# On the Performance of Cache-Enabled Hybrid Wireless Networks

# On the Performance of Cache-Enabled Hybrid Wireless Networks

Tong Zhang, Sudip Biswas, *Member, IEEE* and Tharmalingam Ratnarajah, *Senior Member, IEEE*

***Abstract*—To alleviate the backhaul congestion in future hybrid heterogenous networks, this paper investigates the potential benefits of implementing storages in small base stations (SBSs) operating at frequency range 2 (FR2) bands that co-exist with a tier of massive multiple input multiple output (MIMO) macro BSs (MBSs) operating at frequency range 1 (FR1) bands. We develop a unified analytical framework and derive theoretical bounds for such a cache-enabled FR1–FR2 hybrid network under limited backhaul scenario to analyze the exact and approximate latency, average success probability of file delivery, and average data rate considering two open-access user association policies: i) location-based and ii) content-based association. Numerical results demonstrate that wireless edge caching (WEC) can improve the performance of hybrid wireless networks, albeit certain trade-offs, *e.g.,* increasing cache-enabled SBSs cannot always improve the network performance and there exists an optimal SBS density that provides the best latency and throughput performance. Furthermore, we compare the performance of the network with respect to other key network design parameters such as cache size, content popularity, backhaul capacity, and blockages for both the user associations. Our results show that latency under content-based user association is less than that of location-based user association, and although the difference in the average rates under the two user associations is not obvious, content-based association can extricate more backhaul capacity and thus reduce installation cost significantly.**

***Index Terms*—Wireless edge caching, FR1, FR2, massive MIMO, stochastic geometry, hybrid HetNets.**

## I. Introduction

The rapid proliferation of mobile devices and the emergence of new bandwidth-sensitive applications e.g., augmented reality and internet of things have led to an unprecedented growth in the global mobile data traffic, to which fifth generation (5G) solutions such as network densification through small cells and heterogenous networks (HetNets) [1], moving to frequency range 2 (FR2)–millimeter wave (mmWave) bands [2], and massive multiple-input multiple-output (MIMO) communication in the FR1–sub-6 GHz bands [3] have been proposed. However, while the above solutions are beneficial for access links, they do little to alleviate the burden on backhaul links, which is further exaggerated due to the substantial amount of redundant and repeated requests generated over networks [4]. Thus, pre-fetching certain popular files in the local caches of small cell base stations (BSs) in off-peak hours, also termed as wireless edge caching (WEC), can alleviate network backhaul traffic loads, whereby the requested content will be served directly to the users by one of the neighbouring BSs depending on the availability of the file in its local cache and the association criteria of the users to the BSs.

The above discussion clearly adds up to the fact that the 5G solutions currently being considered for access along with WEC strive towards fulfilling common goals of improving the quality of service (QoS) of networks and the quality of experience (QoE) for users, which makes it imperative to investigate the performance of these technologies in a co-existed network model. While WEC have gained significant attention of late with many undergoing studies on the design and analysis of cache-enabled wireless networks [1], [4]–[6], most studies do not consider a realistic 5G network scenario. For example, while a framework for femto caching with distributed caching helpers and low-rate backhaul capacity but high storage capacity was proposed in [4], outage probability and average delivery rate were analyzed in cache-enabled small cell networks in [5]. Further, in [1] cache-enabled HetNets consisting of a tier of multi-antenna MBSs overlaid with a tier of caching helpers was considered and accordingly the optimal caching placement was analyzed. In [6], the authors analyzed the average ergodic rate, outage probability, throughput, and delay under a general three-tier cache-enabled HetNet consisting of the traditional cellular BSs, cache-enabled relays, and cache-enabled user equipment forming device-to-device (D2D) communications. None of the above frameworks consider any of the standard 5G technologies, except for [1] and [6], which at best implement network densification. Nevertheless, some recent works have integrated caching with either FR2 [7]–[9] or massive MIMO [10]–[12] communication, though no work till date has integrated both technologies with caching in a hybrid network framework.

Motivated by the above, this paper explores and analyzes an ambitious but unified cache-assisted hybrid FR1–FR2 5G network model, whereby a HetNet involving multiple FR2 small base stations (SBSs) equipped with storage memory for caching popular files and assisted by FR1 massive MIMO macro base stations (MBSs) serves multiple users in the downlink. This 5G framework has been leveraged from the concept of femto caching with the assumption that only SBSs are equipped with storage (caching helpers in femto caching), while MBSs are the traditional base stations (BSs) without any storage memory[1]. Further, all BSs are connected to the core network via capacity-limited backhaul, whereby in the event of a cache miss and depending on the user association policy, the non-cached files are retrieved through the backhaul link.

The work of Tong Zhang and Tharmalingam Ratnarajah was supported in part by the U.K. Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/P009549/1, in part by the U.K.-India Education and Research Initiative Thematic Partnerships under Grant DST UKIERI-2016-17-0060, and in part by the SPARC Project 148. The work of Sudip Biswas was supported by Indian Institute of Information Technology (IIIT) Guwahati through TEQUIP-III of Govt. of India. (Corresponding author: Tharmalingam Ratnarajah (t.ratnarajah@ed.ac.uk))

Tong Zhang and Tharmalingam Ratnarajah are with University of Edinburgh, UK. Email: {t.zhang, t.ratnarajah}@ed.ac.uk.

Sudip Biswas is with Indian Institute of Information Technology Guwahati, India. Email: sudip.biswas@iiitg.ac.in.

[1]Early 5G deployment will ensure that FR2 SBSs co-exist with FR1 macro BSs to provide seamless connectivity.

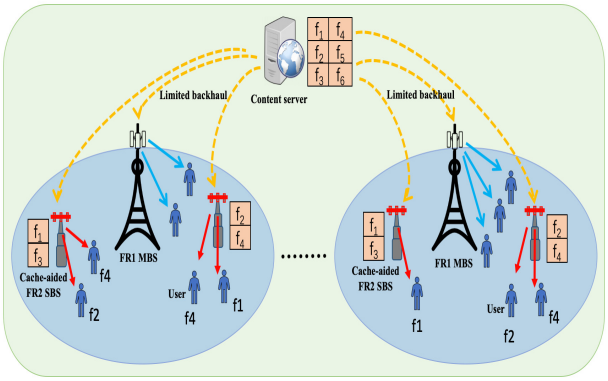



Fig. 1: An illustration of a hybrid FR1–FR2 cache enabled heterogenous network. The SBSs are equipped with storage memory, where UC and MC WEC schemes are implemented from a probabilistic point of view.

At this point, it is worth noting that user association plays a significant role in the performance analysis of the cache-aided hybrid network. In fact, various association strategies, ranging from cache-aware association, where users are only associated with the feasible SBSs that store the requested files and provide the best received signal power [13]–[15], to traditional location-based user association where users are only associated with the nearest BSs, independent of the availability of files in the local caches [6] have been proposed in literature. However, the performance gains and trade-offs between cache-aware and traditional location-based user association strategies under a unified 5G hybrid framework with limited backhaul have not yet been investigated and thus is not well understood. Accordingly, this work analyzes two different user associations: content-based and location-based considering two of the most common proactive caching placement schemes: caching most popular contents (MC) and uniform caching (UC), to give a holistic analysis with respect to three quintessential performance metrics, namely transmission latency, ASP of file delivery, and the average data rate of the typical user. By resorting to stochastic geometric tools, the association probability, the distribution of the serving distance, and the cell load under the two user associations are derived. Next, approximated data rates for the user considering FR2 transmission with hybrid beamforming and massive MIMO with pilot contamination are provided. Finally, to obtain valuable insights on network design, the performance of the cache-enabled hybrid network is analyzed in terms of several network parameters such as SBS density, cache size, content popularity distribution, backhaul capacity, different caching placement strategies, and blockage density.

## II. System Model

In this section, we introduce the hybrid network topology, propagation models for both FR1 and FR2 communications, and the WEC framework. For better illustration, some of the important notations used in the paper are summarized in Table I.

*1) Network topology:* We consider a two-tier hybrid HetNet, where cache-enabled FR2 SBSs are assisted by FR1 MBSs to enable an open access user association policy[2]. The SBSs and MBSs are modelled as two independent PPPs $\Phi_{\text{FR2}}$ with density $\lambda_{\text{FR2}}$ and $\Phi_{\text{FR1}}$ with density $\lambda_{\text{FR1}}$, respectively. All BSs are connected to the core-network through limited backhaul to retrieve any non-cached files[3] which result in both delay and rate limitations. We assume that each BS is allocated with equal backhaul capacity, such that the backhaul capacity of each BS is given by $C_b = \frac{c_{b_1}}{\lambda_{\text{FR1}}+\lambda_{\text{FR2}}} + c_{b_2}$, with $c_{b_1} > 0$ and $c_{b_2} \geq 0$ being arbitrary values to model the backhaul rate limitation for different scenarios [16], [17]. The users in the network follow another homogeneous PPP $\Phi_u$ with density $\lambda_u$. Further, both FR1 and FR2 BSs are assumed to be equipped with multiple antennas $n_t^{\text{FR1}}$ and $n_t^{\text{FR2}}$ with transmit power $\text{P}_{\text{FR1}}$ and $\text{P}_{\text{FR2}}$, respectively. Due to two different transmission frequencies, the users are assumed to be equipped with two different sets of radio frequency (RF) chain modules with antennas $n_r^{\text{FR1}}$ and $n_r^{\text{FR2}}$ to independently receive FR1 and FR2 signals, respectively[4]. Furthermore, based on Slivnyak's theorem, the analysis hereinafter is performed for the typical user located at the origin, denoted by the subscript 0. The users are associated with their serving MBSs or cache-enabled SBSs based on two different user association policies as will be described in Section III. Each FR2 SBS serves its associated users through a fully connected *hybrid beamforming architecture*, while each FR1 MBS is equipped with *massive MIMO antennas* to serve its associated users. Since FR1 and FR2 transmissions occur in different frequency bands, they do not interfere each other.

**Remark 1.** *The idea of a dense heterogenous hybrid network is that each small cell will cater towards a small number of users in the FR2 frequency band. Though beyond the scope of this paper, content placement phase is an important aspect of WEC, which leans on the efficiency of content prediction algorithms. Since the number of users associated with each SBS is expected to be much smaller (owing to network densification) than that of a MBS, hence, designing prediction algorithms catering towards a few users is much more efficient in terms of computational complexity and storage capacity. Accordingly, in this work we consider only the SBSs to be equipped with storage memory.*

*2) Propagation model:* With regards to FR2 transmission, we assume a geometric FR2 channel model similar to [18] based on the steering vector of arrival and departure angles, where each path gains are assumed to follow complex Gaussian distribution. The propagation model includes both large-scale path loss and small scale fading. Accordingly, considering both line-of-sight (LOS) and non-LOS (NLOS) transmissions, the path loss model for FR2 communication is given as $r_{xny}^{-\alpha_{\mathfrak{L}}}$ and $r_{xny}^{-\alpha_{\mathfrak{N}}}$, where $\alpha_{\mathfrak{L}}$ and $\alpha_{\mathfrak{N}}$ denote path loss exponents for LOS and NLOS, respectively. Next, we consider a statistical approach to model the blockages since it can accommodate varying blockage parameters including their density. In particular, a two-state stationary probabilistic "exponential blockage model" proposed and validated in [19],

[2]Open access user association is considered to overcome the short propagation and blockage-sensitivity constraints of FR2 signals, whereby traditional MBSs assist the SBSs to provide seamless and reliable coverage to the users.

[3]These files are the ones requested by the users that are not stored in the SBSs.

[4]We assume $n_r^{\text{FR1}} = 1$ and $n_r^{\text{FR2}} >> 1$ due to the intrinsic relation between the signal wavelength and antenna separation.

TABLE I: Notation Summary

| Notations | Physical meaning |
|---|---|
| $\Phi_{\text{FR1}}$, $\Phi_{\text{FR2}}$, $\Phi_u$ | PPP distributed locations of FR1 MBSs, FR2 SBSs, and UEs |
| $\lambda_{\text{FR1}}$, $\lambda_{\text{FR2}}$, $\lambda_u$, $\lambda_g$ | Spatial densities of FR1 MBSs, FR2 SBSs, UEs, and gateways |
| $n_t^{\text{FR1}}$, $n_t^{\text{FR2}}$ | Number of transmit antennas at each FR1 MBS and FR2 SBS |
| $N_{\text{RF}}$ | Number of RF chains |
| $n_r^{\text{FR1}}$, $n_r^{\text{FR2}}$ | Number of receive antennas at each UE to receive FR1 and FR2 signals |
| $\text{P}_{\text{FR1}}$, $\text{P}_{\text{FR2}}$ | Transmit power of each FR1 MBS and FR2 SBS |
| $\mathcal{F}$ (*i.e.,* $\lvert\mathcal{F}\rvert = F$) | The limited file set with a total of $F$ files |
| $q_i$ for $\forall i \in [1, F]$ | The probability of each user requesting the $i$th file, denoted as $f_i$ |
| $\omega_{\text{FR2}_i}$ for $\forall i \in [1, F]$ | The probability of each SBS storing the $i$th file |
| $C_{\text{FR2}}$, $S$ | Cache sizes of each FR2 SBS, File size in bits |
| $C_b$ | Backhaul capacity per BS (either FR2 or FR1) |
| $x_{\text{FR2}}^{\mathfrak{L}}$, $x_{\text{FR2}}^{\mathfrak{N}}$ | Locations of the associated FR2 SBS with LOS and NLOS transmissions |
| $x_{\text{FR1}}$ | Location of the associated FR1 MBS |
| $\beta$ | Blockage density |
| $B_{\text{FR2}}$, $B_{\text{FR1}}$ | Biased factors |
| $p_{\mathfrak{L}}(\cdot)$, $p_{\mathfrak{N}}(\cdot)$ | LOS and NLOS probabilities of each path |
| $\mathcal{U}_{(\cdot)}$ (*i.e.,* $U_{(\cdot)} = \lvert\mathcal{U}_{(\cdot)}\rvert$) | Set of users that can be served by a certain BS (either FR2 or FR1) |
| $\eta_j$ with $j \in \{\mathfrak{L}, \mathfrak{N}\}$ | Number of FR2 LOS and NLOS transmission paths |
| $\phi$, $\theta$ | AOD and AOA |
| $h_{(*)}^{\text{FR1}}$, $h_{(*)}^{\text{FR2}}$ | FR1 and FR2 channel fading coefficients |

[20] is used to give the probabilities of occurrence $p_{\mathfrak{L}}(\cdot)$ and $p_{\mathfrak{N}}(\cdot)$ of LOS and NLOS paths as $p_{\mathfrak{L}}(r) = e^{-\beta r}$ and $p_{\mathfrak{N}}(r) = 1 - e^{-\beta r}$, respectively, where $\beta$ is the blockage density and $r$ is the link length. As for the small scale fading, the geometric channel matrix between the FR2 BS at $x$ and the user $u_{ny}$ is given as

$$\mathbf{H}_{xny} = \sqrt{\frac{n_r^{\text{FR2}} n_t^m}{r_{xny}^{\alpha_m} \eta_{xny}}} \sum_{k=1}^{\eta_{xny}} h_{kxny}^{\text{FR2}} \mathbf{a}_u(\theta_{kxny}) \mathbf{a}_m^H(\phi_{kxny}), \quad (1)$$

where $h_{kxny}^{\text{FR2}} \sim CN(0,1)$ is the small scale fading coefficient[5] on the $k$th path [18], [21]–[23]. $\alpha_m \in \{\alpha_{\mathfrak{L}}, \alpha_{\mathfrak{N}}\}$, $\eta_{xny} \in \{\eta_{\mathfrak{L}}, \eta_{\mathfrak{N}}\}$ is the number of scatters depending on LOS or NLOS path such that $1 < \eta_{\mathfrak{L}} < \eta_{\mathfrak{N}}$, and $\mathbf{a}_u(\theta)$ and $\mathbf{a}_{\text{FR2}}(\phi)$ are steering vectors of users and SBSs, respectively. In view of the sparsity of FR2 channels, this work assumes that all scatters take place in the azimuth plane and are uniformly distributed in [0, $2\pi$]. Therefore, the steering vectors are described by uniform linear array (ULA) of size $n_t^{\text{FR2}}$ and $n_r^{\text{FR2}}$, respectively, where $\mathbf{a}_{\text{FR2}}(\phi) = \frac{1}{n_t^{\text{FR2}}}[1\ e^{jkd\sin(\phi)} \cdots e^{(n_t^{\text{FR2}}-1)jkd\sin(\phi)}]$ and $\mathbf{a}_u(\theta) = \frac{1}{n_r^{\text{FR2}}}[1\ e^{jkd\sin(\theta)} \cdots e^{(n_r^{\text{FR2}}-1)jkd\sin(\theta)}]$ with $\phi$ and $\theta$ denoted respectively as the angles of departure (AOD) and the angles of arrival (AOA). Further, $k = \frac{2\pi}{\lambda_c}$, $\lambda_c$ is the wavelength and $d$ is the distance between antenna elements.

Similarly, the path loss model for FR1 communication is given as $r_{xny}^{-\alpha_{\text{FR1}}}$, where $\alpha_{\text{FR1}}$ is the path loss exponent and $r_{xny}$ is the distance between the BS at $x$ and the user $n$ served by the BS at $y$, denoted as $u_{ny}$. The small scale fading of the link between the $k$th-antenna of the FR1 MBS at $x$ and the user $u_{ny}$, denoted as $h_{kxny}^{\text{FR1}}$ is modelled as an independent and identical distributed (i.i.d) quasi-static Rayleigh fading $h_{kxny}^{\text{FR1}} \sim \mathcal{CN}(0,1)$, which is widely used in the analysis of massive MIMO systems [12].

**Remark 2.** *The considered small scale fading model for FR2 transmission is widely used in a bulk of mmWave literature, such as [21], [22], [24], [25]. The consideration of a more general Nakagami-$M$ fading may be found in some mmWave literature. However, the choice of fading does not make a significant difference in the performance analysis of FR2 networks. This is primarily because the performance trends are somewhat robust to the choice of fading distribution as long as the distance-dependent channel components are included.*

*3) Caching model:* This work considers a finite file set $\mathcal{F} = \{f_1, f_2, \ldots, f_F\}$, where $F$ is the total number of files and each file is of size $S$ bits[6]. Further, we assume that each user independently requests the $i$th file in $\mathcal{F}$ with probability $q_i$, modelled as Zipf distribution and given as $q_i = (i^{\upsilon} \sum_{j=1}^{F} j^{-\upsilon})^{-1}$. Here, $i \in \{1, 2, \ldots, F\}$ and $\upsilon$ is the tuning parameter that controls the skewness of the content popularity distribution. Further, each FR2 SBS is equipped with a storage capacity of $C_{\text{FR2}}S$ bits, such that $C_{\text{FR2}} \leq F$. To cache popular contents in the SBSs, we consider off-line caching (i.e., proactive caching by prefetching contents in advance before the user requests are revealed) rather than on-line caching (i.e., reactive caching, where some cached files are discarded after specific time intervals according to reactive caching algorithms). By considering a generic content caching placement policy in a probabilistic manner, we define a caching probability set $\Omega_{\text{FR2}} = \{\omega_{\text{FR2}_1}, \omega_{\text{FR2}_2}, \ldots, \omega_{\text{FR2}_F}\}$, such that $0 \leq \omega_{\text{FR2}_i} \leq 1$ and $\sum_{i=1}^{F} \omega_{\text{FR2}_i} \leq C_{\text{FR2}}$ to satisfy the storage capacity constraint in an average sense. Accordingly, we consider two commonly used proactive caching strategies that are primarily influenced by content popularity: $i$) *uniformly caching all files* (UC) with caching probability $\omega_{\text{FR2}_i} = \frac{C_{\text{FR2}}}{F}$ $\forall i \in \{1, 2, \ldots, F\}$ and $ii$) *caching most popular files* (MC) with caching probability $\omega_{\text{FR2}_i} = 1$ for $1 \leq i \leq C_{\text{FR2}}$ and $\omega_{\text{FR2}_i} = 0$ $\forall i > C_{\text{FR2}}$.

