Cache-enabled HetNets With Millimeter Wave Small Cells

Wenqiang Yi, *Student Member, IEEE,* Yuanwei Liu, *Member, IEEE,* and Arumugam Nallanathan, *Fellow, IEEE,*

***Abstract*—In this paper, we consider a novel cache-enabled heterogeneous network (HetNet), where macro base stations (BSs) with traditional sub-6 GHz are overlaid by dense millimeter wave (mmWave) pico BSs. These two-tier BSs, which are modeled as two independent homogeneous Poisson Point Processes, cache multimedia contents following the popularity rank. High-capacity backhauls are utilized between macro BSs and the core server. In contrast to the simplified flat-top antenna pattern analyzed in previous articles, we employ an actual antenna model with the uniform linear array at all mmWave BSs. To evaluate the performance of our system, we introduce two distinctive user association strategies: 1) maximum received power (Max-RP) scheme; and 2) maximum rate (Max-Rate) scheme. With the aid of these two schemes, we deduce new theoretical equations for success probabilities and area spectral efficiencies (ASEs). Considering a special case with practical path loss laws, several closed-form expressions for coverage probabilities are derived to gain several insights. Monte Carlo simulations are presented to verify the analytical conclusions. We show that: 1) the proposed HetNet is an interference-limited system and it outperforms the traditional HetNets in terms of the success probability; 2) there exists an optimal pre-decided rate threshold that contributes to the maximum ASE; and 3) Max-Rate achieves higher success probability and ASE than Max-RP but it needs the extra information of the interference effect.**

***Index Terms*—Caching, heterogeneous networks, millimeter wave, stochastic geometry, user association**

I. Introduction

With the rapid development of the traditional cellular networks and novel Internet-enabled applications, such as multimedia sensors [2] and electric vehicles [3, 4], the total throughput of mobile networks in 2020 is expected to become 1000-fold larger than that in 2010 [5]. To support the explosive data traffic of future fifth-generation (5G) cellular networks, numerous researches [6–9] have paid attention to an innovative framework that densifies the traditional networks with massive small base stations (BSs). However, the improvement of these heterogeneous networks (HetNets) is mainly restricted to the capacity of the backhauls. Although the high-speed optical fiber provides a theoretical solution, in practice, connecting the core server to all BSs with fibers is arduous and costly [10]. Moreover, microwave backhauls may pessimistically weaken the throughput gain fetched by the network densification [11]. A recent study [12] has shown that only 5%-10% of multimedia contents are required by the majority of user equipments (UEs). Additionally, the storage capacity of cache-enabled devices expands promptly at a fairly low cost. Stimulated by such facts, equipping caches at all BSs for storing the most popular contents becomes a promising method to offload the data traffic rather than continuing increasing the networks' density [13, 14].

Lately, the aforementioned cache-enabled HetNets have been studied in various papers. Authors in [15] analyzed the energy efficiency and throughput of cellular networks with caches, but they only considered the small cell networks (SCNs) and BSs were modeled following a regular hexagonal grid. Since stochastic geometry is a useful tool to acquire the networks' randomness [16], modeling a tier of BSs in SCNs or HetNets with a homogeneous Poisson Point Process (HPPP) is more accurate than the traditional hexagonal scenario [17–19]. Under this condition, the throughput of multi-tier cache-enabled HetNets was discussed in [20], where BSs were modeled as mutually independent PPPs. However, the high-capacity backhauls were employed at all nodes including the macro BSs and relays, which is uneconomical in reality. Then the limitation was relaxed by assuming that only macro BSs connected the core networks through backhauls, while BSs in small cells cached the contents via wireless broadcasting [10]. Unfortunately, the further analysis on the impact of backhaul capacity was omitted, which is the key parameter when comparing with the conventional HetNets.

In addition to the network densification, another key capacity-increasing technology for boosting the throughput of future cellular networks is exploiting new spectrum bands, such as millimeter wave (mmWave) [21–23]. Comparing with the traditional sub-6 GHz networks in 4G, two distinctive characteristics of mmWave are small wavelength and the sensitivity to blockages [24]. Thanks to the short wavelength, steerable antennas with huge scales can be employed at devices to enhance the directional array gain [25]. On the other side, the sensitivity gives rise to severe penetration loss for mmWave signals when passing through building exteriors [26]. Therefore, the path loss law of non-line-of-sight (NLOS) links is substantially different from that of line-of-sight (LOS) links in mmWave communications [25, 27], and it is unrealistic to expect an outdoor-to-indoor coverage from macro mmWave BSs. To compensate the blockage-dependent loss, an ingenious hybrid network is created, where mmWave transmitters contribute to the ultra-fast data rate in short-range small cells, and sub-6 GHz BSs provide the universal coverage [28].

There exist numerous studies concentrating on the performance of mmWave communications. As discussed in cache-enabled HetNets, stochastic geometry has also been widely utilized in mmWave networks, where the locations of

W. Yi, Y. Liu, and A. Nallanathan are with Queen Mary University of London, London, UK (email: {w.yi, yuanwei.liu, a.nallanathan}@qmul.ac.uk).

Part of this work was presented in IEEE International Conference on Communications (ICC), May, USA, 2018 [1].

transceivers were modeled following PPPs [24, 29]. With the aid of such structure, the primary article [24], which employed a simplified flat-top antenna pattern, introduced a stochastic blockage model to represent the actual mmWave communication environment. In fact, this simplified model has limited ability to exactly depict several parameters of a practical antenna, such as beamwidth, front-back ratio and nulls [30]. Therefore, the authors in [31] proposed an actual antenna pattern for traditional mmWave networks. Considering the hybrid HetNets, a tractable framework with sub-6 GHz macro cells and mmWave small cells was analyzed under two user association strategies in [28]. However, the Rayleigh fading assumption is not accurate for mmWave communications because of the poor scattering feature [30]. Recent works [24, 32] presented a realistic channel model with Nakagami fading to improve the theoretical accuracy.

A. Motivation and Contribution

Although HetNets with caches have been analyzed under a variety of scenarios with traditional sub-6 GHz networks, there is still lack of articles on a hybrid system with mmWave small cells. Since mmWave has a large range of available bandwidth [33, 34] and it is able to provide fast data rate in short-distance networks [35], adopting mmWave into a dense pico tier of HetNets is a promising way to increase the throughput of 5G cellular networks. Additionally, utilizing low-cost caches at all macro and pico BSs is capable of offloading the backhaul traffic efficiently and hence providing further improvement regarding the quality of service. The other benefit of such hybrid HetNets is no mutual interferences because each tier uses totally distinctive carrier frequency. These advantages motivate us to create this paper.

In contrast to [10], we introduce fiber-connections between macro BSs and the multimedia server to evaluate the impact of backhaul capacity in cache-enabled HetNets. Then, due to the employment of mmWave, the propagation environment and antenna beamforming pattern in the small cells are replaced by Nakagami fading and actual antenna arrays, respectively. Load balancing problems in mmWave-enabled HetNets have been studied in [36, 37]. However, the optimal solutions are based on a simplified framework which ignores the randomness of network nodes. In order to enhance the generality, we use the stochastic geometry to model the locations of transceivers. Regarding the user association scheme in hybrid HetNets (mmWave plus sub-6 GHz), authors in [38] assumed that the typical user is served by the BS which offers the minimum path loss. In most instances, even mmWave transmissions have more severe path loss than sub-6 GHz scenarios, they are still able to provide faster data rate due to the huge transmit bandwidth. As a result, the maximum data rate is also an important criterion for user association, in addition to the minimum path loss. We consider both criteria in this paper. The main contributions are summarized as below:

- The success probability and area spectral efficiency (ASE) of our hybrid cache-enabled HetNets are discussed under two user association strategies: 1) *Maximum Received Power (Max-RP)*, where the requesting user chooses the macro BS or pico BS offering the maximum average biased received power[1] from all BSs containing the requested file; and 2) *Maximum Rate (Max-Rate)*, where the typical UE selects the BS, which provides the highest biased transmitting rate, from all BSs caching the desired content.
- We analyze the cache-related coverage performance of traditional sub-6 GHz macro cells and mmWave small cells with the actual antenna pattern. Furthermore, closed-form coverage probability equations for the mmWave tier and an interference-limited case with sub-6 GHz are derived. Our analytical expressions can be directly applied into other mmWave or sub-6 GHz scenarios with negligible changes.
- Different association probabilities for two considered schemes are introduced to calculate final algorithms of success probabilities. We theoretically demonstrate that the success probability of Max-RP has a positive correlation with the serving tier's biased transmit power. However, the success probability of Max-Rate scheme is independent of the two tier's transmit power and the density of macro BSs. Finally, expressions of ASEs are deduced for analyzing.
- We conclude that: 1) our cache-enabled hybrid HetNets outperform the traditional HetNets where macro BSs have no caching capacity, and Max-Rate achieves a better performance than Max-RP in terms of the success probability and ASE; 2) the proposed network is an interference-limited system due to the nature of sub-6 GHz networks and the high density of mmWave small cells; 3) there is an optimum value of rate requirement for obtaining the maximum ASE; and 4) 73 GHz is the best mmWave carrier frequency for two user association strategies because of possessing the largest antenna scale.

B. Organization

We organize the rest of our treatise as follows: In Section II, we present the system model where two-tier BSs and users in the proposed cache-enabled hybrid HetNets are modeled as three independent HPPPs. In Section III, the expressions of signal-to-interference-plus-noise-ratio (SINR) coverage probabilities for two distinctive tiers are derived with the aid of the random content placement scheme. In Section IV, we discuss two different user association strategies, based on which the algorithms of success probabilities and ASEs are deduced. In Section V, the simulation and numerical results are presented for corroborating the analytical conclusions and providing further analysis, respectively. In Section VI, we draw our conclusions.

II. System Model

A. Network Architecture

In this paper, we present a cache-enabled hybrid HetNet with two-tier BSs as shown in Fig. 1. Macro BSs, pico BSs,

[1]The motivation for considering the average received power is that the network designer is interested in the average metric at the requested user for the universal coordination [39].

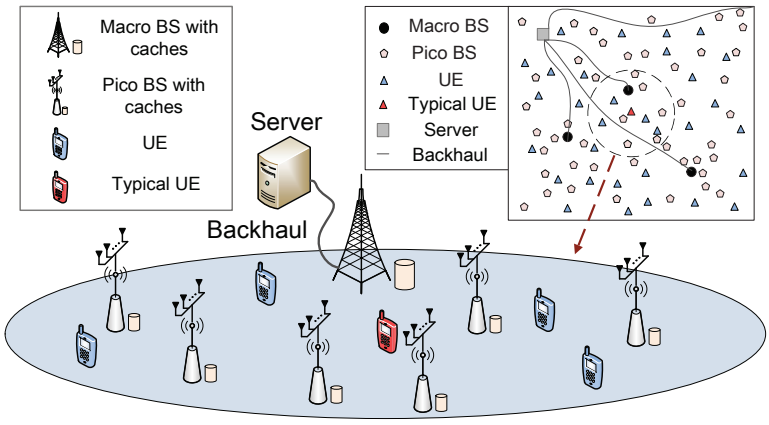


Fig. 1: Layouts of the proposed cache-enabled hybrid HetNet with traditional macro cells and mmWave small cells.

and multiple pieces of UE (UEs) are distributed following three independent HPPPs with density λ_1, λ_2, and λ_u, denoted by Φ_1, Φ_2, and Φ_u, respectively. A randomly selected *typical UE* is fixed at the origin such that the probability density function (PDF) of the distance from the typical UE to its nearest BS in the i-th tier is given by $p_i(r) = 2\pi\lambda_i r \exp(-\pi\lambda_i r^2)$, where $i \in \{1, 2\}$. Apparently, the number of pico BSs in real HetNets is much more than that of macro BSs and thus we consider $\lambda_2 \gg \lambda_1$. In order to compare the performance of the proposed network with traditional HetNets, we provide a server to supply the less-popular contents. Note that deploying wired connections between the core server and all pico BSs is wasted and arduous. We assume the server only connects to each macro BS through a high-capacity wired backhaul.

