Flexible Design of Millimeter-Wave Cache Enabled Fog Networks

Mostafa Emara*, Hesham ElSawy*, Sameh Sorour†, Samir Al-Ghadhban*, Mohamed-Slim Alouini‡ and Tareq Y. Al-Naffouri‡

*Department of Electrical Engineering, King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia
Emails: {mostafaemara, hesham.elsawy, samir}@kfupm.edu.sa
†Department of Electrical and Computer Engineering, University of Idaho, USA
Email: samehsorour@uidaho.edu
‡CEMSE Division, EE program, King Abdullah University of Science and Technology (KAUST), Saudi Arabia
Emails: {slim.alouini, tareq.alnaffouri}@kaust.edu.sa

***Abstract*—Ultra-densification, millimeter wave (mmW) communications, and proactive network-edge caching, utilized within mmW fog networks (mmFNs), are foreseen to provide tangible gains for broadband access, network capacity, and latency. However, caching implementation in mmFN imposes high capital expenditure (CAPEX) due to the ultra-high density of base stations (BSs). For a given caching CAPEX, it may be more efficient to install higher capacity caches in a fraction of the BSs than installing smaller capacity caches in every BSs. In the former case, wireless self-backhauling of mmW systems can be exploited to share the cache contents stored in a given cache enabled BSs (CE-BSs) with other BSs in the network. In this regards, this paper develops a mathematical model, based on stochastic geometry, to study the tradeoff between the cache size and intensity of CE-BSs on the probability that requested popular contents are retrieved from the network edge, denoted as the hit probability. Assuming a power-law inverse relationship between the cache size and intensity of CE-BSs, an optimization problem is formulated and solved for the intensity of CE-BSs and probabilistic file placement in caches such that the hit probability is maximized. The results show that neither installing small caches in every BS nor having sufficiently high capacity caches (i.e., that confine all popular files) installed in small number of BSs exploit the full potential of mmFN. Instead, there exists an optimal balance between the cache size and intensity of CE-BSs, which depends on the network parameters such as the applied caching strategy, required rate, total intensity of BSs, popular content distribution, and cache size/intensity relationship.**

***Keywords*—Caching System; Stochastic Geometry; mmW, self-backhauling, ultra-dense networks.**

I. Introduction

The fifth generation (5G) cellular networks are challenged to offer multi-fold improvement in broad-band access, network capacity, and end-to-end latency. For instance, augmented/virtual reality applications require data rates in the order of Gbps with latency less than 20ms [1]. To concurrently improve such contradictory key performance indicators (KPIs), a paradigm shift in the network deployment and design is required. In this regards, the evolving millimeter wave (mmW) fog networks (mmFN) offers a promising solution to materialize the 5G vision. Particularly, mmFN brings three novel, and complementary, ingredients to cellular networks, namely, mmW communications, ultra-densification, and proactive edge-caching [2].

The mmW communications exploit high bandwidth channels within the underutilized band of 20-100 GHz, which significantly improves transmission rates compared to conventional cellular micro-wave (0.8-2.1 GHz) bands. Ultra-densification mitigates the poor propagation properties of mmW communications (e.g., high path-loss and blockage) by reducing the communication link distances and increasing the probability of line of sight [3]. Furthermore, ultra-densification improves the network capacity through increasing the spatial spectral efficiency and reducing devices load per base station (BS). Proactive caching brings popular contents to the network edge (i.e., nearer to the devices), which reduces communication latency. Hence, the mmFN provides the foreseen low-latency broadband access for dense devices, which is crucial for several 5G applications such as augmented/virtual reality, online gaming, social networking, and video streaming [4].

The potential gains of densification, caching, and mmW communications have attracted tremendous interest from academia and industry. The anticipated performance of ultra-dense networks is characterized in [5]. Experimental proof of concept for mmW communications is conducted in [6], [7]. On the scale of networks, stochastic geometry is utilized in [8]–[11] to characterize the coverage probability and rate in mmW cellular networks. Self-backhauling in dense mmW cellular networks is studied in [9]. For caching systems, optimal file placement that maximizes the probability that requested files are retrieved from the BSs' caches, denoted as the hit probability, is derived in [12]–[14]. The performance gain due to opportunistic spectrum access in cache enabled network is characterized in [15], [16]. However, none of [12]–[16] studies caching in mmW networks. The authors in [17], [18] and [19] study caching in mmW networks when, respectively, all BSs and a fraction of BSs are cache enabled. However, none of the above research studies the interactions between the cache size and intensity of cache enabled BSs (CE-BSs) in self-backhauled mmFN (i.e., ultra-dense) networks.

Caching implementation and design for mmFN networks is a challenging task due to the ultra high density of BSs. For a given capital expenditure (CAPEX) to install caches in BSs, there is an obvious inverse relationship between cache

size and the intensity of BSs to be augmented with caches. Augmenting all BSs with caches may impose stringent constraints on the cache size. Hence, small portion of the popular files library can be cached at the network edge, and hence, proactive caching is not exploited to its full potential. For the same caching CAPEX, larger cache size can be utilized at the network edge, however, at the cost of implementing the caches at a fraction of the BSs. Note that, exploiting the high density of BSs and self-backhauling of mmW systems, the large caches implemented in CE-BSs can be accessed with cache-free BSs (CF-BSs). The tradeoff between the cache size and density, for the same CAPEX, is a crucial problem that has never been addressed before.