## III. Association probability and rate characterization

*1) Association policy:* Based on an open access user association scheme [6], [26], we consider two association policies that will be interchangeably considered throughout the rest of the paper.

[5]The performance trends are somehow robust to the selection of small scale fading distribution whereas the distance-dependent channel components are included. Since Nakagami or complex Gaussian consumption does not make a significant difference in the performance analysis, we apply the latter for the sake of tractability in stochastic geometric framework.

[6]In the event of unequal file size, each file can be divided into small partitions of the same size, with each partition being treated as an individual file.

**Location-based user association (LBUA):** Users associate with either MBSs or SBSs depending on the least biased path loss seen at users, but independent of the availability of files in the local cache. Such an association sacrifices caching gain to acquire the best coverage for users, while offloading traffic to a feasible extent from MBSs to SBSs. The biased path loss seen at the typical user from the tagged FR2 SBS in LOS (NLOS) transmission located at $x^{\mathfrak{L}}_{\text{FR2}} \in \Phi^{\mathfrak{L}}_{\text{FR2}}$ ($x^{\mathfrak{N}}_{\text{FR2}} \in \Phi^{\mathfrak{N}}_{\text{FR2}}$) and the tagged FR1 MBS located at $x_{\text{FR1}} \in \Phi_{\text{FR1}}$ are respectively given as $B_{\text{FR2}} r^{-\alpha_{\mathfrak{L}}}_{x^{\mathfrak{L}}_{\text{FR2}} 0 x^{\mathfrak{L}}_{\text{FR2}}}$ ($B_{\text{FR2}} r^{-\alpha_{\mathfrak{N}}}_{x^{\mathfrak{N}}_{\text{FR2}} 0 x^{\mathfrak{N}}_{\text{FR2}}}$) and $B_{\text{FR1}} r^{-\alpha_{\text{FR1}}}_{x_{\text{FR1}} 0 x_{\text{FR1}}}$, where $B_{\text{FR2}}$ and $B_{\text{FR1}}$ are the bias factors controlling the cell range, such that $B_{\text{FR2}} < B_{\text{FR1}}$ to offload more traffic from MBSs to SBSs. Now, according to the thinning theorem, the two inhomogeneous sub-PPPs $\Phi^{\mathfrak{L}}_{\text{FR2}}$ and $\Phi^{\mathfrak{N}}_{\text{FR2}}$ have densities $\lambda_{\text{FR2}} p_{\mathfrak{L}}(\cdot)$ and $\lambda_{\text{FR2}} p_{\mathfrak{N}}(\cdot)$, respectively.

**Content-based user association (CBUA):** In cache-enabled hybrid HetNets, caching may change the way users associate themselves with BSs. This association policy ensures benefits, such as minimal latency for access, backhaul load alleviation, and reduction in backhaul installation cost. Here, users associate with either cache-hit SBSs (*i.e.,* feasible SBSs that store the requested files) or wide-coverage provider MBSs based on the least biased path loss. However, in certain instances CBUA tends to sacrifice coverage performance since users might associate with far-away cache-hit feasible SBSs rather than the true least-path-loss SBS. Therefore, the design strategy and henceforth acquiring the benefits of caching in SBSs in open access networks is one of the primary challenges of this strategy. By slight abuse of notations, the biased path loss seen at the typical user requesting the $i$th file from the tagged cache-hit SBS under LOS (NLOS) transmission located at $x^{\mathfrak{L}}_{\text{FR2}_i} \in \Phi^{\mathfrak{L}}_{\text{FR2}_i}$ ($x^{\mathfrak{N}}_{\text{FR2}_i} \in \Phi^{\mathfrak{N}}_{\text{FR2}_i}$) and the tagged FR1 MBS located at $x_{\text{FR1}} \in \Phi_{\text{FR1}}$ are respectively given as $B_{\text{FR2}} r^{-\alpha_{\mathfrak{L}}}_{x^{\mathfrak{L}}_{\text{FR2}_i} 0 x^{\mathfrak{L}}_{\text{FR2}_i}}$ ($B_{\text{FR2}} r^{-\alpha_{\mathfrak{N}}}_{x^{\mathfrak{N}}_{\text{FR2}_i} 0 x^{\mathfrak{N}}_{\text{FR2}_i}}$) and $B_{\text{FR1}} r^{-\alpha_{\text{FR1}}}_{x_{\text{FR1}} 0 x_{\text{FR1}}}$. The two inhomogeneous sub-PPPs $\Phi^{\mathfrak{L}}_{\text{FR2}_i}$ and $\Phi^{\mathfrak{N}}_{\text{FR2}_i}$ have densities $\lambda_{\text{FR2}} p_{\mathfrak{L}}(\cdot)\omega_{\text{FR2}_i}$ and $\lambda_{\text{FR2}} p_{\mathfrak{N}}(\cdot)\omega_{\text{FR2}_i}$, respectively. In particular, if the SBS is the serving BS, the requested file will always be served without the need for backhaul connection. Thus, CBUA helps to conserve the backhaul bandwidth and accordingly, the backhaul capacity of each MBS is given as $C_b = \frac{c_{b_1}}{\lambda_{\text{FR1}}} + c_{b_2}$.

*2) Association probabilities:* Let $\mathcal{A}^{\mathfrak{L}}_{\text{FR2}}(\mathcal{A}^{\mathfrak{L}}_{\text{FR2}_i})$, $\mathcal{A}^{\mathfrak{N}}_{\text{FR2}}(\mathcal{A}^{\mathfrak{N}}_{\text{FR2}_i})$, and $\mathcal{A}_{\text{FR1}}(\mathcal{A}_{\text{FR1}_i})$ denote the probabilities of the user associated with the FR2 LOS and NLOS SBS and FR1 MBS under LBUA (CBUA [for the typical user requesting the $i$th file]), respectively. Leveraging the results from [27], the relative association probabilities under LBUA for the considered hybrid HetNet can be given as

$$\mathcal{A}^{\mathfrak{L}}_{\text{FR2}} = \int_0^\infty \exp\left[-\pi\lambda_{\text{FR1}}\left(k_1 R^{\alpha_{\mathfrak{L}}}\right)^{\frac{2}{\alpha_{\text{FR1}}}} - 2\pi\lambda_{\text{FR2}}\left(\hat{Z}\left(R^{\frac{\alpha_{\mathfrak{L}}}{\alpha_{\mathfrak{N}}}}\right) + Z\left(R\right)\right)\right] \times 2\pi\lambda_{\text{FR2}} p_{\mathfrak{L}}(R) R \mathrm{d}R, \tag{2}$$

$$\mathcal{A}^{\mathfrak{N}}_{\text{FR2}} = \int_0^\infty \exp\left[-\pi\lambda_{\text{FR1}}\left(k_1 R^{\alpha_{\mathfrak{N}}}\right)^{\frac{2}{\alpha_{\text{FR1}}}} - 2\pi\lambda_{\text{FR2}}\left(Z\left(R^{\frac{\alpha_{\mathfrak{N}}}{\alpha_{\mathfrak{L}}}}\right) + \hat{Z}\left(R\right)\right)\right] \times 2\pi\lambda_{\text{FR2}} p_{\mathfrak{N}}(R) R \mathrm{d}R, \tag{3}$$

$$\mathcal{A}_{\text{FR1}} = \int_0^\infty \exp\left[-2\pi\lambda_{\text{FR2}}\left(Z\left(\left(k_1^{-1} R^{\alpha_{\text{FR1}}}\right)^{\frac{1}{\alpha_{\mathfrak{L}}}}\right) + \hat{Z}\left(\left(k_1^{-1} R^{\alpha_{\text{FR1}}}\right)^{\frac{1}{\alpha_{\mathfrak{N}}}}\right)\right)\right] \times \exp\left[-\pi\lambda_{\text{FR1}} R^2\right] 2\pi\lambda_{\text{FR1}} R \mathrm{d}R, \tag{4}$$

where $k_1 = \frac{B_{\text{FR1}}}{B_{\text{FR2}}}$, $Z(x) = -\frac{1}{\beta} x e^{-\beta x} - \frac{1}{\beta^2}(e^{-\beta x} - 1)$ and $\hat{Z}(x) = \frac{x^2}{2} + \frac{1}{\beta} x e^{-\beta x} + \frac{1}{\beta^2}(e^{-\beta x} - 1)$. Further, the probability that the typical user is associated with a FR2 BS can be obtained as $\mathcal{A}_{\text{FR2}} = \mathcal{A}^{\mathfrak{L}}_{\text{FR2}} + \mathcal{A}^{\mathfrak{N}}_{\text{FR2}}$ such that $\mathcal{A}_{\text{FR2}} + \mathcal{A}_{\text{FR1}} = 1$. Similarly, by substituting $\hat{\lambda}_{\text{FR2}_i} = \lambda_{\text{FR2}} \omega_{\text{FR2}_i}$ for $\lambda_{\text{FR2}}$ in (2) – (4) (as the FR2 SBSs in CBUA stores the $i$th file), the relative association probabilities for the $i$th file are given by

$$\mathcal{A}^{\mathfrak{L}}_{\text{FR2}_i} = \int_0^\infty \exp\left[-\pi\lambda_{\text{FR1}}\left(k_1 R^{\alpha_{\mathfrak{L}}}\right)^{\frac{2}{\alpha_{\text{FR1}}}} - 2\pi\hat{\lambda}_{\text{FR2}_i}\left(\hat{Z}\left(R^{\frac{\alpha_{\mathfrak{L}}}{\alpha_{\mathfrak{N}}}}\right) + Z\left(R\right)\right)\right] \times 2\pi\hat{\lambda}_{\text{FR2}_i} p_{\mathfrak{L}}(R) R \mathrm{d}R, \tag{5}$$

$$\mathcal{A}^{\mathfrak{N}}_{\text{FR2}_i} = \int_0^\infty \exp\left[-\pi\lambda_{\text{FR1}}\left(k_1 R^{\alpha_{\mathfrak{N}}}\right)^{\frac{2}{\alpha_{\text{FR1}}}} - 2\pi\hat{\lambda}_{\text{FR2}_i}\left(Z\left(R^{\frac{\alpha_{\mathfrak{N}}}{\alpha_{\mathfrak{L}}}}\right) + \hat{Z}\left(R\right)\right)\right] \times 2\pi\hat{\lambda}_{\text{FR2}_i} p_{\mathfrak{N}}(R) R \mathrm{d}R, \tag{6}$$

$$\mathcal{A}_{\text{FR1}_i} = \int_0^\infty \exp\left[-2\pi\hat{\lambda}_{\text{FR2}_i}\left(Z\left(\left(k_1^{-1} R^{\alpha_{\text{FR1}}}\right)^{\frac{1}{\alpha_{\mathfrak{L}}}}\right) + \hat{Z}\left(\left(k_1^{-1} R^{\alpha_{\text{FR1}}}\right)^{\frac{1}{\alpha_{\mathfrak{N}}}}\right)\right)\right] \times \exp\left[-\pi\lambda_{\text{FR1}} R^2\right] 2\pi\lambda_{\text{FR1}} R \mathrm{d}R. \tag{7}$$

Accordingly, the probability that the typical user is associated with a FR2 SBS is $\mathcal{A}_{\text{FR2}_i} = \mathcal{A}^{\mathfrak{L}}_{\text{FR2}_i} + \mathcal{A}^{\mathfrak{N}}_{\text{FR2}_i}$ such that $\mathcal{A}_{\text{FR2}_i} + \mathcal{A}_{\text{FR1}_i} = 1$.

*3) Statistical distribution of the serving distance:* The relative probability density function (PDF) of the serving distances $R^*_{x^{\mathfrak{L}}_{\text{FR2}} 0 x^{\mathfrak{L}}_{\text{FR2}}}$, $R^*_{x^{\mathfrak{N}}_{\text{FR2}} 0 x^{\mathfrak{N}}_{\text{FR2}}}$, and $R^*_{x_{\text{FR1}} 0 x_{\text{FR1}}}$ for LBUA are given as [26]

$$\hat{f}_{R^*_{x^{\mathfrak{L}}_{\text{FR2}} 0 x^{\mathfrak{L}}_{\text{FR2}}}}(R) = \frac{1}{\mathcal{A}^{\mathfrak{L}}_{\text{FR2}}} \exp\left[-\pi\lambda_{\text{FR1}}\left(k_1 R^{\alpha_{\mathfrak{L}}}\right)^{\frac{2}{\alpha_{\text{FR1}}}} - 2\pi\lambda_{\text{FR2}}\left(\hat{Z}\left(R^{\frac{\alpha_{\mathfrak{L}}}{\alpha_{\mathfrak{N}}}}\right) + Z(R)\right)\right] 2\pi\lambda_{\text{FR2}} p_{\mathfrak{L}}(R) R, \tag{8}$$

$$\hat{f}_{R^*_{x^{\mathfrak{N}}_{\text{FR2}} 0 x^{\mathfrak{N}}_{\text{FR2}}}}(R) = \frac{1}{\mathcal{A}^{\mathfrak{N}}_{\text{FR2}}} \exp\left[-\pi\lambda_{\text{FR1}}\left(k_1 R^{\alpha_{\mathfrak{N}}}\right)^{\frac{2}{\alpha_{\text{FR1}}}} - 2\pi\lambda_{\text{FR2}}\left(Z\left(r^{\frac{\alpha_{\mathfrak{N}}}{\alpha_{\mathfrak{L}}}}\right) + \hat{Z}(R)\right)\right] 2\pi\lambda_{\text{FR2}} p_{\mathfrak{N}}(R) R, \tag{9}$$

$$\hat{f}_{R^*_{x_{\text{FR1}} 0 x_{\text{FR1}}}}(R) = \frac{1}{\mathcal{A}_{\text{FR1}}} \exp[-\pi\lambda_{\text{FR1}} R^2] 2\pi\lambda_{\text{FR1}} \times \exp\left[-2\pi\lambda_{\text{FR2}}\left(\hat{Z}\left(\left(k_1^{-1} R^{\alpha_{\text{FR1}}}\right)^{\frac{1}{\alpha_{\mathfrak{N}}}}\right) + Z\left(\left(k_1^{-1} R^{\alpha_{\text{FR1}}}\right)^{\frac{1}{\alpha_{\mathfrak{L}}}}\right)\right)\right] R, \tag{10}$$

where $k_1 = \frac{B_{\text{FR1}}}{B_{\text{FR2}}}$. Similarly, the PDF of the serving distances $R^*_{x^{\mathfrak{L}}_{\text{FR2}_i} 0 x^{\mathfrak{L}}_{\text{FR2}_i}}$, $R^*_{x^{\mathfrak{N}}_{\text{FR2}_i} 0 x^{\mathfrak{N}}_{\text{FR2}_i}}$, and $R^*_{x_{\text{FR1}} 0 x_{\text{FR1}}}$ under CBUA are given as

$$\bar{f}_{R^*_{x^{\mathfrak{L}}_{\text{FR2}_i} 0 x^{\mathfrak{L}}_{\text{FR2}_i}}}(R) = \frac{1}{\mathcal{A}^{\mathfrak{L}}_{\text{FR2}_i}} \exp\left[-\pi\lambda_{\text{FR1}}(k_1 R^{\alpha_{\mathfrak{L}}})^{\frac{2}{\alpha_{\text{FR1}}}} - 2\pi\hat{\lambda}_{\text{FR2}_i}\left(\hat{Z}(R^{\frac{\alpha_{\mathfrak{L}}}{\alpha_{\mathfrak{N}}}}) + Z(R)\right)\right] 2\pi\hat{\lambda}_{\text{FR2}_i} p_{\mathfrak{L}}(R) R, \tag{11}$$

$$\bar{f}_{R^*_{x^{\mathfrak{N}}_{\text{FR2}_i} 0 x^{\mathfrak{N}}_{\text{FR2}_i}}}(R) = \frac{1}{\mathcal{A}^{\mathfrak{N}}_{\text{FR2}_i}} \exp\left[-\pi\lambda_{\text{FR1}}(k_1 R^{\alpha_{\mathfrak{N}}})^{\frac{2}{\alpha_{\text{FR1}}}} - 2\pi\hat{\lambda}_{\text{FR2}_i}\left(Z(R^{\frac{\alpha_{\mathfrak{N}}}{\alpha_{\mathfrak{L}}}}) + \hat{Z}(R)\right)\right] 2\pi\hat{\lambda}_{\text{FR2}_i} p_{\mathfrak{N}}(R) R, \tag{12}$$

$$\bar{f}_{R^*_{x_{\text{FR1}} 0 x_{\text{FR1}}}}(R) = \frac{1}{\mathcal{A}_{\text{FR1}_i}} \exp[-\pi\lambda_{\text{FR1}} R^2] 2\pi\lambda_{\text{FR1}} R \times \exp\left[-2\pi\hat{\lambda}_{\text{FR2}_i}\left(\hat{Z}\left(\left(k_1^{-1} R^{\alpha_{\text{FR1}}}\right)^{\frac{1}{\alpha_{\mathfrak{N}}}}\right) + Z\left(\left(k_1^{-1} R^{\alpha_{\text{FR1}}}\right)^{\frac{1}{\alpha_{\mathfrak{L}}}}\right)\right)\right]. \tag{13}$$

*4) Mean cell load:* We assume that the maximum number of users that are simultaneously served by a FR2 BS and a FR1 BS in each resource block are limited to $M_{\text{FR2}}$ and $M_{\text{FR1}}$, respectively, such that $M_{\text{FR2}} \leq N_{\text{RF}}$ and $M_{\text{FR1}} \leq n_t^{\text{FR1}}$, where $N_{\text{RF}}$ is the number of RF chains.. Now let $\mathcal{U}_x$ denote the set of all users in $\Phi_u$, which are scheduled by the BS at $x$ in one resource block. The cardinality of $\mathcal{U}_x$ is expressed as $U_x = \min(M, N_x)$, where $M \in \{M_{\text{FR2}}, M_{\text{FR1}}\}$ and $N_x$ is the number of associated users of the BS at $x$. Due to LOS and NLOS transmissions, the coverage area of each BS no longer forms weighted Voronoi cells because a user can associate with a far-away BS with LOS path instead of a near BS with NLOS path. Thus, it becomes quite complicated to compute the exact cell distribution. For analytical tractability, we consider the average number of users served by each BS, which follows the same assumption as in [21] and is given by assuming the same mean cell area as that of the Poisson-Voronoi cell area. To summarize, the average numbers of served users by the tagged and non-tagged FR2 and FR1 BSs under LBUA are given as $U_{\text{FR2}} = \min(M_{\text{FR2}}, N_{\text{FR2}})$, $U_{\text{FR1}} = \min(M_{\text{FR1}}, N_{\text{FR1}})$, $\bar{U}_{\text{FR2}} = \min(M_{\text{FR2}}, \bar{N}_{\text{FR2}})$, and $\bar{U}_{\text{FR1}} = \min(M_{\text{FR1}}, \bar{N}_{\text{FR1}})$, respectively, where $N_{\text{FR2}} = 1 + 1.28\frac{\lambda_u \mathcal{A}_{\text{FR2}}}{\lambda_{\text{FR2}}}$, $N_{\text{FR1}} = 1 + 1.28\frac{\lambda_u \mathcal{A}_{\text{FR1}}}{\lambda_{\text{FR1}}}$, $\bar{N}_{\text{FR2}} = \frac{\lambda_u \mathcal{A}_{\text{FR2}}}{\lambda_{\text{FR2}}}$, and $\bar{N}_{\text{FR1}} = \frac{\lambda_u \mathcal{A}_{\text{FR1}}}{\lambda_{\text{FR1}}}$ [21], [28]. Similarly, by a slight abuse of notations, the number of users (requesting the $i$th file) served by the tagged and non-tagged FR2 and FR1 BSs under CBUA are $U_{\text{FR2}_i} = \min(M_{\text{FR2}}, N_{\text{FR2}_i})$, $U_{\text{FR1}_i} = \min(M_{\text{FR1}}, N_{\text{FR1}_i})$, $\bar{U}_{\text{FR2}_i} = \min(M_{\text{FR2}}, \bar{N}_{\text{FR2}_i})$, and $\bar{U}_{\text{FR1}_i} = \min(M_{\text{FR1}}, \bar{N}_{\text{FR1}_i})$, respectively, where $N_{\text{FR2}_i} = 1 + 1.28\frac{\lambda_u \mathcal{A}_{\text{FR2}_i}}{\lambda_{\text{FR2}}\omega_{\text{FR2}_i}}$, $N_{\text{FR1}_i} = 1 + 1.28\frac{\lambda_u \mathcal{A}_{\text{FR1}_i}}{\lambda_{\text{FR1}}}$, $\bar{N}_{\text{FR2}_i} = \frac{\lambda_u \mathcal{A}_{\text{FR2}_i}}{\lambda_{\text{FR2}}\omega_{\text{FR2}_i}}$, and $\bar{N}_{\text{FR1}_i} = \frac{\lambda_u \mathcal{A}_{\text{FR1}_i}}{\lambda_{\text{FR1}}}$.