In order to avoid inter-tier interference, hybrid carrier frequencies are employed in our system. When communicating with UEs, the macro BSs adopt sub-6 GHz, while the pico BSs utilize mmWave. Note that various multiple-access techniques enable the macro BSs to serve multiple users in one time slot. We assume the quantity of UEs is large enough, namely $\lambda_u > \lambda_2 \gg \lambda_1$, to ensure all BSs are active when the typical UE is served.

B. Blockage Model

In the first tier, when the communication distance is r, the path loss law $L_1(r)$ for sub-6 GHz signals is same as that in traditional cellular networks, which is given by

$$L_1(r) = C_1 r^{-\alpha_1}, \tag{1}$$

where α_1 is the path loss exponent and C_1 is the intercept for the macro tier.

In the second tier, the effect of blockages is important due to the employment of mmWave. Therefore, we adopt a LOS ball model[2] [24, 41], as shown in Fig. 2(a), to depict the blockage environment. The radius R_L for the LOS ball represents the departure from nearby obstacles. The probability of LOS links is one inside the ball and zero outside that area. A recent study [42] has advocated that when the density of BSs is large, this blockage pattern has a negligible difference with the commonly used random shape theory model [43]. Note that we consider a dense pico tier. The simplified LOS ball model is capable of providing enough analytical accuracy. Regarding NLOS links, various articles [24, 44] have demonstrated that the impact of NLOS signals can be ignored in mmWave networks due to their severe path loss. As a result, only LOS signals are analyzed in this paper[3]. Accordingly, the path loss law in the second tier $L_2(r)$ can be expressed as follows

$$L_2(r) = \mathbf{U}(R_L - r) C_2 r^{-\alpha_2}, \tag{2}$$

where α_2 and C_2 are the path loss exponent and the intercept of LOS links, respectively. $\mathbf{U}(x)$ is the unit step function, which is defined as

$$\mathbf{U}(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}. \tag{3}$$

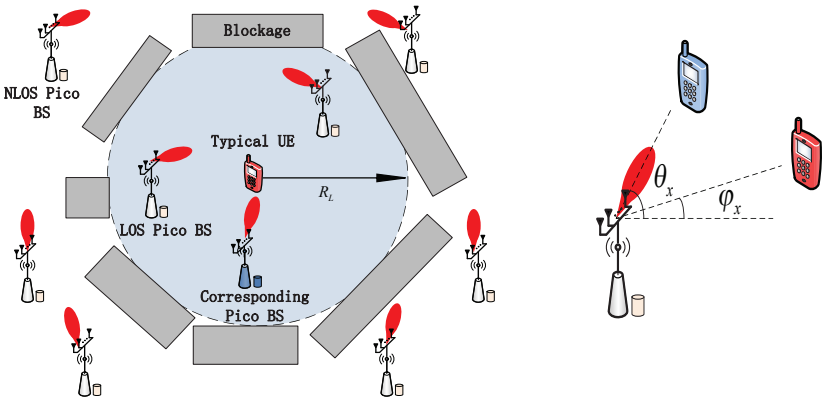
C. Cache-enabled Content Access Protocol

In this paper, we assume that a static multimedia content catalog containing N_c files is stored at the server and all files have the same size with E bits. Each macro BS and pico BS have restricted storage capacities with $(M_1 \times E)$ and $(M_2 \times E)$ bits, respectively, which means the maximum storage capacity of the proposed HetNet obeys $M_c = \max(M_1, M_2) \leq N_c$. High-speed fiber backhauls are employed for connecting the core server to macro BSs like traditional HetNets. The backhaul capacity is C_{bh}. When the data traffic load becomes low, the contents are broadcast to all BSs following the random content placement scheme as discussed in [46] until the storage is fully occupied. On this basis, we introduce the requesting probability, content placement, association strategy, and access protocol in the following part.

Requesting Probability: We apply the Zipf distribution to represent the probability of content being requested [12, 47, 48]. If files are indexed according to the popularity, namely the first and the N_c-th files are the most and the least popular

[2]In most urban scenarios, the considered LOS ball model has negligible deviation with the more accurate multi-slope LOS probability scheme, especially when the altitude of pico BSs is less than the average height of obstacles [40].

[3]Regarding the interference via NLOS transmissions, it can still be ignored since the high density of pico BSs enhances not only the interference from NLOS links but also the counterpart from LOS links [24, 32, 44, 45].



(a) Illustration of the LOS ball model utilized in the pico tier network.

(b) Illustration of two typical spatial AoDs for interfering mmWave BSs.

Fig. 2: LOS Ball Model and Directional Beamforming in The Second Tier.

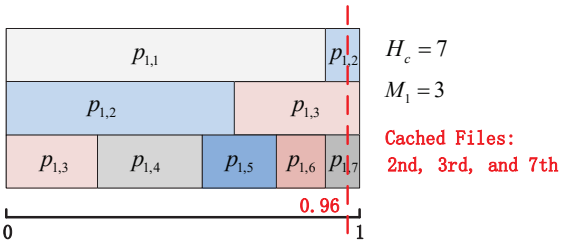


Fig. 3: An example of content placement in the macro tier, with $M_1 = 3$ and $H_c = 7$. The cache memory of one macro BS is uniformly divided into $M_1 = 3$ unit blocks (each row is one block). All blocks are sequentially filled with the probabilities from $p_{1,1}$ to p_{1,H_c}. After that, a random number between zero to one is selected to decide the cached files. In this example, the decision value is 0.96 such that the first, the third, and the 7-th files are cached in the macro tier.

contents, respectively, the requesting probability of the f-th file P_f is given by

$$P_f = \frac{f^{-\delta}}{\sum_{n=1}^{N_c} n^{-\delta}}, \quad (1 \le f \le N_c), \tag{4}$$

where f is an integer and $\delta \ge 0$ is the skew parameter of the popularity distribution.

Content Placement: We assume that the first H_c files are cached in the considered HetNet. The probability that the f-th ranked file is cached at the i-th tier is denoted by $p_{i,f}$. Based on the optimal solution presented in [49], such probability obeys:

$$\sum_{f=1}^{H_c} p_{i,f} = M_i, \ (M_c \le H_c \le N_c, 0 \le p_{i,f} \le 1). \tag{5}$$

An example in the macro tier is illustrated in Fig. 3. Note that the pico tier has the same content placement strategy but different storage capacities.

Association Strategy: We consider association strategies depending on both cached files and channel conditions, which is essentially different with traditional HetNets without caches [10]. In the i-th tier, the locations of BSs containing the f-th file $(1 \le f \le H_c)$ form a set $\Phi_{i,f}$ $(\Phi_{i,f} \subset \Phi_i)$. When the typical UE requests this file, two association strategies are used in the considered HetNet: 1) *Max-RP*, where the typical UE communicates with the BS at x_0 $(x_0 \in \{\Phi_{i,f}\}_{i=1,2})$ that provides the maximum biased average received power; and 2) *Max-Rate*, where the typical UE connects to the BS at x_0 $(x_0 \in \{\Phi_{i,f}\}_{i=1,2})$ that provides the maximum biased received data rate.

Access Protocol: If the desired f-th file is cached at the two-tier HetNet, which obeys $(1 \le f \le H_c)$, the typical UE communicates with the macro or pico BSs following the aforementioned association strategies. However, if the demanded content is absent from the BSs due to limited storage capacities, namely $(H_c < f \le N_c)$, the typical UE turns to the core server via the nearest macro BS.

D. Directional Beamforming

In the i-th tier, we employ antenna arrays composed of N_i elements at all cache-enabled BSs and the transmit power is assumed to be a constant P_i. Due to the small wavelength of mmWave signals, the uniform linear array (ULA) pattern can be deployed at all pico BSs [30]. Directional antenna arrays deployed at the pico BSs supply substantial beamforming gains to compensate the path loss. However, we only consider an omnidirectional antenna model at macro BSs and UEs for tractability of the analysis [28]. When the typical UE requests the f-th file from the i-th tier, the received signal can be expressed as follows

$$\begin{aligned} y_{i,f} =& \sqrt{P_i L_i(||x_0||)} \mathbf{h}_i^{x_0} \mathbf{w}_i^{x_0} s_{x_0} \\ &+ \sum_{x \in \Phi_i \backslash x_0} \sqrt{P_i L_i(||x||)} \mathbf{h}_i^{x} \mathbf{w}_i^{x} s_x + \sigma_i, \end{aligned} \tag{6}$$

where the typical UE is receiving the message from *the corresponding BS* at x_0 $(x_0 \in \Phi_{i,f})$. The locations of all interfering BSs forms a set $\Phi_i \backslash x_0$ and each element in such set is donated by the variable x. The channel vector from the

BS to the typical UE and the beamforming vector of the BS in the i-th tier are denoted by $\mathbf{h}_i^x$ and $\mathbf{w}_i^x$, respectively. σ_i represents the thermal noise.

Combining with the aforementioned assumptions, the product of the fading gain and beamforming gain of the BS located at x in the i-th tier is shown as below [30]

$$H_i^x \triangleq |\mathbf{h}_i^x \mathbf{w}_i^x|^2 = N_i |h_i|^2 G_i(\varphi_x - \theta_x), \tag{7}$$

where h_i is the small fading term. φ_x is the spatial angle of departure (AoD) from the interfering BS to the typical UE, and θ_x is the spatial AoD between the BS at location x and its corresponding receiver, see Fig. 2(b). $G_i(.)$ is the array gain function. More specifically, the actual array pattern is employed at pico BSs so that $G_2(\omega) \triangleq \frac{\sin^2(\pi N_i \omega)}{N_i^2 \sin^2(\pi\omega)}$ [30], where ω is a uniformly distributed random variable over $[-\frac{d}{\lambda}, \frac{d}{\lambda}]$. d and λ are the antenna spacing and wavelength, respectively [50]. On the other hand, the array gain function for macro BSs is $G_1(\omega) \triangleq 1$ due to the omnidirectional antenna pattern.

Since sophisticated beam training protocols [33] can be used at BSs to acquire the location information of the typical UE, we assume the corresponding BS provides the maximum directivity gain $G_0 = 1$ by aligning the antenna beam towards the typical UE.

E. Propagation Model

1) Channel Model: In the proposed HetNet, since the corresponding BS is interfered by other active BSs located in the same tier, the received SINR at the typical UE for requesting the f-th file from the i-th tier can be expressed as follows

$$\Upsilon_{i,f} = \frac{P_i L_i(||x_0||) G_0 N_i |h_i|^2}{\sigma_i^2 + I_{i,f} + I_{i,f'}}, \tag{8}$$

where $I_{i,f} = \sum_{x \in \Phi_{i,f} \setminus x_0} P_i L_i(||x||) H_i^x$ and $I_{i,f'} = \sum_{x \in \Phi_{i,f'}} P_i L_i(||x||) H_i^x$. $\Phi_{i,f'}$ is the set of locations of BSs that do not store the f-th file and it obeys $\Phi_i = \Phi_{i,f} \bigcup \Phi_{i,f'}$, $\Phi_{i,f} \bigcap \Phi_{i,f'} = \emptyset$. h_2 follows independent Nakagami fading due to utilizing mmWave and the parameter of Nakagami fading N_2^p is considered to be a positive integer for simplifying the analysis [24]. Therefore, $|h_2|^2$ is a normalized Gamma random variable. On the other side, we assume a Rayleigh fading model for the macro tier so that the fading parameter $N_1^p \triangleq 1$.