To the best of the authors knowledge, this paper is the first to study the cache size/intensity tradeoff in self-backhauled mmFN for optimized and most popular caching strategies. Particularly, we study a Poisson point process (PPP) mmFN with self-backhauling capabilities where the cache size implemented in the BSs is inversely proportional to the CE-BSs intensity. Without loss of generality, an inverse power-law relationship between the fraction ζ of CE-BSs and the cache capacity $M \propto \frac{1}{\zeta^{\epsilon}}$ at each.[1] Due to the poor propagation characteristics of mmW signals, users are always associated to their closest BSs that offer minimum path loss. If the serving BS does not store the requested file (i.e., CE-BS that caches a different set of files or CF-BS), then a single-hop wireless backhauling is utilized to retrieve the requested file from the closest CE-BS that stores it. Our results manifest the tradeoff between caching size and intensity of CE-BSs and show that there exists an optimal balance that maximizes the hit probability at the network edge. Such optimal balance depends on the network parameters such as the applied caching strategy, required rate, total intensity of BSs, popular content distribution, and cache size/intensity relationship.

II. System Model

A. Network Model

We consider a dense mmFN where the BSs are distributed in $\mathbb{R}^2$ according to a PPP $\mathbf{\Psi}$ with intensity λ. To utilize the tradeoff between the cache size and intensity of CE-BSs, caches are not installed in all BSs. Instead, a BS is assumed to have a cache of size $M \propto \frac{1}{\zeta^{\epsilon}}$ with probability $0 < \zeta \leq 1$, where $\epsilon \geq 1$ due to the high cost of cache casing and installation. The cache implementation probability ζ is assumed to be independent across all BSs. Utilizing the independent thinning of the PPP, the CE-BSs constitute a PPP $\mathbf{\Psi}^c \subseteq \mathbf{\Psi}$ with intensity $\zeta\lambda$.

We focus on a mmW downlink network with universal frequency reuse for network access (i.e., BS to UE) and backhauling (i.e., BS to BS) links [9]. All BSs transmit with the same power level of P_t. Without loss of generality, we focus on a UE located at the origin, which becomes the typical user after spatial averaging [20].

[1]Due to the significant advancement in storage devices along with the high cost of casing, N separately cased caches of a certain capacity (e.g., M Gigabytes) is much more expensive than a single cache with the aggregated (i.e., NM Gigabytes) capacity. Hence, the inverse proportionality is non-linear. While any relationship between the cache sizes and intensity of CE-BSs can be incorporated to the developed model, without loss of generality we choose the power-law inverse proportionality with an exponent ($\epsilon \geq 1$).

B. Directional Beamforming

The small wavelength of mmW signals enables conveniently sized and highly directive antenna arrays, which can be included into small BSs as well as users equipments (UEs). From the wireless link level, high antenna directivity is critical for compensating the poor propagation properties of mmW signals. At the network level, narrow beamwidth communications mitigate the mutual interference between BSs, which is crucial in ultra-dense deployment [5].

In this work, we assume that both the BSs and UEs are equipped with antenna arrays but with different sizes. The antenna array gains of the BSs and UEs are, respectively, denoted as G_{bs} and G_u, where $G_{bs} > G_u$ due to the more stringent size constraints of UEs. For simplicity, we follow [8]–[11] and assume a two-state beam pattern for the employed antenna arrays, which is expressed as:

$$G_j(\phi) = \begin{cases} G_j^{\max}, & \text{if } |\phi| \leq \triangle\phi_j, \\ G_j^{\min}, & \text{if } |\phi| > \triangle\phi_j, \end{cases} \tag{1}$$

where $j \in \{bs, u\}$, $\phi \in [-\pi, \pi)$ is the deviation angle from the antenna boresight, $\triangle\phi_j$ denotes the beamwidth of the mainlobe, $G_j^{\max}$ and $G_j^{\min}$ are the antenna gains of main and side lobes, respectively. During UE to BS association, perfect antenna alignment between the UE and intended BS is assumed. However, the beams of all interfering links are received at random orientation with respect to each other as well as with respect to the intended receiver.

C. Propagation Model

For an arbitrary distance r between the transmitting and receiving nodes, the received power is given by $P_t\,\psi\,\ell(r)$ where ψ is the effective directivity gain due to the relative orientations of the transmit and receive antennas, and $\ell(\text{dB})$ is the distance dependent path loss. The path-loss is expressed as $\ell(\text{dB}) = \delta + 10\alpha\log(r) + \vartheta$, where δ is the path loss at a close-in reference distance, α is the path loss exponent, and $\vartheta \sim \mathcal{N}(0, \eta^2)$ is the normally distributed (i.e., log-normal in the absolute scale) shadowing with variance η^2. Motivated by the studies in [9] and [10], which assume each of the access and backhaul links has its own propagation parameters denoted, respectively, via the tuples $(\delta_a, \alpha_a, \eta_a^2)$ and $(\delta_b, \alpha_b, \eta_b^2)$.

D. Blockage Model

Following the field measurements and stochastic blockage models in [6], [7], we assume that the probability that a link is line of sight (LOS) is $e^{-\rho r}$, where the decaying rate ρ depends on the building parameters and density. For analytical tractability, the LOS probability function can be approximated by step functions, where the irregular geometry of the LOS region is replaced via a LOS ball that would lead to an equivalent distribution for the signal-to-interference-plus-noise-ratio (SINR). It is worth pointing out that such step functions approximation is utilized and validated in [21], [22]. In this paper, we adopt a single-ball

model with transition radius of R_l such that all link distances of $r < R_l$ can be LOS with probability $\mathcal{P}_l$ and NLOS with probability $(1 - \mathcal{P}_l)$. On the other hand, the transmission link of distance $r \geq R_l$ is considered non-LOS (NLOS) with probability one.