*5) SINR characterization: (5.1) FR2 network*: The propagation from FR2 SBSs to each user is through fully connected hybrid precoders that are composed of RF and baseband (BB) precoders. For simplicity, we assume that each BS serves multiple users with one data stream per user. Subsequently, each user applies analog beamforming with a single RF chain, which is sufficient to receive the signal.

**FR2 received signal:** Let the BB and RF precoder matrices of the tagged FR2 BS at $x_{\text{FR2}}^j$ with $j \in \{\mathfrak{L}, \mathfrak{N}\}$ based on LOS and NLOS paths be denoted as $\mathbf{V}_{x_{\text{FR2}}^j}^{\text{BB}} = [\mathbf{v}_{x_{\text{FR2}}^j 0 x_{\text{FR2}}^j}^{\text{BB}}, \mathbf{v}_{x_{\text{FR2}}^j 1 x_{\text{FR2}}^j}^{\text{BB}}, \ldots, \mathbf{v}_{x_{\text{FR2}}^j (U_{\text{FR2}}-1) x_{\text{FR2}}^j}^{\text{BB}}]$ and $\mathbf{V}_{x_{\text{FR2}}^j}^{\text{RF}} = [\mathbf{v}_{x_{\text{FR2}}^j 0 x_{\text{FR2}}^j}^{\text{RF}}, \mathbf{v}_{x_{\text{FR2}}^j 1 x_{\text{FR2}}^j}^{\text{RF}}, \ldots, \mathbf{v}_{x_{\text{FR2}}^j (U_{\text{FR2}}-1) x_{\text{FR2}}^j}^{\text{RF}}]$, respectively. The signal received at the typical user from the FR2 BS located at $x_{\text{FR2}}^j$ is given in (14) on the top of next page, where $\bar{\mathbf{h}}_{x_{\text{FR2}}^j 0 x_{\text{FR2}}^j} = (\mathbf{w}_{0 x_{\text{FR2}}^j}^{\text{RF}})^H \mathbf{H}_{x_{\text{FR2}}^j 0 x_{\text{FR2}}^j} \mathbf{V}_{x_{\text{FR2}}^j}^{\text{RF}}$ is the effective channel coefficient and $\mathbf{w}_{0 x_{\text{FR2}}^j}^{\text{RF}}$ is the RF combiner of the typical user. Further, $n_0^{\text{FR2}} \sim \mathcal{CN}(0, \sigma_{\text{FR2}}^2)$ denotes the noise at the receiver, $s_{(\cdot)}$ is the transmitted symbol with unit power, $\frac{\text{P}_{\text{FR2}}}{U_{\text{FR2}}}$ is the average received signal power where the total power enforces $||\mathbf{V}_{x_{\text{FR2}}^j}^{\text{RF}} \mathbf{v}_{x_{\text{FR2}}^j 0 x_{\text{FR2}}^j}^{\text{BB}}||_F^2 = 1$, with $\mathbf{v}_{x_{\text{FR2}}^j 0 x_{\text{FR2}}^j}^{\text{BB}}$ denoting the BB precoder designed for the typical user to cancel the IUI.

**Remark 3.** *This work does not focus on the design of optimal hybrid precoders. Hence, we follow the sub-optimal approach proposed in [29] to obtain the hybrid precoders, whereby* $\mathbf{w}_{0 x_{FR2}^j}^{\text{RF}} = \mathbf{a}_u(\theta_{i_{\max} x_{FR2}^j 0 x_{FR2}^j})$ *and* $\mathbf{v}_{x_{FR2}^j 0 x_{FR2}^j}^{\text{RF}} = \mathbf{a}_{FR2}(\phi_{i_{\max} x_{FR2}^j 0 x_{FR2}^j})$. *$\theta_{i_{\max} x_{FR2}^j 0 x_{FR2}^j}$ and $\phi_{i_{\max} x_{FR2}^j 0 x_{FR2}^j}$ are chosen such that the maximum channel gain is achieved on the $i_{\max}$ path,* i.e., $i_{\max} = \arg\max_i h_{i x_{FR2}^j 0 x_{FR2}^j}^{FR2}$. *Hereinafter, we ignore the subscript $i_{\max}$ for notational simplicity. In order to eliminate IUI, zero forcing (ZF) is applied at the BB of transmitter such that* $\mathbf{v}_{x_{FR2}^j 0 x_{FR2}^j}^{\text{BB}} = (\bar{\mathbf{h}}_{x_{FR2}^j 0 x_{FR2}^j})^H (\bar{\mathbf{h}}_{x_{FR2}^j 0 x_{FR2}^j} (\bar{\mathbf{h}}_{x_{FR2}^j 0 x_{FR2}^j})^H)^{-1}$.

**SINR model:** Based on (14), the SINR of the typical user from the FR2 BS at $x_{\text{FR2}}^j$ with $j \in \{\mathfrak{L}, \mathfrak{N}\}$ is formulated as $\text{SINR}_{x_{\text{FR2}}^j}^{\text{FR2}} =$

$$\frac{\frac{\text{P}_{\text{FR2}}}{U_{\text{FR2}}} |\bar{\mathbf{h}}_{x_{\text{FR2}}^j 0 x_{\text{FR2}}^j} \mathbf{v}_{x_{\text{FR2}}^j 0 x_{\text{FR2}}^j}^{\text{BB}}|^2}{\sum_{u_{n x_{\text{FR2}}^j} \in \mathcal{U}_{x_{\text{FR2}}^j},\, u_{n x_{\text{FR2}}^j} \neq u_{0 x_{\text{FR2}}^j}} \frac{\text{P}_{\text{FR2}}}{U_{\text{FR2}}} |\bar{\mathbf{h}}_{x_{\text{FR2}}^j n x_{\text{FR2}}^j} \mathbf{v}_{x_{\text{FR2}}^j n x_{\text{FR2}}^j}^{\text{BB}}|^2 + \sigma_{\text{FR2}}^2 + \sum_{b \in \Phi_{\text{FR2}} \setminus \{x_{\text{FR2}}^j\}} \sum_{u_{nb} \in \mathcal{U}_b} \frac{\text{P}_{\text{FR2}}}{U_{\text{FR2}}} |\bar{\mathbf{h}}_{bnb} \mathbf{v}_{bnb}^{\text{BB}}|^2}}, \tag{15}$$

where the first term in the denominator denoting IUI tends to zero after ZF precoding. For tractability, we approximate the SINR in (15) according to [30] with the assumptions: 1) $n_t^{\text{FR2}}$ and $n_r^{\text{FR2}}$ are sufficiently large and, 2) $n_t^{\text{FR2}} \gg U_{\text{FR2}}$ to obtain

$$\text{SINR}_{x_{\text{FR2}}^j}^{\text{FR2}} \approx \frac{\frac{\text{P}_{\text{FR2}}}{U_{\text{FR2}}} \frac{n_t^{\text{FR2}} n_r^{\text{FR2}}}{\eta_j} |h_{x_{\text{FR2}}^j 0 x_{\text{FR2}}^j}^{\text{FR2}}|^2 r_{x_{\text{FR2}}^j 0 x_{\text{FR2}}^j}^{-\alpha_j} p_{\text{ZF}}}{I_{x_{\text{FR2}}^j} + \sigma_{\text{FR2}}^2}, \tag{16}$$

where $p_{\text{ZF}}$ is the ZF precoding penalty[7] defined as

$$p_{\text{ZF}} = \begin{cases} 1 & \text{w.p.} \quad (1 - \frac{1}{n_t^{\text{FR2}}})^{U_{\text{FR2}}-1} \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

Next, after cancelling the IUI, the second term in the denominator of (15) representing the ICI, denoted as $I_{x_{\text{FR2}}^j}$ can be given as

$$I_{x_{\text{FR2}}^j} = \sum\nolimits_{\hat{\jmath} \in \{\mathfrak{L}, \mathfrak{N}\}} \sum_{\substack{b \in \Phi_{\text{FR2}}^{\hat{\jmath}}, \\ b \neq x_{\text{FR2}}^j}} \frac{\text{P}_{\text{FR2}}}{\bar{U}_{\text{FR2}}} \frac{n_r^{\text{FR2}} n_t^{\text{FR2}}}{\eta_{\hat{\jmath}}} r_{b 0 x_{\text{FR2}}^j}^{-\alpha_{\hat{\jmath}}} \sum\nolimits_{u_{nb} \in \mathcal{U}_b} \tag{18}$$

$$\Big| \sum\nolimits_{\hat{k}=1}^{\eta_{\hat{\jmath}}} h_{\hat{k} b 0 x_{\text{FR2}}^j}^{\text{FR2}} \underbrace{\mathbf{a}_u^H(\theta_{x_{\text{FR2}}^j 0 x_{\text{FR2}}^j}) \mathbf{a}_u(\theta_{\hat{k} b 0 x_{\text{FR2}}^j}) \mathbf{a}_{\text{FR2}}^H(\phi_{\hat{k} b 0 x_{\text{FR2}}^j}) \mathbf{a}_{\text{FR2}}(\phi_{bnb})}_{\gamma_{\hat{k},n,b}} \Big|^2.$$

Based on the modified ON/OFF model approximation[8], we have

$$\gamma_{\hat{k},n,b} = \begin{cases} 1, & \text{if } \theta_{x_{\text{FR2}}^j 0 x_{\text{FR2}}^j} = \theta_{\hat{k} b 0 x_{\text{FR2}}^j}, \phi_{\hat{k} b 0 x_{\text{FR2}}^j} = \phi_{bnb} \\ \rho_{\text{BS}}, & \text{if } \theta_{x_{\text{FR2}}^j 0 x_{\text{FR2}}^j} \neq \theta_{\hat{k} b 0 x_{\text{FR2}}^j}, \phi_{\hat{k} b 0 x_{\text{FR2}}^j} = \phi_{bnb} \\ \rho_{\text{UE}}, & \text{if } \theta_{x_{\text{FR2}}^j 0 x_{\text{FR2}}^j} = \theta_{\hat{k} b 0 x_{\text{FR2}}^j}, \phi_{\hat{k} b 0 x_{\text{FR2}}^j} \neq \phi_{bnb} \\ \rho_{\text{BS}} \rho_{\text{UE}}, & \text{otherwise,} \end{cases} \tag{19}$$

where $\rho_{\text{BS}} < 1$ and $\rho_{\text{UE}} < 1$. Hereinafter, (19) will be used to reduce (18) into a simple expression in Section IV.

*(5.2) FR1 network*:

[7]For more details on the ZF penalty, please refer [30].

[8]For tagged BS, the IUI is cancelled via ZF precoding. Thus, we use the ON/OFF approximation model: $\mathbf{a}_*^H(\theta_1)\mathbf{a}_*(\theta_2) = 1$ if $\theta_1 = \theta_2$; otherwise, it is zero. However, this assumption has its limitation when IUI is incorporated. When we analyze the interference from other BSs to the tagged BS, it is not accurate to directly consider $\mathbf{a}_*^H(\theta_1)\mathbf{a}_*(\theta_2) = 0$ for $\theta_1 \neq \theta_2$, as it underestimates the interference. Therefore, we consider the inner product of array response vectors as a non-zero value instead of zero as shown in [25].

$$y_0 = \underbrace{\sqrt{\frac{\mathrm{P}_{\mathrm{FR2}}}{U_{\mathrm{FR2}}}}\bar{\mathbf{h}}_{x^j_{\mathrm{FR2}}0x^j_{\mathrm{FR2}}}\mathbf{v}^{\mathrm{BB}}_{x^j_{\mathrm{FR2}}0x^j_{\mathrm{FR2}}}s_{x^j_{\mathrm{FR2}}0x^j_{\mathrm{FR2}}}}_{\text{Desired signal}} + \underbrace{\sum_{\substack{u_{nx^j_{\mathrm{FR2}}}\in\mathcal{U}_{x^j_{\mathrm{FR2}}},\\ u_{nx^j_{\mathrm{FR2}}}\neq u_{0x^j_{\mathrm{FR2}}}}}\sqrt{\frac{\mathrm{P}_{\mathrm{FR2}}}{U_{\mathrm{FR2}}}}\bar{\mathbf{h}}_{x^j_{\mathrm{FR2}}nx^j_{\mathrm{FR2}}}\mathbf{v}^{BB}_{x^j_{\mathrm{FR2}}nx^j_{\mathrm{FR2}}}s_{x^j_{\mathrm{FR2}}nx^j_{\mathrm{FR2}}}}_{\text{Intra-user interference (IUI)}}$$

$$+ \underbrace{\sum_{b\in\Phi_{\mathrm{FR2}}\setminus\{x^j_{\mathrm{FR2}}\}}\sum_{u_{nb}\in\mathcal{U}_b}\sqrt{\frac{\mathrm{P}_{\mathrm{FR2}}}{\bar{U}_{\mathrm{FR2}}}}\bar{\mathbf{h}}_{b0x^j_{\mathrm{FR2}}}\mathbf{v}^{\mathrm{BB}}_{bnb}s_{bnb}}_{\text{Inter-cell interference (ICI)}} + \underbrace{n^{\mathrm{FR2}}_0}_{\text{Noise}}, \tag{14}$$

**Uplink training and channel estimation:** As previously mentioned the FR1 networks uses massive MIMO in the MBSs for transmission. One of the main predicaments in the performance of massive MIMO systems is pilot contamination due to the reuse of pilot sequences in adjacent cells. We assume that each MBS assigns orthogonal pilots of length $\tau$ symbols for the $U \in \{U_{\mathrm{FR1}}, \bar{U}_{\mathrm{FR1}}\}$ users in its cell such that $\tau \geq U$. For analytical tractability, we further assume that each FR1 BS has at least one user to serve[9] and the mean number of scheduled users of each MBS is the same as the tagged MBS regardless of the difference between the tagged and non-tagged BS. Hereinafter, we use $U_{\mathrm{FR1}}$ to denote the users served by each FR1 BS. Subsequently, we assume that the $n$th user in cell $i$ has pilot sequence identical to the $n$th user in the cell $l$. Let the pilot sequence used by the user $u_{nl}$ be denoted by a $\tau \times 1$ vector $\boldsymbol{\epsilon}_{nl}$, which satisfies the following: 1) $\boldsymbol{\epsilon}^H_{nl}\boldsymbol{\epsilon}_{cl} = \delta(n-c)$ with $\delta(\cdot)$ being the Kronecker delta function, and 2) $\boldsymbol{\epsilon}_{ni} = \boldsymbol{\epsilon}_{nl}$ for $\forall i \neq l$. By transmitting these pilot sequences over $\tau$ symbols in the uplink, the collective received pilot signal at the FR1 BS $i$ can be expressed as

$$\mathbf{Y}_i = \sqrt{\tau \mathrm{P}_p}\sum_{l\in\Phi_{\mathrm{FR1}}}\sum_{n=1}^{U_{\mathrm{FR1}}}\mathbf{g}_{inl}\boldsymbol{\epsilon}_{nl} + \mathbf{N}^{\mathrm{FR1}}_i, \tag{20}$$

where $\mathrm{P}_p$ is the pilot power and $\mathbf{g}_{inl} = [h^{\mathrm{FR1}}_{1inl}\sqrt{r^{-\alpha_{\mathrm{FR1}}}_{inl}}, \ldots, h^{\mathrm{FR1}}_{n^{\mathrm{FR1}}_t inl}\sqrt{r^{-\alpha_{\mathrm{FR1}}}_{inl}}]^T$ is the channel coefficient vector from the user $u_{nk}$ to the FR1 BS $i$. The $n^{\mathrm{FR1}}_t \times \tau$ additive white Gaussian noise (AWGN) matrix $\mathbf{N}^{\mathrm{FR1}}_i$ has i.i.d zero-mean elements and variance $\sigma^2_{\mathrm{FR1}}$. Now, each BS estimates a user channel by multiplying the received pilot signal and the corresponding pilot sequence used by that user. We consider minimum mean squared error (MMSE) estimation, and thus the estimated channel vector $\hat{\mathbf{g}}_{ini}$ can be given as

$$\hat{\mathbf{g}}_{ini} = \eta^{\mathrm{FR1}}_{ini}\frac{1}{\sqrt{\tau \mathrm{P}_p}}\mathbf{Y}_i\boldsymbol{\epsilon}^H_{ni}, \tag{21}$$

where $\eta^{\mathrm{FR1}}_{ini} = r^{-\alpha_{\mathrm{FR1}}}_{ini}/\sum_{l\in\Phi_{\mathrm{FR1}}} r^{-\alpha_{\mathrm{FR1}}}_{inl} + \frac{\sigma^2_{\mathrm{FR1}}}{\tau \mathrm{P}_p}$. The channel estimation error is denoted as $\tilde{\mathbf{g}}_{ini} = \mathbf{g}_{ini} - \hat{\mathbf{g}}_{ini}$, whose elements follow $\mathcal{CN}(0, r^{-\alpha_{\mathrm{FR1}}}_{ini}(1-\eta^{\mathrm{FR1}}_{ini}))$. Further, the elements of the estimated channel $\hat{\mathbf{g}}_{ini}$ follow $\mathcal{CN}(0, r^{-\alpha_{\mathrm{FR1}}}_{ini}\eta^{\mathrm{FR1}}_{ini})$.

**FR1 received signal:** With the estimated channel information obtained from the uplink training phase, each FR1 MBS constructs the downlink precoding vector with it. Accordingly, the received signal at the typical user from the tagged FR1 MBS at $x_{\mathrm{FR1}}$ can be given in (22) on the top of next page, where $\mathbf{v}^{\mathrm{FR1}}_{lnl}$ is the $n^{\mathrm{FR1}}_t \times 1$ precoding vector of the FR1 BS $l$ towards the user $u_{nl}$, $s_{lnl}$ is the transmitted signal from the BS $l$ to the user $u_{nl}$, and $n^{\mathrm{FR1}}_0$ is the AWGN with zero mean and variance $\sigma^2_{\mathrm{FR1}}$. With regards to precoding, we consider maximum-ratio-transmission (MRT) precoding due to fact that the channel responses associated with different users tend to be nearly orthogonal in massive MIMO systems [31]. Accordingly, the precoding vector $\mathbf{v}^{\mathrm{FR1}}_{lnl}$ is given as

$$\mathbf{v}^{\mathrm{FR1}}_{lnl} = \mathcal{K}^{\mathrm{FR1}}_l\frac{\hat{\mathbf{g}}_{lnl}}{\sqrt{\mathbb{E}\{||\hat{\mathbf{g}}_{lnl}||^2\}}}, \tag{23}$$

where the expectation in the denominator is taken over the fast fading channel coefficients and $\mathcal{K}^{\mathrm{FR1}}_l$ is the power normalization factor that maintains the power constraint

$$\mathbb{E}\{\mathrm{tr}(\mathbf{V}^{\mathrm{FR1}}_l(\mathbf{V}^{\mathrm{FR1}}_l)^H\} = 1, \tag{24}$$

where $\mathbf{V}^{\mathrm{FR1}}_l = [\ldots, \mathbf{v}^{\mathrm{FR1}}_{lnl}, \ldots]$ is the precoding matrix. Therefore, we have $\mathcal{K}^{\mathrm{FR1}}_l = \sqrt{\frac{1}{U_{\mathrm{FR1}}}}$.