2) Association Criteria: When the typical UE requests the f-th file. For Max-RP, the biased average received power is defined as follows [8]

$$\bar{P}_{i,f} = b_i^P P_i L_i(||x_0||) N_i G_0, \tag{9}$$

where b_i^P is a bias factor that aims to balance the load between two tiers under the Max-RP scheme [10]. Then, we consider the biased received data rate for Max-Rate, which is given by

$$R_{i,f} = b_i^B B_i \log_2(1 + \Upsilon_{i,f}), \tag{10}$$

where B_i is the bandwidth per resource block. b_i^B is another bias factor to control the data traffic under Max-Rate.

When the typical UE requests a file from the core server, the throughput can be limited by both the macro-tier conditions and the backhaul capacity [15]. Therefore, the instantaneous downlink data rate from the core server can be expressed as $R_1 = \min(B_1 \log_2(1 + \Upsilon_{1,f}), C_{bh})$.

III. Cache-Related SINR Coverage Probability

Cache-related SINR coverage probability is the proportion of the received SINR that surpasses the requested SINR threshold $\tau \in \mathbb{R}$ depending on the content distributions. We separately discuss the cache-related SINR coverage probabilities for two tiers in this section, which is the theoretical basis for analyzing the final performance considering the association strategies.

Based on the content placement, $\Phi_{i,f}$ can be regarded as an independent non-HPPP with the density $p_{i,f}\lambda_i$. Therefore, the PDF of the distance between the typical UE and its nearest BS that contains the f-th file is given by

$$P_{i,f}(r) = 2\pi p_{i,f} \lambda_i r \exp(-\pi p_{i,f} \lambda_i r^2), \quad (r \geq 0). \tag{11}$$

In the following part, we first analyze the cache-related SINR coverage performance in the pico tier and then we consider the macro tier.

A. SINR Coverage Analysis in The Second Tier

If the typical UE is associated with the pico tier, the corresponding cache-related coverage probability can be derived with the aid of *Laplace Transform of Interference*.

1) Laplace Transform of Interference: Since the path loss exponent of LOS links α_2 is no less than 2 for practical scenarios, we divide the analysis on the Laplace transform of interference into two conditions ($\alpha_2 > 2$ and $\alpha_2 = 2$) in order to achieve closed-form expressions.

Lemma 1. When requesting the f-th content, under the condition $\alpha_2 > 2$, the Laplace transform of interference with the pre-decided SINR threshold τ in the pico tier is as follows

$$\mathcal{L}_2(s,\tau) = \exp\Big(-\pi\lambda_2(R_L^2 - p_{2,f} r^2) - \frac{\pi^2 \lambda_2}{2u_1} \sum_{k_1=1}^{u_1} W_f\big(\frac{x_{k_1} d}{\lambda}, s, \tau\big) \sqrt{(1 - x_{k_1}^2)}\Big), \tag{12}$$

where

$$W_f(\omega, s, \tau) = p_{2,f} S_2^0\big(\frac{sG_2(\omega)\tau}{N_2^p r^{\alpha_2}}\big) r^2 + (1 - p_{2,f})\, \Delta_2\left(\frac{sG_2(\omega)\tau}{N_2^p}\right) - S_2^0\big(\frac{sG_2(\omega)\tau}{N_2^p R_L^{\alpha_2}}\big) R_L^2, \tag{13}$$

$x_{k_1} = cos(\frac{2k_1 - 1}{2u_1}\pi), k_1 = 1, 2, ..., u_1$, are Gauss-Chebyshev nodes over $[-1, 1]$, and u_1 is a trade-off parameter between the accuracy and complexity [16, 51]. When $u_1 \to \infty$, the equality is established. $S_i^k(z) = {}_2F_1\big(k - \frac{2}{\alpha_i}, k + N_i^p; k + 1 - \frac{2}{\alpha_i}; -z\big)$ and ${}_2F_1(.)$ denotes Gauss hypergeometric function. $\Delta_i(z) = \Gamma\left(1 - \frac{2}{\alpha_i}\right)\Gamma\left(N_i^p + \frac{2}{\alpha_i}\right)\Gamma(N_i^p)^{-1} z^{\frac{2}{\alpha_i}}$ and $\Gamma(.)$ is the gamma function.

Numerous actual channel measures [52, 53] have indicated that the LOS path loss exponent is 2 for various carrier

frequencies, e.g. 28 GHz, 38 GHz and 73 GHz. Under such condition ($\alpha_2 = 2$), the equation (13) is changed to

$$W_f(\omega, s, \tau) = \frac{sG_2(\omega)\tau}{N_2^p}\Big(F_y\big(\frac{sG_2(\omega)\tau}{N_2^p R_L^2}\big) - p_{2,f}F_y\big(\frac{sG_2(\omega)\tau}{N_2^p r^2}\big)\Big), \quad (14)$$

where

$$F_y(y) = N_2^p \ln\big(1+\frac{1}{y}\big) - \frac{1}{y(1+y)^{N_2^p-1}} - \sum_{m=1}^{N_2^p-1} \frac{N_2^p}{(1+y)^{N_2^p-m}(N_2^p-m)}. \quad (15)$$

Proof: See Appendix A. ∎

Remark 1. *The analytical expressions in* **Lemma 1** *show that $\mathcal{L}_2(s,\tau)$ is independent of the transmit power P_2 and the intercept C_2.*

Remark 2. *With the aid of* (12) *and* (14)*, we conclude that $\mathcal{L}_2(s,\tau)$ is a monotonic increasing function with $p_{2,f}$.*

2) Cache-Related Coverage Probability: Considering the biased received power, we define the cache-related coverage probability $\dot{P}_{\Upsilon_{2,f}}(\tau)$, when requesting the f-th file from the second tier, as follows

$$\dot{P}_{\Upsilon_{2,f}}(\tau) = \mathbb{P}\big[\Upsilon_{2,f} > \tau\big]. \quad (16)$$

where $\mathbb{P}(.)$ represents the probability function. With the aid of **Lemma 1**, the closed-form coverage probability of pico tier is calculated as below.

Theorem 1. When the typical UE requests the f-th ranked content from the pico tier, the cache-related SINR coverage probability $\dot{P}_{\Upsilon_{2,f}}$ in this dense mmWave network can be expressed as follows

$$\dot{P}_{\Upsilon_{2,f}}(\tau) \approx \frac{\pi R_L}{2u_2}\sum_{n=1}^{N_2^p}(-1)^{n+1}\binom{N_2^p}{n} \times \sum_{k_2=1}^{u_2}\dot{F}_D\Big(\frac{(x_{k_2}+1)R_L}{2},\tau\Big)\sqrt{(1-x_{k_2}^2)}, \quad (17)$$

where

$$\dot{F}_D(r,\tau) = \mathcal{L}_2\big(\frac{n\eta_L r^{\alpha_2}}{G_0},\tau\big)\exp\big(-\frac{n\eta_L r^{\alpha_2}\tau\sigma_2^2}{P_2C_2N_2G_0}\big)P_{2,f}(r), \quad (18)$$

and $\eta_L = N_2^p(N_2^p!)^{-1/N_2^p}$.

Proof: Note that in equation (16)*, $|h_2|^2$ is a normalized Gamma random variable with parameter N_2^p. With the aid of* **Lemma 1** *and the tight upper bound equation in Appendix A from [24] ($\mathbb{P}\Big[|h_2|^2 < \gamma\Big] < (1-\exp(-\eta_L\gamma))^{N_2^P}, \gamma > 0$), the cache-related SINR coverage probability is given by*

$$\dot{P}_{\Upsilon_{2,f}}(\tau) \approx \sum_{n=1}^{N_2^p}(-1)^{n+1}\binom{N_2^p}{n}\int_0^{R_L}\mathcal{L}_2\big(\frac{n\eta_L r^{\alpha_2}}{G_0},\tau\big) \times \exp\big(-\frac{n\eta_L\tau\sigma_2^2 r^{\alpha_2}}{P_2C_2N_2G_0}\big)P_{2,f}(r)dr. \quad (19)$$

By applying Gauss-Chebyshev Quadrature, we obtain **Theorem 1**. ∎

Remark 3. *When $p_{2,f} = 1$, $\dot{P}_{\Upsilon_{2,f}}(\tau)$ in* **Theorem 1** *represents the coverage probability of this dense mmWave network without caching ability.*

Remark 4. *If the pico tier is assumed to be a noise-limited system, the signal-to-noise-ratio (SNR) coverage probability can be effortlessly deriving from* **Theorem 1** *by deleting the interference part $\mathcal{L}_2(.)$ in equation* (18)*.*

Due to the long communicating distance and high path loss, traditional cellular networks with mmWave are noise-limited [28]. However, recent articles [29, 32] have shown that with a high BS density, such systems become interference-limited. Under this condition, we present the first assumption below and the corroboration is provided in Section V.

Assumption 1. The dense mmWave network in the pico tier is assumed to be an interference-limited system, $\sigma_2^2 = 0$.

Corollary 1. Under **Assumption 1**, the corresponding cache-related SINR coverage probability $P_{\Upsilon_{2,f}}$ in the dense mmWave network can be simplified as follows

$$P_{\Upsilon_{2,f}}(\tau) \approx \frac{\pi R_L}{2u_2}\sum_{n=1}^{N_2^p}(-1)^{n+1}\binom{N_2^p}{n} \times \sum_{k_2=1}^{u_2}F_D\Big(\frac{(x_{k_2}+1)R_L}{2},\tau\Big)\sqrt{(1-x_{k_2}^2)}, \quad (20)$$

where

$$F_D(r,\tau) = \mathcal{L}_2\big(\frac{n\eta_L r^{\alpha_2}}{G_0},\tau\big)P_{2,f}(r). \quad (21)$$

Proof: By deleting the part $\big(\exp\big(-\frac{n\eta_L\tau\sigma_2^2 r^{\alpha_2}}{P_2C_2N_2G_0}\big)\big)$ in **Theorem 1***, which represents the thermal noise effect on the coverage performance, we obtain the equations for this corollary.* ∎

Remark 5. *Based on* **Remark 1***, we conclude that $P_{\Upsilon_{2,f}}(\tau)$ is independent of P_2 and C_2. Moreover, note that* **Corollary 1** *has a negligible difference with the exact simulations (as illustrated in Section V). We utilize such simplified expression as a replacement of the exact one in the rest of this paper.*

Since the association probability of Max-Rate is deduced on the basis of the derivative of the coverage probability [28], we deduce the derivative of $W_f(.)$ described in **Lemma 1**, based on which the PDF of the coverage probability can be figured out.

