving the maximal network capacity $\mu(C_{\mathrm{s}}, \varphi)^* = \hat{\lambda}_{\mathrm{MR}} + \hat{\lambda}_{\mathrm{SR}}$, as long as the SBS cache size is large enough to compensate backhaul deficiency, i.e., $C_{\mathrm{s}} \geq C_{\min}/\rho_{\mathrm{s}}$. Among these capacity-optimal solutions, those with higher content hit rates can further improve user experience by reducing content fetching delay. Thus, we aim to find the solution $[C_{\mathrm{s}}^*, \varphi^*]$, which can maximize content hit rate while guaranteeing network capacity $\hat{\lambda}_{\mathrm{MR}} + \hat{\lambda}_{\mathrm{SR}}$. Although increasing the MBS cache size can improve content hit rate, larger MBS cache size results in heaver traffic load at MBSs, which can degrade the transmission rate at MBS radio access, especially when the cache budget is large. Thus, the optimal cache deployment and traffic steering depend on the cache budget, given by Proposition 3 and 4.

Proposition 3. If $C_{\min} \leq C < C_{\max}$, the solution $[C_{\mathrm{s}}^*, \varphi^*]$ satisfying

$$C_{\mathrm{s}}^* = C_{\min}/\rho_{\mathrm{s}}, \tag{31a}$$

$$(1 - P_{\mathrm{Hit}}^*)\varphi^* = \sum_{f=C_{\mathrm{s}}^*+C_{\mathrm{m}}^*+1}^{F} q_f \varphi^* = \frac{\hat{\lambda}_{\mathrm{SBH}}}{\hat{\lambda}_{\mathrm{MR}} + \hat{\lambda}_{\mathrm{SR}}}, \tag{31b}$$

can maximize content hit rate while maximizing network capacity, where $C_{\mathrm{m}}^* = (C - \rho_{\mathrm{s}} C_{\mathrm{s}}^*)/\rho_{\mathrm{m}}$ and P_{Hit}^* is the corresponding aggregated content hit rate.

Proof: Please refer to Appendix E. ∎

Proposition 4. If $C \geq C_{\max}$, the solution $[C_{\mathrm{s}}^*, \varphi^*]$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2017.2724547, IEEE Transactions on Vehicular Technology

TABLE III: Simulation parameters

Parameter	Value	Parameter	Value
$D_{\rm m}$	500 m	$\rho_{\rm s}$	50 /km^2
$P_{\rm TM}$	10 W	$P_{\rm TS}$	2 W
$W_{\rm M}$	100 MHz	$W_{\rm S}$	10 MHz
$\alpha_{\rm m}$	3.5	$\alpha_{\rm s}$	4
σ^2	-105 dBm/MHz	$U_{\rm MBH}$	100 Gbps
$\theta_{\rm m}$	1000	$\theta_{\rm s}$	1000
$\check{R}_{\rm RAN}$	5 Mbps	$\check{R}_{\rm BH}$	50 Mbps
F	1000	ν	0.56

satisfying

$$\varphi^* = 1 \tag{32a}$$

$$P_{\rm Hit}^{({\rm m})^*} = \sum_{f=C_{\rm s}^*+1}^{C_{\rm s}^*+C_{\rm m}^*} = \frac{\hat{\lambda}_{\rm MR}}{\hat{\lambda}_{\rm MR}+\hat{\lambda}_{\rm SR}}, \tag{32b}$$

can maximize content hit rate while maximizing network capacity, where $C_{\rm m}^* = (C - \rho_{\rm s}C_{\rm s}^*)/\rho_{\rm m}$ and $P_{\rm Hit}^{({\rm m})}$ is the corresponding content hit rate at MBSs.

Proof: Substituting Eq. (32) into (28) and (29), the network capacity can achieve $\hat{\lambda}_{\rm MR}+\hat{\lambda}_{\rm SR}$. As $\varphi^*=1$, $\mu_{\rm MR}$ will decrease as $C_{\rm m}$ increases, degrading network capacity. Therefore, $[C_{\rm s}^*, \varphi^*]$ achieves maximal content hit rate among the capacity-optimal schemes. ∎

Based on the above analysis, we summarize the propose capacity-optimal cache deployment scheme for SBS-backhaul-constrained HetNets as follows:

- **Case-1:** If $C \le C_{\min}$, all cache instances should be deployed at the SBS tier;
- **Case-2:** If $C_{\min} < C \le C_{\max}$, the cache size of each SBS is $C_{\min}/\rho_{\rm s}$, and the remaining cache budget should be deployed at the MBS tier (i.e., $C_{\rm m} = (C - C_{\min})/\rho_{\rm m}$);
- **Case-3:** If $C > C_{\max}$, the optimal cache deployment should guarantee that MBS-tier content hit rate satisfies Eq. (32b).

Meanwhile, traffic steering ratio should be adjusted with caching deployment, to balance inter-tier traffic load. In addition, the analytical results of thresholds $C_{\min}$ and $C_{\max}$ are derived as Eq. (30), which depend on backhaul and radio resource provisions.

VII. Simulation and Numerical Results

In this section, simulations are conducted to validate the obtained analytical results, and numerical results are provided to offer insights into practical network deployment. The file popularity is considered to follow Zipf distribution [39]:

$$q_f = \frac{1/f^\nu}{\sum_{h=1}^{F} 1/h^\nu}, \tag{33}$$

where $\nu \ge 0$ indicates the skewness of popularity distribution. In this simulation, ν is set as 0.56, featuring video streaming services [39]. Important parameters are listed in Table III [12].

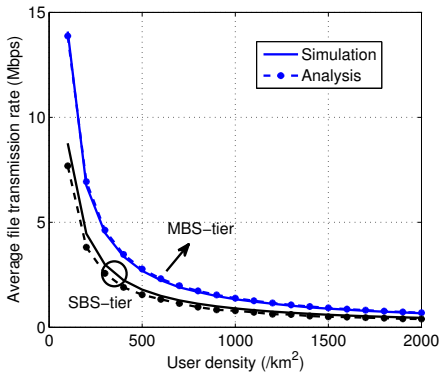


Fig. 3: Evaluation of derived file tranmsission rate at radio access parts.