**SINR model:** By incorporating the channel estimation error into (22), the received signal can be written as (25) on the top of next page, where the user $u_{0l}$ in $l$th cell has the same pilot as the typical user. Based on (25), the downlink SINR is given as

$$\mathrm{SINR}^{\mathrm{FR1}}_{x_{\mathrm{FR1}}} = \frac{\mathrm{P}_{\mathrm{FR1}}|(\hat{\mathbf{g}}_{x_{\mathrm{FR1}}0x_{\mathrm{FR1}}})^H\mathbf{v}^{\mathrm{FR1}}_{x_{\mathrm{FR1}}0x_{\mathrm{FR1}}}|^2}{I_{x_{\mathrm{FR1}}} + \sigma^2_{\mathrm{FR1}}}, \tag{26}$$

where $I_{x_{\mathrm{FR1}}}$ is given in (27) on the top of next page.

*6) Rate characterization:* The effective achievable information rate (bits per second) for FR2 and FR1 transmissions can accordingly be given as

$$R_{x^j_{\mathrm{FR2}}} = W_{\mathrm{FR2}}\log(1+\mathrm{SINR}^{\mathrm{FR2}}_{x^j_{\mathrm{FR2}}}), \tag{28}$$

$$R_{x_{\mathrm{FR1}}} = W_{\mathrm{FR1}}\log(1+\mathrm{SINR}^{\mathrm{FR1}}_{x_{\mathrm{FR1}}}), \tag{29}$$

where $W_{\mathrm{FR1}}$ and $W_{\mathrm{FR2}}$ are the bandwidths used for transmission for FR1 and FR2 signals, respectively. $\mathrm{SINR}^{\mathrm{FR2}}_{x^j_{\mathrm{FR2}}}$ and $\mathrm{SINR}^{\mathrm{FR1}}_{x_{\mathrm{FR1}}}$ are given in (16) and (26), respectively. Similar to the calculation of the above general expressions for LBUA, we can calculate the SINR and rate expressions for CBUA as well.

## IV. Performance analysis

### A. Average latency

WEC is considered as a promising solution to reduce the latency of file delivery. Accordingly, we adopt average latency of file delivery as one of the performance metrics to evaluate the considered hybrid HetNet. When the requested file is

[9]This assumption is valid due to the fact that when the user density is sufficiently large *i.e.,* $\lambda_u \gg \lambda_{\mathrm{FR1}}$, every MBS has at least $U_{\mathrm{FR1}} = \bar{U}_{\mathrm{FR1}} = M_{\mathrm{FR1}}$ users in its coverage area.

$$
\begin{aligned}
y_0 =& \sqrt{\mathrm{P}_{\mathrm{FR1}}} \sum_{l\in\Phi_{\mathrm{FR1}}} \sum_{u_{nl}\in\mathcal{U}_l} (\mathbf{g}_{l0x_{\mathrm{FR1}}})^H \mathbf{v}_{lnl}^{\mathrm{FR1}} s_{lnl} + n_0^{\mathrm{FR1}} \\
=& \underbrace{\sqrt{\mathrm{P}_{\mathrm{FR1}}} (\mathbf{g}_{x_{\mathrm{FR1}}0x_{\mathrm{FR1}}})^H \mathbf{v}_{x_{\mathrm{FR1}}0x_{\mathrm{FR1}}}^{\mathrm{FR1}} s_{x_{\mathrm{FR1}}0x_{\mathrm{FR1}}}}_{\text{Desired signal}} + \underbrace{\sqrt{\mathrm{P}_{\mathrm{FR1}}} \sum_{u_{nx_{\mathrm{FR1}}}\in\mathcal{U}_{x_{\mathrm{FR1}}}\setminus\{u_{0x_{\mathrm{FR1}}}\}} (\mathbf{g}_{x_{\mathrm{FR1}}0x_{\mathrm{FR1}}})^H \mathbf{v}_{x_{\mathrm{FR1}}nx_{\mathrm{FR1}}}^{\mathrm{FR1}} s_{x_{\mathrm{FR1}}nx_{\mathrm{FR1}}}}_{\text{IUI}} \\
&+ \underbrace{\sqrt{\mathbf{P}_{\mathrm{FR1}}} \sum_{l\in\Phi_{\mathrm{FR1}}\setminus\{x_{\mathrm{FR1}}\}} \sum_{u_{nl}\in\mathcal{U}_l} (\mathbf{g}_{l0x_{\mathrm{FR1}}})^H \mathbf{v}_{lnl}^{\mathrm{FR1}} s_{lnl}}_{\text{ICI}} + \underbrace{n_0^{\mathrm{FR1}}}_{\text{Noise}}, \qquad (22)
\end{aligned}
$$

$$
\begin{aligned}
y_0 =& \sqrt{\mathrm{P}_{\mathrm{FR1}}} (\hat{\mathbf{g}}_{x_{\mathrm{FR1}}0x_{\mathrm{FR1}}})^H \mathbf{v}_{x_{\mathrm{FR1}}0x_{\mathrm{FR1}}}^{\mathrm{FR1}} s_{x_{\mathrm{FR1}}0x_{\mathrm{FR1}}} + \sqrt{\mathrm{P}_{\mathrm{FR1}}} \sum_{u_{nx_{\mathrm{FR1}}}\in\mathcal{U}_{x_{\mathrm{FR1}}}\setminus\{u_{0x_{\mathrm{FR1}}}\}} (\hat{\mathbf{g}}_{x_{\mathrm{FR1}}0x_{\mathrm{FR1}}})^H \mathbf{v}_{x_{\mathrm{FR1}}nx_{\mathrm{FR1}}}^{\mathrm{FR1}} s_{x_{\mathrm{FR1}}nx_{\mathrm{FR1}}} \\
&+ \sqrt{\mathrm{P}_{\mathrm{FR1}}} \sum_{u_{nx_{\mathrm{FR1}}}\in\mathcal{U}_{x_{\mathrm{FR1}}}} (\tilde{\mathbf{g}}_{x_{\mathrm{FR1}}0x_{\mathrm{FR1}}})^H \mathbf{v}_{x_{\mathrm{FR1}}nx_{\mathrm{FR1}}}^{\mathrm{FR1}} s_{x_{\mathrm{FR1}}nx_{\mathrm{FR1}}} \\
&+ \sqrt{\mathrm{P}_{\mathrm{FR1}}} \sum_{l\in\Phi_{\mathrm{FR1}}\setminus\{x_{\mathrm{FR1}}\}} \Big( (\mathbf{g}_{l0x_{\mathrm{FR1}}})^H \mathbf{v}_{l0l}^{\mathrm{FR1}} s_{l0l} + \sum_{u_{nl}\in\mathcal{U}_l\setminus\{u_{0l}\}} (\mathbf{g}_{l0x_{\mathrm{FR1}}})^H \mathbf{v}_{lnl}^{\mathrm{FR1}} s_{lnl} \Big) + n_0, \qquad (25)
\end{aligned}
$$

$$
\begin{aligned}
I_{x_{\mathrm{FR1}}} =& \mathrm{P}_{\mathrm{FR1}} \sum_{u_{nx_{\mathrm{FR1}}}\in\mathcal{U}_{x_{\mathrm{FR1}}}\setminus\{u_{0x_{\mathrm{FR1}}}\}} |(\hat{\mathbf{g}}_{x_{\mathrm{FR1}}0x_{\mathrm{FR1}}})^H \mathbf{v}_{x_{\mathrm{FR1}}nx_{\mathrm{FR1}}}^{\mathrm{FR1}}|^2 + \mathrm{P}_{\mathrm{FR1}} \sum_{u_{nx_{\mathrm{FR1}}}\in\mathcal{U}_{x_{\mathrm{FR1}}}} |(\tilde{\mathbf{g}}_{x_{\mathrm{FR1}}0x_{\mathrm{FR1}}})^H \mathbf{v}_{x_{\mathrm{FR1}}nx_{\mathrm{FR1}}}^{\mathrm{FR1}}|^2 \\
&+ \mathrm{P}_{\mathrm{FR1}} \sum_{l\in\Phi_{\mathrm{FR1}}\setminus\{x_{\mathrm{FR1}}\}} (|(\mathbf{g}_{l0x_{\mathrm{FR1}}})^H \mathbf{v}_{l0l}^{\mathrm{FR1}}|^2 + \sum_{u_{nl}\in\mathcal{U}_l\setminus\{u_{0l}\}} |(\mathbf{g}_{l0x_{\mathrm{FR1}}})^H \mathbf{v}_{lnl}^{\mathrm{FR1}}|^2) + \sigma_{\mathrm{FR1}}^2. \qquad (27)
\end{aligned}
$$

cached in the SBSs, the latency is mainly due to access links. Otherwise, two-hop links involving both access and backhaul contribute towards the average latency. The considered latency model is explained below.

- **Wired backhaul latency**: Since the backhaul network is composed of links connecting BSs with central schedulers or gateways, multi-hop links are required depending on the type of backhaul technology being used. In conjunction to [32], the mean latency due to backhaul processing is given as

$$\bar{T}_{bj} \approx \Big( (1 + 1.28\frac{\lambda_j}{\lambda_g}) c_1 + (\bar{n}_b - 1) c_2 \Big)(a + S c_3), \quad (30)$$

where $j \in \{\mathrm{FR2}, \mathrm{FR1}\}$ denotes the FR2 and FR1 entities, $c_1, a$ and $c_3$ are constants reflecting the processing capability of nodes, $n_b$ is the number of hops with one gateway and $n_b - 1$ hubs, $S$ is the file size, $c_2$ is a constant reflecting the delay of each hub, $\bar{n}_b \approx \frac{1}{r_b\sqrt{2\lambda_g}}$ is the average number of hops in the backhaul [32] and $r_b$ is the transmission range of one hop in the backhaul link. As for the transmission delay, it is heavily associated with backhaul capacity of each user. In particular, it is worth mentioning that only cache-miss users will be allocated backhaul bandwidth under LBUA. However, under limited backhaul scenario, since the access rate is always restricted by the backhaul capacity, the analytical trend will not change (assuming that all served users in a cell are provided with equal backhaul capacity). Accordingly, the backhaul capacity of the typical user is given as $\frac{C_b}{U}$. For the tagged FR1 MBS, $U = U_{\mathrm{FR1}}$ and for the tagged FR2 SBS, $U = U_{\mathrm{FR2}}$.
- **Wireless access latency**: The latency on wireless access links primarily arises from the time required for transmission, which is dependent on the average access rate. Accordingly, we formulate the average delay for the typical user under LBUA as [33]

$$
\begin{aligned}
T =& \sum_{i=1}^{F} q_i \Big\{ \mathcal{A}_{\mathrm{FR2}}^{\mathfrak{L}} \big( \omega_{\mathrm{FR2}_i} T_{\mathfrak{L}}^{(1)} + (1 - \omega_{\mathrm{FR2}_i}) T_{\mathfrak{L}}^{(2)} \big) \qquad (31) \\
&+ \mathcal{A}_{\mathrm{FR2}}^{\mathfrak{N}} \big( \omega_{\mathrm{FR2}_i} T_{\mathfrak{N}}^{(1)} + (1 - \omega_{\mathrm{FR2}_i}) T_{\mathfrak{N}}^{(2)} \big) + \mathcal{A}_{\mathrm{FR1}} T_{\mathrm{FR1}} \Big\},
\end{aligned}
$$

where $T_{\mathfrak{L}}^{(1)} = \frac{S}{\bar{R}_{x_{\mathrm{FR2}}^{\mathfrak{L}}}}$, $T_{\mathfrak{L}}^{(2)} = \frac{S}{\bar{R}_{x_{\mathrm{FR2}}^{\mathfrak{L}}}} + \frac{SU_{\mathrm{FR2}}}{C_b} + \bar{T}_{bm}$, $T_{\mathfrak{N}}^{(1)} = \frac{S}{\bar{R}_{x_{\mathrm{FR2}}^{\mathfrak{N}}}}$, $T_{\mathfrak{N}}^{(2)} = \frac{S}{\bar{R}_{x_{\mathrm{FR2}}^{\mathfrak{N}}}} + \frac{SU_{\mathrm{FR2}}}{C_b} + \bar{T}_{bm}$, and $T_{\mathrm{FR1}} = \frac{S}{\bar{R}_{x_{\mathrm{FR1}}}} + \frac{SU_{\mathrm{FR1}}}{C_b} + \bar{T}_{b\mathrm{FR1}}$. Similarly, the average delay for the typical user under CBUA is given as

$$T = \sum_{i=1}^{F} q_i \Big\{ \mathcal{A}_{\mathrm{FR2}_i}^{\mathfrak{L}} T_i^{\mathfrak{L}} + \mathcal{A}_{\mathrm{FR2}_i}^{\mathfrak{N}} T_i^{\mathfrak{N}} + \mathcal{A}_{\mathrm{FR1}_i} T_{\mathrm{FR1}_i} \Big\}, \quad (32)$$

where $T_i^{\mathfrak{L}} = \frac{S}{\bar{R}_{x_{\mathrm{FR2}_i}^{\mathfrak{L}}}}$, $T_i^{\mathfrak{N}} = \frac{S}{\bar{R}_{x_{\mathrm{FR2}_i}^{\mathfrak{N}}}}$, $T_{\mathrm{FR1}_i} = \frac{S}{\bar{R}_{x_{\mathrm{FR1}}}} + \frac{SU_{\mathrm{FR1}_i}}{C_b} + \bar{T}_{b\mathrm{FR1}}$.

In the following, we first propose a lower bound of the average success probability (ASP) of file delivery for FR2 transmission, using which we give and show how to derive the average rate $\bar{R}_{x_{\mathrm{FR2}}^j}$ for LBUA and $\bar{R}_{x_{\mathrm{FR2}_i}^j}$ for CBUA. After that, we derive the average data rate $\bar{R}_{x_{\mathrm{FR1}}}$ for both user association strategies.

*1) Average rate for transmission involving FR2 SBSs:*

**Proposition 1.** *The ASP of file delivery for FR2 SBSs located at $x_{FR2}^{\mathfrak{L}}$ and $x_{FR2}^{\mathfrak{N}}$ under LBUA are lower bounded as*

$$
\begin{aligned}
\mathcal{P}_s^j(\nu) \geq& (1 - \frac{1}{n_t^{FR2}})^{(U_{FR2}-1)} \int_0^\infty \exp\Big( \frac{-Q_{FR2}\sigma_{FR2}^2}{G_{x_{FR2}^j} R^{-\alpha_j}} \Big) \hat{f}_{R^*_{x_{FR2}^j 0 x_{FR2}^j}}(R) \\
&\times \exp\Big[ - \sum_{\hat{j}\in\{\mathfrak{L},\mathfrak{N}\}} \int_{\bar{R}_{j\hat{j}}}^\infty \Big[ 1 - \Big( 1 - s_j \mathrm{P}_{FR2} n_t^{FR2} n_r^{FR2} r^{-\alpha_{\hat{j}}} \Big)^{-\eta_{\hat{j}}} \Big] \\
&\times 2\pi\lambda_{FR2} p_{\hat{j}}(r) r \mathrm{d}r \Big] \mathrm{d}R, \qquad (33)
\end{aligned}
$$

*where $\nu$ is the target rate, $G_{x^j_{FR2}} = \frac{\mathrm{P}_{FR2}}{U_{FR2}} \frac{n_r^{FR2} n_t^{FR2}}{\eta_j}$, $s_j = \frac{-Q_{FR2}}{G_{x^j_m} R^{-\alpha_j}}$, $Q_m = 2^{\frac{\nu}{W_m}} - 1$ with $j \in \{\mathfrak{L}, \mathfrak{N}\}$, $\tilde{R}_{j\hat{j}} = R^{\frac{\alpha_j}{\alpha_{\hat{j}}}}$, and $\hat{f}_{R^*_{x^j_m 0 x^j_{FR2}}}(\cdot)$ is given in the* (8) *and* (9).

*Proof.* The proof is given in Appendix A. □

Using the ASP of file delivery, the average data rate can now be given as

$$\begin{aligned}\bar{R}_{x^j_{\mathrm{FR2}}} &= W_{\mathrm{FR2}} \mathbb{E}\Big\{\log\Big(1+\mathrm{SINR}^{\mathrm{FR2}}_{x^j_{\mathrm{FR2}}}\Big)\Big\} \\ &= W_{\mathrm{FR2}} \int_0^\infty \mathbb{P}\Big[\log\Big(1+\mathrm{SINR}^{\mathrm{FR2}}_{x^j_{\mathrm{FR2}}}\Big) \geq \nu\Big] \mathrm{d}\nu, \end{aligned} \quad (34)$$

where $j \in \{\mathfrak{L}, \mathfrak{N}\}$.

**Proposition 2.** *The ASP of file delivery of the $i$th file by the FR2 SBSs located at $x^{\mathfrak{L}}_{FR2_i}$ and $x^{\mathfrak{N}}_{FR2_i}$ under CBUA are lower bounded as*

$$\begin{aligned}\mathcal{P}^j_{s_i}(\nu) \geq & \int_0^\infty (1-\frac{1}{n_t^{FR2}})^{(U_{FR2_i}-1)} \exp\Big(\frac{-Q_{FR2}\sigma^2_{FR2}}{G_{x^j_{FR2_i}} R^{-\alpha_j}}\Big) \bar{f}_{R^*_{x^j_{FR2_i} 0 x^j_{FR2_i}}}(R) \\ & \times \exp\Big[-\int_R^\infty \Big[1-\Big(1-s_j \mathrm{P}_{FR2} n_t^{FR2} n_r^{FR2} r^{-\alpha_j}\Big)^{-\eta_j}\Big] A_i p_j(r) r \mathrm{d}r\Big] \\ & \times \exp\Big[-\int_R^\infty {}^{\frac{\alpha_j}{\alpha_{j'}}}\Big[1-\Big(1-s_j \mathrm{P}_{FR2} n_t^{FR2} n_r^{FR2} r^{-\alpha_{j'}}\Big)^{-\eta_{j'}}\Big] A_i p_{j'}(r) r \mathrm{d}r\Big] \\ & \times \prod_{\hat{j}\in\{\mathfrak{L},\mathfrak{N}\}} \exp\Big[-\int_0^\infty \Big[1-\Big(1-s_j \mathrm{P}_{FR2} n_t^{FR2} n_r^{FR2} r^{-\alpha_{\hat{j}}}\Big)^{-\eta_{\hat{j}}}\Big] \\ & \times B_i p_{\hat{j}}(r) r \mathrm{d}r\Big] \mathrm{d}R, \end{aligned} \quad (35)$$

*where $A_i = 2\pi\lambda_{FR2}\omega_{FR2_i}$, $B_i = 2\pi\lambda_{FR2}(1-\omega_{FR2_i})$, $G_{x^j_{FR2_i}} = \frac{\mathrm{P}_{FR2}}{U_{FR2_i}} \frac{n_r^{FR2} n_t^{FR2}}{\eta_j}$ and $s_j = \frac{-Q_{FR2}}{G_{x^j_{FR2_i}} R^{-\alpha_j}}$. $Q_{FR2} = 2^{\frac{\nu}{W_{FR2}}} - 1$ with $j \in \{\mathfrak{L}, \mathfrak{N}\}$ and $j'$ is the complementary counterpart of $j$.*

*Proof.* The proof is given in Appendix B. □

Now, similar to (34), we can obtain the approximated rate $\bar{R}_{x^j_{\mathrm{FR2}_i}}$ for CBUA.