E. Content Popularity and Caching Models

We consider a finite library of popular files (contents),[2] denoted by $\mathbf{J} = \{c_1, c_2, ..., c_J\}$. It is assumed that all files have the same length. However, this analysis can be still applied for files of different sizes by chopping each file into equal length packets. We assume that the files popularity is fully known a priori, via big data analytics or machine learning techniques, for the network operator.[3] The content popularity is assumed to follow the Zipf distribution due to its practical relevance [27]. The files popularity distribution is expressed as:

$$a_j = \frac{j^{-\gamma}}{\sum_{i=1}^{J} i^{-\gamma}}, \tag{2}$$

where a_j is the probability that a UE requests the file c_j, and γ is the Zipf parameter that governs the popularity distribution skewness. Larger (smaller) γ increases (decreases) the discrepancies among the files popularity and implies that fewer (more) files are frequently requested. Without loss of generality, it is assumed that the files of the library are enumerated in a descending order of their popularity, i.e., $a_1 \geq a_2 \geq \cdots \geq a_J$.

As discussed before, the cache size is assumed to be inversely proportional to the intensity of CE-BSs $0 < \zeta\lambda \leq \lambda$. The cache size is given as $M = \left\lfloor \frac{M_o}{\zeta^\epsilon} \right\rfloor$ files, where M_o is the proportionality constant (i.e., the smallest cache capacity for a given CAPEX when $\zeta = 1$), and $\lfloor x \rfloor$ is the floor operator. Note that the extreme case of $\zeta = 0$ is eliminated from this paper because it depicts a scenario where no caches are installed in the network, which is out of the scope of this paper. Instead, we assume that the minimum fraction of CE-BSs ζ corresponds to the case where the capacity of each installed cache is large enough to store all the popular J-files. Consequently, $(\frac{M_o}{J})^{1/\epsilon} \leq \zeta \leq 1$ and $M_o \leq M \leq J$.

We adopt a probabilistic file placement technique, where the probability that a CE-BS stores the file c_j is $b_j (0 \leq b_j \leq 1)$, $\forall c_j \in \mathbf{J}$. To avoid duplicate file caching, we adopt the probabilistic caching strategy proposed in [12]. For any tuple (M, J, b_j), M equal rows of unit size are sequentially filled with the probabilities b_j. If the full capacity of a row is reached before encompassing a given b_j, the remaining portion of that b_j is continuously filled in the next row. Finally, a random number within $[0, 1]$ is selected to determine the selected file combination. Since $\sum_{j=1}^{J} b_j = M$, all popular files are considered within such caching strategy. Fig. 1 illustrates an example of the adopted probabilistic content placement strategy for $M = 4$ and $J = 10$, where a vertical line uniformly placed in the region [0,1] is used to choose the $M \leq J$ files to be stored in the CE-BS. It is clear from the illustrative example that the event of storing a file c_j occurs according to the probability b_j and that file duplication is alleviated.

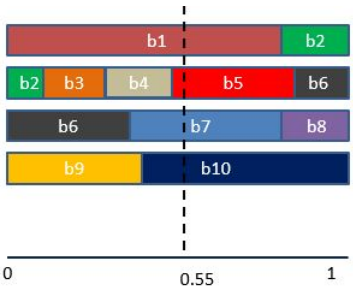


Fig. 1: An illustration of the probabilistic caching method [12] with $M = 4$ and $J = 10$. The set of stored files is $\{c_1, c_5, c_7, c_{10}\}$

F. Association model and self backhauling

Due to the poor propagation properties of mmW signals, each UE is associated to its nearest BS that offers minimum path loss irrespective of its cache status (i.e., CE-BS or CF-BS). Let x_a^* denote the serving BS for a UE that requests the popular file c_j, where the subscript a denotes the access link. Let the thinned PPP $\mathbf{\Psi}_j^c \subseteq \mathbf{\Psi}^c$ of intensity $\lambda\zeta b_j$ denote the set of CE-BSs that store c_j. Based on the proposed system model, the UE requesting c_j is served with one of the following alternatives:

(i) If $x_a^* \in \mathbf{\Psi}_j^c$ is CE-BS and stores the file c_j, then the requested file is served to the UE from the serving BS cache. This event occurs with probability ζb_j.

(ii) The serving BS does not store the file c_j if i) x_a^* is CF-BS, which occurs with probability $(1 - \zeta)$; or ii) x_a^* is a CE-BS but does not store c_j, which occurs with probability $\zeta(1 - b_j)$. In either of these cases, the requested file is retrieved via the wireless backhaul from the CE-BS that stores c_j and offers minimum path loss at x_a^*. Such backhauling CE-BS is denoted as x_b^*. Hence, when the serving BS does not store the requested file, a two hop transmission (i.e., backhaul from x_b^* to x_a^* then access from x_a^* to the UE) is utilized to serve the UE request.

Fig. 2 illustrates the association and self-backhauling policies employed in this paper.

G. Performance Metric

The hit probability, defined as the probability that the typical UE successfully downloads the requested file from the network edge (i.e., through the CE-BS in one or two transmission hops), is the considered performance metric. Note that a successful file transmission requires that the received SINR to be above a certain threshold β. Based on the two aforementioned alternatives (i.e., single and two

[2]Popular content can be popular videos, highly accessed websites, augmented/virtual reality content, software updates, or mobile applications

[3]This assumption is commonly used in the literature [12]–[14], [23]–[26]. Such assumption is consistent with the fact that the rate at which the file popularity changes is most likely much slower than the rate they are requested with, otherwise there is no notion of popularity. Addressing estimation errors and time-varying characteristics of the file popularity is beyond the scope of this work.

