Computation Offloading and Service Caching in Heterogeneous MEC Wireless Networks

Yongqiang Zhang, *Graduate Student Member, IEEE*,
Mustafa A. Kishk, *Member, IEEE*, and Mohamed-Slim Alouini, *Fellow, IEEE*

Abstract—Mobile edge computing (MEC) can dramatically promote the computation capability and prolong the lifetime of mobile users (MUs) by offloading computation-intensive tasks to edge cloud. In this paper, a spatial-random two-tier heterogeneous network (HetNet) is modelled to feature random node distribution, where the small-cell base stations (SBSs) and the macro base stations (MBSs) are cascaded with servers with different levels of computing and storage capacity. Only a certain type of application services and finite number of offloaded tasks can be cached and processed in the resource-limited edge server. For that setup, we investigate the performance of two offloading strategies corresponding to integrated access and backhaul (IAB)-enabled MEC networks and traditional cellular MEC networks. Using tools from stochastic geometry and queuing theory, we derive the average delay for the two different strategies, in order to better understand the influence of IAB on MEC networks. Simulations results are provided to verify the derived expressions and to reveal various system-level insights.

Index Terms—Edge computing, stochastic geometry, HetNet, Integrated access and backhaul

1 Introduction

With the realization of the Internet of Things (IoT), it is an irreversible trend that more mobile devices will be connected to the wireless access system. This paradigm has motivated developers to create more interactive applications that require low-latency communication and computation, such as face recognition, natural language processing, and augmented reality (AR). As a growing computing paradigm, computation offloading can break such an obstacle. By migrating part or all of the data processing tasks of mobile applications from resource-limited mobile devices to the cloud, computation offloading can effectively reduce the energy consumption of mobile devices. The idea behind MEC is to enhance the computing potentials of mobile devices by placing cloud computing platforms at the edges of the networks [1], [2]. Since the computing resources are placed in the vicinity of the MUs, MEC enables mobile devices to execute the highly-demanding computing tasks under the strict delay constraint [3].

As a key-enabled network architecture of MEC, base stations (BSs) cascaded with edge-cloud servers were widely considered in existing works. Moreover, BSs are also able to access the central cloud in case of running out their computing or storage capacity, which results in a hierarchical computation offloading architecture. Different from the consideration of computation offloading architecture, the heterogeneity and diversity of computation services are often overlooked in many recent works. In general, different services require non-identical databases or libraries cached at the BS. For example, the type and data size of object databases for different AR services are different. While the central cloud server has more resources, the limited computing and storage capacity of the edge-cloud server allows only a small set of services to be cached during a typical period. Thus, the type of computation services cached at the BS not only determines the type of tasks that can be offloaded but also affects the network performance.

On the other hand, the density of BSs has grown to $50/\mathrm{km}^2$ in 5G deployments [4]. However, the capital cost of site and fiber deployments is one of the biggest challenges in large-scale networks densification [5]. Therefore, wireless backhaul is a promising replacement solution for wired-fiber backhaul, providing almost the same transmitting rate as fiber optic, but with considerable lower cost and more flexible/timely deployment (e.g., no intrusion) [6]. In this context, integrated access and backhaul (IAB) networks, where the operator can use part of radio resources to do wireless backhauling, has recently attracted more research attention [7].

Based on these observations, in this paper, we develop a tractable analytical framework for a two-tier heterogeneous MEC network, and obtain some useful design guidelines for the deployment studies on IAB-enabled MEC networks. Making use of stochastic geometry and queuing theory, the proposed framework is helpful in designing of the large-scale MEC networks. Moreover, both the heterogeneity of services and diversity of the MEC servers are considered, which enhances the practicality of the analytical framework. Although there are some researches on service caching and computation offloading in two-tier heterogeneous MEC networks [8], [9], [10], the connectivity among the BSs in the

- *The authors are with the Computer, Electrical and Mathematical Science and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia. E-mail: {yongqiang.zhang.2, mustafa.kishk, slim.alouini}@kaust.edu.sa.*

Manuscript received 26 January 2021; revised 29 November 2021; accepted 13 December 2021. Date of publication 20 December 2021; date of current version 5 May 2023.
This work was supported by King Abdullah University of Science and Technology (KAUST).
(Corresponding author: Mustafa A. Kishk.)
Digital Object Identifier no. 10.1109/TMC.2021.3136595

1536-1233 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Maynooth University Library. Downloaded on September 18,2023 at 15:07:55 UTC from IEEE Xplore. Restrictions apply.

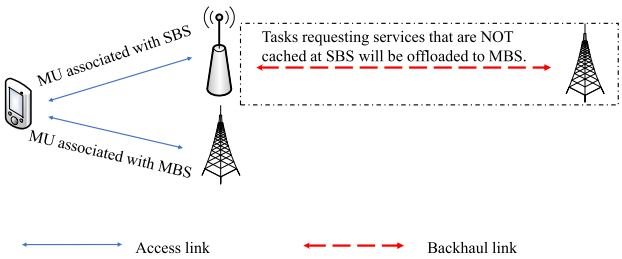


Fig. 1. Illustration of Strategy I.

same tier under the heterogeneity of computing services was not taken into account. The main contributions of the paper can be summarized as following.

1) We construct a tractable and realistic model for analytically characterizing the performance of a heterogeneous MEC network in terms of average delays. We assume that only the MBSs have access to the fiber backhaul while the SBSs are wirelessly backhauled by the MBSs. Due to the limited storage resources, we consider that only finite computation services and limited set of tasks can be cached and offloaded at a typical SBS. We derive the average delay achieved by a typical MU for two strategies:
   - Strategy I: MUs offload the tasks to the nearest BS. If the MU choose to connect with a SBS, and the tagged SBS has not cached the required service, the SBS will offload this task to the nearest MBS.
   - Strategy II: The MUs offload the tasks to the nearest BS that has already cached the required service, whether it is a SBS or a MBS. In other words, the operation inside the dash-lined box in Fig. 1 is not considered for strategy II.
2) Using numerical results, we draw multiple useful system-level insights, which are summarized below.
   - There exists an optimal association bias factor that minimizes the average delay, which is located between the the optimal bias factors for average transmission time and average response time.
   - The average delay for the two considered strategies decreases as the density of SBSs or MBSs increases. However, it is noticed that the increase in the MBSs' density has a stronger influence on the system performance than that of the SBSs.
   - When the SBS service rate or buffer size exceeds a certain level, the average delay for the two considered strategies will achieve convergence.
   - For the low and middle range of BS density and SBS service rate, Strategy I outperforms Strategy II in terms of the average delay. Thus, Strategy I is a more cost-effective solution.

2 Related Studies

In this section, we will provide a brief summary of the related works in four main directions of interest: (i) computation offloading in edge computing, (ii) service caching in edge computing, (iii) IAB networks, and (iv) stochastic geometry-based analysis of edge computing networks.

2.1 Computation Offloading in Edge Computing

Computation offloading has been the central theme of MEC network studies for the past decade. There have been a lot of existing works on designing policies for computation offloading. In [11], You *et al.* studied the trade-off between energy consumption and computation latency for a multi-users MEC system in cases of infinite and finite cloud computation capacities. The authors in [12] addressed the computation offloading and resource allocation problem with the consideration of queuing delay and energy consumption. In [8], considering the scenario of two-tier heterogeneous IoT networks, a game-theoretic greedy scheme was proposed to solve the multi-task multi-access computation offloading problem. The energy consumption minimization problem for the multi-task multi-access MEC system via non-orthogonal multiple access (NOMA) was addressed in [13]. In [14] and [15], based on deep Q-learning, the authors derived the optimal offloading decisions for the vehicular MEC Networks. The authors in [16] studied the computation offloading problem in the unmanned aerial vehicle (UAV) based MEC network, with an objective to optimize the offloading decision, radio resource allocation, and the UAV trajectory. Mao *et al.* [17] incorporated energy harvesting technology into the MEC system and proposed an energy-efficient scheme that considered latency and offloading failure. However, the MEC servers in all the above-mentioned studies were assumed to capable of processing all types of tasks offloaded by MUs. In general, the computing platforms first need to cache the corresponding program data for executing a specific type of application. To bridge this gap, in this paper, we consider the scenario that the service program pool at the SBSs cannot provide all the services required by the MUs.

2.2 Service Caching in Edge Computing

Compared with computation offloading, service caching for the MEC networks has received little attention until recent few years. The authors in [18] designed an online service caching mechanism for edge-severs to determine which service to download from remote data centers without any information about the requests of future arrival tasks. In [19], the authors considered that each service is associated with a cache cost, and addressed the overall cache cost minimization problem via a greedy algorithm. The authors in [10] studied the service caching problem under a service market with multiple service providers which are competing for both communication and computation resources, and proposed a game-theoretical mechanism to minimize the cache cost of all service providers. By taking into account the availability of service in edge cloud servers, the joint optimal service caching and task offloading decisions were presented in [20]. The authors in [21] addressed the joint service caching and user association problem. With the consideration of cooperative edge-cloud servers, the optimal service placement and computation task routing problem was addressed in [22]. In [23], the authors modeled a service caching and resource allocation problem as a mixed-integer nonlinear programming. To maximize the economic profits of the MEC network, the authors in [24] proposed a Lyapunov optimization-based algorithm to obtain the

Authorized licensed use limited to: Maynooth University Library. Downloaded on September 18,2023 at 15:07:55 UTC from IEEE Xplore. Restrictions apply.

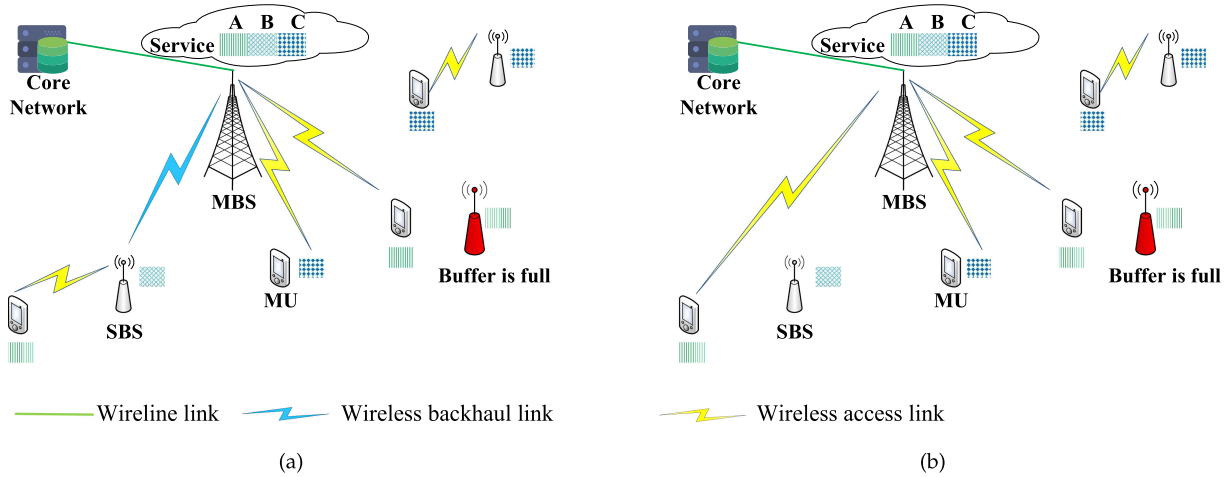


Fig. 2. System model of : (a) Strategy I, (b) Strategy II.

optimal service caching and task offloading policies. In [25], the joint optimization of service caching design, offloading decision, and time allocation was studied in order to minimize the overall energy consumption. To extend the coverage and on-demand deployment capability, the service caching issue in the UAV-assisted MEC networks was explored in [26]. However, all the above-mentioned studies assumed that the number of MUs and BSs is fixed, which may not be realistic for the large-scale network. Although the diversity of services was considered, the heterogeneity of both the computation and storage capacity was neglected.