A. Analytical Results Evaluation

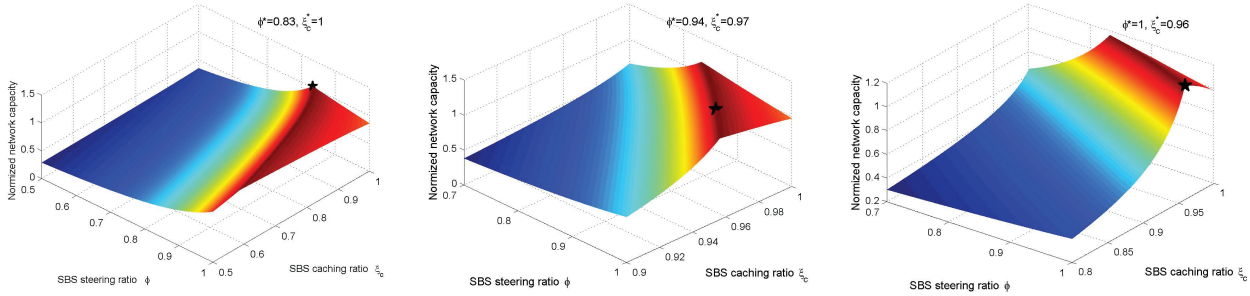
The analytical results of file transmission rates for MBS/SBS radio access are validated in Fig. 3, where 15% traffic is served by the MBS tier and the remaining is steered to the SBS tier. Monte Carlo method is applied in simulation, with SBS topology, user location and channel fading generated according to the corresponding PDFs. The simulation results is averaged over 10000 samples. The analytical results are calculated based on Lemmas 1 and 2. As the analytical and simulation results are shown to be very close, Lemmas 1 and 2 can be applied to approximate transmission rate analysis for radio access.

B. Optimal Hierarchical Caching

To validate the theoretical analysis, Fig. 4 shows network capacity with respect to different cache deployment and traffic steering ratios, where the analytical results obtained by Propositions 2-4 are marked as the star points. The per user rates for radio access and backhaul are set as $\check{R}_{\rm RAN}$=5 Mbps and $\check{R}_{\rm BH}$=50 Mbps, respectively. $\xi_{\rm c}$ is the ratio of cache budget deployed at SBSs, i.e., $\xi_{\rm c} = C_{\rm s}\rho_{\rm s}/C$. According to Eq. (30), $C_{\min}$=870 files/km^2, $C_{\max}$=930 files/km^2. Thus, the three subfigures correspond to the cases of Propositions 2-4, respectively. In Fig. 4(a), the star point is shown to achieve the maximal network capacity, validating the analysis of Proposition 2. In Figs. 4(b) and (c), there are multiple solutions that can achieve the maximal network capacity, including the star points. Furthermore, the star points also minimize the SBS cache size (i.e., minimal $\xi_{\rm c}$) among all the capacity-optimal schemes, indicating high content hit rate. When C=900 files/km^2, the SBS backhaul will become the bottleneck when $\xi_{\rm c}$ is lower than $\xi_{\rm c}^*$, degrading network capacity as shown in Fig. 4(b). Instead, in Fig. 4(c), the performance bottleneck is due to the MBS radio access, and thus the network capacity will decrease when $\xi_{\rm c}$ is lower than $\xi_{\rm c}^*$. The numerical results of Figs. 4(a)-(c) are consistent with Propositions 2-4, validating the theoretical analysis.

Fig. 5 further demonstrates the relationship between the cache budget and the optimal cache deployment, obtained by

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2017.2724547, IEEE Transactions on Vehicular Technology



(a) C=50 files/km^2 (b) C=900 files/km^2 (c) C=1000 files/km^2

Fig. 4: Optimal hierarchical cache.

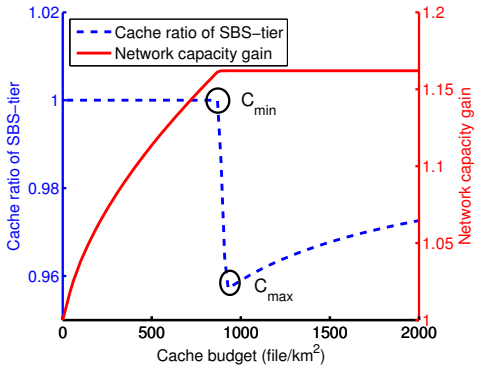


Fig. 5: Optimal cache spliting with respect to cache budget.

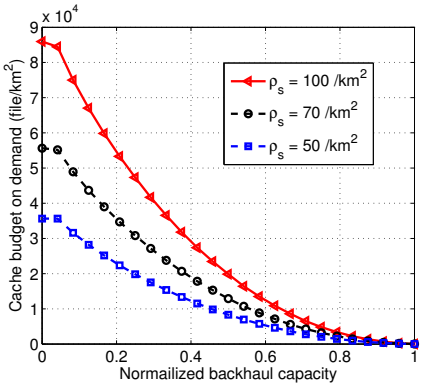


Fig. 7: Cache budget demand on different backhaul capacity.

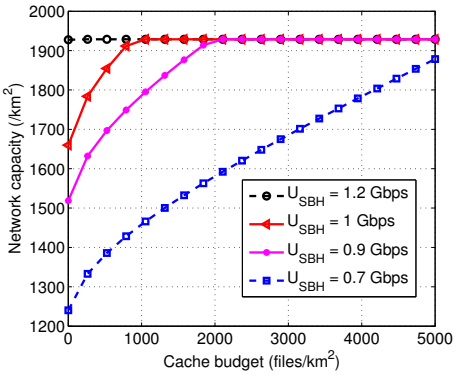


Fig. 6: Cache-enabled network capacity.