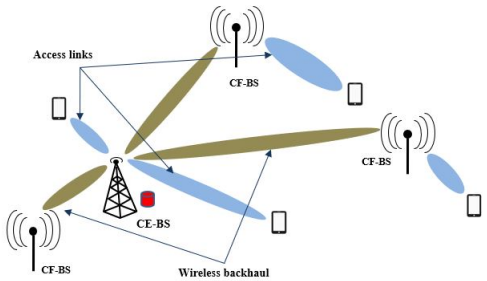


Fig. 2: Self-backhauled network with the CE-BS providing wireless backhaul to the tagged BSs and access link to the tagged users

hops) for serving the typical UE, the hit probability can be expressed as

$$\mathcal{H} = \sum_{j=1}^{J} a_j \Big(\zeta b_j \mathbb{P}[\text{SINR}_a > \beta] + (1 - \zeta b_j) \mathbb{P}[\text{SINR}_b > \beta,\, \text{SINR}_a > \beta] \Big),$$
$$\overset{(a)}{=} \sum_{j=1}^{J} a_j \left(\zeta b_j S_a(\beta) + (1 - \zeta b_j) S_a(\beta) S_b(\beta) \right), \quad (3)$$

where SINR_a and SINR_b correspond to the SINR of the access and backhaul links, respectively; $S_a(\beta) = \mathbb{P}[\text{SINR}_a > \beta]$ and $S_b(\beta) = \mathbb{P}[\text{SINR}_b > \beta]$ are the successful transmissions probabilities, denoted hereafter as the coverage probabilities, of the access and backhaul links, respectively. The equality (a) is obtained by the fact that $S_a(\beta)$ and $S_b(\beta)$ are independent events.

III. The Hit Probability Analysis

This section characterizes the hit probability for the self-backhauled mmFN, where the tranmission success probabilities $S_a(\beta)$ and $S_b(\beta)$ are derived in Section III-A. Then, the cache size and file placement probabilities b_j are optimized in Section III-B.

A. SNR distribution

Thanks to the narrow-beam mmW transmissions, the mmFN is most likely noise-limited [3], [9]–[11]. Therefore, the SINR in (3) is replaced with the signal-to-noise-ratio (SNR). To characterize the SNR distribution, we define the one dimensional point process seen by the typical UE for the access network. Such 1D point process is constructed by mapping $\mathbf{\Psi}$ to $\mathbb{R}$, where the mapping function is the access path loss function. Hence, the access 1D point process is defined as $\mathcal{N}_a = \{\ell_a(x) = \frac{|x|^\alpha}{s}\}$, where $s = \{e^{-0.1\delta \ln 10 - 0.1\vartheta \ln 10}\}_{x \in \mathbf{\Psi}}$ which is log normal distributed random variable, i.e., $s \sim \ln\mathcal{N}(m, \sigma)$ with mean $m = -0.1\delta \ln 10$ and variance $0.1\eta \ln 10$. Using the mapping and displacement theorems [28, Section 2.7], the intensity measure for the point process $\mathcal{N}_a$, denoted by $\Lambda_a([0,t))$ is characterized by the following lemma.

Lemma 1: The distribution of the path loss of the access link between the typical UE and its serving BS x_a^*, defined as $\mathbb{P}\left[\ell_a(x^*) > t\right] = \exp\left(-\Lambda_a([0,t))\right)$, is obtained using the avoid property of the PPP process. The intensity measure $\Lambda_a([0,t))$ is given by (4) at the top of next page, with $m_j = -0.1\delta_j \ln 10$ and $\sigma_j = 0.1\eta_j \ln 10$, with $j \equiv l$ for LOS and $j \equiv n$ for NLOS link and $Q(.)$ is the Q-function.

Proof: Please refer to Appendix A ■

Therefore, the coverage probability for both the access and backhaul links are given by

$$S_a(\beta) = \mathbb{P}\left[SNR_a > \beta\right] = \mathbb{P}\left[\frac{P_t \psi_a \ell^{-1}(x_a^*)}{\sigma_n^2} > \beta\right] = 1 - \exp\left(-\Lambda_a\left(\left[0, \frac{P_t\psi_a}{\sigma_n^2\beta}\right)\right)\right)$$
$$S_b(\beta) = \mathbb{P}\left[SNR_b > \beta\right] = 1 - \exp\left(-\Lambda_b\left(\left[0, \frac{P_t\psi_b}{\sigma_n^2\beta}\right)\right)\right) = 1 - \exp\left(-\zeta b_j F_b(\beta)\right) \quad (5)$$

Note that $\Lambda_b([0,t))$ is obtained following a similar methodology as in Lemma 1 but by considering the backhaul path loss mapping function, where the mapped point process is $\mathbf{\Psi}_j^c$ instead of $\mathbf{\Psi}$. Hence, the intensity measure $\Lambda_b([0,t))$ is similar to (4) with replacing λ by $\zeta b_j \lambda$ and the access link parameters by the backhaul link parameters.

B. Cache Size & Caching Probabilities

1) Optimal caching scheme: The joint optimal caching distribution $b_j^*, \forall j \in \mathbf{J}$ and cache-enabled fraction ζ^* that maximizes the hit probability can be obtained via solving the following formulation:

$$\max_{\mathbf{b},\zeta} \; S_a(\beta) \sum_{j=1}^{J} a_j \left[\zeta b_j + (1 - \zeta b_j)\left(1 - e^{-\zeta b_j F_b(\beta)}\right)\right] \quad (6a)$$
$$\text{s.t.} \quad 0 \le b_j \le 1, \quad j = 1, 2, \ldots, J \quad (6b)$$
$$\sum_{j=1}^{J} b_j = M = \left\lfloor \frac{M_o}{\zeta^\epsilon} \right\rfloor \quad (6c)$$
$$\left(\frac{M_o}{J}\right)^{1/\epsilon} \le \zeta \le 1 \quad (6d)$$

The condition in (6c) leads to a discrete optimization problem. To guarantee an integer cache size M, the cache-enabled fraction $\zeta \in [\left(\frac{M_o}{J}\right)^{1/\epsilon}, 1]$ is chosen such that $\frac{M_o}{\zeta^\epsilon} \in \mathbb{Z}^+$. We propose solving this problem according to the following iterative algorithm.

i) Initialize ζ within the range $\left(\frac{M_o}{J}\right)^{1/\epsilon} \le \zeta \le 1$.
ii) Given ζ, find the optimal caching distribution $b_j^*, \forall j \in \mathbf{J}$ that maximizes the hit probability.
iii) For the optimal caching distribution b_j^*, find the optimal fraction ζ^*, within the range $\left(\frac{M_o}{J}\right)^{1/\epsilon} \le \zeta \le 1$, that maximizes the hit probability.
iv) Iterate between points ii) and iii) until convergence.