*2) Average rate for transmission involving FR1 MBSs:* Before deriving the average data rate for FR1 transmission, we provide the following assumptions and lemma.

**Assumption 1.** *We approximate the coverage region of the MBS located at $x$ as a ball centered at $x$ with radius $C_v = \frac{1}{\sqrt{\pi\lambda_{FR1}}}$ [28].*

**Assumption 2.** *The point process $\Theta_n$ formed by taking the locations of the user $n$ in each macro cell as the reference point is the the perturbation of the original PPP $\Phi_{FR1}$, which is no longer a PPP. However, for tractability, by considering that all the MBSs that the user $n$ is not associated to are interferers, it is approximated as an inhomogeneous PPP with density $\lambda_{\Theta_n}(r) = \lambda_{FR1}(1-p_{aFR1})$ [34]. For $n_1 \neq n_2$, $\Theta_{n_1}$ and $\Theta_{n_2}$ are independent with the same density. Further, in the FR1 network the probability that the user is associated with the tagged MBS is given as $p_{aFR1} = \mathbb{P}[R^{-\alpha_{FR1}} > r^{-\alpha_{FR1}}] = \exp[-\pi\lambda_{FR1}R^2]$.*

**Lemma 1.** *For arbitrary non-negative random variables $\{x_i | i = 1, 2, \ldots, N\}$ and $\{y_j | j = 1, 2, \ldots, M\}$, the following equation can be obtained [35]*

$$\mathbb{E}\Big\{\ln\Big(1+\frac{\sum_{i=1}^N x_i}{\sum_{j=1}^M y_j + 1}\Big)\Big\} = \int_0^\infty \frac{\mathcal{M}_y(z) - \mathcal{M}_{x,y}(z)}{z} e^{-z} \mathrm{d}z, \quad (36)$$

*where $\mathcal{M}_y(z) = \mathbb{E}\{e^{-z\sum_{j=1}^M y_j}\}$ and $\mathcal{M}_{x,y}(z) = \mathbb{E}\{e^{-z(\sum_{i=1}^N x_i + \sum_{j=1}^M y_j)}\}$.*

Using the above assumptions and Lemma, the average downlink rate of the typical user served by the FR1 massive MIMO-aided MBS located at $x_{\mathrm{FR1}}$ under LBUA is given as [36]

$$\begin{aligned}\bar{R}_{x_{\mathrm{FR1}}} = & \int_0^\infty \int_0^\infty \frac{W_{\mathrm{FR1}}}{\ln 2} \frac{e^{-z}}{z} \mathcal{B}(R) \big(\exp(-z\mathcal{C}_1(R)) - \exp(-z\mathcal{C}_2(R))\big) \\ & \times \hat{f}_{R^*_{x_{\mathrm{FR1}} 0 x_{\mathrm{FR1}}}}(R) \mathrm{d}z \mathrm{d}R, \end{aligned} \quad (37)$$

where

$$\mathcal{B}(R) = \exp\Big[-\int_R^\infty \Big(1-\exp\Big(-z\frac{1}{\rho_1} r^{-\alpha_{\mathrm{FR1}}}\Big)\Big) 2\pi\lambda_{\mathrm{FR1}} r \mathrm{d}r\Big], \quad (38)$$

$$\mathcal{C}_1(R) = -\frac{\mathrm{FR1}}{R^{-2\alpha_{\mathrm{FR1}}}} U_{\mathrm{FR1}} \rho_1 \Big(R^{-\alpha_{\mathrm{FR1}}} + \mathrm{FR1}_{14} + \frac{\sigma^2_{\mathrm{FR1}}}{\tau \mathrm{P}_p}\Big) + \frac{R^{-\alpha_{\mathrm{FR1}}}}{\rho_1}, \quad (39)$$

$$\mathcal{C}_2(R) = \frac{n_t^{\mathrm{FR1}} R^{-2\alpha_{\mathrm{FR1}}}}{U_{\mathrm{FR1}} \rho_1 \Big(R^{-\alpha_{\mathrm{FR1}}} + \mathrm{FR1}_{14} + \frac{\sigma^2_{\mathrm{FR1}}}{\tau \mathrm{P}_p}\Big)} + \frac{1}{\rho_1} R^{-\alpha_{\mathrm{FR1}}}. \quad (40)$$

and $\mathrm{FR1}_{14} = \int_{C_v}^\infty r^{-\alpha_{\mathrm{FR1}}} 2\pi\lambda_{\mathrm{FR1}}(1 - p_{a\mathrm{FR1}}) r \mathrm{d}r$, $\rho_1 = \mathrm{FR1}_1 + \frac{\sigma^2_{\mathrm{FR1}}}{\tau \mathrm{P}_{\mathrm{FR1}}}$, $\mathrm{FR1}_1 = \frac{n_t^{\mathrm{FR1}} \mathrm{FR1}_{13}}{U_{\mathrm{FR1}}\left(\mathrm{FR1}_{12} + \mathrm{FR1}_{11} + \frac{\sigma^2_{\mathrm{FR1}}}{\tau \mathrm{P}_p}\right)}$, $\mathrm{FR1}_{11} = \int_{C_v}^\infty r^{-\alpha_{\mathrm{FR1}}} 2\pi r \ \lambda_{\Theta_{\tilde{0}}}(r) \mathrm{d}r = \int_{C_v}^\infty r^{-\alpha_{\mathrm{FR1}}} 2\pi\lambda_{\mathrm{FR1}}(1 - p_{a\mathrm{FR1}}) r \mathrm{d}r$, $\mathrm{FR1}_{12} = \int_0^\infty r^{-\alpha_{\mathrm{FR1}}} \hat{f}_{R^*_{x_{\mathrm{FR1}} 0 x_{\mathrm{FR1}}}}(r) \mathrm{d}r$, $\mathrm{FR1}_{13} = \int_{C_v}^\infty (r^{-\alpha_{\mathrm{FR1}}})^2 2\pi\lambda_{\mathrm{FR1}} r \mathrm{d}r$. $\hat{f}_{R^*_{x_{\mathrm{FR1}} 0 x_{\mathrm{FR1}}}}(\cdot)$ is given in (10). Similarly, the average data rate for FR1 transmission under CBUA can be computed by substituting the expressions related to CBUA (*i.e.,* the PDF of serving distance) in the above.

### *B. ASP of file delivery of FR2 SBSs*

Since the MBS tier is independent of caching, we consider the ASP of file delivery in the access links of SBSs only. In particular, under LBUA, we also consider the effect of backhaul on the ASP of file delivery of SBSs. The backhaul ASP of file delivery is considered in the event of cache miss. By assuming a minimum rate $\nu$, the maximum number of users served by the backhaul links are fixed as $N_b = \frac{C_b}{\nu}$. If the number of cache-miss users is greater than $N_b$, the backhaul fails to support all cache-miss users and will randomly pick $N_b$ users with equal probability. Therefore, considering that the typical user is a cache-miss user, the probability of the typical user to be served is given as [17]

$$\mathcal{P}_b^m(\nu)=\sum_{n=0}^{U_{\text{FR2}}-1}\binom{U_{\text{FR2}}-1}{n}p_{\text{hit}}^{U_{\text{FR2}}-n-1}(1-p_{\text{hit}})^n\min\left\{1,\frac{N_b}{n+1}\right\}, \tag{41}$$

where $p_{\text{hit}}=\sum_{i=1}^{F}q_i\omega_{\text{FR2}_i}$ is the average probability that the requested files are cached in the local caches of SBSs. Thus, the ASP of file delivery of SBSs under LBUA with LOS and NLOS transmissions are given as

$$\mathcal{P}_{x_{\text{FR2}}^j}(\nu)=\sum_{i=1}^{F}q_i\Big\{\omega_{\text{FR2}_i}\mathbb{P}[R_{x_{\text{FR2}}^j}\geq\nu] + (1-\omega_{\text{FR2}_i})\mathbb{P}[R_{x_{\text{FR2}}^j}\geq\nu]\mathcal{P}_b^{\text{FR2}}(\nu)\Big\}, \tag{42}$$

where $j\in\{\mathfrak{L},\mathfrak{N}\}$. When $\mathcal{P}_b^{\text{FR2}}(\nu)=1$, the above ASP of file delivery will be affected by the access link of SBSs only.

Similarly, based on Proposition 2, the average ASP of file delivery in access for CBUA is given as

$$\mathcal{P}_{x_{\text{FR2}}^j}(\nu)=\sum_{i=1}^{F}q_i\mathbb{P}[R_{x_{\text{FR2}_i}^j}\geq\nu]. \tag{43}$$

### C. Average data rate for the typical user

The average data rate for the typical user is defined as the total average throughput achieved by the typical user. Now assuming equal allocation of backhaul capacity to all users, the average data rate for the typical user under LBUA is given as

$$\begin{aligned}\text{T}_l=&\sum_{i=1}^{F}q_i\Big[\mathcal{A}_{\text{FR2}}^{\mathfrak{L}}\Big(\omega_{\text{FR2}_i}\bar{R}_{x_{\text{FR2}}^{\mathfrak{L}}}+(1-\omega_{\text{FR2}_i})\min\Big\{\bar{R}_{x_{\text{FR2}}^{\mathfrak{L}}},\frac{C_b}{U_{\text{FR2}}}\Big\}\Big)\\&+\mathcal{A}_m^{\mathfrak{N}}\Big(\omega_{\text{FR2}_i}\bar{R}_{x_{\text{FR2}_i}^{\mathfrak{N}}}+(1-\omega_{\text{FR2}_i})\min\Big\{\bar{R}_{x_{\text{FR2}}^{\mathfrak{N}}},\frac{C_b}{U_{\text{FR2}}}\Big\}\Big)\\&+A_\mu\min\Big\{\bar{R}_{x_{\text{FR1}}},\frac{C_b}{U_{\text{FR1}}}\Big\}\Big].\end{aligned} \tag{44}$$

Similarly, for CBUA

$$\text{T}_c=\sum_{i=1}^{F}q_i\Big[\mathcal{A}_{\text{FR2}_i}^{\mathfrak{L}}\bar{R}_{x_{\text{FR2}_i}^{\mathfrak{L}}}+\mathcal{A}_{\text{FR2}_i}^{\mathfrak{N}}\bar{R}_{x_{\text{FR2}_i}^{\mathfrak{N}}}+A_{\text{FR1}_i}\min\Big\{\bar{R}_{x_{\text{FR1}}},\frac{C_b}{U_{\text{FR1}}}\Big\}\Big]. \tag{45}$$

## V. 2–TERM APPROXIMATION

The aforementioned section deals with deriving integral expressions to evaluate the performance metrics of the hybrid HetNet. In this section, however, by considering the instance of latency, we provide an approximation for (31) that is analytically and computationally more tractable, but albeit certain trade-offs in performance. In particular, we resort to 2–term approximation by applying Taylor series expansion to association probabilities and rates, i.e., $\{\mathcal{A}_{\text{FR2}}^{\mathfrak{L}},\ \mathcal{A}_{\text{FR2}}^{\mathfrak{N}},\ \mathcal{A}_{\text{FR1}}\}$ and $\{\bar{R}_{x_{\text{FR2}}^{\mathfrak{L}}},\ \bar{R}_{x_{\text{FR2}}^{\mathfrak{N}}},\ \bar{R}_{x_{\text{FR1}}}$[10]$\}$, respectively, and then derive the latency under LBUA (the CBUA case can be obtained in a similar way). We consider the first two terms[11] with respect to the variable (i.e., SBS density) at the initial point $\lambda_{\text{FR2}}^0$ over the exponential function in the integral. For simplicity, we consider the full load case, whereby the cell load is independent of SBS density.

We begin by applying the approximation to $\mathcal{A}_{\text{FR2}}^{\mathfrak{L}}$ and denoting the exponential term in (2) by $F_{\mathfrak{L}}[\lambda_{\text{FR2}}]$. Now, applying Taylor's series expansion at $\lambda_{\text{FR2}}^0$ we obtain $F_{\mathfrak{L}}[\lambda_{\text{FR2}}]=$

$$\begin{aligned}&\exp\left[-\pi\lambda_{\text{FR1}}\big(k_1R^{\alpha_{\mathfrak{L}}}\big)^{\frac{2}{\alpha_{\text{FR1}}}}-2\pi\lambda_{\text{FR2}}\left(\hat{Z}\big(R^{\frac{\alpha_{\mathfrak{L}}}{\alpha_{\mathfrak{N}}}}\big)+Z\big(R\big)\right)\right]\\&\approx F_{\mathfrak{L}}[\lambda_{\text{FR2}}^0]+F_{\mathfrak{L}}^{'}[\lambda_{\text{FR2}}^0](\lambda_{\text{FR2}}-\lambda_{\text{FR2}}^0)+\frac{F_{\mathfrak{L}}^{''}[\lambda_{\text{FR2}}^0]}{2}(\lambda_{\text{FR2}}-\lambda_{\text{FR2}}^0)^2,\end{aligned} \tag{46}$$

where the first and second derivative of $F_{\mathfrak{L}}[\lambda_{\text{FR2}}]$ at $\lambda_{\text{FR2}}^0$ are $F_{\mathfrak{L}}^{'}[\lambda_{\text{FR2}}^0]=$

$$\begin{aligned}&\exp\left[-\pi\lambda_{\text{FR1}}\big(k_1R^{\alpha_{\mathfrak{L}}}\big)^{\frac{2}{\alpha_{\text{FR1}}}}-2\pi\lambda_{\text{FR2}}^0\left(\hat{Z}\big(R^{\frac{\alpha_{\mathfrak{L}}}{\alpha_{\mathfrak{N}}}}\big)+Z\big(R\big)\right)\right]\\&\times\left[-2\pi\left(\hat{Z}\big(R^{\frac{\alpha_{\mathfrak{L}}}{\alpha_{\mathfrak{N}}}}\big)+Z\big(R\big)\right)\right],\end{aligned} \tag{47}$$

$$\text{and } F_{\mathfrak{L}}^{''}[\lambda_{\text{FR2}}^0]=F_{\mathfrak{L}}^{'}[\lambda_{\text{FR2}}^0]\left[-2\pi\left(\hat{Z}\big(R^{\frac{\alpha_{\mathfrak{L}}}{\alpha_{\mathfrak{N}}}}\big)+Z\big(R\big)\right)\right]. \tag{48}$$

After rearranging them, we have the relative association probabilities under LBUA as

$$\mathcal{A}_{\text{FR2}}^{\mathfrak{L}}=\lambda_{\text{FR2}}C_1^{\mathfrak{L}}+\lambda_{\text{FR2}}^2C_2^{\mathfrak{L}}+\lambda_{\text{FR2}}^3C_3^{\mathfrak{L}}, \tag{49}$$

where

$$C_1^{\mathfrak{L}}=\int_0^\infty\Big(F_{\mathfrak{L}}[\lambda_{\text{FR2}}^0]-F_{\mathfrak{L}}^{'}[\lambda_{\text{FR2}}^0]\lambda_{\text{FR2}}^0+\frac{F_{\mathfrak{L}}^{''}[\lambda_{\text{FR2}}^0]}{2}(\lambda_{\text{FR2}}^0)^2\Big)2\pi p_{\mathfrak{L}}(R)R\text{d}R, \tag{50}$$

$$C_2^{\mathfrak{L}}=\int_0^\infty\Big(F_{\mathfrak{L}}^{'}[\lambda_{\text{FR2}}^0]-F_{\mathfrak{L}}^{''}[\lambda_{\text{FR2}}^0]\lambda_{\text{FR2}}^0\Big)2\pi p_{\mathfrak{L}}(R)R\text{d}R, \tag{51}$$

$$C_3^{\mathfrak{L}}=\int_0^\infty\frac{F_{\mathfrak{L}}^{''}[\lambda_{\text{FR2}}^0]}{2}2\pi p_{\mathfrak{L}}(R)R\text{d}R. \tag{52}$$

Similarly, we derive the approximated $\mathcal{A}_{\text{FR2}}^{\mathfrak{N}}$, $\mathcal{A}_{\text{FR1}}$ with parameters $C_1^{\mathfrak{N}}, C_2^{\mathfrak{N}},\ C_3^{\mathfrak{N}}, C_{\text{FR1}_1}, C_{\text{FR1}_2}, C_{\text{FR1}_3}$ (omitted here due to space constraints). For simplicity, we make $\lambda_{\text{FR2}}$ that appears in the rate $\bar{R}_{x_{\text{FR1}}}$ to be a constant (by choosing an initial point). This is due to the fact that $\lambda_{\text{FR2}}$ plays a prominent role for the case of small cells only. In the following, we take Taylor serious expansion at $\lambda_{\text{FR1}}^0$ for the ASP of file delivery (given in Proposition 1) then derive the approximated $\bar{R}_{x_{\text{FR2}}^{\mathfrak{L}}}$, and $\bar{R}_{x_{\text{FR2}}^{\mathfrak{N}}}$. Let the exponential term in (33) be denoted by $B_j[\lambda_{\text{FR1}}]$ with $j\in\{\mathfrak{L},\mathfrak{N}\}$. Then applying Taylor's series expansion at $\lambda_{\text{FR1}}^0$, we obtain $B_j[\lambda_{\text{FR1}}]=$

$$\begin{aligned}&\exp\left[-\sum_{\hat{j}\in\{\mathfrak{L},\mathfrak{N}\}}\int_{\bar{R}_{j\hat{j}}}^\infty\Big[1-\Big(1-s_j\text{P}_{\text{FR1}}n_t^{\text{FR1}}n_r^{\text{FR1}}r^{-\alpha_{\hat{j}}}\Big)^{-\eta_{\hat{j}}}\right]\\&2\pi\lambda_{\text{FR1}}p_{\hat{j}}(r)r\text{d}r\Big]\\&\approx B_j[\lambda_{\text{FR1}}^0]+B_j^{'}[\lambda_{\text{FR1}}^0](\lambda_{\text{FR1}}-\lambda_{\text{FR1}}^0)+\frac{B_j^{''}[\lambda_{\text{FR1}}^0]}{2}(\lambda_{\text{FR1}}-\lambda_{\text{FR1}}^0)^2,\end{aligned} \tag{53}$$

where the first and second derivatives of $B_j[\lambda_{\text{FR1}}]$ at $\lambda_{\text{FR1}}^0$ are $B_j^{'}[\lambda_{\text{FR1}}^0]$ and $B_j^{''}[\lambda_{\text{FR1}}^0]$, respectively. Now, by substituting the above equations into the probability equation, we have

[10]For tractability, we consider $\lambda_{\text{FR2}}$ that appears in the rate $\bar{R}_{x_{\text{FR1}}}$ as constant (in order to choose the initial point). This is due to the fact that $\lambda_{\text{FR2}}$ plays a prominent role for the case of small cells only.