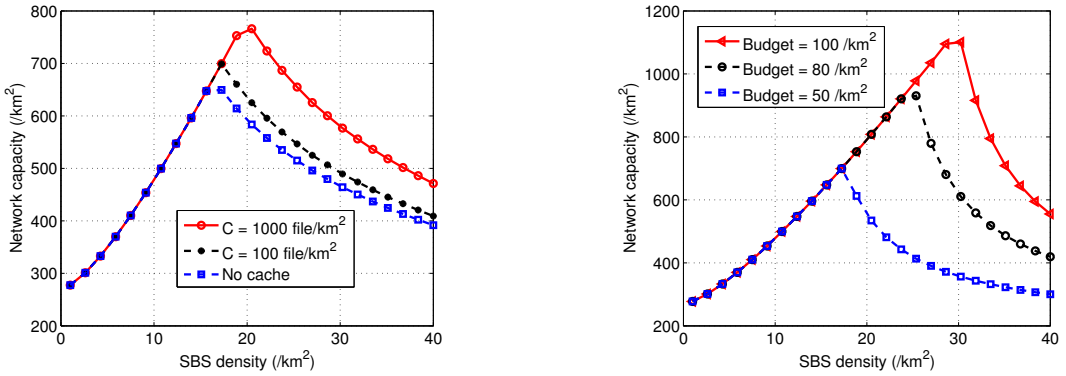
exhaustive search. As shown by the dash line, the optimal cache deployment can be divided into three cases. Firstly, all cache budget should be allocated to the SBS tier when the cache budget is insufficient, i.e., $C < C_{\min}$. As the cache budget achieves $C_{\min}$ and is lower than $C_{\max}$, it is shown that the ratio of cache budget allocated to the SBS tier begins to decrease. Furthermore, the ratio of cache budget allocated to the SBS tier increases again when the cache budget exceeds $C_{\max}$.

C. Cache-Backhaul Trading

The solid line of Fig. 5 presents the network capacity with respect to cache budget, which is normalized by the capacity without cache. As the cache budget increases, the network capacity firstly increases and then levels off as a constant. The reason is that the SBS backhaul is no longer the bottleneck when the cache budget achieves $C_{\min}$, and the network performance is constrained by the radio resources. Fig. 6 further illustrates the cache-enabled network capacity gain under different SBS backhaul capacities. Similarly, the network capacities firstly increase and then level off, and the turning points $C_{\min}$ depend on cache budgets. Furthermore, a larger backhaul capacity results in a smaller turning point, as shown in Fig. 6. Specifically, no cache budget is needed when the backhaul capacity is $U_{\mathrm{SBH}} = 1.2$ Gbps, since such backhaul capacity is sufficiently large compared with SBS radio access resources.

Fig. 6 reveals the trading relationship between backhaul capacity and cache budget demands. Specifically, networks with insufficient backhaul capacity can be compensated by deploying cache, and the backhaul deficiency determines the amount of cache budget needed. The relationship between required cache budget and backhaul capacity is illustrated in Fig. 7, where the backhaul capacity is normalized by the capacity of radio access. It is shown that less cache budget is needed as the backhaul capacity increases, and denser networks demand higher cache budget. The trading

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2017.2724547, IEEE Transactions on Vehicular Technology



(a) Cache-enabled small cells (b) Caching station

Fig. 8: Cost-effective network deployment.

relationship between backhaul capacity and cache budget demand can be applied to cost-effective network deployment, which determines the optimal combination of backhaul capacity and cache budget.

D. Case Studies on Cost-Effective Network Deployment

Finally, we provide case studies on cost-effective network deployment, by applying the results of Fig. 7. Fig. 8(a) illustrates the cost-optimal SBS deployment for the given backhaul deployment cost, under different caching budget C. The backhaul deployment cost is considered to increase with SBS density as well as backhaul capacity: $\rho_{\mathrm{s}}(1 + K_{\mathrm{BH}}U_{\mathrm{SBH}}^{\zeta_{\mathrm{BH}}})$[4]. For illustration, $K_{\mathrm{BH}} = 0.001$ and $\zeta_{\mathrm{BH}} = 0.5$[5]. When budget of the backhaul cost is 100 /km^2, the network capacity with respect to different SBS density is shown as Fig. 8(a). The network capacity is demonstrated to firstly increase and then decrease with SBS densities, falling into two regions. On one hand, the capacity of radio access increases with SBS density. On the other hand, the backhaul capacity per SBS decreases with SBS density due to the constrained deployment cost. Accordingly, the performance of SBS tier is constrained by the radio access resources when the SBS density is low, and becomes backhaul-constrained when the SBS density exceeds some threshold. In fact, the optimal SBS density achieves the best match between radio and backhaul resource settings.

Fig. 8(b) further demonstrates the cost-optimal deployment of caching stations, which is a special case when the SBSs have no backhaul and all content miss users are served by MBSs. The deployment cost is considered to increase with SBS density as well as SBS cache size, i.e., $\rho_{\mathrm{s}}(1 + K_{\mathrm{c}}C_{\mathrm{s}}^{\zeta_{\mathrm{C}}})$[6]. K_{C} and ζ_{C} are system parameters, set as $K_{\mathrm{C}} = 0.1$ and $\zeta_{\mathrm{C}} = 0.5$ for illustration[7]. The network capacity with respect to SBS density is demonstrated in Fig. 8(b). The cost-optimal SBS density reflects the tradeoff between radio access capacity and content hit rate. The capacity for radio access increases with SBS density, whereas the cache size decreases due to the deployment budget. With low density, the SBSs are overloaded due to the constrained radio resources. However, the high-dense SBSs can only serve few users due to low content hit rate, causing radio resource underutilized. Thus, the optimal SBS density should balance the capacity of SBS radio access and content hit rate, to maximize network capacity. The results of Fig. 8 can provide insightful design criteria for cost-effective network deployment.

[4] K_{BH} denotes the ratio of backhaul deployment cost to the SBS equipment cost, and ζ_{BH} reflects how the backhaul deployment cost scales with capacity.

[5] With this setting, the backhaul deployment cost is comparable to the SBS equipment cost when backhaul capacity is 1 Gbps.

[6] K_{c} denotes the ratio of storage cost to the cost of other modules, and ζ_{C} reflects how storage cost scales with cache size.