2.3 Studies of IAB Networks

With the development of network densification, backhaul link will be a primary bottleneck of wireless cellular networks. In order to break such an obstacle, IAB was proposed and received substantial interest [27], [28]. The authors investigated the joint resource allocation and IAB-nodes deployment problem for a MIMO orthogonal frequency-division multiple access (OFDMA) based IAB network in [29]. Authors in [30] presented a simulated annealing algorithm to optimize scheduling and power allocation. In [31], using end-to-end simulations, the authors studied the feasibility of mmWave-based IAB networks. In [32], the authors evaluated the benefit of adopting IAB in a fixed wireless access system. By using stochastic geometry, the coverage probability and several bandwidth allocation schemes of large-scale IAB-enabled mmWave heterogeneous networks were investigated in [33], [34], [35].

2.4 Stochastic Geometry-Based Analysis of Edge Computing Networks

Stochastic geometry has been established as a strong tool for modeling, analysing, and designing large scale wireless networks [36], [37], [38]. There have been several studies on stochastic geometry-based analysis of MEC-enabled networks. In particular, single-tier MEC networks were considered in [39], [40], [41] while heterogeneous MEC networks were considered in [9], [42]. MUs with identical computation tasks were considered in [40] and [41], with a spatial model characterized by Poisson point processes (PPPs) and Poisson cluster processes (PCPs), respectively. Although the heterogeneity of computation tasks was considered in [42] and [9], the heterogeneity of computation services was not captured. The load migration from the limited computation capacity edge server to the central cloud in a cell-free network was investigated in [9]. However, the detailed process of this load balancing scheme was not further studied.

3 System Model

We consider a two-tier hierarchical MEC network. There are multiple MBSs and SBSs. Each BS is co-located with a computing server in order to provide computing services for MUs. The locations of MBSs, SBSs and MUs are modeled by three independent PPPs, Φ_m with intensity λ_m, Φ_s with intensity λ_s, and Φ_u with intensity λ_u, respectively. One possible realization of the locations of the MBSs, SBSs, and MUs is illustrated in Fig. 3.

MUs are capable of offloading their time-consuming and computation-intensive tasks to the MBSs or the SBSs, and then receiving the output of the task through the wireless network. The time consumption of computation offloading in the considered system consists of three sequential phases: transmitting, computing and receiving phases.

As shown in Fig. 2, we consider two different computation offloading strategies for the considered system. For strategy II, a typical MU is only able to offload its task to a MBS or an *available* SBS. A SBS is *available* if it has already cached the required service for the computing task generated at the MU, and it has not run out of the limited storage space at its buffer. Compared with Strategy II, Strategy I is more flexible. For instance, MU can offload its task to a SBS

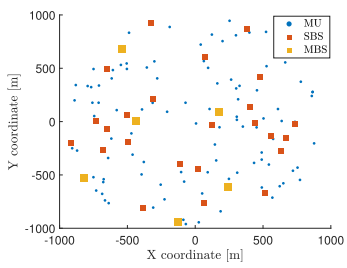


Fig. 3. Simulated locations of the MBSs, SBSs and MUs.

Authorized licensed use limited to: Maynooth University Library. Downloaded on September 18,2023 at 15:07:55 UTC from IEEE Xplore. Restrictions apply.

even though this SBS has not cached the requested service. With the consideration of IAB, a MBS provides cloud computing service for MUs and SBSs in its voronoi cell through access link and backhaul link, respectively. In other words, SBSs in strategy I can act as relays to assist MUs in offloading their computation tasks to the MBS.

3.1 Communication Model

We assume MBSs and SBSs have different transmitting powers (which are denoted by p_m and p_s), and the signals experience path loss with the same path loss exponent α. The received power at a receiver located at Y from a transmitter located at X is $p_X H \|X-Y\|^{-\alpha}$, where H denotes the channel power gain, p_X is the transmitting power. The random channel gains are Rayleigh distributed with unitary average power, i.e., $H \sim \exp(1)$. Besides, we assume that the MUs in the same small cell use different orthogonal resource blocks, as well as the SBSs in the same macro cell.

We consider using the maximum biased average power-based cell association rule. Let D_m and D_s denote the distance of a typical MU from the nearest MBS and the nearest SBS, respectively. The typical MU located at X will choose to connect to tier $\mathcal{K}$ if

$$\mathcal{K}_X = \arg \max_{k \in \{s,m\}} p_k B_k D_k^{-\alpha}, \tag{1}$$

where B_k denotes the association bias factor for tier k BSs. For example, if $p_k B_k > p_j B_j$, then more traffic load will be routed through tier k BSs as compared to tier j BSs.

The association cell of a BS of type $k \in \{s,m\}$ located at X is given by

$$\mathcal{C}_X = \left\{ Z \in \mathbb{R}^2 : p_k B_k \|Z-X\|^{-\alpha} \geq p_j B_j {D_j}^{-\alpha},\ \forall j \in \{s,m\} \right\}. \tag{2}$$

Lemma 1 (Mobile user association probability). *The probability that a typical MU connects with tier k is given by*

$$\mathcal{A}_k = \frac{\lambda_k}{\lambda_k + \lambda_j [(p_j B_j)/(p_k B_k)]^{2/\alpha}}. \tag{3}$$

Proof.

$$\begin{aligned}
\mathcal{A}_k &= \Pr\left(p_k B_k D_k^{-\alpha} > p_j B_j D_j^{-\alpha}, k \neq j \right) \\
&= \int_0^\infty \left(1 - F_{D_j}\left(\left(\frac{p_j B_j}{p_k B_k} \right)^{1/\alpha} x \right) \right) f_{D_k}(x) \mathrm{d}x \\
&= \int_0^\infty 2\pi \lambda_k x \exp\left(-\pi x^2 \left(\lambda_k + \lambda_j \left(\frac{p_j B_j}{p_k B_k} \right)^{2/\alpha} \right) \right) \mathrm{d}x \\
&= \frac{\lambda_k}{\lambda_k + \lambda_j (p_j B_j / p_k B_k)^{2/\alpha}}.
\end{aligned} \tag{4}$$

□

Lemma 1 indicates that a MU prefers to associate with the tier with higher BS density, transmit power, and bias. In the considered system model, $\frac{2}{\alpha} < 1$ ($\alpha > 2$) which leads to that the BS density is more dominant in determining $\mathcal{A}_k$ than BS transmit power or bias factor. Lemma 1 is useful for quantifying the number of connected MUs of each tier.

As for the backhaul link in the Strategy I, the typical SBS associates with its nearest MBS.

3.2 Service and Computation Model

There is a set of computation tasks $\{\mathcal{T}_i | i = 1, 2, \ldots, \mathcal{M}\}$, each $\mathcal{T}_i$ can be described by a 4-tuple, i.e., $\mathcal{T}_i = (\mathcal{S}_i, \mathcal{L}_i, \mathcal{D}_i^{ul}, \mathcal{D}_i^{dl})$. Here, $\mathcal{S}_i$ is the required type of services to execute $\mathcal{T}_i$, $\mathcal{L}_i$ is the required central processing unit (CPU) cycles to complete $\mathcal{T}_i$, $\mathcal{D}_i^{ul}$ and $\mathcal{D}_i^{dl}$ denote the input and output data size of $\mathcal{T}_i$. A typical user generates a task $\mathcal{T}_i$ with probability q_i. Moreover, due to the limited storage capacity at the SBS, we assume that each SBS only cached one specific $\mathcal{S}_i$, with probability q_i. In the following, we refer to the SBS which has cached $\mathcal{S}_i$ as the type-i SBS.

Since both MUs and SBSs are distributed as PPP, the arrivals of offloading tasks at a computing server follow a Poisson process with a specific arrival rate. The arrival rate for at a type-i SBS in two considered strategies are given respectively by

$$\Lambda_{s,i}^1 = \frac{q_i \mathcal{A}_s \lambda_u}{\lambda_s}, \tag{5a}$$

$$\Lambda_{s,i}^2 = \frac{\mathcal{A}_{s,i} \lambda_u}{\lambda_s}, \tag{5b}$$

where $\mathcal{A}_{s,i}$ is the probability that a typical MU with type-i task associates with a SBS in Strategy II.

Moreover, due to the limited computation and communication capacity at the SBSs, we assume that the computing buffer at each SBS is bounded by $\mathcal{N}$. The SBS can not provide service for additional MUs when its computing buffer is full. The probability that the SBS has a full buffer is provided next.

The probability that the buffer is not full for the type-i SBS is defined as

$$P_{\text{ofld}}^i = 1 - \Pr(\text{The SBS}'\text{s buffer is full}) = 1 - \frac{(1-\rho_i)\rho_i^{\mathcal{N}}}{1-\rho_i^{\mathcal{N}+1}}, \tag{6}$$

where $\rho_i = \Lambda_{s,i}/\mu_s$, and $\Lambda_{s,i}$ is the arrival rate given by (5a) and (5b) for Strategy I and Strategy II, respectively. Unlike SBSs, the MBSs are supposed to be equipped with an abundant computing buffer.

For Strategy I, an SBS needs to offload a task to the MBS if the MU requests a task that is not cached at the SBS. The fraction of SBSs that are offloading tasks to MBSs using wireless backhaul can be computed using the below probability.

Definition 1. *The typical SBS is backhaul-active (BH-actv) if it receives at least one task that its required service is not cached in this SBS. The probability for a typical SBS to be BH-actv is*

$$\begin{aligned}
P_{\text{actv}} &= \Pr(\text{SBS received } \mathcal{T}_i, \text{SBS cached service } j, i \neq j) \\
&= \mathbb{E}_{N_s^u}\left[\sum_{i=1}^{\mathcal{M}} (1-q_i) \sum_{k=1}^{N_s^u} q_i^k (1-q_i)^{N_s^u - k} \right],
\end{aligned} \tag{7}$$

where N_s^u is the number of MUs served by typical SBS.

According to [43], the probability mass function (PMF) of N_s^u is given by

$$\Pr(N_s^u = n) = \frac{\Gamma(n+3.5) \cdot 3.5^{3.5}}{\Gamma(3.5)(n-1)!} \left(\frac{\lambda_u \mathcal{A}_s}{\lambda_s} \right)^{(n-1)} \left(3.5 + \frac{\lambda_u \mathcal{A}_s}{\lambda_s} \right)^{-(n+3.5)}. \tag{8}$$

Authorized licensed use limited to: Maynooth University Library. Downloaded on September 18,2023 at 15:07:55 UTC from IEEE Xplore. Restrictions apply.