Lemma 2. As s in $W_f(\omega, s, \tau)$ is the transform variable of r in our calculation, we introduce $s = n\eta_L r^{\alpha_2}/G_0$ into the equations to make the notation straightforward. When $\alpha_2 > 2$, the derivative of $W_f(\omega, r, \tau)$ is given by

$$w_f(\omega, r, \tau) = \frac{2N_2^p Z(\omega)}{(\alpha_2-2)}\left(S_2^1\left(Z(\omega)\tau\right)r^2 - p_{2,f}S_2^1\left(\frac{Z(\omega)\tau r^{\alpha_2}}{R_L^{\alpha_2}}\right)\frac{r^{\alpha_2}}{R_L^{\alpha_2-2}}\right) + (1-p_{2,f})\,\mathrm{Z}(\omega)\,\Lambda_2\left(\mathrm{Z}(\omega)\tau\right)r^2 \quad (22)$$

and when $\alpha_2 = 2$, such derivative can be expressed as

$$w_f(\omega, r, \tau) = Z(\omega) r^2 \times \left(f_y\left(\frac{Z(\omega)\tau r^2}{R_L^2}, r \right) - p_{2,f} f_y\left(Z(\omega)\tau, R_L \right) \right), \tag{23}$$

where

$$f_y(y,z) = F_y(y) + Z(\omega)\tau \frac{z^2}{R_L^2} \Big(\frac{N_2^p y + 1}{y^2 (1+y)^{N_2^p}} - \frac{N_2^p}{y(1+y)} + \sum_{m=1}^{N_2^p - 1} \frac{N_2^p}{(1+y)^{N_2^p - m + 1}} \Big), \tag{24}$$

$\Lambda_i(z) = \frac{2}{\alpha_i} \Gamma\left(1 - \frac{2}{\alpha_i}\right) \Gamma\left(N_i^p + \frac{2}{\alpha_i}\right) \Gamma(N_i^p)^{-1} z^{\frac{2}{\alpha_i} - 1}$ and $Z(\omega) = \frac{n \eta_L G_2(\omega)}{G_0 N_2^p}$.

Proof: With the fact that $\frac{d}{dz} S_i^0(z) = \frac{2 N_i^P S_i^1(z)}{(\alpha_i - 2)}$ *and* $\frac{d}{dz} \Delta_i(z) = \Lambda(z)$*, we deduce the derivative of equations* (13) *and* (14) *under two conditions* ($\alpha_2 > 2$ *and* $\alpha_2 = 2$). ∎

Corollary 2. With the aid of **Lemma 2**, when the typical UE requires the f-th ranked content, the PDF of cache-related SINR coverage probability for the second tier $p_{\Upsilon_{2,f}}$ is as follows

$$p_{\Upsilon_{2,f}}(\tau) \approx \frac{\pi R_L}{2 u_3} \sum_{n=1}^{N_2^p} (-1)^{n+1} \binom{N_2^p}{n} \times \sum_{k_3=1}^{u_3} f_D\Big(\frac{R_L(x_{k_3} + 1)}{2}, \tau \Big) \sqrt{(1 - x_{k_3}^2)}, \tag{25}$$

where

$$f_D(r,\tau) = \frac{\pi^2 \lambda_2 F_D(r,\tau)}{2 u_2} \sum_{k_2=1}^{u_2} w_f\Big(\frac{x_{k_2} d}{\lambda}, r, \tau \Big) \sqrt{(1 - x_{k_2}^2)}. \tag{26}$$

Proof: See Appendix B. ∎

B. SINR Coverage Analysis in The First Tier

In the macro tier, we utilize the Rayleigh fading channel for sub-6 GHz signals. The exact expression of cache-related SINR coverage probability can be achieved in this part.

1) Cache-Related Coverage Probability: As discussed in the previous part, we keep the SINR threshold τ in the sub-6 GHz tier. With the similar analysis in the pico tier, the cache-related coverage probability in the second tier for requiring the f-th ranked content can be expressed as follows

$$\dot{P}_{\Upsilon_{1,f}}(\tau) = \mathbb{P}\left[\Upsilon_{1,f} > \tau \right]. \tag{27}$$

Due to the Rayleigh fading channel, it is effortless to derive the Laplace transform of interference for the first tier. Therefore, we directly provide the cache-related coverage probability in the following paragraph.

Theorem 2. When the typical UE requests the f-th ranked file from the macro tier, the exact cache-related SINR coverage probability $\dot{P}_{\Upsilon_{1,f}}$ is given by

$$\dot{P}_{\Upsilon_{1,f}}(\tau) = \int_0^\infty \exp\big(-\frac{\tau \sigma_1^2 r^{\alpha_1}}{P_1 C_1 N_1 G_0} - \pi \lambda_1 r^2 \times \big(p_{1,f}(S_1^0(\tau) - 1) + (1 - p_{1,f}) \Delta_1(\tau) \big) \big) P_{1,f}(r) dr. \tag{28}$$

Proof: As shown in equation (27), $\dot{P}_{\Upsilon_{1,f}}(\tau) = \mathbb{P}\left[|h_1|^2 > \frac{\tau (I_{1,f} + I_{1,f'} + \sigma_1^2) r^{\alpha_1}}{C_1 N_1 G_0} | r = ||x_0|| \right]$ *where* $|h_1|^2 \sim \exp(1)$ *due to Rayleigh fading assumption. Thus* $\dot{P}_{\Upsilon_{1,f}}(\tau)$ *can be expressed as* $\dot{P}_{\Upsilon_{1,f}}(\tau) = \mathbb{E}\left[\exp(-\frac{\tau (I_{1,f} + I_{1,f'} + \sigma_1^2) r^{\alpha_1}}{C_1 N_1 G_0}) | r = ||x_0|| \right]$*. Using the same method of* **Lemma 1**, *the coverage probability can be figured out as shown above.* ∎

Special Case 1: We assume $\alpha_2 = 4$, as it is valid for most sub-6 GHz networks [54].

Corollary 3. Under *Special Case 1*, the closed-form cache-related coverage probability in the first tier can be expressed as follows

$$\tilde{P}_{\Upsilon_{1,f}}(\tau) = \frac{1}{2} \pi p_{1,f} \lambda_1 \sqrt{\frac{\pi}{B(\tau)}} \exp\big(\frac{C^2(\tau)}{4 B(\tau)} \big) \operatorname{erfc}\big(\frac{C(\tau)}{2\sqrt{B(\tau)}} \big), \tag{29}$$

where $B(\tau) = \frac{\tau \sigma_1^2}{P_1 C_1 N_1 G_0}$, $C(\tau) = \pi \lambda_1 (p_{1,f} S_1^0(\tau) + (1 - p_{1,f}) \Delta_1(\tau))$, and $\operatorname{erfc}(.)$ is the complementary error function.

Proof: Note that $\alpha_1 = 4$ *as mentioned in Special Case 1, the equation of* **Theorem 2** *can be simplified as* $\tilde{P}_{\Upsilon_{1,f}}(\tau) = 2\pi p_{1,f} \lambda_1 \int_0^\infty \exp(-B(\tau) r^4 - C(\tau) r^2) r dr$*. Deploying (3.462-1) in [55], we have a closed-form expression* $\tilde{P}_{\Upsilon_{1,f}}(\tau) = \pi p_{1,f} \lambda_1 (2B(\tau))^{-\frac{1}{2}} \exp(\frac{C^2(\tau)}{8B(\tau)}) D_{-1}(\frac{C(\tau)}{\sqrt{2B(\tau)}})$*, where* $D_n(z)$ *is the Parabolic cylinder functions and* $D_{-1}(z) = \exp(\frac{z^2}{4}) \sqrt{\frac{\pi}{2}} \operatorname{erfc}(\frac{z}{\sqrt{2}})$. ∎

Assumption 2. *Since in various articles [10, 28, 54], the noise can be ignored in the traditional cellular networks with sub-6 GHz, we assume only the signal-to-interference-ratio (SIR) is considered in the first tier, namely,* $\sigma_1^2 = 0$.

Corollary 4. Under **Assumption 2**, the cache-related coverage probability for the first tier in **Theorem 2** can be simplified as follows

$$P_{\Upsilon_{1,f}}(\tau) = \frac{p_{1,f}}{p_{1,f} S_1^0(\tau) + (1 - p_{1,f}) \Delta_1(\tau)}. \tag{30}$$

Proof: By removing the noise part $\big(\exp(-\frac{\tau \sigma_1^2 r^{\alpha_1}}{P_1 C_1 N_1 G_0}) \big)$ *from* **Theorem 2**, *the SIR coverage probability can be deduced with the fact that* $\int_0^\infty r \exp(-a r^2) dr = \frac{1}{2a}, a > 0$. ∎

Remark 6. $P_{\Upsilon_{1,f}}(\tau)$ *is a monotonic increasing function with* $p_{1,f}$*. Moreover,* $P_{\Upsilon_{1,f}}(\tau)$ *is independent of* λ_1*,* P_1*,* N_1*, and* C_1*. Due to the negligible difference with the simulation shown in Section V, we use this closed-form equation as a proxy of the exact expression in the remainder of this paper.*

With the aid of the closed-form expression in **Corollary 4**, we are able to figure out the closed-form derivative of cache-related coverage probability for the first tier effortlessly.

Corollary 5. Under **Assumption 2**, the PDF $p_{\Upsilon_{1,f}}(\tau)$ of cache-related coverage probability when requesting the f-th ranked file is shown as follows

$$p_{\Upsilon_{1,f}}(\tau)=\frac{2p_{1,f}^2S_1^1(\tau)+(\alpha_1-2)\left(p_{1,f}-p_{1,f}^2\right)\Lambda_1(\tau)}{(\alpha_1-2)\left(p_{1,f}S_1^0(\tau)+(1-p_{1,f})\Delta_1(\tau)\right)^2}. \tag{31}$$

Proof: As the PDF of coverage probability for the second tier is $p_{\Upsilon_{1,f}}(\tau)=-\frac{dP_{\Upsilon_{1,f}}(\tau)}{d\tau}$*, we are able to calculate the expression based on* **Corollary 4**. ∎

IV. Success Probability and Area Spectral Efficiency Analysis

From the perspective of customers, the success probability is an important parameter to appraise the quality of service. In our cache-enabled HetNet, the data rate at the typical UE exceeding the pre-decided rate threshold R_{th} contributes to the success probability [10].

As discussed in the previous sections, we conclude that the considered system has two different processes in sending multimedia contents: 1) *Association Mode*, when the requested f-th file obeys $(1\le f\le H_c)$, the typical UE chooses the suitable BS as the corresponding BS depending on two association strategies; and 2) *Server Mode*, when the demanded f-th content only exists in the server due to limited storage capacity at BSs, which means $(H_c<f\le N_c)$, the typical UE requests such content from the server via the nearest macro BS. We detailedly discuss these two modes below.

A. Association Mode

In this mode, since two association strategies (Max-RP and Max-Rate) have different judgment standards to decide the corresponding BS, we study them separately.

1) Maximum Received Power Scheme: The Max-RP scheme has been utilized in numerous HetNets proposed in recent articles, for example, the traditional cache-enabled HetNets [10] and the hybrid HetNets with mmWave [28]. Under this scheme, the association procedure is fast and at a low cost due to ignoring the interference effects. We define the Max-RP association probability, when the typical UE connects to the i-th tier BS for the f-th file, as follows

$$\mathcal{A}_{i,f}^P=\mathbb{P}\left[\bar{P}_{i,f}>\bar{P}_{j,f}\right], \tag{32}$$

where $j\neq i$ and $j\in[1,2]$.