[7] With this setting, the storage cost is comparable to the other modules when cache size is 100 files, i.e., 10% of all contents

VIII. Conclusions and Future Work

In this paper, the cost-effective cache deployment problem has been investigated for a large-scale two-tier HetNet, aiming at maximizing network capacity while meeting file transmission rate requirements. By conducting stochastic geometry analysis, the capacity-optimal cache sizes have been derived, which is threshold-based with respect to cache budget under the SBS-backhaul-constrained case. The analytical results of cache budget threshold have been obtained, which characterize the backhaul deficiency and the cache-backhaul trading relationship. The proposed cache deployment schemes can be applied to practical network upgrades as well as capacity enhancement. When the existing networks upgrade with storage units for edge caching, the optimal cache sizes of different BSs can be directly determined with the obtained cache budget threshold, based on system parameters such as base station density, radio resources, backhaul capacity, and content popularity. When more cache-enabled MBSs or SBSs are deployed for capacity enhancement, the proposed method can be applied to determine the optimal cache sizes and simplify the optimization of other system parameters. For future work, we will optimize cache deployment based on cooperative caching scheme, where multiple SBSs can cooperate to serve users.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2017.2724547, IEEE Transactions on Vehicular Technology

APPENDIX A
PROOF OF LEMMA 1

The average transmission rate of MBS radio access is given by:

$$\begin{aligned}\mathbb{E}[R_{\mathrm{MR}}] &= W_{\mathrm{m}}\mathbb{E}\left[\frac{1}{1+N_{\mathrm{MR}}}\right]\mathbb{E}\left[\log_2(1+\gamma_{\mathrm{m}})\right]\\ &= \frac{W_{\mathrm{m}}}{\ln 2}\mathbb{E}\left[\frac{1}{1+N_{\mathrm{MR}}}\right]\mathbb{E}[\ln(1+\gamma_{\mathrm{m}})]\end{aligned} \tag{34}$$

As N_{MR} follows Poisson distribution of mean $\lambda_{\mathrm{MR}}/\rho_{\mathrm{m}}$,

$$\mathbb{E}\left[\frac{1}{1+N_{\mathrm{MR}}}\right] = \frac{\rho_{\mathrm{m}}}{\lambda_{\mathrm{MR}}}\left(1-e^{-\frac{\lambda_{\mathrm{MR}}}{\rho_{\mathrm{m}}}}\right), \tag{35}$$

which can be derived in the same way as Eq. (12) by replacing λ_{MBH} by λ_{MR}. Furthermore,

$$\begin{aligned}\mathbb{E}\left[\ln(1+\gamma_{\mathrm{m}})\right] &= \frac{D_{\min}^2}{D_{\mathrm{m}}^2}\ln\left(1+\frac{P_{\mathrm{TM}}D_{\min}^{-\alpha_{\mathrm{m}}}}{(1+\theta_{\mathrm{m}})\sigma^2}\right)\\ &+\int_{D_{\min}}^{D_{\mathrm{m}}}\frac{2d}{D_{\mathrm{m}}^2}\ln\left(1+\frac{P_{\mathrm{TM}}d^{-\alpha_{\mathrm{m}}}}{(1+\theta_{\mathrm{m}})\sigma^2}\right)\mathrm{d}d\\ &\geq \frac{D_{\min}^2}{D_{\mathrm{m}}^2}\ln\left(\frac{P_{\mathrm{TM}}D_{\min}^{-\alpha_{\mathrm{m}}}}{(1+\theta_{\mathrm{m}})\sigma^2}\right)-\frac{2\alpha_{\mathrm{m}}}{D_{\mathrm{m}}^2}\int_{D_{\min}}^{D_{\mathrm{m}}} d\ln d\mathrm{d}d\\ &+\left(1-\frac{D_{\min}^2}{D_{\mathrm{m}}^2}\right)\ln\left(\frac{P_{\mathrm{TM}}}{(1+\theta_{\mathrm{m}})\sigma^2}\right) \qquad (36a)\\ &= \ln\frac{P_{\mathrm{TM}}D_{\mathrm{m}}^{-\alpha_{\mathrm{m}}}}{(1+\theta_{\mathrm{m}})\sigma^2}+\frac{\alpha_{\mathrm{m}}}{2}\left(1-\frac{D_{\min}^2}{D_{\mathrm{m}}^2}\right),\end{aligned}$$

where $\theta_{\mathrm{m}}\sigma^2 = \mathbb{E}[I_{\mathrm{m}}]$, and the equality of (36a) holds when $\frac{\sigma^2}{P_{\mathrm{TM}}} \to 0$. Substituting Eqs. (35) and (36) into (34), Lemma 1 can be proved.

APPENDIX B
PROOF OF LEMMA 2

The average transmission rate of SBS radio access is given by:

$$\mathbb{E}\left[R_{\mathrm{SR}}\right] = \frac{W_{\mathrm{m}}}{\ln 2}\mathbb{E}\left[\frac{1}{1+N_{\mathrm{SR}}}\right]\mathbb{E}\left[\ln(1+\gamma_{\mathrm{s}})\right]. \tag{37}$$

Similar to Eq. (15),

$$\mathbb{E}\left[\frac{1}{1+N_{\mathrm{SR}}}\right] = \frac{\kappa\rho_{\mathrm{s}}}{\lambda_{\mathrm{SR}}}\frac{\Gamma(\kappa-1)}{\Gamma(\kappa)}\left(1-\frac{1}{\left(1+\frac{\lambda_{\mathrm{SR}}}{\kappa\rho_{\mathrm{s}}}\right)^{\kappa-1}}\right). \tag{38}$$

As SBSs follows PPP of density ρ_{s}, the PDF of transmission distance d_{s} follows

$$f_{d_{\mathrm{s}}}(d) = \frac{\mathrm{d}}{\mathrm{d}d}\left(1-e^{-\pi\rho_{\mathrm{s}}d^2}\right). \tag{39}$$