Similar to (5a) and (5b), the arrival rates at MBS for Strategy I and Strategy II respectively are

$$\Lambda_M^1 = \frac{\mathcal{A}_m \lambda_u + P_{\text{actv}} \lambda_s}{\lambda_m}, \tag{9}$$

and

$$\Lambda_M^2 = \sum_{i=1}^{\mathcal{M}} \frac{\mathcal{A}_{m,i} \lambda_u}{\lambda_m}, \tag{10}$$

where $\mathcal{A}_{m,i}$ is the probability that a MU with type-i task associates a MBS in Strategy II.

The computing capacity of the server at tier k is denoted by F_k, which is measured in CPU cycles per second. It is assumed that the computing servers are operated in the first-come-first-serve manner. For type-i task, the service rate at the tier k server is determined as $\mu_k^i = F_k/\mathcal{L}_i$. The distribution of the service time for the type-i task at tier k server is followed by the exponential distribution with $1/\mu_k^i$. In this context, we are able to adopt M/G/1 and M/M/1/N queueing models to analyze the computing server cascaded with MBSs and SBSs, respectively.

TABLE 1
List of Key Notations

Symbols	Description
Φ_m, λ_m, p_m	PPP of MBSs, the corresponding density, and the corresponding transmission power
Φ_s, λ_s, p_s	PPP of SBSs, the corresponding density, and the corresponding transmission power
Φ_u, λ_u, p_u	PPP of MUs, the corresponding density, and the corresponding transmission power
α; b	Path loss exponent; $2/\alpha$
W	bandwidth
ϵ	Power control fraction
$\mathcal{T}$	Set of computation tasks
$\mathcal{N}$	Capacity of the buffer at SBSs
N_s^u	The number of MUs served by the SBS
$\Lambda_{s,i}$, μ_s^i	Arrival rate at type-i SBS, corresponding serving rate
F_k	Computing capacity of tier k server
$\mathcal{A}_k$	Probability of MUs association with tier k
$\mathcal{C}_B$	Service region of base station located at B
$\mathcal{K}_X$	Serving tier of the MU located at X
H	Small scale fading gain
B_k	The association bias factor of tier k BSs

4 Strategy I

We consider using the Orthogonal Resource Allocation (*ORA*) as the bandwidth partitioning scheme for Strategy I [34]. In *ORA*, there exists a fraction η_a of bandwidth reserved for access links and the rest is used for backhaul links. Therefore, the backhaul link will not experience interference from the access link and vice versa.

4.1 Uplink Coverage Analysis

Let $\Phi_{s^1_{\text{ofld}}}$ denote the point process that models the locations offloadable SBSs in Strategy I, which can be regarded as an independent thinning of Φ_s. By using (6) and law of total probability, the thinning probability is given by $P_{\text{ofld}} = \mathbb{E}[P^i_{\text{ofld}}] = \sum_{i=1}^{\mathcal{M}} q_i P^i_{\text{ofld}}$. Thus, the intensity of $\Phi_{s^1_{\text{ofld}}}$ is $P_{\text{ofld}}\lambda_s$.

If the typical MU is connected to the nearest SBS/MBS located at $\mathbf{B_a}$, assuming that the distance between the user and $\mathbf{B_a}$ is $R_{a,1}$, the signal-to-interference-plus-noise ratio (SINR) at the tagged SBS/MBS for the uplink access link is given by

$$\text{SINR}^u_{a,1} = \frac{p_u H_u R_{a,1}^{\alpha(\epsilon-1)}}{\sigma^2 + p_u I_{u,1}}, \tag{11}$$

with

$$I_{u,1} = \sum_{\mathbf{x} \in \Phi^a_{u,1}} (R_x^\alpha)^\epsilon H_x \|\mathbf{x} - \mathbf{B_a}\|^{-\alpha}, \tag{12}$$

where R_x denotes the serving link distance for interfering MU, $\Phi^a_{u,1}$ denotes the point process for MUs that use the same resource block as the typical user in Strategy I, and the rest of the notations are provided in Table 1.

If the tagged SBS has not already cached the service, the SBS will transmit the offloaded task to the nearest MBS (located at $\mathbf{B}_b$), the SINR at the tagged MBS for the uplink backhaul link is given by

$$\text{SINR}^u_b = \frac{p_s H_s R_2^{-\alpha}}{\sigma^2 + p_s I_s}, \tag{13}$$

with

$$I_s = \sum_{\mathbf{x} \in \Phi^b_{S_{\text{actv}}}} H_j \|\mathbf{x} - \mathbf{B_b}\|^{-\alpha}, \tag{14}$$

where $\Phi^b_{S_{\text{actv}}}$ and R_b denote the point process that models the locations of the active SBSs that are using the same resource block as the typical SBS and the serving distance from typical SBS to its nearest MBS, respectively.

According to the law of total probability, the uplink coverage probability in Strategy I is given by

$$\begin{aligned}\mathcal{P}^u_1(\tau) &= \sum_{i=1}^{\mathcal{M}} q_i \Pr(\text{SINR}^{\text{u}}_1 > \tau | \text{Typical user has task } \mathcal{T}_i) \\ &= \sum_{i=1}^{\mathcal{M}} q_i \Big\{ \Pr(\mathcal{K}=s) \big[q_i \mathcal{P}^u_{s,1}(\tau) \\ &\quad + (1-q_i)\mathcal{P}^u_{s,1}(\tau)\mathcal{P}^u_b(\tau) \big] + \Pr(\mathcal{K}=m)\mathcal{P}^u_{m,1}(\tau) \Big\}, \end{aligned} \tag{15}$$

in which

$$\begin{aligned}\mathcal{P}^u_{k,1}(\tau) &= \Pr(\text{SINR}^u_{a,1} > \tau | \mathcal{K}=k, k \in \{s,m\}) \\ &= \mathbb{E}\left[\exp\left(-\frac{\tau(\sigma^2 + p_u I_{u,1})}{p_u R_{a,1}^{\alpha(\epsilon-1)}} \right) | \mathcal{K}=k \right] \\ &= \mathbb{E}\Bigg[\mathcal{L}_{u,1|\mathcal{K}=k}\left(R_{a,1}^{\alpha(1-\epsilon)} \tau \right) \\ &\quad \times \exp\left(-\frac{R_{a,1}^{\alpha(1-\epsilon)} \tau \sigma^2}{p_u} \right) | \mathcal{K}=k \Bigg], \end{aligned} \tag{16}$$

Authorized licensed use limited to: Maynooth University Library. Downloaded on September 18,2023 at 15:07:55 UTC from IEEE Xplore. Restrictions apply.

and

$$\begin{aligned}\mathcal{P}_b^u(\tau) &= \Pr(\mathrm{SINR}_b^u > \tau)\\ &= \mathbb{E}\left[\exp\left(-\frac{\tau(\sigma^2+p_sI_s)}{p_sR_b^{-\alpha}}\right)\right]\\ &= \mathbb{E}\left[\mathcal{L}_s\left(R_2^{\alpha}\tau\right)\times\exp\left(\frac{-R_2^{\alpha}\tau\sigma^2}{p_s}\right)\right],\end{aligned} \tag{17}$$

where $\mathcal{L}_s$ and $\mathcal{L}_{u,1|\mathcal{K}=k}$ are the Laplace transform of I_s and $I_{u,1}$ conditional on tier k being the serving tier, respectively. In the following, we use $k=1$ or $k=2$ instead of $k=m$ or $k=s^1_{\text{ofld}}$ for simplification.

In order to derive the uplink coverage probability, we first need to characterize the path loss distribution for the desired link of the typical MU and the Laplace transform of the interference. The distribution of the path loss between a typical MU and its serving BS is given in the Lemma below.

Lemma 2. *The probability density function (PDF) of path loss at MU to its serving BS (i.e., $L_{a,1}=R_{a,1}^{\alpha}$) is given by*

$$f_{L_{a,1}|\mathcal{K}=k}(l) = \frac{ba_k}{\mathcal{A}_k}l^{b-1}\exp(-G_kl^b), \tag{18}$$

where $b=\frac{2}{\alpha}$, $a_k=\lambda_k\pi$, $G_k=\sum_{j=1}^{2}a_j(p_jB_j/p_kB_k)^b$, and $\mathcal{A}_k$ is defined in Lemma 1.

Proof. The complementary cumulative distribution function (CCDF) of $L_{a,1}$ conditioned on serving tier being k is

$$\begin{aligned}\Pr(L_{a,1}>l|\mathcal{K}=k) &= \frac{\Pr(D_k^{\alpha}>l, p_kB_kD_k^{-\alpha}>p_jB_jD_j^{\alpha})}{\mathcal{A}_k}\\ &= \frac{\mathbb{E}_{D_k^{\alpha}>l}\left[\exp\left(-\pi\lambda_j\left(\frac{p_jB_j}{p_kB_k}\right)^{2/\alpha}D_k^2\right)\right]}{\mathcal{A}_k}\\ &= \frac{ba_k}{\mathcal{A}_k}\int_l^{\infty}x^{b-1}\exp\left(-x^bG_k\right)\mathrm{d}x.\end{aligned} \tag{19}$$

Taking the derivative of the CCDF leads to the final expression in Lemma 2. □

Lemma 2 indicates that the path loss distribution for the MU to its desired BS is related to the BS density, transmit power, path loss exponent, and bias factor. In case of $\alpha=1$, Lemma 2 will be the PDF of the distance between the typical MU and its serving BS, which is same as a result proved in [44, Lemma 3].

Similarly, we can obtain the PDF the path loss between the SBS and its serving MBS, i.e., $L_b=R_b^{\alpha}$ as

$$f_{L_b}(l) = ba_1l^{b-1}\exp(-a_1l^b). \tag{20}$$

The path loss distribution of the interfering MU at a typical BS is the conditional distribution given that the interfering MU is not associated with the tagged BS. Therefore, the distribution of the path loss between an interfering MU and the tagged BS is not identical to the distribution given in Lemma 2. This correlation is formalized in the following.

The PDF of the path loss of an interfering MU located at U served by a tier j BS, conditioned on it not lying in the association cell of the tagged tier k BS located at Y, is given by

$$\begin{aligned}&f_{L_U}(l|\mathcal{K}_U=j, Y\in\text{tier } k, U\notin\mathcal{C}_Y, \|U-Y\|^{\alpha}=y)\\ &=\frac{bG_j}{1-\exp(-G_ky^b)}l^{b-1}\exp(-G_jl^b),\quad 0\le l\le\frac{p_jB_j}{p_kB_k}y.\end{aligned} \tag{21}$$

Since different MUs associated with the same BS transmit on different resource blocks, there is only one MU from each cell can act as interferer to other cellular cells. Thereby $\Phi_{u,1}^a$ is not a PPP but a Poisson-Voronoi perturbed lattice. For tractability purpose, we adopt the method to approximate $\Phi_{u,1}^a$ as an inhomogeneous PPP [45]. In the following, we will characterize the Laplace transform of the interference.

Remark 1. Suppose a typical MU is located at X, and a tier k BS is located at Y, the probability that this MU is associated with the base station is $\exp(-G_k\|X-Y\|^{\alpha})$. Then, the intensity measure function for interference from MUs associated with tier j BS ($\Phi_{u_j,1}^a$) can be written as

$$\lambda_{u,1|k}^j = ba_jx^{b-1}[1-\exp(-G_k\|X-Y\|^{\alpha})](\mathrm{d}x). \tag{22}$$

.