For known ζ, the optimization problem in (6) turns to the following formulation:

$$\max_{\mathbf{b}} \quad S_a(\beta) \sum_{j=1}^{J} a_j \left[1 - (1 - \zeta b_j) e^{-\zeta b_j F_b(\beta)}\right] \quad (7a)$$
$$\text{s.t.} \quad 0 \le b_j \le 1, \quad j = 1, 2, \ldots, J \quad (7b)$$
$$\sum_{j=1}^{J} b_j = \left\lfloor \frac{M_o}{\zeta^\epsilon} \right\rfloor = M \quad (7c)$$

$$\Lambda_a([0,t)) = \pi\lambda\left\{t^{2/\alpha_n}\exp\left(2(\frac{\sigma_n}{\alpha_n})^2+\frac{2m_n}{\alpha_n}\right)\left(1-\mathcal{P}_l\,Q\left(\frac{\sigma_n^2+m_n-\ln\left(\frac{R_l^{\alpha_n}}{t}\right)}{\sigma_n}\right)\right)-\mathcal{P}_l R_l^2 Q\left(\frac{\ln\left(\frac{R_l^{\alpha_n}}{t}\right)-m_n}{\sigma_n}\right)\right.$$
$$\left.+\,\mathcal{P}_l\left[t^{2/\alpha_l}\exp\left(2(\frac{\sigma_l}{\alpha_l})^2+\frac{2m_l}{\alpha_l}\right)Q\left(\frac{\sigma_l^2+m_l-\ln\left(\frac{R_l^{\alpha_l}}{t}\right)}{\sigma_l}\right)+R_l^2 Q\left(\frac{\ln\left(\frac{R_l^{\alpha_l}}{t}\right)-m_l}{\sigma_l}\right)\right]\right\} \quad (4)$$

It is easy to show the concavity of the objective function $\mathcal{H}$ by confirming that the first derivative is ≥ 0 and the second derivative is ≤ 0. Also, the constraints are linear, which imply that the necessity and sufficiency conditions for optimality exist. The KKT Lagrangian function of this problem is given by:

$$L(\mathbf{b},\mathbf{w},\mu,v) = S_a\sum_{j=1}^{J}a_j\left[1-(1-\zeta b_j)\,e^{-\zeta b_j F_b}\right]-\sum_{j=1}^{J}\mu_j\ b_j$$
$$+\sum_{j=1}^{J}w_j(b_j-1)+v(M-\sum_{j=1}^{J}b_j). \quad (8)$$

For brevity, we omit the details of finding the optimal caching distribution b_j^*. Thus, b_j^* at a given cache-enabled fraction ζ is given by

$$b_j^* = \begin{cases} 0 & ,v^* < a_j S_a \zeta e^{-\zeta F_b}\left(1+F_b\left(1-\zeta\right)\right) \\ 1 & ,v^* > a_j S_a \zeta (1+F_b) \\ \Omega(v^*) & ,\text{otherwise} \end{cases}, \quad (9)$$

where $\Omega(v^*)$ is the solution of $v^* = a_j S_a \zeta e^{-\zeta b_j^* F_b}\Big[1+F_b\left(1-\zeta b_j^*\right)\Big]$ that satisfies $\sum_{j=1}^{J} b_j^* = \left\lfloor \frac{M_o}{\zeta^\epsilon} \right\rfloor$.

Note that the optimal caching b_j^* can be obtained using bisection method in similar to [12, Algorithm].

2) Most popular files caching scheme: Another widely accepted and simpler file placement strategy is the most popular file caching (MPC) scheme. Given the descending popularity order of the files indices, the MPC scheme stores all files with indices $j \leq M$ with probability one. Hence, files with indices $j > M$ are never cached at the network edge. The caching probabilities for the MPC scheme are given by

$$b_j = \begin{cases} 1, & j = 1 : \lfloor \frac{M_o}{\zeta^\epsilon} \rfloor \\ 0, & j = \lfloor \frac{M_o}{\zeta^\epsilon} \rfloor + 1 : J \end{cases} \quad (10)$$

Therefore the optimal cache size for the MPC can be obtained via the following one parameter optimization problem

$$\max_{\zeta}\ \ S_a\left(\beta\right)\sum_{j=1}^{\lfloor \frac{M_o}{\zeta^\epsilon} \rfloor} a_j\left[\zeta+(1-\zeta)\left(1-e^{-\zeta F_b(\beta)}\right)\right] \quad (11a)$$

$$\text{s.t.} \quad \left(\frac{M_o}{J}\right)^{1/\epsilon} \leq \zeta \leq 1 \quad (11b)$$