Thus,

$$\begin{aligned}\mathbb{E}\left[\ln(1+\gamma_{\mathrm{s}})\right] &= \int_0^\infty \ln\left(1+\frac{P_{\mathrm{TS}}d^{-\alpha_{\mathrm{s}}}}{(1+\theta_{\mathrm{s}})\sigma^2}\right)f_{d_{\mathrm{s}}}(d)\mathrm{d}d\\ &\geq \ln\frac{P_{\mathrm{TS}}}{(1+\theta_{\mathrm{s}})\sigma^2}-\alpha_{\mathrm{s}}\int_0^\infty 2\pi\rho_{\mathrm{s}}de^{-\pi\rho_{\mathrm{s}}d^2}\ln d\mathrm{d}d \qquad (40a)\\ &= \ln\frac{P_{\mathrm{TS}}(\pi\rho_{\mathrm{s}})^{\frac{\alpha_{\mathrm{s}}}{2}}}{(1+\theta_{\mathrm{s}})\sigma^2}-\alpha_{\mathrm{s}}\int_0^\infty e^{-x}\ln x\mathrm{d}x\\ &= \ln\frac{P_{\mathrm{TS}}(\pi\rho_{\mathrm{s}})^{\frac{\alpha_{\mathrm{s}}}{2}}}{(1+\theta_{\mathrm{s}})\sigma^2}+\frac{1}{2}\alpha_{\mathrm{s}}\gamma, \qquad (40b)\end{aligned}$$

where $\theta_{\mathrm{s}}\sigma^2 = \mathbb{E}[I_{\mathrm{s}}]$, and the equality of (40b) holds when $\frac{\sigma^2}{P_{\mathrm{TS}}} \to 0$. Substituting Eqs. (38) and (40) into (37), Lemma 2 can be proved.

APPENDIX C
PROOF OF TABLE II

When φ increases, the ratio of traffic load steered to MBS backhaul and radio access both decrease according to Eq. (28a) and (28b), hence increasing μ_{MR} and μ_{MBH}. On the contrary, μ_{SR} and μ_{SBH} both decrease, according to Eq. (28c) and (28d).

Suppose the SBS cache size C_{s} increases to $C_{\mathrm{s}}' = C_{\mathrm{s}} + \Delta_{\mathrm{s}}$, and the MBS cache size C_{m} decreases to $C_{\mathrm{m}}' = C_{\mathrm{m}} + \Delta_{\mathrm{m}}$, where $\rho_{\mathrm{s}}\Delta_{\mathrm{s}} + \rho_{\mathrm{m}}\Delta_{\mathrm{m}} = 0$. As $\rho_{\mathrm{s}} > \rho_{\mathrm{m}}$ in practical networks, $\Delta_{\mathrm{s}} + \Delta_{\mathrm{m}} < 0$. Denote by $\Delta P_{\mathrm{Hit}}^{(\mathrm{m})}$ and $\Delta P_{\mathrm{Hit}}^{(\mathrm{s})}$ the corresponding variations of MBS-tier and SBS-tier content hit rates, respectively. Apparently, $\Delta P_{\mathrm{Hit}}^{(\mathrm{m})} < 0$ and $\Delta P_{\mathrm{Hit}}^{(\mathrm{s})} > 0$, and $\Delta P_{\mathrm{Hit}}^{(\mathrm{m})} + \Delta P_{\mathrm{Hit}}^{(\mathrm{s})} < 0$ since $\Delta_{\mathrm{s}} + \Delta_{\mathrm{m}} < 0$. Thus, the total content hit rate decreases. Therefore, μ_{MBH} and μ_{SBH} both decrease, according to Eqs. (28b) and (28d). In addition, μ_{SR} also decreases as $P_{\mathrm{Hit}}^{(\mathrm{s})}$ increases, according to Eq. (28c). For μ_{MR},

$$\begin{aligned}&\Delta P_{\mathrm{Hit}}^{(\mathrm{m})} - (\Delta P_{\mathrm{Hit}}^{(\mathrm{m})} + \Delta P_{\mathrm{Hit}}^{(\mathrm{s})})(1-\varphi)\\ &= -\Delta P_{\mathrm{Hit}}^{(\mathrm{s})} + (\Delta P_{\mathrm{Hit}}^{(\mathrm{m})} + \Delta P_{\mathrm{Hit}}^{(\mathrm{s})})\varphi < 0\end{aligned} \tag{41}$$

and thus μ_{MR} increases with C_{s}, according to Eq. (28a).

APPENDIX D
PROOF OF PROPOSITION 2

Set $C_{\mathrm{s}} = C/\rho_{\mathrm{s}}$, $C_{\mathrm{m}} = 0$, $\varphi = \frac{\hat{\lambda}_{\mathrm{SBH}}}{\hat{\lambda}_{\mathrm{MR}}+\hat{\lambda}_{\mathrm{SBH}}}$. Accordingly, $P_{\mathrm{Hit}}^{(\mathrm{s})} = \sum_{f=1}^{C/\rho_{\mathrm{s}}} q_f$, $P_{\mathrm{Hit}}^{(\mathrm{m})} = 0$, and $P_{\mathrm{Hit}} = \sum_{f=1}^{C/\rho_{\mathrm{s}}} q_f$. Substituting $P_{\mathrm{Hit}}^{(\mathrm{m})}$ and φ into Eqs. (28a) and (28d), we have $\mu_{\mathrm{MR}} = \mu_{\mathrm{SBH}} = \frac{\hat{\lambda}_{\mathrm{MR}}+\hat{\lambda}_{\mathrm{SBH}}}{1-P_{\mathrm{Hit}}}$, and $\mu_{\mathrm{SR}} = \frac{\hat{\lambda}_{\mathrm{SR}}}{1-\frac{\hat{\lambda}_{\mathrm{MR}}}{\hat{\lambda}_{\mathrm{MR}}+\hat{\lambda}_{\mathrm{SBH}}}(1-P_{\mathrm{Hit}})}$. As $C < C_{\min}$,

$$\begin{aligned}&P_{\mathrm{Hit}} = \sum_{f=1}^{C/\rho_{\mathrm{s}}} q_f < \frac{\hat{\lambda}_{\mathrm{SR}}-\hat{\lambda}_{\mathrm{SBH}}}{\hat{\lambda}_{\mathrm{MR}}+\hat{\lambda}_{\mathrm{SR}}}\\ \Longleftrightarrow &P_{\mathrm{Hit}}(\hat{\lambda}_{\mathrm{MR}}+\hat{\lambda}_{\mathrm{SR}}) < \hat{\lambda}_{\mathrm{SR}}-\hat{\lambda}_{\mathrm{SBH}}\\ \Longleftrightarrow &P_{\mathrm{Hit}}\hat{\lambda}_{\mathrm{MR}}+\hat{\lambda}_{\mathrm{SBH}} < (1-P_{\mathrm{Hit}})\hat{\lambda}_{\mathrm{SR}}\end{aligned} \tag{42}$$