Note that the path loss law for the pico tier $L_2(r)$ has a step character. With the aid of similar proof in [17], the PDF $f_{i,f}^P(r)$ of the distance r between the typical UE and its corresponding i-th tier BS with containing f-th ranked file under Max-RP scheme is given by

$$f_{1,f}^P(r)=$$
$$2\pi p_{1,f}\lambda_1 r\exp\left(-\pi\sum_{j=1}^{2}p_{j,f}\lambda_j P_{j,1}^{r\,\frac{2}{\alpha_2}}r^{\frac{2}{\alpha_{j,1}^r}}\right)\mathbf{U}(R_r-r)$$
$$+2\pi p_{1,f}\lambda_1 r\exp\left(-\pi p_{2,f}\lambda_2 R_L^2-\pi p_{1,f}\lambda_1 r^2\right)\mathbf{U}(r-R_r), \tag{33}$$

$$f_{2,f}^P(r)=$$
$$2\pi p_{2,f}\lambda_2 r\exp\left(-\pi\sum_{j=1}^{2}p_{j,f}\lambda_j P_{j,2}^{r\,\frac{2}{\alpha_1}}r^{\frac{2}{\alpha_{j,2}^r}}\right)\mathbf{U}(R_L-r), \tag{34}$$

where $P_{j,i}^r=\frac{b_j^P P_j C_j N_j}{b_i^P P_i C_i N_i}$, $\alpha_{j,i}^r=\frac{\alpha_j}{\alpha_i}$, and $R_r=\left(P_{1,2}^r R_L^{\alpha_2}\right)^{\frac{1}{\alpha_1}}$.

2) Maximum Rate Scheme: Since the path loss laws and bandwidth for two tiers are dissimilar, the received data rates are totally different even they have the same average received power [28]. Compared with Max-RP, Max-Rate is able to provide higher data rate, but the extra knowledge of channel state information is indispensable. For Max-Rate, the association probability of the typical UE being associated with the i-th tier BS for requesting the f-th file is defined as

$$\mathcal{A}_{i,f}^R=\mathbb{P}\left[R_{i,f}>R_{j,f}\right]. \tag{35}$$

Instead of analyzing the relationship between the PDF of the corresponding distance as discussed in Max-RP, we are able to directly derive the PDF of considered coverage probability with the SINR threshold τ.

Lemma 3. When requesting the f-th ranked content, the PDF of the cache-related coverage probability for the i-th tier under the Max-Rate strategy is shown as follows

$$f_{i,f}^R(\tau)=p_{\Upsilon_{i,f}}(\tau)\Big(1-P_{\Upsilon_{j,f}}\big((1+\tau)^{\frac{b_i^B B_i}{b_j^B B_j}}-1\big)\Big). \tag{36}$$

Proof: See Appendix C. ∎

Remark 7. *Since* $P_{\Upsilon_{j,f}}(\infty)=0$*, if* $(b_i^B B_i\gg b_j^B B_j)$*, the PDF of this coverage probability is same as* $p_{\Upsilon_{i,f}}(\tau)$*, which means the typical UE is associated with the* i*-th tier BS invariably. As a consequence, if the bandwidth of one tier is far more than the other, the typical UE always connects to the tier with large bandwidth.*

Average Load Approximation: When all UEs are associated with the HetNet, the average number of UEs served by the i-th tier BSs N_i^{load} can be approximated by[4] [28] $\bar{N}_i^{load}\approx 1+\frac{1.28\lambda_u\mathcal{A}_i}{\lambda_i}$, where $\mathcal{A}_i^P=\sum_{f=1}^{H_c}P_f\int_0^\infty f_{i,f}^P(r)dr$, $\mathcal{A}_i^R=\sum_{f=1}^{H_c}P_f\int_{-\infty}^{+\infty}f_{i,f}^R(\tau)d\tau$, and $\mathcal{A}_i\in\{\mathcal{A}_i^P,\mathcal{A}_i^R\}$.

Remark 8. *Since* $\mathcal{A}_i$ *is a monotonic increasing function with the corresponding bias factor* (b_i^P *or* b_i^B)*, the small value of*

[4]This approximation is valid for sub-6 GHz scenarios [38] and mmWave scenarios [56]. Due to the low probability of requesting the content from the core server, we ignore the corresponding load in the server mode for simplifying the analysis.

such bias factor is able to offload the data traffic in the i-th tier. As a result, by adjusting the bias factors b_i^P in the Max-RP scheme and b_i^B in the Max-Rate scheme, we are able to control the average number of UEs for each tier, thereby balancing the load of the proposed HetNet.

B. Server Mode

We present the backhaul capacity in order to compare our system with the traditional HetNets [10] in which the macro BSs have no caching ability. The comparison is illustrated in Section V. In the server mode, the backhaul capacity restricts the performance of our system. More specifically, if the required rate R_{th} exceeds the backhaul capacity C_{bh}, no content can be sent successfully due to low system rate. On the other hand, if C_{bh} is larger than R_{th}, the success probability under this case is limited by the received data rate from the relay macro BS. As a result, we provide the success probability in this mode as follows.

Lemma 4. The success probability $P_S(R_{th})$ in the server mode is given by

$$P_S(R_{th}) = \mathbf{U}(C_{bh} - R_{th})\mathbf{U}(N_c - H_c - 1) \times \sum_{f=H_c+1}^{N_c} P_f P_{\Upsilon_1}(2^{\frac{R_{th}}{B_1}} - 1), \quad (37)$$

where $P_{\Upsilon_1}(.)$ is the coverage probability of the first tier without cache capacity, which equals $P_{\Upsilon_{1,f}}(.)$ with the condition that $p_{1,f} = 1$.

Proof: As discussed above, if $C_{bh} < R_{th}$, $P_S(R_{th}) = 0$, while if $C_{bh} \geq R_{th}$, the success probability for requesting the f-th file is decided by the wireless capacity of the macro tier, which is $P_f P_{\Upsilon_1}(2^{\frac{R_{th}}{B_1}} - 1)$. Then, we consider the request probability for the files from $(H_c + 1)$ to N_c to obtain this lemma (note that when $H_c = N_c$, the typical UE is able to acquire all files from the HetNet, namely $P_S(R_{th}) = 0$). ■

C. Success Probability

As mentioned in the beginning of this section, we present the success probability to evaluate the performance of our system. This parameter can be calculated with the aid of the cache-related SINR coverage probability discussed in Section III. We first define the universal success probability as below

$$\mathcal{P}_s(R_{th}) = \sum_{f=1}^{H_c}\sum_{i=1}^{2} P_f \mathcal{A}_{i,f}^{\kappa} \mathbb{P}\left[B_i \log_2(1 + \Upsilon_{i,f}) > R_{th}\right] + P_S(R_{th}), \quad (38)$$

where $\kappa \in \{P, R\}$. We calculate the success probability under two association strategies in the following part.

1) Maximum Received Power Scheme: As the system rate can be derived from the received SINR, we first deduce the SINR coverage probability for the i-th tier under Max-RP.

Lemma 5. When requiring the f-th ranked content, the SINR coverage probability for the i-th tier $\Theta_{i,f}(\tau)$ under Max-RP scheme is shown as

$$\Theta_{1,f}(\tau) = \int_0^{\infty} \exp\left(-\pi\lambda_1 r^2 \big(p_{1,f}(S_1^0(\tau) - 1) + (1 - p_{1,f})\Delta_1(\tau)\big)\right) f_{1,f}^P(r) dr, \quad (39)$$

$$\Theta_{2,f}(\tau) \approx \frac{\pi R_L}{2u_2} \sum_{n=1}^{N_2^p} (-1)^{n+1} \binom{N_2^p}{n} \times \sum_{k_2=1}^{u_2} F_R\Big(\frac{(x_{k_2} + 1)R_L}{2}, \tau\Big)\sqrt{(1 - x_{k_2}^2)}, \quad (40)$$

where

$$F_R(r, \tau) = \mathcal{L}_2\left(\frac{n\eta_L r^{\alpha_2}}{G_0}, \tau\right) f_{2,f}^P(r). \quad (41)$$

Proof: Since the probability of distance under Max-RP scheme is changed into equation (33)*, $P_{i,f}(r)$ is replaced by $f_{i,f}^R(r)$ in* **Corollary 1** *and* **Theorem 2** *for the second and first tiers, respectively. Moreover, with the aid of* **Assumption 1** *and* **Assumption 2***, we also ignore the noise effect. Therefore, these probabilities are calculated as above.* ■

Remark 9. *Since $f_{i,f}^R(.)$ has a negative correlation with $P_{j,i}^r$ in equation* (33)*, $\Theta_{i,f}(.)$ is a monotonic increasing function with the product of b_i^P, P_i, and C_i.*

Then, based on the cache-related coverage probability under the Max-RP scheme, we present the corresponding success probability in the following part.

Theorem 3. With the aid of **Lemma 4** and **Lemma 5**, the success probability for Max-RP scheme can be expressed as follows

$$\mathcal{P}_P(R_{th}) \approx \sum_{f=1}^{H_c}\sum_{i=1}^{2} P_f \Theta_{i,f}(2^{\frac{R_{th}}{B_i}} - 1) + P_S(R_{th}). \quad (42)$$

Proof: When the required rate is R_{th}, the SINR threshold is $(2^{\frac{R_{th}}{B_i}} - 1)$, so the success probability for requesting the f-th file from the first H_c contents is $P_f \Theta_{i,f}(2^{\frac{R_{th}}{B_i}} - 1)$. Considering the whole multimedia contents, the success probability will be calculated as above. ■

2) Maximum Rate Scheme: In contrast to the discussion in Max-RP scheme, we have already figured out PDF of the coverage probability for the i-th tier in **Lemma 3**. The success probability can be calculated based on this result.

Theorem 4. With the aid of **Lemma 3** and **Lemma 4**, the success probability in the Max-Rate scheme is given by

$$\mathcal{P}_R(R_{th}) \approx \sum_{f=1}^{H_c}\sum_{i=1}^{2} P_f \int_{2^{\frac{R_{th}}{B_i}} - 1}^{\infty} f_{i,f}^R(\tau) d\tau + P_S(R_{th}). \quad (43)$$

Proof: Note that the received SINR should be larger than the SINR threshold $(2^{\frac{R_{th}}{B_i}} - 1)$. For the Max-Rate scheme, the success probability for requesting the f-th file is changed to $P_f \int_{2^{\frac{R_{th}}{B_i}} - 1}^{\infty} f_{i,f}^R(\tau) d\tau$. After considering both the association mode and the server mode, universal success probability can be deduced in this theorem. ■

Remark 10. *Combining* **Remark 5** *and* **Remark 6**, *we conclude that the success probability under Max-Rate is independent of λ_1, N_1, C_i and P_i.*

D. Area Spectral Efficiency

The ASE is the average data rate transmitted in unit bandwidth and unit area, which can be represented in the unit of bps/Hz/m^2. Assuming that Gaussian Codebooks are utilized for all transmissions, we are able to define ASE with the aid of Shannon's Capacity Formula. It is expressed as follows $ASE = \lambda \log_2(1+\tau) P_\tau$, where P_τ is the SINR coverage probability of the considered networks and λ denotes the active BSs' density [54].

Proposition 1. The ASEs for two user association strategies share the same equation as below

$$ASE = \sum_{f=1}^{H_c} \sum_{i=1}^{2} \frac{P_f p_{i,f} \lambda_i R_{th}}{B_i} P_a^{\kappa}(2^{\frac{R_{th}}{B_i}} - 1) + \frac{\lambda_1 R_{th}}{B_1} P_S(R_{th}), \tag{44}$$

where $P_a^P(\tau) = \Theta_{i,f}(\tau)$ is the coverage probability when requiring the f-th file from the i-th tier under Max-RP and $P_a^R(\tau) = \int_\tau^\infty f_{i,f}^R(t)dt$ is that under Max-Rate.