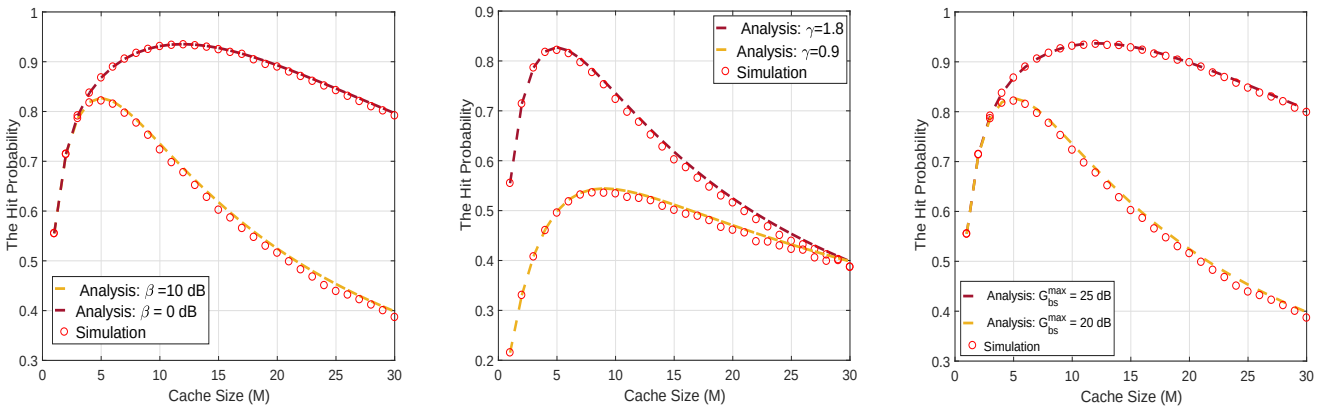
which can be solved numerically via looking through all feasible values of ζ.

IV. Numerical Results

This section validates the developed mathematical model via Monte Carlo simulations. In each simulation run, a PPP of intensity $\lambda = 200$ BSs/Km^2 is generated in a 10×10 km^2 area. Caches with capacities $M = \left\lfloor \frac{M_o}{\zeta^\epsilon} \right\rfloor$ are installed independently at the BSs with probability ζ. CE-BSs independently cache the popular files according to the considered scenarios, namely the optimized caching and MPC schemes. The UEs requests are realized using a Zipf distribution with parameter γ. Each UE is associated to the nearest BS that offers minimum path loss. If the serving BS does not store the requested file by the test UE at the origin, the serving BS utilizes the backhaul to retrieve the requested file from the closest CE-BS that stores it. The simulation is repeated 10^3 times and the hit probability is recored for the test UE. Unless otherwise stated, the mmW network parameters are selected according to [9, Table I] as follows. The mmW frequency and bandwidth are set to $f_c = 73$ GHz and $B = 2$ GHz, respectively. The BSs transmit power is $P_t = 30$ dBm, the standard deviation of the path-loss for the access link is $\eta^a_{\{l,n\}} = \{5.2, 7.6\}$, for the backhaul link is $\eta^b_{\{l,n\}} = \{4.2, 7.9\}$, the path loss exponent of the access link is $\alpha^a_{\{l,n\}} = \{2, 3.3\}$, for the backhaul link is $\alpha^b_{\{l,n\}} = \{2, 3.5\}$, where the subscripts l and n denote the LOS and NLOS parameters, respectively. The path loss at reference distance $d_o = 1$m is $\delta_{dB} = 20\log(\frac{4\pi d_o}{\nu}) = 70$ dB, where ν denotes the wavelength. The antennas parameter for the BSs is $\{G_{bs}^{max} = 20$ dB, $G_{bs}^{min} = -2$ dB and $\triangle\phi_{bs} = 10^\circ\}$ and for the UEs is $\{G_u^{max} = 10$ dB, $G_u^{min} = -5$ dB and $\triangle\phi_u = 20^\circ\}$. The single-ball blockage model parameters are $R_l = 200$ m and $\mathcal{P}_l = 0.1$. The caching parameters are chosen as follows; The library size $J = 30$, the smallest cache capacity when $\zeta = 1$, i.e., $M_o = 1$, and the Zipf exponent $\gamma = 1.8$.

Fig. 3 shows the hit probability versus the cache size $M = \left\lfloor \frac{M_o}{\zeta^\epsilon} \right\rfloor$ for the MPC caching scheme. It is worth noting that the matching between the simulation and analytical results verifies the proposed mathematical framework. Fig. 3 also manifests the cache size/intensity trade-off and indicates that there exists an optimal ζ^* that maximizes the hit probability. In Fig. 3a, it can be shown that a higher SNR threshold β requires high intensity of caches, despite their lower capacity of each, due to the low success probability for retrieving files through self-backhauling, and vice versa. Fig. 3b shows that decreasing the Zipf exponent requires higher capacities of caches, despite their lower implementation intensities, to satisfy the high number of frequently requested files. Fig. 3c illuminates the fact that increasing the BSs antenna gain improves the backhauling efficiency to share files at high SNR thresholds, which allows lower intensity of caches with higher sizes to improve the hit probability.

The optimal caching compared with the MPC is shown in Fig.4. The figure both validates the developed mathematical framework and shows that diversifying the file placement



(a) $\gamma = 1.8$ and $G_{bs}^{max} = 20$ dB. (b) $\beta = 10$ dB and $G_{bs}^{max} = 20$ dB. (c) $\gamma = 1.8,$ and $\beta = 10$ dB.

Fig. 3: The hit probability vs. the cache size ($M = \lfloor \frac{M_o}{\zeta} \rfloor$) adopting MPC caching scheme with ($G_u^{max} = 10$ dB, $J = 30$, $M_o = 1$, $\lambda = 200$ BSs/Km^2, and $\epsilon = 1$).

through optimal caching leads to higher hit probability. However, the gain of the optimal caching diminishes with large cache size as the effective popular contents are already stored by both schemes. Fig. 4 also shows that the optimal cache size (i.e., hence the optimal CE-BSs intensity) can be obtained through a plot over a single parameter over the range $M_o \leq M \leq 1$.

V. Conclusion

This paper develops a mathematical framework to study the cache capacity/intensity tradeoff, for a given CAPEX, in a dense self-backhauled mmW fog network. Instead of implementing a cache with a lower capacity at each BS, it may be better to implement larger caches in a fraction of the BSs and share the caches content across BSs through wireless backhaul. To study such tradeoff, the hit probability in the depicted self-backhauled mmW fog network is characterized via stochastic geometry. Assuming an inverse power-law relationship between the cache size and intensity, an optimization problem is formulated to find the optimal cache intensity and file placement in order to maximize the hit probability. The results show that there exists an optimal balance between the cache size and the intensity of cache enabled BSs, which depends on the network parameters, for the optimized and most popular file caching schemes.