Notice that

$$\begin{aligned}&\frac{\hat{\lambda}_{\mathrm{SR}}}{1-\frac{\hat{\lambda}_{\mathrm{MR}}}{\hat{\lambda}_{\mathrm{MR}}+\hat{\lambda}_{\mathrm{SBH}}}(1-P_{\mathrm{Hit}})}-\frac{\hat{\lambda}_{\mathrm{MR}}+\hat{\lambda}_{\mathrm{SBH}}}{1-P_{\mathrm{Hit}}}\\ &= \frac{\hat{\lambda}_{\mathrm{MR}}+\hat{\lambda}_{\mathrm{SBH}}}{1-P_{\mathrm{Hit}}}\left[\frac{\hat{\lambda}_{\mathrm{SR}}(1-P_{\mathrm{Hit}})}{\hat{\lambda}_{\mathrm{SBH}}+P_{\mathrm{Hit}}\hat{\lambda}_{\mathrm{MR}}}-1\right] > 0.\end{aligned} \tag{43}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2017.2724547, IEEE Transactions on Vehicular Technology

Thus, $\mu_{\rm SR} > \mu_{\rm MR} = \mu_{\rm SBH}$. According to Table II, $\mu_{\rm MR}$ and $\mu_{\rm SBH}$ cannot be simultaneously improved. Therefore, $C_{\rm s}^* = C/\rho_{\rm s}, \varphi^* = \frac{\hat{\lambda}_{\rm SBH}}{\hat{\lambda}_{\rm MR}+\hat{\lambda}_{\rm SBH}}$ is the optimal solution to problem (P2).

APPENDIX E
PROOF OF PROPOSITIONS 3

Firstly, we prove that $[C_{\rm s}^*, \varphi^*]$ satisfying Eq. (31) is feasible to constraint (27f) in problem (P2). As $C \geq C_{\min}$, $C_{\rm s}^* = C_{\min}/\rho_{\rm s}$ is feasible to (27f). When $C_{\rm s}^* = C_{\min}/\rho_{\rm s}$, the SBS-tier content hit rate is $P_{\rm Hit}^{({\rm s})^*} = \frac{\hat{\lambda}_{\rm SR}-\hat{\lambda}_{\rm SBH}}{\hat{\lambda}_{\rm MR}+\hat{\lambda}_{\rm SR}}$, according to the definition of $C_{\min}$ in Eq. (30). As $C < C_{\max}/\rho_{\rm s}$, $P_{\rm Hit}^{({\rm m})^*} < \frac{\hat{\lambda}_{\rm MR}}{\hat{\lambda}_{\rm MR}+\hat{\lambda}_{\rm SR}}$. Thus, $1 - {P_{\rm Hit}}^* > \frac{\hat{\lambda}_{\rm SBH}}{\hat{\lambda}_{\rm MR}+\hat{\lambda}_{\rm SR}}$, and $\varphi \in (0,1)$ is feasible to (27f) in (P2).

Then, we prove that the network capacity achieves the maximum under $[C_{\rm s}^*, \varphi^*]$. Substituting Eq. (31) into (28), we have $\mu_{\rm MR}^* = \mu_{\rm SR}^* = \mu_{\rm SBH}^* = \hat{\lambda}_{\rm MR} + \hat{\lambda}_{\rm SR}$. Thus, the network capacity is $\hat{\lambda}_{\rm MR} + \hat{\lambda}_{\rm SR}$. Adding constraints (27b) and (27d), we can prove that the maximal network capacity cannot exceed $\hat{\lambda}_{\rm MR} + \hat{\lambda}_{\rm SR}$. Therefore, Eq. (31) guarantees the optimality of capacity.

In addition, we prove that the network capacity is smaller than $\hat{\lambda}_{\rm MR} + \hat{\lambda}_{\rm SR}$ if $C_{\rm s} \leq C_{\rm s}^*$, by contradiction. Assume there exist a solution $[C_{\rm s}', \varphi']$ with network capacity of $\hat{\lambda}_{\rm MR} + \hat{\lambda}_{\rm SR}$, where $C_{\rm s}' \leq C_{\min}/\rho_{\rm s}$. According to Eqs. (31a) and (31d), we have

$$P_{\rm Hit}^{({\rm M})\prime} + (1 - P_{\rm Hit}')(1 - \varphi') \leq \frac{\hat{\lambda}_{\rm MR}}{\hat{\lambda}_{\rm MR} + \hat{\lambda}_{\rm SR}}, \tag{44a}$$

$$(1 - P_{\rm Hit}')\varphi' \leq \frac{\hat{\lambda}_{\rm SBH}}{\hat{\lambda}_{\rm MR} + \hat{\lambda}_{\rm SR}}. \tag{44b}$$

In addition

$$\begin{aligned} & P_{\rm Hit}^{({\rm M})\prime} + (1 - P_{\rm Hit}')(1 - \varphi') \\ & \geq 1 - P_{\rm Hit}^{({\rm s})\prime} - \frac{\hat{\lambda}_{\rm SBH}}{\hat{\lambda}_{\rm MR} + \hat{\lambda}_{\rm SR}} && (45a) \\ & > 1 - P_{\rm Hit}^{({\rm s})^*} - (1 - P_{\rm Hit}^*)\varphi^* && (45b) \\ & = \frac{\hat{\lambda}_{\rm MR}}{\hat{\lambda}_{\rm MR} + \hat{\lambda}_{\rm SR}}, && (45c) \end{aligned}$$

where (45a) is based on (44b), (45b) is due to condition (31b), (45c) holds as $P_{\rm Hit}^{({\rm s})}$ increases with $C_{\rm s}$, and (45d) comes from (31). As (45) is contradictory with (44a), there exists no $C_{\rm s}' \leq C_{\min}/\rho_{\rm s}$ to achieve the maximal network capacity. As content hit rate decreases with SBS cache size, $[C_{\rm s}^*, \varphi^*]$ achieves the maximal content hit rate among the capacity-optimal solutions.