Lemma 3. *We assume that the point process of interfering MUs from each tier are independent, and the path loss distributions of interfering MUs are independent. In that case, the Laplace transform of MU interference is given by*

$$\begin{aligned}\mathcal{L}_{u,1|\mathcal{K}=k}(s) = \exp\Bigg(&-\frac{bs}{1-b}\sum_{j=1}^{2}\left(\frac{p_jB_j}{p_kB_k}\right)^{1-b}a_j\\ &\times\mathbb{E}_{L_a|\mathcal{K}=j}\left[L_a^{b-(1-\epsilon)}C_b\left(\frac{sp_jB_j}{p_kB_kL_a^{1-\epsilon}}\right)\right]\Bigg),\end{aligned} \tag{23}$$

where $C_b(x) = {}_2F_1(1,1-b;2-b;-x)$ and ${}_2F_1(\cdot,\cdot;\cdot;\cdot)$ is the hypergeometric function.

Proof. We know that $\mathcal{L}_{u,1|\mathcal{K}=k}(s)=\prod_{j=1}^{2}\mathcal{L}_{u,1|\mathcal{K}=k}^j(s)$. The Laplace transform of the interference at the typical tier k BS located at $\mathbf{B}$ from MU associated with tier j BSs is given by,

$$\mathcal{L}_{u,1|\mathcal{K}=k}^j(s) = \mathbb{E}\left[\sum_{X\in\Phi_{u_j,1}^a}(R_X^{\alpha})^{\epsilon}H_X\|X-\mathbf{B}\|^{-\alpha}\right]. \tag{24}$$

Since $\{H_X\}$ are i.i.d exponential random variables, (25) can be rewritten as

$$\mathcal{L}_{u,1|\mathcal{K}=k}^j(s) = \mathbb{E}\left[\prod_{X\in\Phi_{u_j,1}^a}\frac{1}{1+sL_X^{\epsilon}\|X-B\|^{-\alpha}}\right]. \tag{25}$$

By using Remark 1 and probability generating functional (PGFL), we can deuce (26) into

Authorized licensed use limited to: Maynooth University Library. Downloaded on September 18,2023 at 15:07:55 UTC from IEEE Xplore. Restrictions apply.

$$\mathcal{L}^{j}_{u,1|\mathcal{K}=k}(s) = \mathbb{E}\left[\prod_{X\in\Phi^{a}_{u_j,1}} \mathbb{E}_{L_X}\left[\frac{1}{1+sL^{\epsilon}_X X^{-1}}\right]\right] = \exp\left(-\int_{x>0}\mathbb{E}_{L_a}\left[\frac{1}{1+(sL^{\epsilon})^{-1}x}|\mathcal{K}_X=j, L<\frac{p_jB_j}{p_kB_k}x\right]\times\lambda^{j}_{u,k}\mathrm{d}x\right). \quad (26)$$

By employing a change of variable $t=\left(\frac{xP_jB_j}{P_kB_kL}\right)^{b}$ and the definition of hypergeometric function, we can rewrite (25) as

$$\mathcal{L}^{j}_{u,1|\mathcal{K}=k}(s) = \exp\left(-\mathbb{E}_{L_a|\mathcal{K}=j}\left[a_j\left(\frac{p_kB_kL_a}{p_jB_j}\right)^{b}\times\int_1^{\infty}\frac{\mathrm{d}t}{1+L_a^{(1-\epsilon)}t^{1/b}p_kB_k/(p_jB_js)}\right]\right) = \exp\left(-\frac{bs}{1-b}\left(\frac{p_jB_j}{p_kB_k}\right)^{1-b}a_j\times\mathbb{E}_{L_a|\mathcal{K}=j}\left[L_a^{b-(1-\epsilon)}C_b\left(\frac{sp_jB_j}{p_kB_kL_a^{1-\epsilon}}\right)\right]\right). \quad (27)$$

□

Lemma 4. *Laplace transform of SBSs interference is*

$$\mathcal{L}_s(s)=\exp\left(-2\pi\lambda_{SI}\int_0^{\infty}\frac{[1-\exp(-a_mx^2)]x}{1+x^{\alpha}/s}\mathrm{d}x\right), \quad (28)$$

where $\lambda_{SI}=\min(\lambda_m,\lambda_sP_{\mathrm{actv}})$ is density of active SBSs that might potentially interfere in the uplink.

Proof. The proof is similar to that of [46, Theorem 1] and hence omitted here. □

By using Lemma 4, (16) and (17) can be recast in the following reduced forms

$$\mathcal{P}^{u}_{s}(\tau)=\frac{ba_2}{\mathcal{A}_2}\int_0^{\infty}x^{b-1}\exp\left(-\frac{b\tau x^{(1-\epsilon)}}{1-b}\sum_{j=1}^{2}\left(\frac{p_jB_j}{p_sB_s}\right)^{1-b}\times a_j\mathbb{E}_{L_a|\mathcal{K}=j}\left[L_a^{b-(1-\epsilon)}C_b\left(\frac{\tau x^{(1-\epsilon)}p_jB_j}{p_sB_sL_a^{1-\epsilon}}\right)\right]-\frac{\tau x^{(1-\epsilon)}\sigma^2}{p_u}-G_2x^{b}\right)\mathrm{d}x, \quad (29)$$

$$\mathcal{P}^{u}_{m}(\tau)=\frac{ba_1}{\mathcal{A}_1}\int_0^{\infty}x^{b-1}\exp\left(-\frac{b\tau x^{(1-\epsilon)}}{1-b}\sum_{j=1}^{2}\left(\frac{p_jB_j}{p_mB_m}\right)^{1-b}\times a_j\mathbb{E}_{L_a|\mathcal{K}=j}\left[L_a^{b-(1-\epsilon)}C_b\left(\frac{\tau x^{(1-\epsilon)}p_jB_j}{p_mB_mL_a^{1-\epsilon}}\right)\right]-\frac{\tau x^{(1-\epsilon)}\sigma^2}{p_u}-G_1x^{b}\right)\mathrm{d}x, \quad (30)$$

$$\mathcal{P}^{u}_{b}(\tau)=ba_1\int_0^{\infty}x^{b-1}\exp\left(-\frac{\tau x\sigma^2}{p_s}-a_1x^{b}-2\pi\lambda_{SI}\int_0^{\infty}\frac{[1-\exp(-a_1y^2)]y}{1+x\tau y^{\alpha}}\mathrm{d}y\right)\mathrm{d}x. \quad (31)$$

By substituting (29), (30), and (31) into (15), we can derive the analytical expression of the uplink coverage probability.

4.2 Downlink Coverage Analysis

For the downlink transmission, the SINR at the typical MU associated with BS located at X for the access link is given by

$$\mathrm{SINR}^{d}_{a,1}=\frac{p_XH_dR_{a,1}^{-\alpha}}{\sigma^2+I_{d,1}}, \quad (32)$$

with

$$I_{d,1}=\sum_{\mathbf{x}\in\Phi^{a}_{d,1}}p_xH_x\|\mathbf{x}\|^{-\alpha}, \quad (33)$$

where $\Phi^{a}_{d,1}$ denotes the point process of SBSs and MBSs using the same resource block as the associated BS.

The SINR at the typical SBS for the downlink backhaul is given by

$$\mathrm{SINR}^{d}_{b}=\frac{p_mH_mR_b^{-\alpha}}{\sigma^2+I_{s'}}, \quad (34)$$

with

$$I_{s'}=\sum_{\mathbf{x}\in\Phi_b^{S_m}}p_mH_x\|\mathbf{x}\|^{-\alpha}, \quad (35)$$

where Φ_b^{m} denotes the MBSs use the same resource block as the typical SBS serving MBS.

The downlink coverage probability is given by

$$\mathcal{P}^{d}_{1}(\tau)=\sum_{i=1}^{\mathcal{M}}q_i\Pr(\mathrm{SINR}^{d}_{1}>\tau|\text{Typical user has task }\mathcal{T}_i)=\sum_{i=1}^{\mathcal{M}}q_i\left\{\Pr(\mathcal{K}=m)\mathcal{P}^{d}_{m,1}(\tau)+\Pr(\mathcal{K}=s)\left[q_i\mathcal{P}^{d}_{s,1}(\tau)+(1-q_i)\mathcal{P}^{d}_{s,1}(\tau)\mathcal{P}^{d}_{b}(\tau)\right]\right\}, \quad (36)$$

in which

$$\mathcal{P}^{d}_{k,1}(\tau)=\Pr(\mathrm{SINR}^{d}_{a,1}>\tau|\mathcal{K}=k,k\in\{s,m\})=\mathbb{E}\left[\exp\left(-\frac{\tau(\sigma^2+I_{d,1})}{p_kR_a^{-\alpha}}\right)|\mathcal{K}=k\right]=\mathbb{E}\left[\mathcal{L}_{d,1|\mathcal{K}=k}\left(\frac{R_a^{\alpha}\tau}{p_k}\right)\times\exp\left(-\frac{R_{a,1}^{\alpha}\tau\sigma^2}{p_k}\right)|\mathcal{K}=k\right], \quad (37)$$

and

$$\mathcal{P}^{d}_{b}(\tau)=\Pr(\mathrm{SINR}^{d}_{b}>\tau)=\mathbb{E}\left[\exp\left(-\frac{\tau(\sigma^2+I'_s)}{p_mR_b^{-\alpha}}\right)\right]=\mathbb{E}\left[\mathcal{L}_{s'}\left(\frac{R_b^{\alpha}\tau}{p_m}\right)\times\exp\left(-\frac{R_b^{\alpha}\tau\sigma^2}{p_m}\right)\right]. \quad (38)$$

Authorized licensed use limited to: Maynooth University Library. Downloaded on September 18,2023 at 15:07:55 UTC from IEEE Xplore. Restrictions apply.

Lemma 5. *The Laplace transform of downlink access link interference ($I_{d,1}$) when the serving BS belongs to tier k and the corresponding path loss is l can be expressed as*

$$\mathcal{L}_{d,1|L_a=l}(s) = \exp\left(-\frac{bs}{1-b}\sum_{j=1}^{2} p_j\left(\frac{p_j B_j l}{p_k B_k}\right)^{b-1} a_j C_b\left(\frac{s p_k B_k}{l B_j}\right)\right). \quad (39)$$

The Laplace transform of downlink backhaul link interference ($I_{s'}$) when the corresponding path loss is l can be expressed as

$$\mathcal{L}_{s'|L_b=l}(s) = \exp\left(-\frac{2s}{b-2} p_m a_m l^{b-1} C_b\left(\frac{s p_m}{l}\right)\right). \quad (40)$$

Proof. The proof is similar to that of [44, Theorem 1] and hence omitted here. □

By using Lemma 5, (37) and (38) can be written in the following reduced forms

$$\mathcal{P}^d_{s,1}(\tau) = \frac{b a_2}{\mathcal{A}_2}\int_0^\infty x^{b-1}\exp\left(-\frac{bx^b\tau}{(1-b)}\sum_{j=1}^{2}\left(\frac{B_j}{B_s}\right)^{b-1}\left(\frac{p_j}{p_s}\right)^b \times a_j C_b\left(\frac{B_s\tau}{B_j}\right) - \frac{\tau x\sigma^2}{p_s} - G_2 x^b\right)\mathrm{d}x, \quad (41)$$

$$\mathcal{P}^d_{m,1}(\tau) = \frac{b a_1}{\mathcal{A}_1}\int_0^\infty x^{b-1}\exp\left(-\frac{bx^b\tau}{(1-b)}\sum_{j=1}^{2}\left(\frac{B_j}{B_m}\right)^{b-1}\left(\frac{p_j}{p_m}\right)^b \times a_j C_b\left(\frac{B_m\tau}{B_j}\right) - \frac{\tau x\sigma^2}{p_m} - G_1 x^b\right)\mathrm{d}x, \quad (42)$$

$$\mathcal{P}^d_b(\tau) = b a_1\int_0^\infty x^{b-1}\exp\left(-\frac{2a_1\tau x^b}{\alpha-2}C_b(\tau) - \frac{\tau x\sigma^2}{p_m} - a_1 x^b\right)\mathrm{d}x. \quad (43)$$

By substituting (41), (42), and (43) into (36), we can derive the analytical expression of the downlink coverage probability.