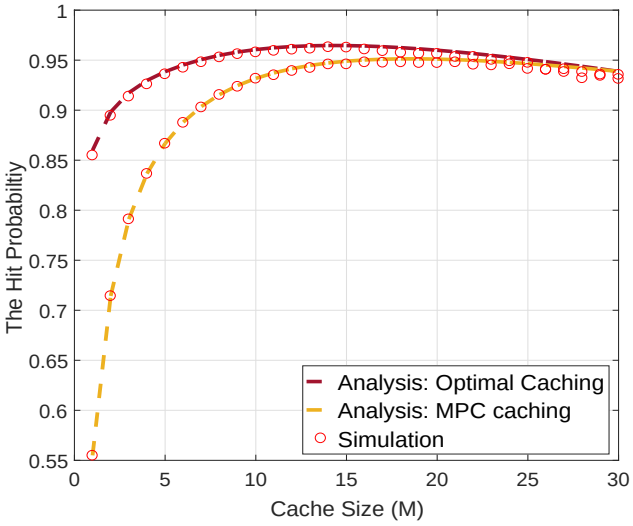


Fig. 4: The hit probability vs. the cache size for both the MPC and optimal caching schemes with ($J = 30$, $M_o = 1$, $\gamma = 1.8$, $\lambda = 200$ BSs/Km^2, $\beta = 10$ dB, and $\epsilon = 2$).

Appendix A
Proof of Lemma 1

Recall the point process $\{\mathcal{N}_a = \ell_a(x) = \frac{|x|^\alpha}{s}\}_{x\in\Psi}$ on $\mathbb{R}$ formed by the path loss from each BS to the typical user at the origin. Using the displacement theorem, the intensity measure $\Lambda_a((0,t])$ can be expressed by

$$\Lambda_a([0,t)) = \int_{\mathbb{R}^2} \mathbb{P}[\ell(x) < t]\mathrm{d}x = 2\pi\lambda \int_{\mathbb{R}} \mathbb{P}\left[\frac{r^{\alpha(r)}}{s(r)} < t\right] r\mathrm{d}r$$
$$= 2\pi\lambda \Bigg\{ \sum_{j\in\{l,n\}} \mathcal{P}_j \int_{\mathbb{R}} \mathbb{P}\left[r < (ts_j)^{1/\alpha_j}\right] \mathbb{1}_{(r<R_l)}\, r\mathrm{d}r$$
$$+ \int_{\mathbb{R}} \mathbb{P}\left[r < (ts_n)^{1/\alpha_n}\right] \mathbb{1}_{(r\geq R_l)}\, r\mathrm{d}r \Bigg\} \qquad (12)$$

where $\mathcal{P}_j = \{\mathcal{P}_l, 1-\mathcal{P}_l\}$ for $j \in \{l,n\}$. Therefore,

$$\Lambda_a([0,t))$$
$$= 2\pi\lambda\, \mathbb{E}_s \Bigg\{ \sum_{j\in\{l,n\}} \mathcal{P}_j \Bigg[\int_0^{(ts_j)^{1/\alpha_j}} \mathbb{1}_{(s_j < R_l^{\alpha_j}/t)}\, r\mathrm{d}r + \int_0^{R_l} r$$
$$\mathbb{1}_{(s_j > R_l^{\alpha_j}/t)}\mathrm{d}r \Bigg] + \int_{R_l}^{(ts_n)^{1/\alpha_n}} \mathbb{1}_{(s_n > R_l^{\alpha_n}/t)}\, r\mathrm{d}r \Bigg\}$$
$$= \pi\lambda \Bigg\{ \sum_{j\in\{l,n\}} \mathcal{P}_j \left[t^{2/\alpha_j} \Upsilon_{s_j,2/\alpha_j}(R_l^{\alpha_j}/t) + R_l^2 \bar{F}_{s_j}(R_l^{\alpha_j}/t) \right]$$
$$+ t^{2/\alpha_n} \bar{\Upsilon}_{s_n,2/\alpha_n}(R_l^{\alpha_n}/t) - R_l^2 \bar{F}_{s_n}(R_l^{\alpha_n}/t) \Bigg\} \qquad (13)$$

where $\Upsilon_{s,n}(x) = \int_0^x s^n f_s(s)\mathrm{d}s$ and $\bar{\Upsilon}_{s,n}(x) = \int_x^\infty s^n f_s(s)\mathrm{d}s$ denote the lower and upper truncated n^{th}

moment of s, respectively. $\bar{F}_s(x) = \int_x^\infty f_s(s)\mathrm{d}s$ represents the counterpart cumulative density function (CCDF) of s at x. Recall that $s \sim \ln\mathcal{N}(m, \sigma)$ with mean $m = -0.1\beta \ln 10$ and variance $\sigma = 0.1\vartheta \ln 10$, $\Upsilon_{s,n}(x)$ and $\bar{F}_s(x)$ are given by

$$\Upsilon_{s,n}(x) = \exp(\sigma^2 n^2/2 + nm)\, Q(\frac{\sigma^2 n + m - \ln x}{\sigma}),$$
$$\bar{\Upsilon}_{s,n}(x) = \exp(\sigma^2 n^2/2 + nm)\, Q(-\frac{\sigma^2 n + m - \ln x}{\sigma}),$$
$$\bar{F}_s(x) = Q(\frac{\ln x - m}{\sigma}). \quad (14)$$

By substituting (14) into (12), lemma 1 can be obtained.

ACKNOWLEDGMENT

This research was funded by a grant from the office of competitive research funding (OCRF) at the King Abdullah University of Science and Technology (KAUST).