[11]Note that in a $x$-term approximation, with $x\in\{1,2,\dots\}$, the choice of $x$ is subject to design criteria and error tolerence.

$$\mathcal{P}_s^j(\nu,\lambda_{\text{FR1}}) \approx D_1^j + \lambda_{\text{FR1}} D_2^j + \lambda_{\text{FR1}}^2 D_3^j. \quad (54)$$

In the above,

$$D_1^j = \int_0^\infty \Big(1-\frac{1}{n_t^{\text{FR1}}}\Big)^{U_{\text{FR1}}-1} \exp\left[-\frac{Q_{\text{FR1}}\sigma_{\text{FR1}}^2}{G_{x_{\text{FR1}}^j} R^{-\alpha_j}}\right] \hat{f}_{R^*_{x_{\text{FR1}}^j 0 x_{\text{FR1}}^j}}(R)$$
$$\times \Big(B_j[\lambda_{\text{FR1}}^0] - B_j^{'}[\lambda_{\text{FR1}}^0]\lambda_{\text{FR1}}^0 + \frac{B_j^{''}[\lambda_{\text{FR1}}^0]}{2}(\lambda_{\text{FR1}}^0)^2\Big)\mathrm{d}R, \quad (55)$$

$$\times D_2^j = \int_0^\infty \Big(1-\frac{1}{n_t^{\text{FR1}}}\Big)^{U_{\text{FR1}}-1} \exp\left[-\frac{Q_{\text{FR1}}\sigma_{\text{FR1}}^2}{G_{x_{\text{FR1}}^j} R^{-\alpha_j}}\right]$$
$$\Big(B_j^{'}[\lambda_{\text{FR1}}^0] - B_j^{''}[\lambda_{\text{FR1}}^0]\lambda_{\text{FR1}}^0\Big)\hat{f}_{R^*_{x_{\text{FR1}}^j 0 x_{\text{FR1}}^j}}(R)\mathrm{d}R, \quad (56)$$

$$D_3^j = \int_0^\infty \Big(1-\frac{1}{n_t^{\text{FR1}}}\Big)^{U_{\text{FR1}}-1} \exp\left[-\frac{Q_{\text{FR1}}\sigma_{\text{FR1}}^2}{G_{x_{\text{FR1}}^j} R^{-\alpha_j}}\right]$$
$$\times \frac{B_j^{''}[\lambda_{\text{FR1}}^0]}{2}\hat{f}_{R^*_{x_{\text{FR1}}^j 0 x_{\text{FR1}}^j}}(R)\mathrm{d}R, \quad (57)$$

Finally, based on (34), we have the average data rate as

$$\bar{R}_{x_{\text{FR1}}^j} = B_1^j + B_2^j \lambda_{\text{FR1}} + B_3^j \lambda_{\text{FR1}}^2, \quad (58)$$

where

$$B_1^j = W_{\text{FR1}} \int_0^\infty D_1^j \mathrm{d}\nu,\ B_2^j = W_{\text{FR1}} \int_0^\infty D_2^j \mathrm{d}\nu,\ B_3^j = W_{\text{FR1}} \int_0^\infty D_3^j \mathrm{d}\nu. \quad (59)$$

Now, by substituting all parameters into (31), we have the approximate latency as

$$\frac{\lambda_{\text{FR2}} C_1^{\mathfrak{L}} + \lambda_{\text{FR2}}^2 C_2^{\mathfrak{L}} + \lambda_{\text{FR2}}^3 C_3^{\mathfrak{L}}}{B_1^{\mathfrak{L}} + B_2^{\mathfrak{L}}\lambda_{\text{FR2}} + B_3^{\mathfrak{L}}\lambda_{\text{FR2}}^2} N_1$$
$$+ (\lambda_{\text{FR2}}^2 C_1^{\mathfrak{L}} + \lambda_{\text{FR2}}^3 C_2^{\mathfrak{L}} + \lambda_{\text{FR2}}^4 C_3^{\mathfrak{L}}) N_2$$
$$+ \lambda_{\text{FR2}} N_4 + (\lambda_{\text{FR2}} C_1^{\mathfrak{L}} + \lambda_{\text{FR2}}^2 C_2^{\mathfrak{L}} + \lambda_{\text{FR2}}^3 C_3^{\mathfrak{L}}) N_3$$
$$+ \frac{\lambda_{\text{FR2}} C_1^{\mathfrak{N}} + \lambda_{\text{FR2}}^2 C_2^{\mathfrak{N}} + \lambda_{\text{FR2}}^3 C_3^{\mathfrak{N}}}{B_1^{\mathfrak{N}} + B_2^{\mathfrak{N}}\lambda_{\text{FR2}} + B_3^{\mathfrak{N}}\lambda_{\text{FR2}}^2} N_1 + \lambda_{\text{FR2}}^2 N_5$$
$$+ (\lambda_{\text{FR2}}^2 C_1^{\mathfrak{N}} + \lambda_{\text{FR2}}^3 C_2^{\mathfrak{N}} + \lambda_{\text{FR2}}^4 C_3^{\mathfrak{N}}) N_2$$
$$+ (\lambda_{\text{FR2}} C_1^{\mathfrak{N}} + \lambda_{\text{FR2}}^2 C_2^{\mathfrak{N}} + \lambda_{\text{FR2}}^3 C_3^{\mathfrak{N}}) N_3 + N_6, \quad (60)$$

where

$$N_1 = \sum_{i=1}^{F} q_i S, \quad (61)$$

$$N_2 = \sum_{i=1}^{F} (1-\omega_{\text{FR2}_i}) q_i \left[\frac{SU_{\text{FR2}}}{C_{b_1}} + 1.28\frac{c_1(a+Sc_3)}{\lambda_g}\right], \quad (62)$$

$$N_3 = \sum_{i=1}^{F} (1-\omega_{\text{FR2}_i}) q_i \left[\frac{SU_{\text{FR2}}}{C_{b_1}}\lambda_{\text{FR1}} + c_1(a+Sc_3) + (\bar{n}_b - 1)c_2(a+Sc_3)\right], \quad (63)$$

$$N_4 = \sum_{i=1}^{F} q_i C_{\text{FR1}_2} T_{\text{FR1}}, \quad (64)$$

$$N_5 = \sum_{i=1}^{F} q_i C_{\text{FR1}_3} T_{\text{FR1}}, \quad (65)$$

$$N_6 = \sum_{i=1}^{F} q_i C_{\text{FR1}_1} T_{\text{FR1}}. \quad (66)$$

**Algorithm I:** Computation of optimal $\lambda_m^*$

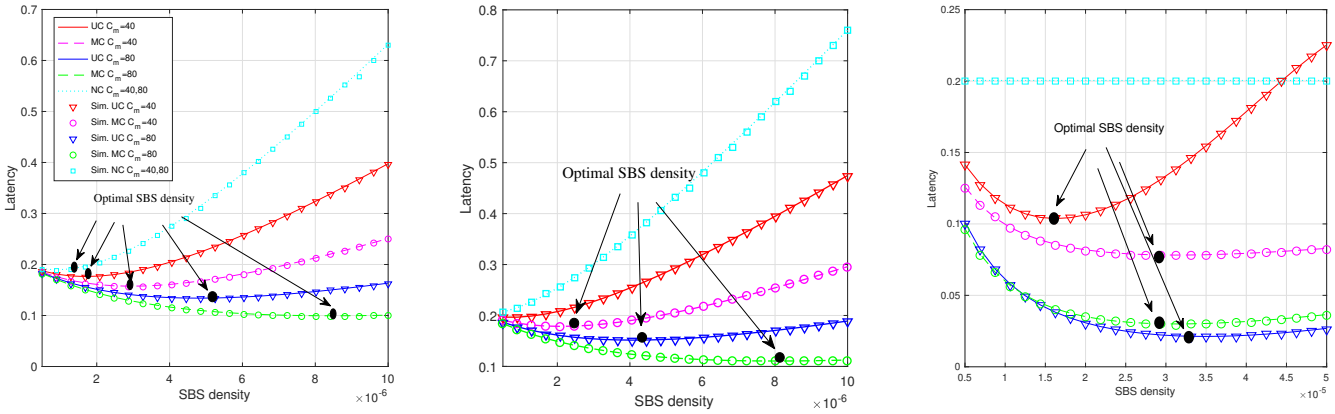
1 : Initialize : $\lambda_{\text{FR2}}, \delta, \Delta$.
2 : Compute : The latency $T(\lambda_{\text{FR2}})$ from (31) and (32).
3 : Update : $\hat{\lambda}_{\text{FR2}} \rightarrow \lambda_{\text{FR2}} + \Delta$.
4 : Repeat steps $2-4$ until convergence $i.e.,|T(\lambda_{\text{FR2}}) - T(\hat{\lambda}_{\text{FR2}})| \leq \delta$.

The 2–term approximations for ASP of file delivery and rate can be obtained similarly.

## VI. Numerical results

In this section, we numerically evaluate the performance of a cache-enabled hybrid HetNet equipped with hybrid beamforming architecture in the FR2 small cell tier and massive MIMO in the FR1 macro cell tier with respect to average latency, average data rate and ASP of file delivery under two commonly used caching strategies (UC and MC) and compare them with baseline traditional BSs without any local storages (termed as no caching (NC) hereinafter). Unless otherwise stated, the following parameters are used for evaluation [32]: $\lambda_{\text{FR1}} = 2\times10^{-6}$, $\lambda_{\text{FR2}} = 10^{-5}$, $\lambda_u = 5\times10^{-4}$, $\text{P}_{\text{FR1}} = 46$ dBm, $\text{P}_{\text{FR2}} = 30$ dBm, $n_t^{\text{FR1}} = 128$, $n_t^{\text{FR2}} = 256$, $n_r^{\text{FR2}} = 16$, $M_{\text{FR1}} = 20$, $M_{\text{FR2}} = N_{\text{RF}} = 10$, $W_{\text{FR1}} = 200$ MHz, $W_{\text{FR2}} = 1$ GHz, $F = 100$, $\beta = 0.008$, $\alpha_{\mathfrak{L}} = 2$, $\alpha_{\mathfrak{N}} = 4$, $\alpha_{\text{FR1}} = 3.5$, $\text{P}_p = 24$ dBm, $S = 10^6$ bits, $r_b = 1000$ m, $\lambda_g = 0.1\lambda_{\text{FR1}}$, $c_1 = 1$, $c_2 = 10$, $a = 10^{-5}$, $c_3 = 10^{-8}$, $c_{b_1} = 100$, $c_{b_2} = 0$, $\delta = 5\times10^{-7}$, $\Delta = 2\times10^{-7}$, $\eta_{\mathfrak{L}} = 3$, and $\eta_{\mathfrak{N}} = 5$.

*1) Latency:* We begin by evaluating the effect of cache size on the optimal SBS density with respect to latency for the two user association policies in Fig. 2. The optimal SBS density is obtained by performing an exhaustive search over all reasonable values of $\lambda_{\text{FR2}}$ that achieves the minimum latency. The algorithm used to obtain $\lambda_{\text{FR2}}^*$ is summarized in Algorithm I, where $\Delta$ and $\delta$ are step size and error threshold, respectively. As cache size increases, the performance of the network improves. This is due to the increase in probability of the requested file to be found in the local cache, thus avoiding utilization of the backhaul. However, there exists an optimal SBS density that minimizes the latency for both user associations. Nevertheless, CBUA always performs better than LBUA. In particular, while the optimal SBS density increases as the cache size increases under LBUA for both UC and MC, under CBUA, it increases for UC but stays constant for MC. This is due to the fact that MC is a deterministic policy and the latency performance with MC under CBUA is related to the interference from SBS, while under LBUA, it is also related to the processing time and backhaul capacity. Further, as for the baseline NC scenario in LBUA, there also exists an optimal SBS density, but is quite less when compared to that of UC and MC. On the contrary, latency does not change for NC in CBUA as seen in Fig. 2c, as no SBS is a serving BS for NC. Now, in order to study the effect of backhaul on delay performance, we consider different backhauls for SBS and MBS and assume that SBSs have lower backhaul processing capacity. Longer processing time at SBSs when compared to MBSs is given by giving lower transmission range ($r_b = 500$ m) for each hop. In Fig. 2b, it is seen that there is no optimal SBS density for UC under small cache size (caching gain is less than processing time loss) in LBUA (unless the cache size is large). Under this condition, it may not be beneficial to use caching helpers. Further, in Fig. 2c, it is also seen that higher




(a) LBUA with $r_b = 1000$m for both SBS and MBS (b) LBUA with $r_b = 500$m for SBS and $r_b = 1000$m for MBS (c) CBUA

Fig. 2: Latency of file delivery v.s. SBS density with different cache sizes

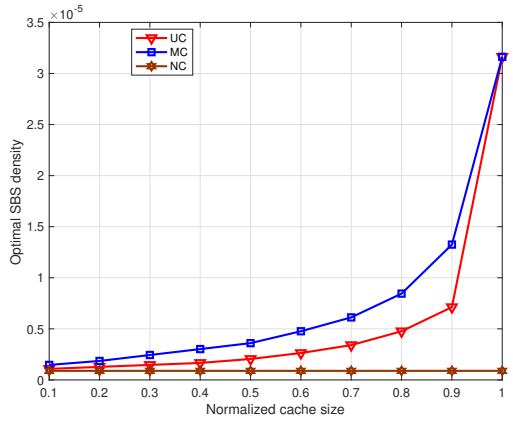
cache size is required in CBUA to obtain higher optimal SBS density compared to LBUA to reduce latency. But even for lower cache size, cache-aided latency performance is superior to NC. In addition, we show a tradeoff, whereby MC is inferior to UC under high cache size with high SBS density. This is due to the fact that UC permits more content diversity gain than MC.

Further, as a complement to Fig. 2 we now illustrate the impact of cache size on the optimal SBS density in the considered hybrid HetNet in Fig. 3. While Fig. 3a a corresponds to LBUA, Fig. 3b corresponds to CBUA. From Fig. 3a it can be seen that as the cache size increases, the optimal SBS density also increases for both UC and MC. Alternatively, from Fig. 3b, it can be seen that in CBUA, the change in cache size doesn't impact the optimal SBS density in MC but for UC, there is an optimal cache size that corresponds to an optimal SBS density. Note that these results complement Fig. 3 in the updated manuscript.

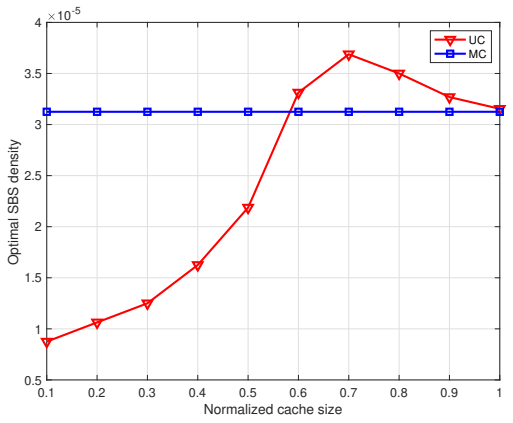
In Fig. 4, we evaluate the effect of content popularity on the optimal SBS density with respect to latency under two different user association strategies. In particular, the latency performance under UC is independent of the skewness parameter since the caching probability is uniformly distributed. However, under MC, the latency performance improves as the skewness increases, which is due to the fact that MC deterministically prefetches the first $C_{\text{FR2}}$ top-ranked files that are highly related to the skewed content popularity. Further, NC is independent of the content popularity and both UC and MC under both user association policies perform better than NC. In particular, the optimal caching placement under LBUA can be obtained by solving a linear optimal caching problem [7], which eventually leads to MC. It can be seen that when $\upsilon = 0$, the curves for MC overlap with UC since UC performs the best for uniformly distributed content popularity. However, under CBUA policy, MC is inferior to UC for lower value of skewness and when $\upsilon$ increases, MC starts to outperform UC. Here again, there exists an optimal SBS density and higher value of $\upsilon$ requires higher SBS density to achieve the optimal latency under MC in LBUA as shown in Fig. 4a. Higher value of $\upsilon$ requires higher optimal SBS density for LBUA but is independent of $\upsilon$ for CBUA as can be seen in Fig. 4b. However, as SBS density further increases, the latency increases due to higher interference, higher processing time, and lower backhaul capacity. For NC, although $i$) there exists an optimal SBS density for LBUA that is related to the backhaul network and $ii$) it is independent of the density of SBSs in CBUA, the latency performance is inferior to cache-aided performance (MC and UC).

*2) Average data rate:* In Fig. 5, we evaluate the effect of backhaul capacity on the average data rate of the typical user. In particular, Fig. 5a shows the gap in average data rate performance between cache-aided and non cache-aided SBSs. Under lower backhaul capacity the gap is significant, whereas it reduces as the backhaul capacity increases. Thus, when the the backhaul capacity is limited, caching can improve the average data rate of the typical user. However, when backhaul capacity is sufficiently large, both UC and MC converge towards a common point. This suggests that caching may not be beneficial in improving the average data rate for networks with very high backhaul capacity. However, this phenomenon is reversed for CBUA as can be seen in Fig. 5b. When the backhaul capacity is limited, the performance under higher cache size is superior to that under lower cache size with MC performing better than UC.

In Fig. 6, we evaluate the average data rate of the typical user with respect to SBS density for different cache sizes and FR1 MBS densities. In particular, for LBUA policy, Fig. 6a demonstrates that caching can improve the average data rate of the typical user since the average data rate increases as the cache size increases. For a particular cache size, lower value of MBS density leads to higher average data rate than that with higher value of MBS density. This is due to the fact that lower value of MBS density leads to higher probability of users associating with the SBS-tier, where caching gain is available. However, when the SBS density increases further, the gap diminishes due to low backhaul capacity and high interference. As such, there exists an optimal SBS density which is independent of the cache size and MBS density. Next, the average data rate for CBUA policy is shown in Fig. 6b, where the trend is similar to LBUA, but the gap in performance for different $\lambda_{\text{FR1}}$ is more significant. This means that higher ratio of SBS and MBS density leads to higher

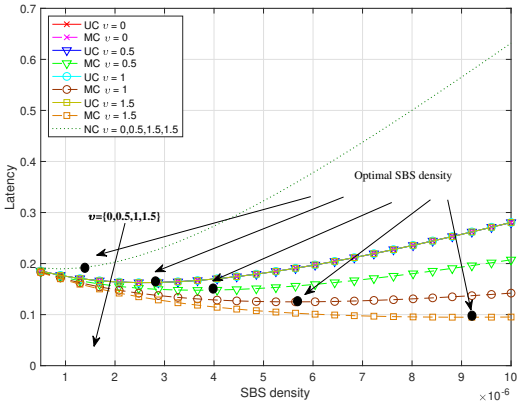



(a) LBUA policy

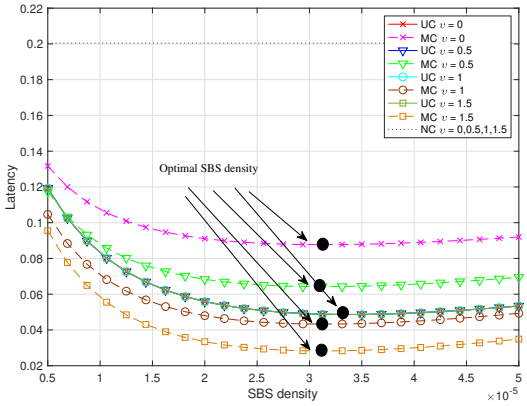



(b) CBUA policy

Fig. 3: Optimal SBS density vs normalized cache size

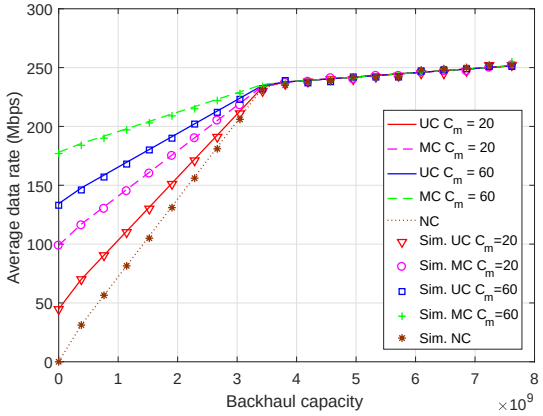



(a) LBUA policy

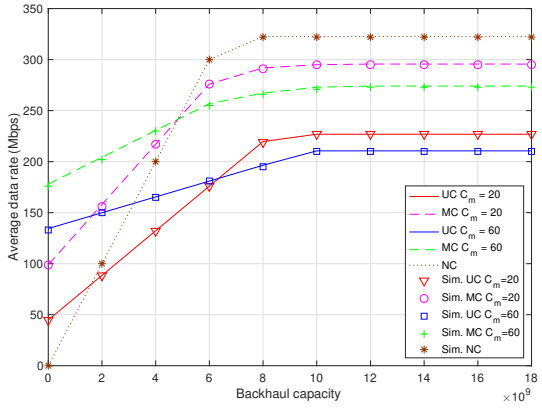



(b) CBUA policy

Fig. 4: Latency of file delivery v.s. SBS density with different skewness




(a) LBUA policy

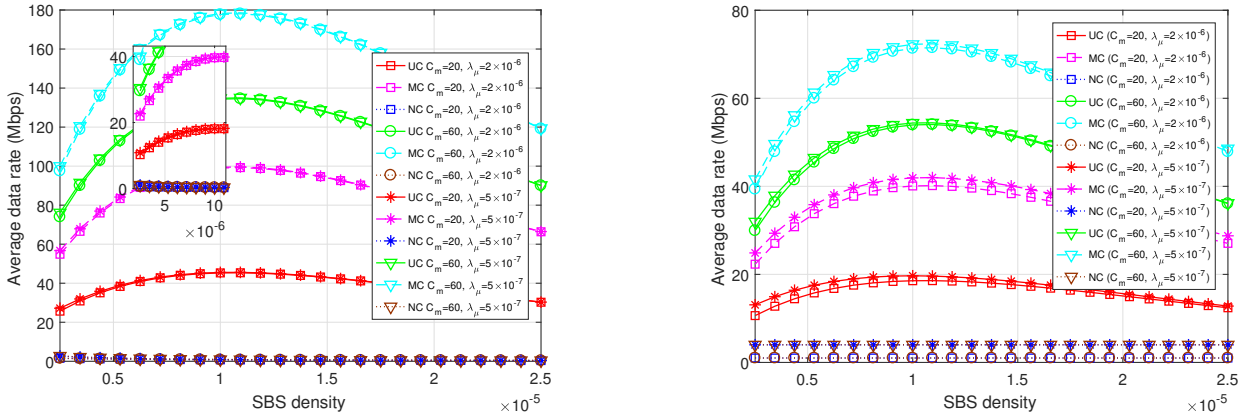



(b) CBUA policy

Fig. 5: Average data rate per user v.s. backhaul capacity with different cache size ($\lambda_{\text{FR2}} = 10^{-5}, \lambda_{\text{FR1}} = 5 \times 10^{-7}$)

performance for CBUA policy. Although the improvement of CBUA over LBUA is not readily visible, CBUA clearly saves more backhaul bandwidth, which is a desirable feature.