4.3 Average Delay

Clearly, the desired average delay can be expressed as the sum of the transmission time and the average response time. In the following, we will first derive the arrival rate at the MEC servers in order to obtain the average response time. Then, the expressions of the average transmission time for the task offloading and result downloading are obtained based on the coverage probability.

Definition 2 (Response Time). *The response time is the time that a computation task spends in server, i.e., the sum of waiting time and service time. $T^r_{m_i,1}$ and $T^r_{s_i,1}$ denote the response time of MBS server and type-i SBS server, respectively.*

The condition for successful offloading is that SINR exceeds a given threshold γ depending on the coding rate [40]. For the $M/G/1$ queuing model applied at the MBS server, the arrivals of type-i computing tasks come from both the SBS servers and type-i MUs. In particular, the probability for a typical SBS to offload the type-i task to the MBS is given by

$$P^i_{\text{arrvl}} = \mathbb{E}_{N^u_s}\left[(1-q_i)\sum_{k=1}^{N^u_s} q_i^k (1-q_i)^{N^u_s-k}\right]. \quad (44)$$

As a result, the arrival rate for arrivals of type-i tasks at the MBS is determined as

$$\Lambda^1_{m,i} = \frac{q_i\lambda_u\mathcal{A}_m\mathcal{P}^u_{m,1}(\gamma)}{\lambda_m} + \frac{\lambda_s P^i_{\text{arrvl}}\mathcal{P}^u_b(\gamma)}{\lambda_m}. \quad (45)$$

The PDF of the service time at MBS for Strategy I is given by

$$f_{T^1_{sv,m}}(t) = \sum_{i=1}^{\mathcal{M}} \mu^i_m \exp(-\mu^i_m t)\hat{\Lambda}^1_{m,i}, \quad (46)$$

where $\hat{\Lambda}^1_{m,i} = \frac{\Lambda^1_{m,i}}{\sum_{j=1}^{\mathcal{M}}\Lambda^1_{m,j}}$.

Using the Pollaczek-Khinchin mean formula [47], the average response time for the type-i task at the MBS can be expressed as

$$T^r_{m_i,1} = \frac{1}{\mu^i_m} + \frac{\left(\sum_{j=1}^{\mathcal{M}}\Lambda^1_{m,j}\right)\times\left(\sum_{j=1}^{\mathcal{M}}\frac{2\hat{\Lambda}^1_{m,j}}{(\mu^j_m)^2}\right)}{2\left[1-\left(\sum_{j=1}^{\mathcal{M}}\Lambda^1_{m,j}\right)\times\left(\sum_{j=1}^{\mathcal{M}}\frac{\hat{\Lambda}^1_{m,j}}{\mu^j_m}\right)\right]}. \quad (47)$$

According to [47], the expected response time at type-i SBS is given by

$$T^r_{s_i,1} = \frac{\bar{Q}_i}{\Lambda^1_{s,i}(1-\sigma_i)}, \quad (48)$$

where $\bar{Q}_i = \frac{\rho_i\left(1-(\mathcal{N}+1)\rho_i^{\mathcal{N}}+\mathcal{N}\rho_i^{\mathcal{N}+1}\right)}{(1-\rho)\left(1-\rho_i^{\mathcal{N}+1}\right)}$, $\sigma_i = \frac{\rho_i^{\mathcal{N}}}{\sum_{n=0}^{\mathcal{N}}\rho_i^n}$, $\rho_i = \Lambda^1_{s,i}/\mu^i_s$, and $\Lambda^1_{s,i} = q_i\mathcal{A}_s\lambda_u\mathcal{P}^u_{s,1}(\gamma)/\lambda_{s^1_{\text{ofld}}}$.

Assumption 1. *For simplicity of analysis, we assume the load (i.e., the number of MUs associated with tagged base station) is equal to its mean.*

Recalling that η_a is the fraction of bandwidth reserved for access links, then for a given η_a, the bandwidth obtained by a MU in the access link W^a or by the SBS in the backhaul link W^b is

$$\begin{cases} W^a_{m,1} = \dfrac{\eta_a W}{N^u_m}, & \text{MU connected to MBS,} \\ W^a_{s,1} = \dfrac{\eta_a W}{N^u_s}, & \text{MU connected to SBS,} \\ W^b = \dfrac{(1-\eta_a)W}{N^s_m}, & \text{MU connected to SBS.} \end{cases} \quad (49)$$

According to (49), for given γ, the average uplink transmission rate can be expressed as

$$\begin{cases} R^u_{m,1} = \mathbb{E}[W^a_{m,1}\log_2(1+\gamma)\mathbb{I}(\text{SINR}^u_{m,1}\geq\gamma)] \\ \quad = \mathcal{P}^u_{m,1}(\gamma)W^a_{m,1}\log_2(1+\gamma), \quad \text{MU-MBS,} \\ R^u_{s,1} = \mathbb{E}[W^a_{s,1}\log_2(1+\gamma)\mathbb{I}(\text{SINR}^a_{s,1}\geq\gamma] \\ \quad = \mathcal{P}^u_{s,1}(\tau)W^a_{s,1}\log_2(1+\gamma), \quad \text{MU-SBS,} \\ R^u_b = \mathbb{E}[W^b\log_2(1+\gamma)\mathbb{I}(\text{SINR}^u_b\geq\gamma)] \\ \quad = \mathcal{P}^u_b(\gamma)W_b\log_2(1+\gamma), \quad \text{SBS-MBS,} \end{cases} \quad (50)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function.

Authorized licensed use limited to: Maynooth University Library. Downloaded on September 18,2023 at 15:07:55 UTC from IEEE Xplore. Restrictions apply.

Therefore, for the type-i task with input data size $\mathcal{D}_i^{ul}$, the average uplink transmission time is

$$\begin{cases} T_{m_i,1}^{t,u} = \frac{\mathcal{D}_i^{ul}}{R_{m,1}^u}, & \text{From MU to MBS,} \\ T_{s_i,1}^{t,u} = \frac{\mathcal{D}_i^{ul}}{R_{s,1}^u}, & \text{From MU to SBS,} \\ T_{b_i}^{t,u} = \frac{\mathcal{D}_i^{ul}}{\min(R_{s,1}^u, R_b^u)}, & \text{From MU to MBS with the aid of SBS.} \end{cases} \tag{51}$$

Similarly, the average downlink transmission rate can be expressed as

$$\begin{cases} R_{m,1}^d = \mathbb{E}[W_{m,1}^a \log_2(1+\gamma)\mathbb{I}(\text{SINR}_{m,1}^d \geq \gamma)] \\ \quad = \mathcal{P}_{m,1}^d(\gamma) W_{m,1}^a \log_2(1+\gamma), \quad \text{MBS-MU,} \\ R_{s,1}^d = \mathbb{E}[W_{s,1}^a \log_2(1+\gamma)\mathbb{I}(\text{SINR}_{s,1}^d \geq \gamma)] \\ \quad = \mathcal{P}_{s,1}^d(\gamma) W_{s,1}^a \log_2(1+\gamma), \quad \text{MU-SBS,} \\ R_b^d = \mathbb{E}[W^b \log_2(1+\gamma)\mathbb{I}(\text{SINR}_b^d \geq \gamma)] \\ \quad = \mathcal{P}_b^d(\gamma) W_b \log_2(1+\gamma), \quad \text{SBS-MBS.} \end{cases} \tag{52}$$

The downlink transmission time for the type-i task is given by,

$$\begin{cases} T_{m_i,1}^{t,d} = \frac{\mathcal{D}_i^{dl}}{R_{m,1}^d}, & \text{From MBS to MU,} \\ T_{s_i,1}^{t,d} = \frac{\mathcal{D}_i^{dl}}{R_{s,1}^d}, & \text{From SBS to MU,} \\ T_{b_i}^{t,d} = \frac{\mathcal{D}_i^{dl}}{\min(R_{s,1}^d, R_b^d)}, & \text{From MBS to MU with the aid of SBS.} \end{cases} \tag{53}$$

The average delay for a MU with $\mathcal{T}_i$ is given by

$$\begin{aligned} T_{\text{avg},1}^i(\gamma) &= \mathbb{E}\left[T_{\text{respon.}} + T_{\text{trans.}}\right] \\ &= \mathcal{A}_m\left[T_{m_i,1}^r + T_{m_i,1}^{t,d}(\gamma) + T_{m_i,1}^{t,u}(\gamma)\right] \\ &\quad + \mathcal{A}_s\Big\{q_i\left[T_{s_i,1}^r + T_{s_i,1}^{t,u}(\gamma) + T_{s_i,1}^{t,d}(\gamma)\right] \\ &\quad + (1-q_i)\left[T_{m_i,1}^r + T_{b_i}^{t,u}(\gamma) + T_{b_i}^{t,d}(\gamma)\right]\Big\}. \end{aligned} \tag{54}$$

5 Strategy II

For Strategy II, the typical MU only offloads its task to the MBS or SBS via the access link directly. Hence, there are no wireless backhaul links between MBSs and SBSs. In this section, we use λ_{s_i} to denote the density of the SBSs that have cached the service database $\mathcal{S}_i$ and λ_{u_i} to denote the density of the MUs with task $\mathcal{T}_i$, where $\sum_{i=1}^{\mathcal{M}} \lambda_{s_i} = \lambda_s$ and $\sum_{i=1}^{\mathcal{M}} \lambda_{u_i} = \lambda_u$.

The average number of MUs connected to a SBS with $\mathcal{S}_i$ is

$$\mathbb{E}[N_{s_i}^u] = 1 + \frac{1.28\lambda_{u_i}\mathcal{A}_{s_i}}{\lambda_{s_i}}, \tag{55}$$

in which

$$\mathcal{A}_{s_i} = \frac{\lambda_{s_i}}{\lambda_{s_i} + \lambda_m (p_m B_m / p_s B_s)^{2/\alpha}}. \tag{56}$$

By using (6), we can obtain P_{ofld}^i in Strategy II. The sets of BSs can be regraded as $\mathcal{M}+1$ tiers, where tier 0 denotes the MBSs and tier i $(1 \leq i \leq \mathcal{M})$ denotes the SBSs which have cached $\mathcal{S}_i$. In the following, we use $\lambda_0 = \lambda_m$ and $\lambda_i = \lambda_{s_i} \times P_{\text{ofld}}^i$ to represent the intensity of point processes for MBSs and offloadable type-i SBSs, respectively.