Proof: When the requested f-th ranked file is located between the first and H_c-th popularity rank, the ASE with pre-decided SINR threshold τ for the i-th tier is $P_f p_{i,f} \lambda_i \log_2(1+\tau) P_a^{\kappa}(\tau)$. On the other side, if $(f > H_c)$, the typical UE is associated with the macro tier to request the content from the server. Therefore, the ASE is $P_f \lambda_1 \log_2(1+\tau) P_{\Upsilon_1}(\tau)$. Then sum them up, we have the final equation above. ∎

Remark 11. *Note that an optimum value of R_{th} exists for achieving the maximum ASE. The reason is that the large R_{th} potentially increase the ASE as shown in* (44)*, but it also decreases the coverage probabilities of both tiers.*

V. Numerical Results

A. Network Settings and Simulations

The general network settings are presented in Table I [10, 24, 28] and the reference distance for the intercept is one meter. Since we employ multi-input single-output (MISO) system at mmWave tier, the actual beamforming of ULA is imposed in Monte Carlo simulations. Additionally, the NLOS transmissions are provided in simulations as well for evaluating the effect of NLOS BSs. The path loss exponent is $\alpha_N = 4$, the intercept is $C_N = C_2$ and the parameter N_N for Nakagami fading channel in NLOS scenario is 2 [24].

In Fig. 4(a), we validate the expressions of cache-related coverage probabilities for requesting the f-th file. For the second tier, the simulation results with NLOS BSs overlap those without NLOS transmissions, which means ignoring all NLOS signals does not impair the analytical accuracy. Our tight approximate equations for the second tier fit the simulation results with an insignificant difference, especially when the path loss exponent $\alpha_2 = 2$. With the increase of α_2, the noise effect is slightly enhanced but our mmWave networks in the pico tier is still dominated by the interference. On the other hand, the theoretical expressions for the first tier match the numerical results perfectly, which means the macro tier with sub-6 GHz is an interference-limited system and the closed-form expression in **Corollary 4** is able to replace the exact equation in **Theorem 2** for simplifying the analysis. Moreover, the mmWave networks achieve a higher SINR coverage probability than the traditional sub-6 GHz networks under the same pre-decided coverage threshold, which indicates that large association probability to the pico tier benefits the universal throughput.

Comparing the analytical results of the success probability with the simulations in Fig. 4(b), we note that they match each other ideally, thereby certifying the analysis. Max-Rate performs better than Max-RP in terms of the success probability when the pre-decided rate threshold R_{th} is high. In the real world, we may use Max-RP as the best scheme for fetching the maximum success probability in the low R_{th} region due to the easy operation with cheap system cost, while with the rise of R_{th}, the best strategy for user association in our system should be changed into Max-Rate although it needs extra system overheads. Additionally, it is obvious that our cache-enabled HetNets outperform the traditional HetNets, especially in the area $R_{th} \geq C_{bh}$.

B. Impact of Noise and Antenna Scales

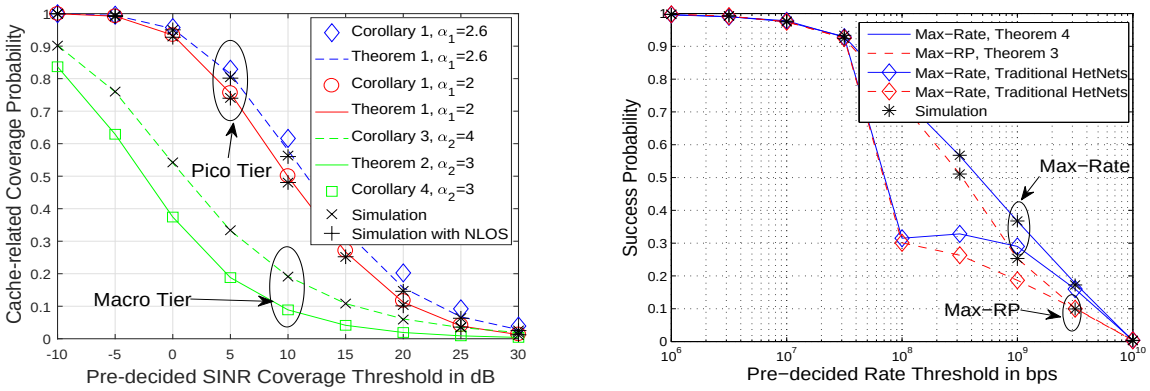
In this part, we corroborate the accuracy of our interference-limited assumption in dense mmWave networks. Fig. 5(a) shows the simulation results of SINR and SIR coverage probabilities with different λ_2. It can be seen from the figure that for $\lambda_2 = 20/(250^2\pi)$ m^{-2} and $\lambda_2 = 30/(250^2\pi)$ m^{-2}, two ratios are almost overlapping and even at $\lambda_2 = 10/(250^2\pi)$ the difference between them is reasonably small. Accompanying with the condition $\lambda_1 \ll \lambda_2$ assumed in Section II, we are able to confirm **Assumption 1** that the noise has a negligible impact on the coverage in the second tier due to high considered densities in our system. Therefore SINR can be approximated by SIR in the pico tier. As the first tier is also dominated by the interference rather than the noise as discussed in the previous part, we are capable of concluding that our cache-enabled hybrid HetNet is an interference-limited system.

C. Impact of Transmit Power and Bias Factors

The success probability is mainly decided by the association probability. Fig. 5(b) illustrates that the association probability under Max-Rate is independent of the transmit power of macro BSs as mentioned in **Remark 10**. Regarding the bias factors, when the biased bandwidth $b_1^B B_1$ in Max-Rate changes from 500 MHz to 20 MHz, $(b_2^B B_2 / b_1^B B_1)$ tends towards infinite and the association probability with the pico tier rises up to nearly one, which means under this condition, the typical UE always connects to the nearest pico BS as we discussed in **Remark 7**. For Max-RP, the probability associating with the second tier has a negative correlation with b_1^P. Therefore, we are able to change the bias factors to balance the traffic load between two tiers as mentioned in **Remark 8**.

TABLE I: General Settings of the Network

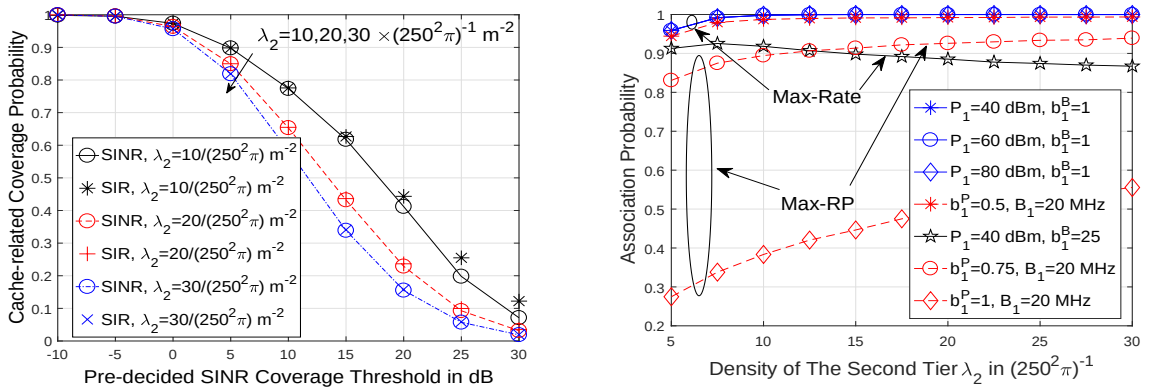
LOS ball range	$R_L = 200$ m
Density of PPP	$\lambda_1 = 1/(250^2\pi)$; $\lambda_2 = 20/(250^2\pi)$; $\lambda_u = 30/(250^2\pi)$ m^{-2}
Bandwidth	$B_1 = 20$ MHz; $B_2 = 1$ GHz
Path loss law	$\alpha_1 = 4$, $N_1^p = 1$; $\alpha_2 = 2$, $N_2^p = 3$
Number of antennas	$N_1 = 1$; $N_2 = 20$
Carrier frequency	$f_{macro} = 2$ GHz; $f_{pico} = 28$ GHz
Transmit Power at BSs	$P_1 = 80$ dBm; $P_2 = 30$ dBm
Transmit Power at the typical UE	$P_0 = 30$ dBm
Backhaul capacity	$C_{bh} = 50$ Mbps
Caching capacity	$M_1 = M_2 = 80$, $H_c = 90$, $N_c = 100$
Skew of the popularity distribution	$\delta = 0.6$
Bias factors	$b_i^P = b_i^B = 1$



(a) Cache-related coverage probability versus pre-decided SINR threshold, with $p_{i,f} = 1$.

(b) Success probability versus pre-decided rate threshold, with $M_2 = 10$, $B_1 = 200$ MHz, and $p_{i,f} = \mathbf{U}(M_i - f)$.

Fig. 4: Simulation and Validation.



(a) Cache-related coverage probability versus pre-decided SINR threshold in the second tier, with $N_2 = 32$ and $p_{2,f} = 1$.

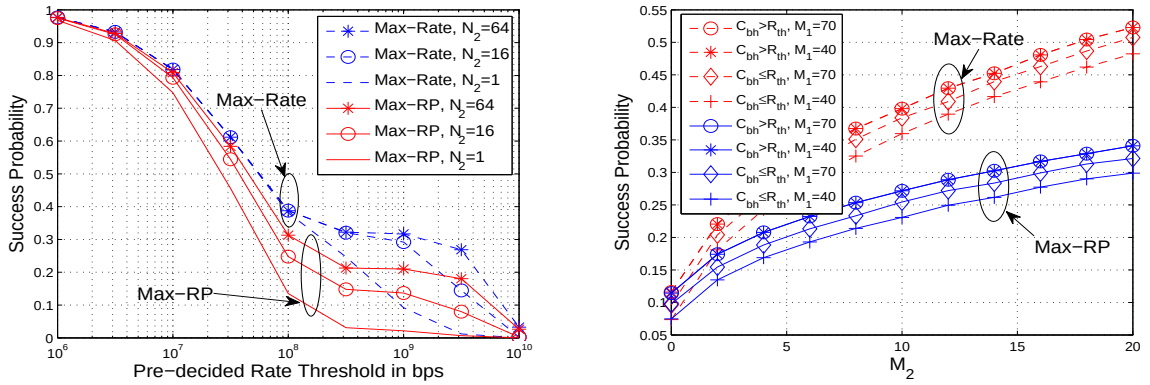
(b) Association probability with the second tier versus the density of the second tier λ_2, with $p_{i,f} = 1$.

Fig. 5: Interference-limited Property and The Impact of Antenna Scales.

D. Impact of Antenna Scales and Cache Capacity

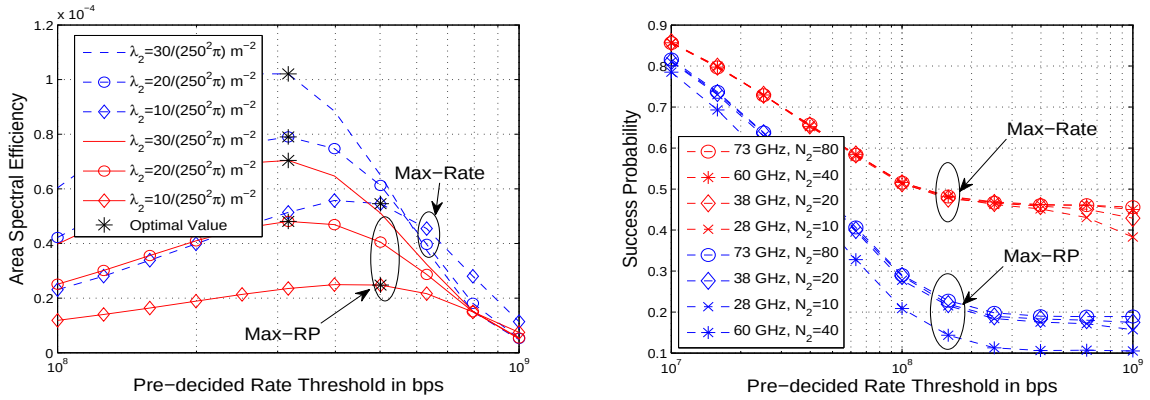
As we adopt ULA with N_2 antenna elements at pico BSs, the increase of antenna scales N_2 enhances the received power and it also narrows the half-power beamwidth (HPBW). Additionally, the smaller HPBW contributes to the less interference in our system. Fig. 6(a) shows that the success probability arises with the increase of N_2, especially in the high pre-decided rate threshold R_{th} region.