*3) ASP of file delivery:* In Fig. 7a, we compare the ASP of file delivery of SBSs in the access link with respect to varying cache size under the two different user association strategies. In particular, as NC scenario under CBUA has no SBSs to serve the users, we consider NC for LBUA as the baseline. The figure on the left is for LOS with the one on the right representing NLOS. In LOS, for fixed SBS density, higher blockage density improves the ASP of file delivery since the interference from reflected paths is mitigated by the blockages. Further, the performance of UC is superior to the performance of MC for given SBS density and blockage density under CBUA. However, the gap decreases as the cache size increases. This is due to the fact that MC is a deterministic caching policy where $F - C_{\text{FR2}}$ files are excluded, while UC provides higher content diversity gain compared to MC. Next, for NLOS, the performance of both UC and MC is strongly dependent on the blockage and SBS densities. When the SBS density is small,




(a) LBUA policy

(b) CBUA policy

Fig. 6: Average data rate of the typical user v.s. SBS density with different cache size and MBS density

MC is superior to UC under lower value of blockage density while UC is superior to MC under higher value of blockage density. Alternatively, when the SBS density is large, MC is inferior to UC for a given small blockage density, while the trend reverses for a higher value of blockage density. This shows that when SBS density is small, it is unlikely to find a cache-hit NLOS SBS, while for large SBS density, it is likely that a cache hit NLOS SBS can be found. Next, content diversity gain is more beneficial for low blockage scenarios, while availing the least path loss association can be useful for scenarios with high blockage densities. In particular, there is a tradeoff between UC and MC in terms of SBS and blockage density. Compared to LBUA which is independent of cache size, CBUA is highly related to cache size and lower cache size results in poor coverage in access, which suggests that it needs the assistance of MBSs. When the cache size is equal to the total number of file size, the performance of both associations converge with each other.

Further, for the sake of completeness, in Fig. 7b, we show the ASP of file delivery under LBUA with respect to the backhaul capacity. Intuitively, when the backhaul capacity increases, the ASP of file delivery improves. It can be seen that MC is superior to UC for both LOS and NLOS and the ASP increases with the increase in cache size. Further, similar to the previous figure, when blockage density increases, the ASP of file delivery also increases.

Finally, in Fig. 8 we validate the 2–term approximation presented in Section V. In particular, we numerically evaluate the 2-term approximated latency and compare it with the exact analytical expression. It can be seen from the figure that the 2-term approximation holds particularly well in the low SBS density region, while some divergence happens in the higher density region. This is due to the fact that interference increases with the increase in SBS density and ignoring the higher order terms underestimates the interference in the network. Increasing the order of approximation will surely lead to better fitting, but at the cost of computing higher order derivatives, that will add more complexity. Alternatively, if reduced complexity is one of the design criteria, it might even be worth considering a 1–term approximation if the number of SBSs is less. Nevertheless, the 2-term approximation can be considered as a "straddling the fence" solution to compute the latency of file delivery for the considered FR1–FR2 hybrid HetNet.

## VII. Conclusion

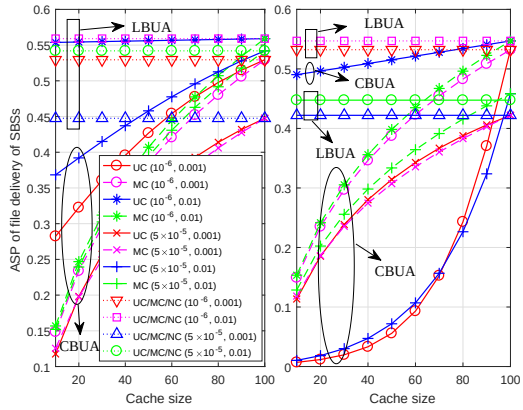
This paper proposed an analytical framework for a cache-enabled hybrid HetNet involving both FR2 SBSs aided by a hybrid precoding architecture and FR1 MBS aided by massive MIMO. We investigated different caching strategies for the proposed hybrid HetNet and analyzed its performance with respect to latency, ASP of file delivery, and average data rate that accounted for blockage effect, pilot contamination, channel fading, and random network topology. Numerical results based on exact and approximate expressions were presented considering various network design parameters. In particular, we compared two different user associations policies and found that CBUA is better than LBUA for reducing the latency of file delivery. Although the average throughput for the two association polices were found to be similar, CBUA has the advantage over LBUA due to the fact that it requires less backhaul. Nevertheless, we showed that there exists an optimal SBS density depending on caching parameters to improve latency and throughput performance. In a nutshell, our results demonstrated the effectiveness of WEC and showed that it can act as a support chain for 5G technologies such as FR2, massive MIMO and HetNets to further improve the network performance.

Furthermore, the current analysis does not focus on optimal caching placement schemes. Nevertheless, while the optimal caching probability in LBUA is actually MC, the design of optimal caching schemes for CBUA is non-trivial and will be considered as an important topic for future research. Similarly, new user association schemes that balances between latency/ caching probability, and data rate/ path loss will be explored in future on top of the current analysis.

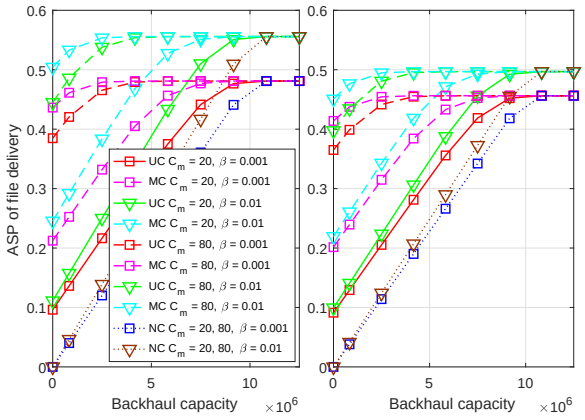
## Appendix A
## Proof of Proposition 1

In this proof, we show the ASP of file delivery in LOS transmission. According to the definition of the ASP of file delivery, we have

$$\mathcal{P}_s^{\mathfrak{L}}(\nu) \approx \mathbb{P}\Big[\frac{G_{x^{\mathfrak{L}}_{\mathrm{FR2}}}|h^{\mathrm{FR2}}_{x^{\mathfrak{L}}_{\mathrm{FR2}}0x^{\mathfrak{L}}_{\mathrm{FR2}}}|^2 r^{-\alpha_{\mathfrak{L}}}_{x^{\mathfrak{L}}_{\mathrm{FR2}}0x^{\mathfrak{L}}_{\mathrm{FR2}}} p_{\mathrm{ZF}}}{I_{x^{\mathfrak{L}}_{\mathrm{FR2}}} + \sigma^2_{\mathrm{FR2}}} \geq Q_{\mathrm{FR2}}\Big]$$




(a) ASP of file delivery of SBSs in access for different $(\lambda_{\text{FR2}}, \beta)$




(b) ASP of file delivery under LBUA policy vs the backhaul capacity

Fig. 7: ASP of file delivery

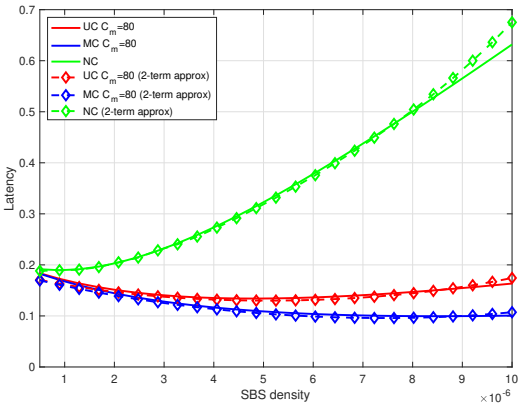



Fig. 8: Comparison between the 2-term approximation and the exact expression for Latency with respect to SBS density.

$$\overset{(a)}{=}(1-\frac{1}{n_t^{\text{FR2}}})^{U_{\text{FR2}}-1}\mathbb{E}\Big\{\mathbb{P}\Big[|h^{\text{FR2}}_{x^{\mathfrak{L}}_{\text{FR2}}0x^{\mathfrak{L}}_{\text{FR2}}}|^2\geq\frac{Q_{\text{FR2}}(I_{x^{\mathfrak{L}}_{\text{FR2}}}+\sigma^2_{\text{FR2}})}{G_{x^{\mathfrak{L}}_{\text{FR2}}}R^{-\alpha_{\mathfrak{L}}}}\Big]\Big\}$$

$$\overset{(b)}{=}(1-\frac{1}{n_t^{\text{FR2}}})^{U_{\text{FR2}}-1}\int_0^{\infty}\underbrace{\mathbb{E}\Big\{\mathbb{P}\Big[|h^{\text{FR2}}_{x^{\mathfrak{L}}_{\text{FR2}}0x^{\mathfrak{L}}_{\text{FR2}}}|^2\geq\frac{Q_{\text{FR2}}(I_{x^{\mathfrak{L}}_{\text{FR2}}}+\sigma^2_{\text{FR2}})}{G_{x^{\mathfrak{L}}_{\text{FR2}}}R^{-\alpha_{\mathfrak{L}}}}\Big]\Big\}}_{A}$$

$$\times\,\hat{f}_{R^*_{x^{\mathfrak{L}}_{\text{FR2}}0x^{\mathfrak{L}}_{\text{FR2}}}}(R)\mathrm{d}R, \tag{A.1}$$

where $Q_{\text{FR2}}=2^{\frac{\nu}{W_{\text{FR2}}}}-1$, $(a)\rightarrow$ we replace $R$ with $r_{x^{\mathfrak{L}}_{\text{FR2}}0x^{\mathfrak{L}}_{\text{FR2}}}$ for notational simplicity throughout the proof, $(b)\rightarrow$ we take average with respect to the distance $R$, $G_{x^{\mathfrak{L}}_{\text{FR2}}}=\frac{\text{P}_{\text{FR2}}n_t^{\text{FR2}}n_r^{\text{FR2}}}{U_{\text{FR2}}\eta_{\mathfrak{L}}}$, and $\hat{f}_{R^*_{x^{\mathfrak{L}}_{\text{FR2}}0x^{\mathfrak{L}}_{\text{FR2}}}}(\cdot)$ is given in (8). Now we derive the expectation term in the integral with regards to the interference term as

$$A\overset{(c)}{=}\mathbb{E}_{I_{x^{\mathfrak{L}}_{\text{FR2}}}}\Big\{\exp\Big[-\frac{Q_{\text{FR2}}(I_{x^{\mathfrak{L}}_{\text{FR2}}}+\sigma^2_{\text{FR2}})}{G_{x^{\mathfrak{L}}_{\text{FR2}}}R^{-\alpha_{\mathfrak{L}}}}\Big]\Big\}$$

$$=\underbrace{\mathbb{E}\Big\{\exp\Big[-\frac{Q_{\text{FR2}}I_{x^{\mathfrak{L}}_{\text{FR2}}}}{G_{x^{\mathfrak{L}}_{\text{FR2}}}R^{-\alpha_{\mathfrak{L}}}}\Big]\Big\}}_{B}\exp\Big[-\frac{Q_{\text{FR2}}\sigma^2_{\text{FR2}}}{G_{x^{\mathfrak{L}}_{\text{FR2}}}R^{-\alpha_{\mathfrak{L}}}}\Big], \tag{A.2}$$

where $(c)$ follows from the CDF of exponential distribution. Now, our aim is to derive the expectation of $B$. Applying thinning theorem, the interference term can be further partitioned into two parts with respect to LOS and NLOS transmissions, which is given as $I_{x^{\mathfrak{L}}_{\text{FR2}}}=I_{\Phi^{\mathfrak{L}}_{\text{FR2}}\setminus\{x^{\mathfrak{L}}_{\text{FR2}}\}}+I_{\Phi^{\mathfrak{N}}_{\text{FR2}}}$. Thus, it can further be expanded as

$$B=\underbrace{\mathbb{E}_{I_{\Phi^{\mathfrak{L}}_{\text{FR2}}\setminus\{x^{\mathfrak{L}}_{\text{FR2}}\}}}\Big\{\exp[-\frac{Q_{\text{FR2}}I_{\Phi^{\mathfrak{L}}_{\text{FR2}}\setminus\{x^{\mathfrak{L}}_{\text{FR2}}\}}}{G_{x^{\mathfrak{L}}_{\text{FR2}}}R^{-\alpha_{\mathfrak{L}}}}]\Big\}}_{B_1}$$

$$\times\,\underbrace{\mathbb{E}_{I_{\Phi^{\mathfrak{N}}_{\text{FR2}}}}\Big\{\exp[-\frac{Q_{\text{FR2}}I_{\Phi^{\mathfrak{N}}_{\text{FR2}}}}{G_{x^{\mathfrak{L}}_{\text{FR2}}}R^{-\alpha_{\mathfrak{L}}}}]\Big\}}_{B_2} \tag{A.3}$$

Now, we take the first term as an example to show how to compute the expectation. The other term can be computed in a similar manner. Here, we give the lower bounded result. The upper bound can be found in a similar way. Let $s_{\mathfrak{L}}=-\frac{Q_{\text{FR2}}}{G_{x^{\mathfrak{L}}_{\text{FR2}}}R^{-\alpha_{\mathfrak{L}}}}$. Accordingly, we have $B_1$ given in (A.4).

In the above, $(a)$ follows from the Cauchy-Schwarz inequality, $(b)$ follows from the fact that $\sum_{\hat{k}=1}^{\eta_{\mathfrak{L}}}|h_{\hat{k}b0x^{\mathfrak{L}}_{\text{FR2}}}|^2$ follows Chi-squre/gamma distribution with parameters $\eta_{\mathfrak{L}}$ and 1, $(c)$ follows from: $i)$ $\gamma^2_{\hat{k},n,b}\leq 1$, $ii)$ the probability generating functional of a PPP and $iii)$ we replace $r$ with $r_{b0x^{\mathfrak{L}}_{\text{FR2}}}$ for notational simplicity. The other expectation term can also be calculated in a similar manner, albeit with different integral lower limits *i.e.,* $R^{\frac{\alpha_{\mathfrak{L}}}{\alpha_{\mathfrak{N}}}}$. Finally, substituting the intermediate results into (A.1), the proof is completed.