REFERENCES

[1] X. Ge, S. Tu, G. Mao, C.-X. Wang, and T. Han, "5g ultra-dense cellular networks," *IEEE Wireless Commun.*, vol. 23, no. 1, pp. 72–79, Feb. 2016.

[2] D. López-Pérez, M. Ding, H. Claussen, and A. H. Jafari, "Towards 1 Gbps/UE in cellular systems: understanding ultra-dense small cell deployments," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2078–2101, 2015.

[3] Y.-J. Yu, W.-C. Tsai, and A.-C. Pang, "Backhaul traffic minimization under cache-enabled CoMP transmissions over 5G cellular systems."

[4] M. Ding, J. Zou, Z. Yang, H. Luo, and W. Chen, "Sequential and incremental precoder design for joint transmission network MIMO systems with imperfect backhaul," *IEEE Trans. Veh. Technol.*, vol. 61, no. 6, pp. 2490–2503, Jul. 2012.

[5] S. Zhou, J. Gong, Z. Zhou, W. Chen, and Z. Niu, "GreenDelivery: proactive content caching and push with energy-harvesting-based small cells," *IEEE Commun. Mag.*, vol. 53, no. 4, pp. 142–149, Apr. 2015.

[6] P. Yang, N. Zhang, Y. Bi, L. Yu, and X. Shen, "Catalyzing cloud-fog interoperation in 5G wireless networks: An SDN approach," *arXiv preprint arXiv:1612.05291*, 2016.

[7] J. Qiao, Y. He, and X. Shen, "Proactive caching for mobile video streaming in millimeter wave 5G networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 7187–7198, Oct. 2016.

[8] R. Tandon and O. Simeone, "Harnessing cloud and edge synergies: Toward an information theory of fog radio access networks," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 44–50, Aug. 2016.

[9] X. Li, X. Wang, and V. C. Leung, "Weighted network traffic offloading in cache-enabled heterogeneous networks," in *IEEE ICC'16*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[10] B. Bharath, K. Nagananda, and H. V. Poor, "A learning-based approach to caching in heterogenous small cell networks," *IEEE Trans. Commun.*, vol. 64, no. 4, pp. 1674–1686, Apr. 2016.

[11] M. Akon, M. T. Islam, X. Shen, and A. Singh, "OUR: Optimal update-based replacement policy for cache in wireless data access networks with optimal effective hits and bandwidth requirements," *Wireless Commun. and Mobile Comput.*, vol. 13, no. 15, pp. 1337–1352, Oct. 2013.

[12] D. Liu and C. Yang, "Energy efficiency of downlink networks with caching at base stations," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 907–922, Apr. 2016.

[13] A. Liu and V. K. N. Lau, "Exploiting base station caching in MIMO cellular networks: Opportunistic cooperation for video streaming," *IEEE Trans. Signal Process.*, vol. 63, no. 1, pp. 57–69, Jan 2015.

[14] N. Zhang, S. Zhang, S. Wu, J. Ren, J. W. Mark, and X. Shen, "Beyond coexistence: Traffic steering in LTE networks with unlicensed bands," *IEEE Wireless Commun.*, vol. 23, no. 6, pp. 40–46, Dec. 2016.

[15] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. Leung, "Cache in the air: exploiting content caching and delivery techniques for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 131–139, Feb. 2014.

[16] 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.

[17] M. Akon, M. T. Islam, X. Shen, and A. Singh, "A bandwidth and effective hit optimal cache scheme for wireless data access networks with client injected updates," *Comput. Netw.*, vol. 56, no. 7, pp. 2080–2095, May 2012.

[18] B. Serbetci and J. Goseling, "On optimal geographical caching in heterogeneous cellular networks," *arXiv preprint arXiv:1601.07322*, 2016.

[19] Y. Chen, M. Ding, J. Li, Z. Lin, G. Mao, and L. Hanzo, "Probabilistic small-cell caching: Performance analysis and optimization," *IEEE Trans. Veh. Technol.*, 2017, DOI: 10.1109/TVT.2016.2606765, to appear.

[20] Z. Chang, Y. Gu, Z. Han, X. Chen, and T. Ristaniemi, "Context-aware data caching for 5G heterogeneous small cells networks," in *IEEE ICC'16*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[21] Y. Cui, F. Lai, S. Hanly, and P. Whiting, "Optimal caching and user association in cache-enabled heterogeneous wireless networks," *arXiv preprint arXiv:1604.05828*, 2016.

[22] J. Liu, B. Bai, J. Zhang, and K. B. Letaief, "Content caching at the wireless network edge: A distributed algorithm via belief propagation," in *IEEE ICC'16*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[23] E. Baştuğ, M. Kountouris, M. Bennis, and M. Debbah, "On the delay of geographical caching methods in two-tiered heterogeneous networks," in *IEEE SPAWC'16*, Edinburgh, UK, Aug. 2016, pp. 1–6.

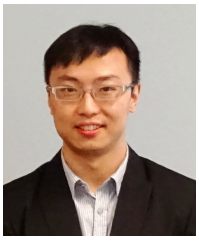
[24] F. Sun, B. Liu, F. Hou, H. Zhou, J. Chen, Y. Rui, and L. Gui, "A QoE centric distributed caching approach for vehicular video streaming in cellular networks," *Wireless Commun. and Mobile Comput.*, vol. 16, pp. 1612–1624, Sep. 2015.

[25] R. Wang, X. Peng, J. Zhang, and K. B. Letaief, "Mobility-aware caching for content-centric wireless networks: modeling and methodology," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 77–83, Aug. 2016.