Cache capacity of BSs deployed in our networks is also an important parameter for analyzing the performance. The success probability is a monotonic increasing function with pico BSs' storage capacity M_2 as shown in Fig. 6(b). Considering the cache capacity of macro BSs M_1, when $C_{bh} > R_{th}$, the success probability has no relationship with M_1 since the less-popular contents which are only contained in the server can be transmitted freely through backhauls. In this case, the proposed cache-enabled HetNet is same with the traditional one. On the other side, when $C_{bh} \leq R_{th}$, the server is blocked so that the success probability can be benefited by the large M_1, which represents that more multimedia files are stored at macro BSs.



(a) Success probability versus pre-decided rate threshold, with with $M_2 = 10$, $\lambda_2 = 20/(250^2\pi)$ m^{-2}, $C_{bh} = 10^8$ bps, and $p_{i,f} =$ with $R_{th} = 10^8$ bps and $p_{i,f} = \mathbf{U}(M_i - f)$. $\mathbf{U}(M_i - f)$. (b) Success probability versus cache capacity M_2 in the second tier,

Fig. 6: Impact of Transmit Power and Bias Factors.



(a) Area spectral efficiency versus pre-decided rate threshold, with $M_1 = 10$, $B_1 = 200$ MHz, and $p_{i,f} = \mathbf{U}(M_i - f)$. (b) Success probability versus pre-decided rate threshold, with $M_2 = 20$ and $C_{bh} = 10^8$ bps, and $p_{i,f} = \mathbf{U}(M_i - f)$

Fig. 7: ASE and Various Carrier Frequencies.

E. Performance of ASE and Various Carrier Frequencies

We first present the performance of ASEs in Fig. 7(a). It illustrates that Max-Rate scheme outperforms Max-RP scheme regarding the ASE. Moreover, the optimum pre-decided rate threshold R_{th} for achieving the maximum ASE can be easily figured out from Fig. 7(a), thereby corroborating **Remark 11**. When the density of pico tier λ_2 increases from $10/(250^2\pi)$ m^{-2} to $30/(250^2\pi)$ m^{-2}, the optimal value of R_{th} decreases. In the real world, this optimum R_{th} can be used to design a network with maximum ASE.

TABLE II: Path Loss Exponent and Antenna Scales for The Second Tier

Carrier frequencies	28G	38G	60G	73G
Path loss exponent α_2 for LOS links	2	2	2.25	2
Number of antenna elements N_2	10	20	40	80

Based on the actual path loss exponents of LOS links [52, 53] and estimated antenna scales [32] shown in Table II, the performance of four different carrier frequencies is illustrated in Fig. 7(b). After comparing the carrier frequencies at 28 GHz, 38 GHz, 60 GHz, and 73 GHz, we conclude that 73 GHz is the best choice for both two user association strategies thanks to the largest antenna scales. Moreover, 28 GHz performs the worst among these four carrier frequencies in Max-Rate scheme due to the limited anti-interference ability, while 60 GHz causes the lowest success probability in Max-RP scheme because of the largest path loss exponent.

F. Performance of Different Content Placement Policies

In the cache-enabled HetNet, the optimal content placement scheme is able to enhance the capacity of networks. Although the content distribution in the previous illustrations is a binary case, the proposed expressions in this work are suitable for other random distribution scenarios as shown in Fig. 8. We propose two content placement policies here for comparison: 1) policy 1: all BSs cache the most popular files, namely $p_{i,f} = \mathbf{U}(M_i - f)$; and 2) policy 2: the files from the first to H_c popular rank has the same probability of being cached at all BSs, namely $p_{i,f} = H_c/M_i$. For policy 1, the success probability is independent of H_c. However, the success probability in policy 2 has a positive correlation with H_c.

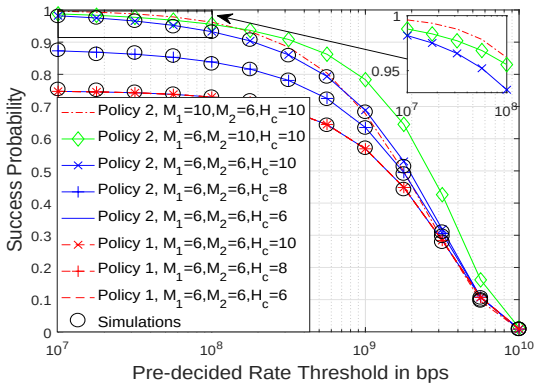


Fig. 8: Success probability versus pre-decided rate threshold in Max-RP, with $B_1 = 500$ MHz, $C_{bh} = 10^6$ bps, and $N_c = 10$.

Regarding the cache capacity, large M_1 for macro BSs slightly increases the success probability in low R_{th} regions, while large M_2 for pico BSs significantly enhance the performance in high R_{th} areas. The reason is that mmWave BSs is capable of providing faster data rate than sub-6 GHz BSs. When $M_1 = M_2 = H_c$, Policy 2 is same as policy 1, otherwise policy 2 outperforms policy 1.

VI. Conclusion

In this treatise, the performance of our cache-enabled hybrid HetNet has been analyzed in details. We have compared two different user association strategies with the aid of the stochastic geometry. More specifically, Max-Rate scheme outperforms Max-RP scheme regarding the success probability and ASE, but the difference between two association strategies can be eliminated by decreasing the transmit power of the macro tier. The proposed network, which performs better than the traditional HetNet, can be regarded as an interference-limited system due to the high density of mmWave tier and the nature of sub-6 GHz tier. We have analytically shown that the success probability of Max-Rate scheme is independent of λ_1, N_1, C_i and P_i. Moreover, our system has a positive correlation with pico BSs' antenna scales and the cache capacity of both tiers. Additionally, there exists an optimum value of pre-decided rate threshold contributing to the maximum ASE. Lastly, for two user association strategies, 73 GHz is the best carrier frequency of mmWave tier. In addition to two considered content placement policies, our future work will focus on the optimization of the content placement schemes

Appendix A: Proof of Lemma 1

The Laplace transform of interference in the second tier is given by (A.1) at the top of the next page. For (A.1), (a) follows the Gamma random variable's moment generating function [24]; (b) is computing the expectation of the second tier antenna gain G_2.

When $\alpha_2 > 2$, (A.1) can be simplified into (A.2). For (A.2), $X(\omega) = \frac{sG_2(\omega)\tau}{N_2^p}$. (c) follows $\Delta_i(z) = \lim_{r\to 0} r^2\left(S_i^0(\frac{z}{r^{\alpha_i}}) - 1\right)$ and Gauss hypergeometric function [10].

When $\alpha_2 = 2$, (A.1) can be simplified into (A.3). For (A.3), (d) follows (2.117-1), (2.117-3) and (2.118-1) in [55]. With the aid of Gauss-Chebyshev Quadrature, we obtain **Lemma 1**. The proof is finished.

Appendix B: Proof of Corollary 2

Under the **Assumption 1**, we ignore the noise effect. The PDF of the coverage probability in the second tier is shown as follows

$$p_{\Upsilon_{2,f}}(\tau) = -\frac{d}{d\tau}\mathbb{P}[\Upsilon_{2,f} > \tau] \approx \sum_{n=1}^{N_2^p}(-1)^{n+1}\binom{N_2^p}{n}\int_0^{R_L} f_D(r,\tau)dr, \quad \text{(B.1)}$$

where $\mathbb{P}[\Upsilon_{2,f} > \tau] = P_{\Upsilon_{2,f}}(\tau)$ in **Corollary 1** and $f_D(r,\tau)$ is the derivative of $-F_D(r,\tau)$. (e) follows the fact that variable τ is only contained in $F_D(r,\tau)$. Then $f_D(r,\tau)$ is given by

$$f_D(r,\tau) = -P_{2,f}(r)\frac{d}{d\tau}\mathcal{L}_2\left(\frac{n\eta_L r^{\alpha_2}}{G_0},\tau\right). \quad \text{(B.2)}$$

With the aid of **Lemma 2**, we obtain

$$f_D(r,\tau) = F_D(r,\tau)\frac{\pi\lambda_2\lambda}{d}\int_{-\frac{d}{\lambda}}^{\frac{d}{\lambda}} w_f(\omega,r,\tau)d\omega. \quad \text{(B.3)}$$

By substituting (B.3) into (B.1) and then applying Gauss-Chebyshev Quadrature, we obtain **Corollary 2**. The proof is finished.

Appendix C: Proof of Lemma 3

The derivative of the coverage probability under Max-Rate scheme can be calculated with the coverage probabilities of two tiers discussed in Section III. We first figure out the probability of the first tier coverage based on Max-Rate strategy as follows

$$\begin{aligned} &F_{1,f}^R(\tau) \\ =&\mathbb{P}\left[\Upsilon_{1,f} > \tau | b_1^B B_1 \log_2(1+\Upsilon_{1,f}) > b_2^B B_2 \log_2(1+\Upsilon_{2,f})\right] \\ =&\int_\tau^\infty p_{\Upsilon_{1,f}}(\Upsilon_{1,f})\mathbb{P}\left[\Upsilon_{2,f} < (1+\Upsilon_{1,f})^{\frac{b_1^B B_1}{b_2^B B_2}} - 1\right] d\Upsilon_{1,f} \\ =&\int_\tau^\infty p_{\Upsilon_{1,f}}(\Upsilon_{1,f})(1 - P_{\Upsilon_{2,f}}((1+\Upsilon_{1,f})^{\frac{b_1^B B_1}{b_2^B B_2}} - 1))d\Upsilon_{1,f}. \end{aligned} \quad \text{(C.1)}$$