5.1 Uplink Coverage Analysis

If the typical MU with $\mathcal{T}_i$ is connected to the nearest offloadable type-i SBS or MBS located at $\mathbf{B}$, denoting the distance between the user and $\mathbf{B}$ as R, the SINR at the tagged SBS/MBS is given by

$$\text{SINR}_{a,2}^u = \frac{p_u H_u R^{\alpha(\epsilon-1)}}{\sigma^2 + p_u I_{u,2}}, \tag{57}$$

with

$$I_{u,2} = \sum_{\mathbf{x}_i \in \Phi_a^{u,2}} (R_{x_i}^{\alpha})^{\epsilon} H_i \|\mathbf{x}_i - \mathbf{B}\|^{-\alpha}. \tag{58}$$

The uplink coverage probability is given by

$$\begin{aligned} \mathcal{P}_2^u(\tau) &= \sum_{i=1}^{\mathcal{M}} q_i \Pr(\text{SINR}_2^u > \tau | \text{Typical MU with task } \mathcal{T}_i) \\ &= \sum_{i=1}^{\mathcal{M}} q_i \Big[\Pr(\mathcal{K}=i)\mathcal{P}_{s_i}^u(\tau) + \Pr(\mathcal{K}=0)\mathcal{P}_{0_i}^u(\tau)\Big], \end{aligned} \tag{59}$$

in which

$$\begin{aligned} \mathcal{P}_{s_i}^u(\tau) &= \Pr(\text{SINR}_2^u > \tau | \mathcal{K}=i) \\ &= \mathbb{E}\left[\exp\left(-\frac{\tau(\sigma^2 p_u + p_u I_{u,2})}{p_u R^{\alpha(\epsilon-1)}}\right) | \mathcal{K}=i\right] \\ &= \mathbb{E}\Bigg[\mathcal{L}_{u,2|\mathcal{K}=i}\left(R^{\alpha(1-\epsilon)}\tau\right) \\ &\quad \times \exp\left(-\frac{R^{\alpha(1-\epsilon)}\tau\sigma^2}{p_u}\right) | \mathcal{K}=i\Bigg], \end{aligned} \tag{60}$$

and

$$\begin{aligned} \mathcal{P}_{0_i}^u(\tau) &= \Pr(\text{SINR}_2^u > \tau | \mathcal{K}=0, \text{MU with } \mathcal{T}_i) \\ &= \mathbb{E}\left[\exp\left(-\frac{\tau(\sigma^2 + p_u I_{u,2})}{p_u R^{\alpha(\epsilon-1)}}\right) | \mathcal{K}=0, \text{MU with } \mathcal{T}_i\right] \\ &= \mathbb{E}\Bigg[\mathcal{L}_{u,2|\mathcal{K}=0}\left(R^{\alpha(1-\epsilon)}\tau\right) \\ &\quad \times \exp\left(-\frac{R^{\alpha(1-\epsilon)}\tau\sigma^2}{p_u}\right) | \mathcal{K}=0, \text{ MU with } \mathcal{T}_i\Bigg]. \end{aligned} \tag{61}$$

The PDF of $L_i = \{R^{\alpha} | \mathcal{K}=i, i > 0\}$ is

$$f_{L_i}(l) = \frac{b a_i}{\mathcal{A}_{s_i}} l^{b-1} \exp(-G_i l^b), \tag{62}$$

where $\forall\ 1 \leq i \leq \mathcal{M}$, $a_i = P_{\text{ofld}}^i \lambda_i \pi$, and $G_i = a_i + a_0\left(\frac{p_m B_m}{p_s B_s}\right)^b$.

The PDF of path loss of MU with task $\mathcal{T}_i$ to MBS is

$$f_{L_{0_i}}(l) = b G_{0_i} l^{b-1} \exp(-G_{0i} l^b), \tag{63}$$

where $G_{0_i} = a_0 + a_j\left(\frac{p_s B_s}{p_m B_m}\right)^b$.

Authorized licensed use limited to: Maynooth University Library. Downloaded on September 18,2023 at 15:07:55 UTC from IEEE Xplore. Restrictions apply.

Corollary 1. *The PDF of the path loss of a MU with task $\mathcal{T}_k$ located at $\boldsymbol{X}$ associated with tier j, conditioned on it not lying in the serving cell of the tagged tier i BS located at B and the corresponding path loss $\|\mathbf{X}-\mathbf{B}\|^{\alpha}=y$, is*

$$f_{L_X}(l|\mathcal{K}_X=j, X\notin\mathcal{C}_B, B\in\text{tier } i, \|\mathbf{X}-\mathbf{B}\|^{\alpha}=y) = \begin{cases} \frac{bG_j}{1-\exp(-G_{0j}y^b)}l^{b-1}\exp(-G_jl^b), \\ \qquad 0\le l\le \frac{p_sB_s}{p_mB_m}y, i=0, j\neq 0, \\ \frac{bG_{0k}}{1-\exp(-G_{0k}y^b)}l^{b-1}\exp(-G_{0k}l^b), \\ \qquad 0\le l\le y, i=j=0, \\ \frac{bG_j}{1-\exp(-G_jy^b)}l^{b-1}\exp(-G_jl^b), \\ \qquad 0\le l\le y,\ i\neq 0, j=i, \\ bG_jl^{b-1}\exp(-G_jl^b), \quad i\neq 0,\ i\neq j, j\neq 0, \\ \frac{bG_{0i}}{1-\exp(-G_iy^b)}l^{b-1}\exp(-G_{0i}l^b), \\ \qquad 0\le l\le \frac{p_mB_m}{p_sB_s}y, i\neq 0,\ j=0, k=i, \\ bG_{0k}l^{b-1}\exp(-G_{0k}l^b), \quad i\neq 0,\ j=0, k\neq i. \end{cases} \tag{64}$$

Proof. 1) For $i=0$ and $j\neq 0$, given the association policy, L_X should be bounded by $\frac{P_s}{P_m}\|\mathbf{X}-\mathbf{B}\|^{\alpha}$. With $G_{0j}\times\left(\frac{P_sB_s}{P_mB_m}\right)^b=G_j$, we can get the result for the first case in the above distribution. 2) Similarly, for $i=j=0$, $\forall k$, we can obtain the result for the second case. 3) If $i\neq 0$ and $j=i$, with the fact that the path loss of an interfering user to its serving BS has to be smaller than its distance to the tagged BS, the distribution is given for the third case. 4) As for $i\neq 0$, $i\neq j$, as well as $j\neq 0$, the user associated with tier j will always be interfering at tagged BS in tier i. Thus, there is no bound for L_X. 5) When $i\neq 0$, $j=0$, and $k=i$, L_X has an upper bound. 6) However, if $k\neq j$, the user will always be the interference at the tagged BS. Thereby, there is no upper bound in the sixth case. □

Remark 2. According to our association policy, conditioned on a BS of tier i located at V, a MU with task $\mathcal{T}_j$ at U associates with V with probability

$$\begin{cases} \exp(-G_{0i}\|\mathbf{U}-\mathbf{V}\|^{\alpha}) & ,i=0, \\ \exp(-G_i\|\mathbf{U}-\mathbf{V}\|^{\alpha}) & ,i=j, \\ 0 & ,i\neq j. \end{cases} \tag{65}$$

Assumption 2. *Conditioned on the tagged BS located at B and of tier i, the propagation process of interfering MUs from tier j ($j\ge 1, i.e. type-jSBS$) to B, $\mathcal{I}_{u,j}:=\{\|\mathbf{X}-\mathbf{B}\|^{\alpha}\}_{X\in\Phi^b_{u,j}}$ with intensity measure function*

$$\lambda_{u,j}(\mathrm{d}x)=\begin{cases} ba_jx^{b-1}\big(1-\exp\big(-G_{0j}x^b\big)\big)(\mathrm{d}x), & i=0, j\neq 0, \\ ba_jx^{b-1}\big(1-\exp\big(-G_jx^b\big)\big)(\mathrm{d}x), & i\neq 0, j=i, \\ ba_jx^{b-1}(\mathrm{d}x), & i\neq 0, i\neq j. \end{cases} \tag{66}$$

If $i=j=0$, the propagation process of interference comes from the MUs with $\mathcal{T}_k$, $\mathcal{I}_0^{u_k}:=\{\|\mathbf{X}-\mathbf{B}\|^{\alpha}\}_{X\in\Phi^b_{u_k,0}}$ with intensity measure function

$$\lambda_0^{u_k}(\mathrm{d}x)=\hat{q}_kba_0x^{b-1}\big(1-\exp(-G_{0k}x^b)\big). \tag{67}$$

If $i\neq 0$ and $j=0$, the propagation process of interference comes from the MUs with $\mathcal{T}_k$, $\mathcal{I}_i^{u_k}:=\{\|\mathbf{X}-\mathbf{B}\|^{\alpha}\}_{X\in\Phi^b_{u_k,0}}$ with intensity measure function

$$\lambda_i^{u_k}(\mathrm{d}x)=\begin{cases} {}_kba_0x^{b-1}\big(1-\exp\big(-G_kx^b\big)\big)(\mathrm{d}x), & k=i, \\ \hat{q}_kba_0x^{b-1}, & k\neq i. \end{cases} \tag{68}$$

In a typical MBS cell, MUs use different resource blocks. If we randomly choose one MU from a MBS cell, the probability that the chosen MU has task $\mathcal{T}_k$ is $\hat{q}_k=\frac{q_k\lambda_u\mathcal{A}_{0k}}{\sum_{n=1}^{\mathcal{M}}\lambda_uq_n\mathcal{A}_{0n}}$, where $\mathcal{A}_{0k}=\frac{\lambda_m}{\lambda_m+\lambda_{s_k}(p_sB_s/p_mB_m)^{2/\alpha}}$ denotes the probability that the typical MU with task $\mathcal{T}_k$ associates with MBS.

Lemma 6. *Laplace transform of MUs interference at the tagged MBS is $\mathcal{L}_{u,2|\mathcal{K}=0}(s)$ is*

$$\mathcal{L}_{u,2|\mathcal{K}=0}(s)=\prod_{j=0}^{\mathcal{M}}\mathcal{L}_{u,2|0,j}(s), \tag{69}$$

where $\mathcal{L}_{u,2|0,j}(s)$ denotes the Laplace transform of interference from tier j MU

$$\mathcal{L}_{u,2|0,j}(s) \tag{70}$$

$$=\begin{cases} \exp\Bigg(-\frac{bs}{1-b}\sum_{k=1}^{\mathcal{M}}\hat{q}_ka_0 \\ \qquad\times\mathbb{E}_{L_{0k}}\left[L_{0k}^{b-(1-\epsilon)}C_b\left(\frac{s}{L_{0k}^{1-\epsilon}}\right)\right]\Bigg),\ j=0, \\ \exp\Bigg(-\frac{bs}{1-b}a_j\left(\frac{p_sB_s}{p_mB_m}\right)^{1-b} \\ \qquad\times\mathbb{E}_{L_j}\left[L_j^{b-(1-\epsilon)}C_b\left(\frac{sp_s}{L_j^{1-\epsilon}p_m}\right)\right]\Bigg),\ j\neq 0. \end{cases} \tag{71}$$

Proof. Let $\mathcal{L}_{I_0^{u_k}}(s)$ denote the interference from type-k MUs associated with MBSs, we can have

$$\mathcal{L}_{u,2|0,0}(s)=\prod_{k=1}^{\mathcal{M}}\mathcal{L}_{I_0^{u_k}}(s). \tag{72}$$

By replacing the intensity measure function and PDF of path loss in the the derivation of (24) with results given in Assumption 2 and (63), the expression of $\mathcal{L}_{I_0^{u_k}}(s)$ can be written as

$$\exp\left(-\frac{bs}{1-b}\hat{q}_ka_0\mathbb{E}_{L_{0k}}\left[L_{0k}^{b-(1-\epsilon)}C_b\left(\frac{s}{L_{0k}^{1-\epsilon}}\right)\right]\right). \tag{73}$$

This proof is concluded by substituting (73) into (72).□

Authorized licensed use limited to: Maynooth University Library. Downloaded on September 18,2023 at 15:07:55 UTC from IEEE Xplore. Restrictions apply.