[26] J. Qiao, X. Shen, J. W. Mark, and L. Lei, "Video quality provisioning for millimeter wave 5G cellular networks with link outage," *IEEE Trans. Wireless Commun.*, vol. 14, no. 10, pp. 5692–5703, Oct. 2015.

[27] M. Gregori, J. Gómez-Vilardebó, J. Matamoros, and D. Gündüz, "Wireless content caching for small cell and D2D networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1222–1234, May 2016.
[28] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6118–6131, Sep. 2016.
[29] S. E. Ghoreishi, V. Friderikos, D. Karamshuk, N. Sastry, and A. H. Aghvami, "Provisioning cost-effective mobile video caching," in *IEEE ICC'16*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.
[30] W. Han, A. Liu, and V. K. Lau, "Tradeoff between phy caching and core network caching in cellular networks," in *IEEE ICC'16*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.
[31] X. Peng, J. Zhang, S. Song, and K. B. Letaief, "Cache size allocation in backhaul limited wireless networks," in *IEEE ICC'16*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.
[32] S. Sesia, I. Toufik, and M. Baker, *LTE: the UMTS long term evolution*. Wiley Online Library, 2009.
[33] M. Ismail, A. Abdrabou, and W. Zhuang, "Cooperative decentralized resource allocation in heterogeneous wireless access medium," *IEEE Trans. Wireless Commun.*, vol. 12, no. 2, pp. 714–724, Feb. 2013.
[34] W. Song and W. Zhuang, "Performance analysis of probabilistic multipath transmission of video streaming traffic over multi-radio wireless devices," *IEEE Trans. Wireless Commun.*, vol. 11, no. 4, pp. 1554–1564, Apr. 2012.
[35] M. Haenggi, J. G. Andrews, F. Baccelli, O. Dousse, and M. Franceschetti, "Stochastic geometry and random graphs for the analysis and design of wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 7, pp. 1029–1046, Sep. 2009.
[36] D. Cao, S. Zhou, and Z. Niu, "Optimal combination of base station densities for energy-efficient two-tier heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 9, pp. 4350–4362, Sep. 2013.
[37] S. Zhang, J. Gong, S. Zhou, and Z. Niu, "How many small cells can be turned off via vertical offloading under a separation architecture?" *IEEE Trans. Wireless Commun.*, vol. 14, no. 10, pp. 5440–5453, Oct. 2015.
[38] "C-RAN: the road towards green RAN," China Mobile, Tech. Rep., 2011.
[39] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "Analyzing the video popularity characteristics of large-scale user generated content systems," *IEEE/ACM Trans. on Netw.*, vol. 17, no. 5, pp. 1357–1370, Oct. 2009.

Shan Zhang (M'16) received her Ph.D. degree in Department of Electronic Engineering from Tsinghua University and B.S. degree in Department of Information from Beijing Institute Technology, Beijing, China, in 2016 and 2011, respectively. She is currently a post doctoral fellow in Department of Electronical and Computer Engineering, University of Waterloo, Ontario, Canada. Her research interests include resource and traffic management for green communication, intelligent vehicular networking, and software defined networking. Dr. Zhang received the Best Paper Award at the Asia-Pacific Conference on Communication in 2013.



Ning Zhang (M'15) received the Ph.D degree from University of Waterloo in 2015. He is now an assistant professor in the Department of Computing Science at Texas A&M University-Corpus Christi. Before that, he was a postdoctoral research fellow at BBCR lab in University of Waterloo. He was the co-recipient of the Best Paper Award at IEEE GLOBECOM 2014 and IEEE WCSP 2015. His current research interests include next generation wireless networks, software defined networking, vehicular networks, and physical layer security.

Peng Yang (S'15) received his B.E. degree from the Department of Electronics and Information Engineering, Huazhong University of Science and Technology (HUST), Wuhan, China, in 2013. Currently, he is pursuing his Ph.D. degree in the School of Electronic Information and Communications, HUST. From Sep. 2015, he is also a visiting Ph.D. student in the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His current research interests include next generation wireless networking, software defined networking and fog computing.

Xuemin (Sherman) Shen (M'97-SM'02-F'09) received the B.Sc.(1982) degree from Dalian Maritime University (China) and the M.Sc. (1987) and Ph.D. degrees (1990) from Rutgers University, New Jersey (USA), all in electrical engineering. He is a University Professor and the Associate Chair for Graduate Studies, Department of Electrical and Computer Engineering, University of Waterloo, Canada. Dr. Shens research focuses on resource management, wireless network security, social networks, smart grid, and vehicular ad hoc and sensor networks. He was an elected member of IEEE ComSoc Board of Governor, and the Chair of Distinguished Lecturers Selection Committee. Dr. Shen served as the Technical Program Committee Chair/Co-Chair for IEEE Globecom16, Infocom14, IEEE VTC10 Fall, and Globecom07, the Symposia Chair for IEEE ICC10, the Tutorial Chair for IEEE VTC'11 Spring and IEEE ICC08, the General Co-Chair for ACM Mobihoc15, Chinacom07 and QShine06, the Chair for IEEE Communications Society Technical Committee on Wireless Communications, and P2P Communications and Networking. He also serves/served as the Editor-in-Chief for IEEE Internet of Things Journal, IEEE Network, Peer-to-Peer Networking and Application, and IET Communications; a Founding Area Editor for IEEE Transactions on Wireless Communications; an Associate Editor for IEEE Transactions on Vehicular Technology, Computer Networks, and ACM/Wireless Networks, etc.; and the Guest Editor for IEEE JSAC, IEEE Wireless Communications, IEEE Communications Magazine, and ACM Mobile Networks and Applications, etc. Dr. Shen received the Excellent Graduate Supervision Award in 2006, and the Premiers Research Excellence Award (PREA) in 2003 from the Province of Ontario, Canada. Dr. Shen is a registered Professional Engineer of Ontario, Canada, an IEEE Fellow, an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering Fellow, a Royal Society of Canada Fellow, and a Distinguished Lecturer of IEEE Vehicular Technology Society and Communications Society.