Then, the PDF of such coverage probability $f_{1,f}^R(\tau)$ is given by

$$\begin{aligned} f_{1,f}^R(\tau) =& -\frac{d}{d\tau}F_{1,f}^R(\tau) \\ =& -\frac{d}{d\tau}\int_\tau^\infty p_{\Upsilon_{1,f}}(\Upsilon_{1,f}) \\ &\times (1 - P_{\Upsilon_{2,f}}((1+\Upsilon_{1,f})^{\frac{b_1^B B_1}{b_2^B B_2}} - 1))d\Upsilon_{1,f} \\ =& p_{\Upsilon_{1,f}}(\tau)(1 - P_{\Upsilon_{2,f}}((1+\tau)^{\frac{b_1^B B_1}{b_2^B B_2}} - 1)). \end{aligned} \quad \text{(C.2)}$$

$$\begin{aligned}\mathcal{L}_2(s,\tau) =& \mathbb{E}\Big[\exp\big(-ns\tau \sum_{x\in\Phi_2\setminus x_0} G_2(\omega)|h_2|^2||x||^{-\alpha_2}\big)\Big] \\ \overset{(a)}{=}& e^{-2\pi p_{2,f}\lambda_2\mathbb{E}_{G_2}\left[\int_r^{R_L}(1-(1+\frac{ns\tau G_2(\omega)}{N_2^p v^{\alpha_2}})^{-N_2^p})vdv\right]-2\pi(1-p_{2,f})\lambda_2\mathbb{E}_{G_2}\left[\int_0^{R_L}(1-(1+\frac{ns\tau G_2(\omega)}{N_2^p v^{\alpha_2}})^{-N_2^p})vdv\right]} \\ \overset{(b)}{=}& e^{-\frac{\pi\lambda_2\lambda}{d}\int_{-\frac{d}{\lambda}}^{\frac{d}{\lambda}}\left(p_{2,f}\int_r^{R_L}(1-(1+\frac{ns\tau G_2(\omega)}{N_2^p v^{\alpha_2}})^{-N_2^p})vdv+(1-p_{2,f})\int_0^{R_L}(1-(1+\frac{ns\tau G_2(\omega)}{N_2^p v^{\alpha_2}})^{-N_2^p})vdv\right)d\omega},\end{aligned} \tag{A.1}$$

$$\mathcal{L}_2(s,\tau) \overset{(c)}{=} e^{-\pi\lambda_2(R_L^2-p_{2,f}r^2)-\frac{\pi\lambda_2\lambda}{2d}\int_{-\frac{d}{\lambda}}^{\frac{d}{\lambda}}\left(p_{2,f}S_2^0\left(\frac{X(\omega)}{r^{\alpha_2}}\right)r^2+(1-p_{2,f})\Delta_2(X(\omega))-S_2^0\left(\frac{X(\omega)}{R_L^{\alpha_2}}\right)R_L^2\right)d\omega}, \tag{A.2}$$

$$\mathcal{L}_2(s,\tau) \overset{(d)}{=} e^{-\pi\lambda_2(R_L^2-p_{2,f}r^2)-\frac{\pi\lambda_2\lambda}{2d}\int_{-\frac{d}{\lambda}}^{\frac{d}{\lambda}}\left(X(\omega)\left(F_y\left(\frac{X(\omega)}{R_L^2}\right)-p_{2,f}F_y\left(\frac{X(\omega)}{r^2}\right)\right)\right)d\omega}. \tag{A.3}$$

Using the same method, the PDF for the probability of the second tier coverage under Max-Rate scheme is shown as below

$$f_{2,f}^R(\tau) = p\Upsilon_{2,f}(\tau)(1-P_{\Upsilon_{1,f}}((1+\tau)^{\frac{b_2^B B_2}{b_1^B B_1}}-1)). \tag{C.3}$$

Combing (C.2) and (C.3), we obtain **Lemma 3**. The proof is finished.

REFERENCES

[1] W. Yi, Y. Liu, and A. Nallanathan, "Modeling and analysis of mmWave communications in cache-enabled HetNets," in *IEEE Proc. of International Commun. Conf. (ICC)*, May 2018.

[2] M. Z. Hasan, H. Al-Rizzo, and F. Al-Turjman, "A survey on multipath routing protocols for QoS assurances in real-time wireless multimedia sensor networks," *IEEE Commun. Surv. Tuts.*, vol. 19, no. 3, pp. 1424–1456, 3rd Quart. 2017.

[3] Y. Cao, O. Kaiwartya, R. Wang, T. Jiang, Y. Cao, N. Aslam, and G. Sexton, "Toward efficient, scalable, and coordinated on-the-move EV charging management," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 66–73, Apr. 2017.

[4] Y. Cao, S. Yang, G. Min, X. Zhang, H. Song, O. Kaiwartya, and N. Aslam, "A cost-efficient communication framework for battery-switch-based electric vehicle charging," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 162–169, May 2017.

[5] C. V. N. I. Cisco, "Global mobile data traffic forecast update, 2013–2018," *white paper*, 2014.

[6] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. Sukhavasi, C. Patel, and S. Geirhofer, "Network densification: the dominant theme for wireless evolution into 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 82–89, Feb. 2014.

[7] J. Zhao, Y. Liu, K. K. Chai, A. Nallanathan, Y. Chen, and Z. Han, "Spectrum allocation and power control for non-orthogonal multiple access in HetNets," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5825–5837, Sep. 2017.

[8] Y. Liu, Z. Qin, M. Elkashlan, A. Nallanathan, and J. A. McCann, "Non-orthogonal multiple access in large-scale heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2667–2680, Dec. 2017.

[9] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, C. L. I, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[10] D. Liu and C. Yang, "Caching policy toward maximal success probability and area spectral efficiency of cache-enabled HetNets," *IEEE Trans. Commun.*, vol. 65, no. 6, pp. 2699–2714, Jun. 2017.

[11] V. Chandrasekhar, J. G. Andrews, and A. Gatherer, "Femtocell networks: A survey," *IEEE Commun. Mag.*, vol. 46, no. 9, pp. 59–67, Sep. 2008.

[12] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system," in *Proc. ACM Int. Conf. Special Interest Group Data Commun. (SIGCOMM)*. ACM, Oct. 2007, pp. 1–14.

[13] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Commun. Mag.*, vol. 51, no. 4, pp. 142–149, Apr. 2013.

[14] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. M. 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.

[15] 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.

[16] Y. Liu, Z. Ding, M. Elkashlan, and H. V. Poor, "Cooperative non-orthogonal multiple access with simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 938–953, Apr. 2016.

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

[18] H. ElSawy, E. Hossain, and M. Haenggi, "Stochastic geometry for modeling, analysis, and design of multi-tier and cognitive cellular wireless networks: A survey," *IEEE Commun. Surv. Tuts.*, vol. 15, no. 3, pp. 996–1019, Jun. 2013.

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

[20] C. Yang, Y. Yao, Z. Chen, and B. Xia, "Analysis on cache-enabled wireless heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 131–145, Jan. 2016.

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

[22] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[23] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, and L. Hanzo, "Nonorthogonal multiple access for 5G and beyond," *Proc. IEEE*, vol. 105, no. 12, pp. 2347–2381, Dec. 2017.

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

[25] T. S. Rappaport, F. Gutierrez, E. Ben-Dor, J. N. Murdock, Y. Qiao, and J. I. Tamir, "Broadband millimeter-wave propagation measurements and models using adaptive-beam antennas for outdoor urban cellular communications," *IEEE Trans. Antennas Propag.*, vol. 61, no. 4, pp. 1850–1859, Apr. 2013.

[26] A. V. Alejos, M. G. Sanchez, and I. Cuinas, "Measurement and analysis of propagation mechanisms at 40 GHz: Viability of site shielding forced

by obstacles," *IEEE Trans. Veh. Technol.*, vol. 57, no. 6, pp. 3369–3380, Nov. 2008.

[27] 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, May 2013.

[28] H. Elshaer, M. N. Kulkarni, F. Boccardi, J. G. Andrews, and M. Dohler, "Downlink and uplink cell association with traditional macrocells and millimeter wave small cells," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6244–6258, Sep. 2016.

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

[30] X. Yu, J. Zhang, M. Haenggi, and K. B. Letaief, "Coverage analysis for millimeter wave networks: The impact of directional antenna arrays," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1498–1512, Jul. 2017.

[31] D. Maamari, N. Devroye, and D. Tuninetti, "Coverage in mmWave cellular networks with base station co-operation," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2981–2994, Apr. 2016.

[32] W. Yi, Y. Liu, and A. Nallanathan, "Modeling and analysis of D2D millimeter-wave networks with Poisson cluster processes," *IEEE Trans. Commun.*, vol. 65, no. 12, pp. 5574–5588, Dec. 2017.

[33] T. S. Rappaport, R. W. Heath Jr, R. C. Daniels, and J. N. Murdock, *Millimeter wave wireless communications*. Pearson Education, 2014.

[34] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 101–107, Jun. 2011.

[35] C. Park and T. S. Rappaport, "Short-range wireless communications for next-generation networks: UWB, 60 GHz millimeter-wave WPAN, and ZigBee," *IEEE Trans. Wireless Commun.*, vol. 14, no. 4, pp. 70–78, Aug. 2007.

[36] O. Semiari, W. Saad, and M. Bennis, "Downlink cell association and load balancing for joint millimeter wave-microwave cellular networks," in *IEEE Proc. of Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1–6.

[37] G. Athanasiou, P. C. Weeraddana, C. Fischione, and L. Tassiulas, "Optimizing client association for load balancing and fairness in millimeter-wave wireless networks," *IEEE/ACM Trans. on Netw.*, vol. 23, no. 3, pp. 836–850, Jun. 2015.

[38] S. Singh, H. S. Dhillon, and J. G. Andrews, "Offloading in heterogeneous networks: Modeling, analysis, and design insights," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 2484–2497, May 2013.

[39] H. S. Dhillon, R. K. Ganti, F. Baccelli, and J. G. Andrews, "Modeling and analysis of K-tier downlink heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 550–560, Apr. 2012.

[40] "Propagation data and prediction methods required for the design of terrestrial broadband radio access systems operating in a frequency range from 3 to 60 GHz," *ITU-R, Tech. Rep.*, 2012.

[41] X. Zhang and J. G. Andrews, "Downlink cellular network analysis with multi-slope path loss models," *IEEE Trans. Commun.*, vol. 63, no. 5, pp. 1881–1894, May 2015.

[42] 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 Trans. Commun.*, vol. 65, no. 1, pp. 403–430, Jan. 2017.

[43] M. D. Renzo, "Stochastic geometry modeling and analysis of multi-tier millimeter wave cellular networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 9, pp. 5038–5057, Sep. 2015.

[44] A. Thornburg, T. Bai, and R. W. Heath, "Performance analysis of outdoor mmWave ad hoc networks," *IEEE Trans. Signal Process.*, vol. 64, no. 15, pp. 4065–4079, Aug. 2016.

[45] Z. Ding, P. Fan, and H. V. Poor, "Random beamforming in millimeter-wave noma networks," *IEEE Access*, vol. 5, pp. 7667–7681, 2017.

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

[47] K. Wang, Z. Chen, and H. Liu, "Push-based wireless converged networks for massive multimedia content delivery," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2894–2905, May 2014.

[48] N. Golrezaei, A. G. Dimakis, and A. F. Molisch, "Scaling behavior for device-to-device communications with distributed caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4286–4298, Jul. 2014.

[49] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in *IEEE Proc. of International Commun. Conf. (ICC)*, Jun. 2015, pp. 3358–3363.

[50] G. Lee, Y. Sung, and J. Seo, "Randomly-directional beamforming in millimeter-wave multiuser MISO downlink," *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1086–1100, Feb. 2016.

[51] Y. Liu, Z. Qin, M. Elkashlan, Y. Gao, and L. Hanzo, "Enhancing the physical layer security of non-orthogonal multiple access in large-scale networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1656–1672, Mar. 2017.

[52] S. Deng, M. K. Samimi, and T. S. Rappaport, "28 GHz and 73 GHz millimeter-wave indoor propagation measurements and path loss models," in *Prof. IEEE Int. Conf.on Commun. Workshop (ICCW)*, Jun. 2015, pp. 1244–1250.

[53] T. S. Rappaport, E. Ben-Dor, J. N. Murdock, and Y. Qiao, "38 GHz and 60 GHz angle-dependent propagation for cellular & peer-to-peer wireless communications," in *IEEE Proc. of International Commun. Conf. (ICC)*, Jun. 2012, pp. 4568–4573.

[54] M. Afshang, H. S. Dhillon, and P. H. J. Chong, "Modeling and performance analysis of clustered device-to-device networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4957–4972, Jul. 2016.

[55] A. Jeffrey and D. Zwillinger, *Table of integrals, series, and products*. Academic press, 2007.

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