## Appendix B
## Proof of Proposition 2

Here, we derive the ASP of the $i$th file requested by the typical user and delivered by the LOS FR2 BS located at $x^{\mathfrak{L}}_{\text{FR2}_i}$. All other cases are computed in a similar manner. Now, according to the definition of the ASP of file delivery, we have

$$\mathbb{P}\Big[W_{\text{FR2}}\log\Big(1+\text{SINR}_{x^{\mathfrak{L}}_{\text{FR2}_i}}\Big)\Big]$$

$$\approx\mathbb{P}\left[\frac{\frac{\text{P}_{\text{FR2}}}{U_{\text{FR2}_i}}\frac{n_r^{\text{FR2}}n_t^{\text{FR2}}}{\eta_{\mathfrak{L}}}\mid h^{\text{FR2}}_{x^{\mathfrak{L}}_{\text{FR2}_i}0x^{\mathfrak{L}}_{\text{FR2}_i}}\mid^2 r^{-\alpha_{\mathfrak{L}}}_{x^{\mathfrak{L}}_{\text{FR2}_i}0x^{\mathfrak{L}}_{\text{FR2}_i}}p_{ZF}}{I_{x^{\mathfrak{L}}_{\text{FR2}_i}}+\sigma^2_{\text{FR2}}}\geq Q_{\text{FR2}}\right]$$

$$=\Big(1-\frac{1}{n_t^{\text{FR2}}}\Big)^{(U_{\text{FR2}_i}-1)}$$

$$\times\,\mathbb{E}\left\{\mathbb{P}\left[\mid h^{\text{FR2}}_{x^{\mathfrak{L}}_{\text{FR2}_i}0x^{\mathfrak{L}}_{\text{FR2}_i}}\mid^2\geq\frac{Q_{\text{FR2}}(\sigma^2_{\text{FR2}}+I_{x_{\text{FR2}_i}})}{G_{x^{\mathfrak{L}}_{\text{FR2}_i}}r^{-\alpha_{\mathfrak{L}}}_{x^{\mathfrak{L}}_{\text{FR2}_i}0x^{\mathfrak{L}}_{\text{FR2}_i}}}\right]\right\}, \tag{B.1}$$

$$
\begin{aligned}
B_1 =& \mathbb{E}\Bigg\{ \exp\Big(s_{\mathfrak{L}} \sum_{b\in\Phi^{\mathfrak{L}}_{\text{FR2}},\, b\neq x^{\mathfrak{L}}_{\text{FR2}}} \frac{\mathrm{P}_{\text{FR2}}}{\bar{U}_{\text{FR2}}} \frac{n_r^{\text{FR2}} n_t^{\text{FR2}}}{\eta_{\mathfrak{L}}} r_{b0x^{\mathfrak{L}}_{\text{FR2}}}^{-\alpha_{\mathfrak{L}}} \sum_{u_{nb}\in\mathcal{U}_b} |\sum_{\hat{k}=1}^{\eta_{\mathfrak{L}}} h^{\text{FR2}}_{\hat{k}b0x^{\mathfrak{L}}_{\text{FR2}}} \gamma_{\hat{k},n,b}|^2\Big)\Bigg\} \\
&\overset{(a)}{\geq} \mathbb{E}\Big\{ \prod_{b\in\Phi^{\mathfrak{L}}_{\text{FR2}},\, b\neq x^{\mathfrak{L}}_{\text{FR2}}} \mathbb{E}\Big\{ \exp\Big(s_{\mathfrak{L}} \frac{\mathrm{P}_{\text{FR2}}}{\bar{U}_{\text{FR2}}} \frac{n_r^{\text{FR2}} n_t^{\text{FR2}}}{\eta_{\mathfrak{L}}} r_{b0x^{\mathfrak{L}}_{\text{FR2}}}^{-\alpha_{\mathfrak{L}}} \big( \sum_{u_{nb}\in\mathcal{U}_b} \sum_{\hat{k}=1}^{\eta_{\mathfrak{L}}} \gamma^2_{\hat{k},n,b}\big) \sum_{\hat{k}=1}^{\eta_{\mathfrak{L}}} |h^{\text{FR2}}_{\hat{k}b0x^{\mathfrak{L}}_{\text{FR2}}}|^2\Big)\Big\}\Big\} \\
&\overset{(b)}{=} \mathbb{E}\Big\{ \prod_{b\in\Phi^{\mathfrak{L}}_{\text{FR2}},\, b\neq x^{\mathfrak{L}}_{\text{FR2}}} \Big(1 - s_{\mathfrak{L}} \frac{\mathrm{P}_{\text{FR2}}}{\bar{U}_{\text{FR2}}} \frac{n_r^{\text{FR2}} n_t^{\text{FR2}}}{\eta_{\mathfrak{L}}} r_{b0x^{\mathfrak{L}}_{\text{FR2}}}^{-\alpha_{\mathfrak{L}}} \big( \sum_{u_{nb}\in\mathcal{U}_b} \sum_{\hat{k}=1}^{\eta_{\mathfrak{L}}} \gamma^2_{\hat{k},n,b}\big)\Big)^{-\eta_{\mathfrak{L}}} \Big\} \\
&\overset{(c)}{\geq} \exp\Big\{ -\int_R^\infty \Big[1 - \Big(\frac{1}{1 - s_{\mathfrak{L}} \mathrm{P}_{\text{FR2}} n_r^{\text{FR2}} n_t^{\text{FR2}} r^{-\alpha_{\mathfrak{L}}}}\Big)^{\eta_{\mathfrak{L}}}\Big] 2\pi\lambda_{\text{FR2}} e^{-\beta r} r \mathrm{d}r \Big\}. \qquad \text{(A.4)}
\end{aligned}
$$

where $G_{x^{\mathfrak{L}}_{\text{FR2}_i}} = \frac{\mathrm{P}_{\text{FR2}}}{U_{\text{FR2}_i}} \frac{n_r^{\text{FR2}} n_t^{\text{FR2}}}{\eta_{\mathfrak{L}}}$ and $Q_{\text{FR2}} = 2^{\frac{\nu}{W_{\text{FR2}}}} - 1$. The expectation terms are computed below.

$$
\begin{aligned}
&\mathbb{E}\Bigg\{ \mathbb{P}\Bigg[ |\, h^m_{x^{\mathfrak{L}}_{\text{FR2}_i} 0 x^{\mathfrak{L}}_{\text{FR2}_i}}|^2 \geq \frac{Q_{\text{FR2}}(\sigma^2_{\text{FR2}} + I_{x^{\mathfrak{L}}_{\text{FR2}_i}})}{G_{x^{\mathfrak{L}}_{\text{FR2}_i}} r^{-\alpha_{\mathfrak{L}}}_{x^{\mathfrak{L}}_{\text{FR2}_i} 0 x^{\mathfrak{L}}_{\text{FR2}_i}}} \Bigg] \Bigg\} \\
&\overset{(a)}{=} \int_0^\infty \mathbb{E}_{I_{x^{\mathfrak{L}}_{\text{FR2}_i}}} \Bigg\{ \exp\Bigg( - \frac{Q_{\text{FR2}}(\sigma^2_{\text{FR2}} + I_{x^{\mathfrak{L}}_{\text{FR2}_i}})}{G_{x^{\mathfrak{L}}_{\text{FR2}_i}} r^{-\alpha_{\mathfrak{L}}}_{x^{\mathfrak{L}}_{\text{FR2}_i} 0 x^{\mathfrak{L}}_{\text{FR2}_i}}} \Bigg) \Bigg\} \\
&\times \bar{f}_{R^*_{x^{\mathfrak{L}}_{\text{FR2}_i} 0 x^{\mathfrak{L}}_{\text{FR2}_i}}} (r_{x^{\mathfrak{L}}_{\text{FR2}_i} 0 x^{\mathfrak{L}}_{\text{FR2}_i}}) \mathrm{d}r_{x^{\mathfrak{L}}_{\text{FR2}_i} 0 x^{\mathfrak{L}}_{\text{FR2}_i}} \\
&= \int_0^\infty \exp\Bigg( - \frac{Q_{\text{FR2}} \sigma^2_{\text{FR2}}}{G_{x^{\mathfrak{L}}_{\text{FR2}_i}} r^{-\alpha_{\mathfrak{L}}}_{x^{\mathfrak{L}}_{\text{FR2}_i} 0 x^{\mathfrak{L}}_{\text{FR2}_i}}} \Bigg) \bar{f}_{R^*_{x^{\mathfrak{L}}_{\text{FR2}_i} 0 x^{\mathfrak{L}}_{\text{FR2}_i}}} (r_{x^{\mathfrak{L}}_{\text{FR2}_i} 0 x^{\mathfrak{L}}_{\text{FR2}_i}}) \\
&\times \mathbb{E}_{I_{x^{\mathfrak{L}}_{\text{FR2}_i}}} \Bigg\{ \exp\Bigg( - \frac{Q_{\text{FR2}} I_{x^{\mathfrak{L}}_{\text{FR2}_i}}}{G_{x^{\mathfrak{L}}_{\text{FR2}_i}} r^{-\alpha_{\mathfrak{L}}}_{x^{\mathfrak{L}}_{\text{FR2}_i} 0 x^{\mathfrak{L}}_{\text{FR2}_i}}} \Bigg) \Bigg\} \mathrm{d}r_{x^{\mathfrak{L}}_{\text{FR2}_i} 0 x^{\mathfrak{L}}_{\text{FR2}_i}}, \qquad \text{(B.2)}
\end{aligned}
$$

where $(a)$ follows from the CCDF of the exponential distribution and $\bar{f}_{R^*_{x^{\mathfrak{L}}_{\text{FR2}_i} 0 x^{\mathfrak{L}}_{\text{FR2}_i}}}(\cdot)$ is the distribution of the serving distance between the serving BS and the typical user given as (11). Now, the aim is to compute the expectation term in the integral with respect to the interference.

Applying thinning theorem, the interference can be partitioned into four terms, namely $I_{x^{\mathfrak{L}}_{\text{FR2}_i}} = I_{\Phi^{\mathfrak{L}}_{\text{FR2}_i} \setminus \{x^{\mathfrak{L}}_{\text{FR2}_i}\}} + I_{\Phi^{\mathfrak{N}}_{\text{FR2}_i}} + I_{\bar{\Phi}^{\mathfrak{L}}_{\text{FR2}_i}} + I_{\bar{\Phi}^{\mathfrak{N}}_{\text{FR2}_i}}$, where the thinned PPPs $\Phi^j_{\text{FR2}_i}$ and $\bar{\Phi}^j_{\text{FR2}_i}$ with $j \in \{\mathfrak{L}, \mathfrak{N}\}$ are respectively thinned from the cache hit and cache miss cases and then thinned again with respect to LOS and NLOS. For notational simplicity, we replace $r_{x^{\mathfrak{L}}_{\text{FR2}_i} 0 x^{\mathfrak{L}}_{\text{FR2}_i}}$ with $R$ in the hereinafter. Accordingly, the expectation term in (B.2) is reduced to

$$
\begin{aligned}
&\mathbb{E}_{I_{x^{\mathfrak{L}}_{\text{FR2}_i}}} \Bigg\{ \exp(-\frac{Q_{\text{FR2}} I_{x^{\mathfrak{L}}_{\text{FR2}_i}}}{G_{x^{\mathfrak{L}}_{\text{FR2}_i}} R^{-\alpha_{\mathfrak{L}}}}) \Bigg\} \\
&= \mathbb{E}_{I_{\Phi^{\mathfrak{L}}_{\text{FR2}_i} \setminus \{x^{\mathfrak{L}}_{\text{FR2}_i}\}}} \Bigg\{ \exp\Bigg( - \frac{Q_{\text{FR2}} I_{\Phi^{\mathfrak{L}}_{\text{FR2}_i}}}{G_{x^{\mathfrak{L}}_{\text{FR2}_i}} R^{-\alpha_{\mathfrak{L}}}} \Bigg) \Bigg\} \\
&\times \mathbb{E}_{I_{\Phi^{\mathfrak{N}}_{\text{FR2}_i}}} \Bigg\{ \exp\Bigg( - \frac{Q_{\text{FR2}} I_{\Phi^{\mathfrak{N}}_{\text{FR2}_i}}}{G_{x^{\mathfrak{L}}_{\text{FR2}_i}} R^{-\alpha_{\mathfrak{L}}}} \Bigg) \Bigg\} \\
&\times \mathbb{E}_{I_{\bar{\Phi}^{\mathfrak{L}}_{\text{FR2}_i}}} \Bigg\{ \exp\Bigg( - \frac{Q_{\text{FR2}} I_{\bar{\Phi}^{\mathfrak{L}}_{\text{FR2}_i}}}{G_{x^{\mathfrak{L}}_{\text{FR2}_i}} R^{-\alpha_{\mathfrak{L}}}} \Bigg) \Bigg\} \\
&\times \mathbb{E}_{I_{\bar{\Phi}^{\mathfrak{N}}_{\text{FR2}_i}}} \Bigg\{ \exp\Bigg( - \frac{Q_{\text{FR2}} I_{\bar{\Phi}^{\mathfrak{N}}_{\text{FR2}_i}}}{G_{x^{\mathfrak{L}}_{\text{FR2}_i}} R^{-\alpha_{\mathfrak{L}}}} \Bigg) \Bigg\}. \qquad \text{(B.3)}
\end{aligned}
$$

Particular care is taken to compute the lower limit of the probability generating functional (PGFL) of the PPP $\bar{\Phi}^j_{m_i}$ with $j \in \{\mathfrak{L}, \mathfrak{N}\}$, which is independent of the serving distance. Below, we only show the expectation for the first term as all other terms follow suit. Following the PGFL of a PPP and the extension of the proof for Proposition 1, the lower bound is given as

$$
\begin{aligned}
\mathbb{E}\Big\{ \exp\Big( s I_{\Phi^{\mathfrak{L}}_{\text{FR2}_i}} \Big) \Big\} \geq \exp\Bigg[ -\int_R^\infty [1 - (1 - s \mathrm{P}_{\text{FR2}} n_r^{\text{FR2}} n_t^{\text{FR2}} r^{-\alpha_{\mathfrak{L}}})^{-\eta_{\mathfrak{L}}}] \\
\times 2\pi\lambda_{\text{FR2}} p_{\mathfrak{L}}(r) r \omega_{\text{FR2}_i} \mathrm{d}r \Bigg], \qquad \text{(B.4)}
\end{aligned}
$$

where $s = \frac{-Q_{\text{FR2}}}{G_{x^{\mathfrak{L}}_{\text{FR2}_i}} R^{-\alpha_{\mathfrak{L}}}}$. Now, by substituting all the terms in (B.2) the proof is obtained.

## REFERENCES


[1] 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.

[2] Z. Pi, J. Choi, and R. Heath, "Millimeter-wave gigabit broadband evolution toward 5G: Fixed access and backhaul," *IEEE Commun. Mag.*, vol. 54, no. 4, pp. 138–144, Apr. 2016.

[3] S. Biswas, C. Masouros, and T. Ratnarajah, "Performance analysis of large multiuser mimo systems with space-constrained 2-D antenna arrays," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3492–3505, May. 2016.

[4] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

[5] E. Bastug, M. Bennis, and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," *EURASIP J. Wireless Commun. Netw.*, vol. 2015, no. 1, p. 41, Feb. 2015.

[6] 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. 2015.

[7] Y. Zhu, G. Zheng, K. Wong, S. Jin, and S. Lambotharan, "Performance analysis of cache-enabled millimeter wave small cell networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 6695–6699, Jul. 2018.

[8] N. Giatsoglou, K. Ntontin, E. Kartsakli, A. Antonopoulos, and C. Verikoukis, "D2D-aware device caching in mmwave-cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 2025–2037, Sep. 2017.

[9] S. Vuppala, T. X. Vu, S. Gautam, S. Chatzinotas, and B. Ottersten, "Cache-aided millimeter wave ad hoc networks with contention-based content delivery," *IEEE Trans. Commun.*, vol. 66, no. 8, pp. 3540–3554, Aug. 2018.

[10] Z. Tan, X. Li, F. R. Yu, H. Ji, and V. C. M. Leung, "Joint resource allocation in cache-enabled small cell networks with massive MIMO and full duplex," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Singapore, Singapore, Dec. 2017, pp. 1–6.

[11] A. Papazafeiropoulos and T. Ratnarajah, "Modeling and performance of uplink cache-enabled massive MIMO heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 8136–8149, Dec. 2018.

[12] L. Wang, K. Wong, S. Lambotharan, A. Nallanathan, and M. Elkashlan, "Edge caching in dense heterogeneous cellular networks with massive MIMO-aided self-backhaul," *IEEE Trans. Wireless Commun.*, vol. 17, no. 9, pp. 6360–6372, Sep. 2018.

[13] 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.

[14] Y. Chen, M. Ding, J. Li, Z. Lin, G. Mao, and L. Hanzo, "Probabilistic small-cell caching: Performance analysis and optimization," *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 4341–4354, May. 2017.

[15] S. Tamoor-ul-Hassan, M. Bennis, P. H. J. Nardelli, and M. Latva-aho, "Caching in wireless small cell networks: A storage-bandwidth tradeoff," *IEEE Commun. Lett.*, vol. 20, no. 6, pp. 1175–1178, Jun. 2016.

[16] Z. Yan, S. Chen, Y. Ou, and H. Liu, "Energy efficiency analysis of cache-enabled two-tier HetNets under different spectrum deployment strategies," *IEEE Access*, vol. 5, pp. 6791–6800, Mar. 2017.

[17] C. Fan, T. Zhang, Z. Zeng, and Y. Chen, "Energy efficiency analysis of cache-enabled cellular networks with limited backhaul," *Wireless Commun. Mobile Comput.*, vol. 2018, Feb. 2018.

[18] M. N. Kulkarni, A. Ghosh, and J. G. Andrews, "A comparison of MIMO techniques in downlink millimeter wave cellular networks with hybrid beamforming," *IEEE Trans. Commun.*, vol. 64, no. 5, pp. 1952–1967, May. 2016.

[19] 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. 2014.

[20] 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.

[21] 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.

[22] 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.

[23] A. K. Gupta, A. Alkhateeb, J. G. Andrews, and R. W. Heath, "Restricted secondary licensing for mmwave cellular: How much gain can be obtained?" in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2016, pp. 1–6.

[24] C. Saha and H. S. Dhillon, "Millimeter wave integrated access and backhaul in 5G: Performance analysis and design insights," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 12, pp. 2669–2684, Dec. 2019.

[25] O. Y. Kolawole, S. Vuppala, and T. Ratnarajah, "Multiuser millimeter wave cloud radio access networks with hybrid precoding," *IEEE Syst. J.*, vol. 12, no. 4, pp. 3661–3672, Dec. 2018.

[26] H. 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.

[27] S. Biswas, T. Zhang, K. Singh, S. Vuppala, and T. Ratnarajah, "An analysis on caching placement for millimeter–micro-wave hybrid networks," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1645–1662, Feb. 2018.

[28] 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.

[29] A. Alkhateeb, R. W. Heath, and G. Leus, "Achievable rates of multi-user millimeter wave systems with hybrid precoding," in *Proc. IEEE Int. Conf. Commun. Workshop (ICCW)*, London, UK, Jun. 2015, pp. 1232–1237.

[30] M. N. Kulkarni, A. Alkhateeb, and J. G. Andrews, "A tractable model for per user rate in multiuser millimeter wave cellular networks," in *Proc. IEEE Asilomar Conf. Signals, Syst. Comput. (ACSSC)*, Pacific Grove, CA, USA, Nov. 2015, pp. 328–332.

[31] F. A. Khan, H. He, J. Xue, and T. Ratnarajah, "Performance analysis of cloud radio access networks with distributed multiple antenna remote radio heads," *IEEE Trans. Signal Process.*, vol. 63, no. 18, pp. 4784–4799, Sep. 2015.

[32] G. Zhang, T. Q. S. Quek, M. Kountouris, A. Huang, and H. Shan, "Fundamentals of heterogeneous backhaul design-Analysis and optimization," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 876–889, Feb. 2016.

[33] Z. Yang *et al.*, "Cache placement in two-tier HetNets with limited storage capacity: Cache or buffer?" *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5415–5429, Nov. 2018.

[34] S. Singh, X. Zhang, and J. G. Andrews, "Joint rate and SINR coverage analysis for decoupled uplink-downlink biased cell associations in hetnets," *IEEE Trans. Wireless Commun.*, vol. 14, no. 10, pp. 5360–5373, Oct. 2015.

[35] K. A. Hamdi, "A useful lemma for capacity analysis of fading interference channels," *IEEE Trans. Commun.*, vol. 58, no. 2, pp. 411–416, Feb. 2010.

[36] Q. Zhang, H. H. Yang, T. Q. S. Quek, and J. Lee, "Heterogeneous cellular networks with LoS and NLoS transmissions-The role of massive MIMO and small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 7996–8010, Dec. 2017.


**Tong Zhang** received his MSc degree in signal processing and communications from the University of Edinburgh, Edinburgh, UK, in 2016. He is currently pursuing his Ph.D at the Institute for Digital Communications at the University of Edinburgh. His research interests mainly lie in the area of wireless edge caching, millimeter-wave communications and stochastic geometry.

**Sudip Biswas** (S'16, M'17) received the Ph.D. degree in Digital Communications in 2017 from the University of Edinburgh (UEDIN), UK and currently works at the Indian Institute of Information Technology Guwahati (IIITG) as an Asst. Professor in the Department of Electronics and Communications Engineering. Prior to this he held the position of research associate from 2017 to 2019 at the Institute of Digital Communications in UEDIN. He also has industrial experience with Tata Consultancy Services, India (Lucknow & Kolkata), where he held the position of Asst. Systems Engineer from 2010 to 2012. Currently, he leads research on signal processing for wireless communications, with particular focus on 5G's long-term evolution.

**Tharmalingam Ratnarajah** (A'96-M'05-SM'05) is currently with the Institute for Digital Communications, the University of Edinburgh, Edinburgh, UK, as a Professor in Digital Communications and Signal Processing. He was the Head of the Institute for Institute for Digital Communications during 2016-2018. His research interests include signal processing and information theoretic aspects of 5G and beyond wireless networks, full-duplex radio, mmWave communications, random matrix theory, interference alignment, statistical and array signal processing and quantum information theory. He has published over 400 papers in these areas and holds four U.S. patents. He has supervised 15 PhD students and 20 post-doctoral research fellows, and raised 11 million+ USD of research funding. Dr Ratnarajah was an associate editor for IEEE Transactions on Signal Processing (2015-2017) and technical co-chair in the 17th IEEE International workshop on Signal Processing advances in Wireless Communications, Edinburgh, UK, 2016.