By substituting (70) into (61), we can have

$$\mathcal{P}_{0i}^{u}(\tau)=bG_{0_i}\int_0^{\infty}x^{b-1}\exp\Bigg(-\frac{x^{1-\epsilon}\tau\sigma^2}{p_u}-G_{0i}x^b+\hat{q}_j a_0\mathbb{E}_{L_{0j}}\left[L_{0j}^{b-(1-\epsilon)}C_b\left(\frac{x^{1-\epsilon}\tau}{L_{0j}^{1-\epsilon}}\right)\right]-\frac{bx^{1-\epsilon}\tau}{1-b}\sum_{j=1}^{\mathcal{M}}a_j\left(\frac{p_sB_s}{p_mB_m}\right)^{1-b}\times\mathbb{E}_{L_j}\left[L_j^{b-(1-\epsilon)}C_b\left(\frac{x^{1-\epsilon}\tau p_sB_s}{L_j^{1-\epsilon}p_mB_m}\right)\right]\Bigg)\mathrm{d}x. \quad (74)$$

Lemma 7. *Laplace transform of MUs interference at the tagged type i SBS is given in (78), as shown at the bottom of the page.*

Proof. Let $\mathcal{L}_{u,2|i,j}(s)$ denote the interference from the MU associated with tier j BS at the typical type-i SBS, we can have

$$\mathcal{L}_{u,2|\mathcal{K}=i}(s)=\prod_{j=0}^{\mathcal{M}}\mathcal{L}_{u,2|i,j}(s). \quad (75)$$

Let $\mathcal{L}_{I_i^{u_k}}(s)$ denote the interference from the type-k MU associated with MBS at the type-i SBS, then

$$\mathcal{L}_{u,2|i,0}(s)=\prod_{k=1}^{\mathcal{M}}\mathcal{L}_{I_i^{u_k}}(s). \quad (76)$$

By using Assumption 2 and following by the derivation of (24), the expression of $\mathcal{L}_{I_i^{u_k}}(s)$ is given by

$$\mathcal{L}_{I_i^{u_k}}(s)=\begin{cases}\exp\Bigg(-\frac{bs}{1-b}\frac{q_k\mathcal{A}_{0k}}{\sum_{n=1}^{\mathcal{M}}q_n\mathcal{A}_{0n}}\left(\frac{p_mB_m}{p_sB_s}\right)^{1-b}a_0\times\mathbb{E}_{L_{0k}}\left[L_{0k}^{b-(1-\epsilon)}C_b\left(\frac{p_ms}{L_{0k}^{1-\epsilon}p_s}\right)\right]\Bigg), & k=i,\\ \exp\left(-\frac{bs}{1-b}\hat{q}_ka_0\mathbb{E}_{L_{0k}}\left[L_{0k}^{b-(1-\epsilon)}C_b\left(\frac{s}{L_{0k}^{1-\epsilon}}\right)\right]\right), & k\neq i.\end{cases}$$

Similarly, the interference from MUs associated with type-k SBSs at the typical type-i SBS is given by

$$\mathcal{L}_{u,2|i,k}(s)=\exp\left(-\frac{bs}{1-b}a_k\mathbb{E}_{L_k}\left[L_k^{b-(1-\epsilon)}C_b\left(\frac{s}{L_k^{1-\epsilon}}\right)\right]\right). \quad (77)$$

This proof is concluded by substituting (76) and (77) into (75). □

By using (78) and (62), the analytical expression of $\mathcal{P}_{s_i}^{u}(\tau)$ is given by

$$\mathcal{P}_{s_i}^{u}(\tau)=bG_i\int_0^{\infty}x^{b-1}\exp\Big(-x^{1-\epsilon}\tau\sigma^2/p_u-G_ix^b+\ln\big[\mathcal{L}_{u,2|\mathcal{K}=i}\big(x^{1-\epsilon}\tau\big)\big]\Big)\mathrm{d}x. \quad (79)$$

5.2 Downlink Coverage Analysis

For the downlink transmission, the SINR at the typical MU associated with BS located at X for the access link is given by

$$\mathrm{SINR}_2^d=\frac{p_XH_dR_X^{-\alpha}}{\sigma^2+I_{d,2}}, \quad (80)$$

with

$$I_{d,2}=\sum_{\mathbf{x}\in\Phi_{d,2}^a}p_xH_x\|\mathbf{x}\|^{-\alpha}, \quad (81)$$

where $\Phi_{d,2}^a$ denotes the point process of SBSs and MBSs using the same resource block as the serving BS of the typical MU.

The downlink coverage probability in Strategy II is given by

$$\mathcal{P}_2^d(\tau)=\sum_{i=1}^{\mathcal{M}}q_i\Pr(\mathrm{SINR}_2^d>\tau|\text{Typical MU with task }\mathcal{T}_i)=\sum_{i=1}^{\mathcal{M}}q_i\Big[\Pr(\mathcal{K}=i)\mathcal{P}_{s_i}^d(\tau)+\Pr(\mathcal{K}=0)\mathcal{P}_{0_i}^d(\tau)\Big], \quad (82)$$

in which

$$\mathcal{P}_{0_i}^d(\tau)=\Pr(\mathrm{SINR}_2^u>\tau|\mathcal{K}=0,\text{Typical MU with task }\mathcal{T}_i)=\mathbb{E}_{L_{0i}}\left[\mathcal{L}_{d,2}^{0i}\left(\frac{L_{0i}\tau}{p_m}\right)\times\exp\left(-\frac{L_{0i}\tau\sigma^2}{p_m}\right)\right], \quad (83)$$

and

$$\mathcal{P}_{s_i}^d(\tau)=\Pr(\mathrm{SINR}_2^u>\tau|\mathcal{K}=i,i>0)=\mathbb{E}_{L_i}\Big[\mathcal{L}_{d,2}^{i}\left(\frac{L_i\tau}{p_s}\right)\times\exp\left(\frac{-L_i\tau\sigma^2}{p_s}\right)\Big]. \quad (84)$$

Lemma 8. *The Laplace transform of downlink interference $I_{d,2}$ at the type-k MU when the serving BS belongs to MBS and type-k SBS are given respectively by*

$$\mathcal{L}_{u,2|\mathcal{K}=i}(s)=\exp\Bigg\{-\frac{bs}{1-b}\Bigg(\hat{q}_i\left(\frac{p_mB_m}{p_sB_s}\right)^{1-b}a_0\mathbb{E}_{L_{0i}}\left[L_{0i}^{b-(1-\epsilon)}C_b\left(\frac{sp_mB_m}{L_{0i}^{1-\epsilon}p_sB_s}\right)\right]-\sum_{k=1,k\neq i}^{\mathcal{M}}\hat{q}_ka_0\mathbb{E}_{L_{0k}}\left[L_{0k}^{b-(1-\epsilon)}C_b\left(\frac{s}{L_{0k}^{1-\epsilon}}\right)\right]-\sum_{j=1}^{\mathcal{M}}a_j\mathbb{E}_{L_j}\left[L_j^{b-(1-\epsilon)}C_b\left(\frac{s}{L_j^{1-\epsilon}}\right)\right]\Bigg)\Bigg\}. \quad (78)$$

Authorized licensed use limited to: Maynooth University Library. Downloaded on September 18,2023 at 15:07:55 UTC from IEEE Xplore. Restrictions apply.

TABLE 2
Simulation Parameters

Parameter	Value	Parameter	Value
W	60 Mhz	B_m	1
p_u	23 dBm	$\mathcal{N}$	5
p_m	43 dBm	p_s	33 dBm
γ	0 dB	α	4
λ_u	$100/km^2$	η	0.6
$\mathcal{D}_i^{ul}$	0.5−0.6 MB	$\mathcal{D}_i^{dl}$	0.3−0.4 MB
F_m	6 Ghz	F_s	2.5 Ghz

$$\mathcal{L}_{d,2|L_{0k}=l}^{0k}(s) = \exp\Bigg\{ -\frac{bsl^{b-1}}{1-b}\Bigg[p_s\left(\frac{B_s p_s}{B_m p_m}\right)^{b-1} a_k C_b\left(\frac{sB_m p_m}{lB_s}\right) + p_m a_0 C_b\left(\frac{sp_m}{l}\right)\Bigg] - 2\pi \sum_{j=1,j\neq k}^{\mathcal{M}} \lambda_j g(sp_s)\Bigg\}, \tag{85}$$

and

$$\mathcal{L}_{d,2|L_k=l}^{k}(s) = \exp\Bigg\{ -\frac{bsl^{b-1}}{1-b}\Bigg[p_m\left(\frac{B_m p_m}{B_s p_s}\right)^{b-1} a_0 C_b\left(\frac{sB_s p_s}{lB_m}\right) + p_s a_k C_b\left(\frac{sp_s}{l}\right)\Bigg] - 2\pi \sum_{j=1,j\neq k}^{\mathcal{M}} \lambda_j g(sp_s)\Bigg\}, \tag{86}$$

where $g(s) = \int_0^\infty \frac{sx}{s+x^\alpha}\mathrm{d}x = \frac{\Gamma\left(\frac{2}{\alpha}\right)\Gamma\left(1+\frac{2}{\alpha}\right)s^{2/\alpha}}{\alpha\Gamma\left(1+\frac{2}{\alpha}\right)}$ *(cf. [48, Eq. (4.11)]).*

Proof. Since all of the tier j SBS ($j \neq k$) contribute to interference at the type-k MU, the Laplace transform of interference results from tier j SBS is easily to be obtained by PGFL. The final result is the product of Laplace transform of SBS-tier and MBS-tier. □

Making use of Lemma 8, we can deuce (84) and (85) into

$$\mathcal{P}_{0i}^d(\tau) = bG_{0_i}\int_0^\infty x^{b-1}\exp\Bigg(-\frac{x\tau\sigma^2}{p_m} - G_{0i}x^b + \ln\left[\mathcal{L}_{d,2|L_{0i}=x}^{0i}\left(\frac{x\tau}{p_m}\right)\right]\Bigg)\mathrm{d}x, \tag{87}$$

and

$$\mathcal{P}_{s_i}^d(\tau) = bG_i\int_0^\infty x^{b-1}\exp\Bigg(-\frac{x\tau\sigma^2}{p_s} - G_i x^b + \ln\left[\mathcal{L}_{d,2|L_i=x}^{i}\left(\frac{x\tau}{p_s}\right)\right]\Bigg)\mathrm{d}x. \tag{88}$$

The downlink coverage probability can be derived by substituting (87) and (88) into (82).