The work was also supported by the Deanship of Scientific Research (DSR) at King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia, through project number KAUST-002.

REFERENCES

[1] M. S. Elbamby, C. Perfecto, M. Bennis, and K. Doppler, "Edge computing meets millimeter-wave enabled vr: Paving the way to cutting the cord," in *Wireless Communications and Networking Conference (WCNC), 2018 IEEE*. IEEE, 2018, pp. 1–6.

[2] L. Wang, K. Wong, S. Jin, G. Zheng, and R. W. H. Jr., "A new look at physical layer security, caching, and wireless energy harvesting for heterogeneous ultra-dense networks," *CoRR*, vol. abs/1705.09647, 2017. [Online]. Available: http://arxiv.org/abs/1705.09647

[3] J. G. Andrews, T. Bai, M. N. Kulkarni, A. Alkhateeb, A. K. Gupta, and R. W. Heath, "Modeling and analyzing millimeter wave cellular systems," *IEEE Transactions on Communications*, vol. 65, no. 1, pp. 403–430, 2017.

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

[5] A. AlAmmouri, J. G. Andrews, and F. Baccelli, "Sinr and throughput of dense cellular networks with stretched exponential path loss," *IEEE Transactions on Wireless Communications*, vol. 17, no. 2, pp. 1147–1160, 2018.

[6] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5g cellular: It will work!" *IEEE access*, vol. 1, pp. 335–349, 2013.

[7] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE journal on selected areas in communications*, vol. 32, no. 6, pp. 1164–1179, 2014.

[8] T. Bai and R. W. Heath, "Coverage and rate analysis for millimeter-wave cellular networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 2, pp. 1100–1114, 2015.

[9] S. Singh, M. N. Kulkarni, A. Ghosh, and J. G. Andrews, "Tractable model for rate in self-backhauled millimeter wave cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 10, pp. 2196–2211, 2015.

[10] M. Di Renzo, "Stochastic geometry modeling and analysis of multi-tier millimeter wave cellular networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 9, pp. 5038–5057, 2015.

[11] E. Turgut and M. C. Gursoy, "Coverage in heterogeneous downlink millimeter wave cellular networks," *IEEE Transactions on Communications*, vol. 65, no. 10, pp. 4463–4477, 2017.

[12] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in *2015 IEEE International Conference on Communications (ICC)*. IEEE, 2015, pp. 3358–3363.

[13] B. Serbetci and J. Goseling, "On optimal geographical caching in heterogeneous cellular networks," in *Wireless Communications and Networking Conference (WCNC), 2017 IEEE*. IEEE, 2017, pp. 1–6.

[14] M. Afshang and H. S. Dhillon, "Optimal geographic caching in finite wireless networks," in *Signal Processing Advances in Wireless Communications (SPAWC), 2016 IEEE 17th International Workshop on*. IEEE, 2016, pp. 1–5.

[15] M. Emara, H. E. Sawy, S. Sorour, S. Al-Ghadhban, M. S. Alouini, and T. Y. Al-Naffouri, "Optimal caching in 5g networks with opportunistic spectrum access," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2018.

[16] M. Emara, H. ElSawy, S. Sorour, S. Al-Ghadhban, M.-S. Alouini, and T. Y. Al-Naffouri, "Stochastic geometry model for multi-channel fog radio access networks," in *Modeling & Optimization in Mobile, Ad Hoc & Wireless Networks (WiOpt), 2017 15th International Symposium on*. IEEE, 2017, pp. 1–6.

[17] W. Yi, Y. Liu, and A. Nallanathan, "Modeling and analysis of mmwave communications in cache-enabled hetnets," *arXiv preprint arXiv:1801.08801*, 2018.

[18] Y. Zhu, G. Zheng, L. Wang, K. Wong, and L. Zhao, "Content placement in cache-enabled sub-6 ghz and millimeter-wave multi-antenna dense small cell networks," *CoRR*, vol. abs/1801.05756, 2018. [Online]. Available: http://arxiv.org/abs/1801.05756

[19] Y. Zhu, G. Zheng, K.-K. Wong, S. Jin, and S. Lambotharan, "Performance analysis of cache-enabled millimeter wave small cell networks," *IEEE Transactions on Vehicular Technology*, 2018.

[20] S. N. Chiu, D. Stoyan, W. S. Kendall, and J. Mecke, *Stochastic geometry and its applications*. John Wiley & Sons, 2013.

[21] W. Lu and M. Di Renzo, "Stochastic geometry modeling of cellular networks: Analysis, simulation and experimental validation," in *Proceedings of the 18th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*. ACM, 2015, pp. 179–188.

[22] M. Ding, P. Wang, D. López-Pérez, G. Mao, and Z. Lin, "Performance impact of los and nlos transmissions in dense cellular networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 3, pp. 2365–2380, 2016.

[23] S. H. Chae and W. Choi, "Caching placement in stochastic wireless caching helper networks: Channel selection diversity via caching," *IEEE Transactions on Wireless Communications*, vol. 15, no. 10, pp. 6626–6637, 2016.

[24] Y. Chen, M. Ding, J. Li, Z. Lin, G. Mao, and L. Hanzo, "Probabilistic small-cell caching: Performance analysis and optimization," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 5, pp. 4341–4354, 2017.

[25] Y. Cui, D. Jiang, and Y. Wu, "Analysis and optimization of caching and multicasting in large-scale cache-enabled wireless networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 7, pp. 5101–5112, 2016.

[26] Y. Cui and D. Jiang, "Analysis and optimization of caching and multicasting in large-scale cache-enabled heterogeneous wireless networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 1, pp. 250–264, 2017.

[27] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM, 2007, pp. 1–14.

[28] M. Haenggi, *Stochastic geometry for wireless networks*. Cambridge University Press, 2012.