5.3 Average Delay

Without the consideration of backhaul link, the total bandwidth W is allocated to the typical MU based the number of MUs associated its serving BS. The bandwidth obtained by a typical MU is given by

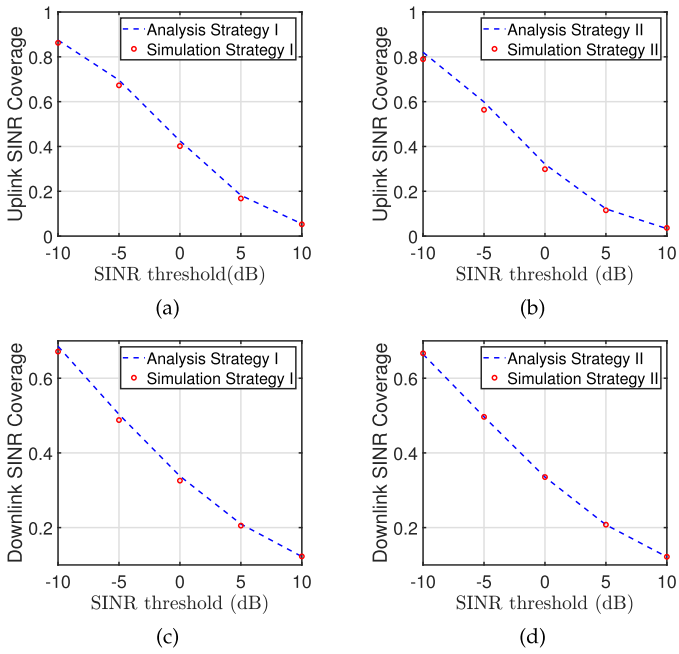


Fig. 4. Validation of uplink and downlink coverage probability.

$$\begin{cases} W_{m,2}^a = \frac{W}{N_m^u}, & \text{MU connected to MBS,} \\ W_{s_i}^a = \frac{W}{N_{s_i}^u}, & \text{MU connected to type-}i\text{ SBS.} \end{cases} \tag{89}$$

The average uplink and downlink transmission time for the type-i task are given by

$$\begin{cases} T_{0i}^{t,u} = \frac{\mathcal{D}_i^{ul}}{\mathcal{P}_{0i}^u(\gamma)W_{m,2}^a\log_2(1+\gamma)}, & \text{Type-}i\text{ MU-MBS,} \\ T_{s_i}^{t,u} = \frac{\mathcal{D}_i^{ul}}{\mathcal{P}_{s_i}^u(\gamma)W_{s_i}^a\log_2(1+\gamma)}, & \text{Type-}i\text{ MU-SBS,} \end{cases} \tag{90}$$

and

$$\begin{cases} T_{0i} = \frac{\mathcal{D}_i^{dl}}{\mathcal{P}_{0i}^d(\gamma)W_{m,2}^a\log_2(1+\gamma)}, & \text{Type-}i\text{ MU-MBS,} \\ T_{s_i}^{t,d} = \frac{\mathcal{D}_i^{dl}}{\mathcal{P}_{s_i}^d(\gamma)W_{s_i}^a\log_2(1+\gamma)}, & \text{Type-}i\text{ MU-SBS.} \end{cases} \tag{91}$$

For Strategy II, the arrival rate of type-i tasks at the MBS and SBS are respectively determined as

$$\Lambda_{m,i}^2 = \frac{\mathcal{A}_{0i}q_i\lambda_u\mathcal{P}_{0i}^u(\gamma)}{\lambda_m}, \tag{92}$$

and

$$\Lambda_{s,i}^2 = \frac{\mathcal{A}_{s_i}\lambda_u q_i\mathcal{P}_{s_i}^u(\gamma)}{\lambda_m}. \tag{93}$$

By replacing $\Lambda_{m,i}^1$ in (48) with $\Lambda_{m,i}^2$ and $\Lambda_{s,i}^1$ in (49) with $\Lambda_{s,i}^2$, we can derive the the average response time for the type-i task at MBS and SBS which are denoted by $T_{m_i,2}^r$ and $T_{s_i,2}^r$.

Thus, the average delay for the typical MU with T_i in Strategy II is given by

$$T_{\text{avg},2}^i(\gamma) = \mathcal{A}_{0i}\left[T_{m_i,2}^r + T_{0i}^{t,u}(\gamma) + T_{0i}^{t,d}(\gamma)\right] \tag{94}$$

$$+ \mathcal{A}_{s_i}\left[T_{s_i,2}^r + T_{s_i}^{t,u}(\gamma) + T_{s_i}^{t,d}(\gamma)\right]. \tag{95}$$

Authorized licensed use limited to: Maynooth University Library. Downloaded on September 18,2023 at 15:07:55 UTC from IEEE Xplore. Restrictions apply.

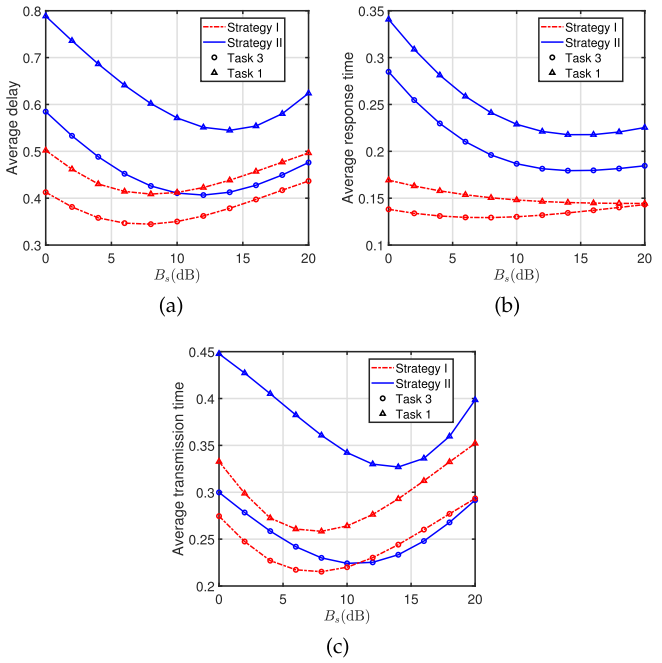


Fig. 5. Effect of bias factor.

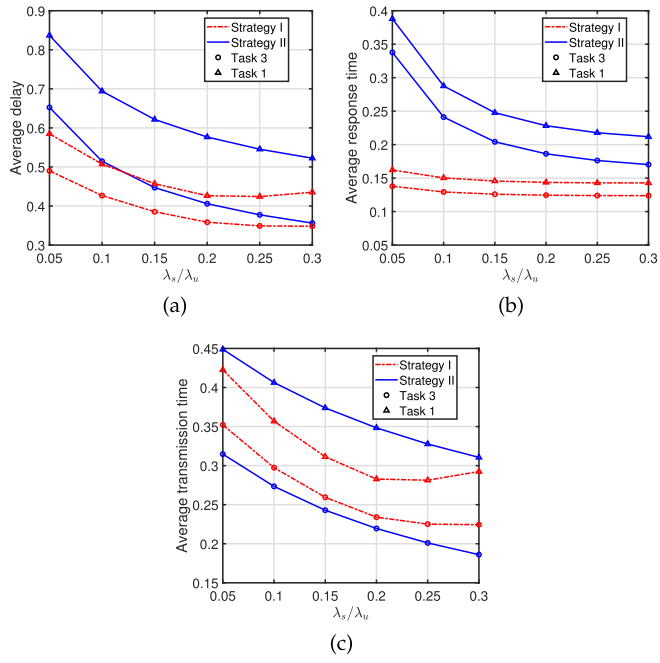


Fig. 7. Effect of SBS density.

6 Simulation Results

In this section, we provide numerical results for the MEC-enabled 2-tier HetNet in the area of 1 km^2. $\mathcal{L}_i$ is ranging from 400 to 700 Megacycles and α is 4. The remaining simulation parameters are given in Table 2 according to [40], [41], [42].

We consider that there is three types of services (i.e., $\{\mathcal{T}_i|i=1,2,3\}$), and corresponding popularity probabilities are 0.2, 0.3, and 0.5, respectively (except for Fig. 4 that considers two types of services). The other parameters are introduced when the corresponding figures are discussed.

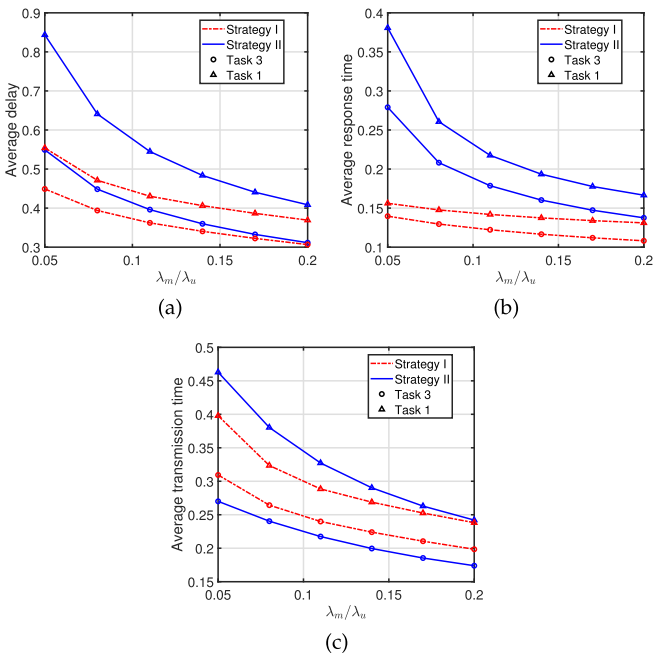


Fig. 6. Effect of MBS density.

6.1 Verification of Accuracy

In this section, we present Monte Carlo simulation to validate our analytical expressions for coverage probability. In each case, we perform at least 10^5 times of the realizations of the channel, the positions of communication nodes, the types of services cached at the BSs, and the required types of services for MUs. For each realization, we compare the SINR threshold with the simulated SINR which is induced by (11), (13), (32), or (34). We consider that there are two types of services with popularity probability 0.3 and 0.7. The density of MBSs and SBSs are $10/km^2$ and $20/km^2$, respectively. In Fig. 4, we plot the uplink and the downlink coverage probability for Strategy I and Strategy II. The results show the accuracy of the derived analytical expressions for the coverage probabilities, which are perfectly matching with simulations.

6.2 Effect of Bias Factor

Fig. 5 depicts how the bias factors in cell association affect the performance of the considered system via numerical results. The average delay first decrease as the B_s increase, then increases with the increasing of B_s. This is justifiable since the load at the MBS servers is heavy when B_s is small. As B_s increases, more computation tasks will be offloaded to the less loaded SBS servers, which are located closer to MUs. However, when B_s exceeds a certain level, the SBS servers become heavily-loaded, and the incremental offloading tasks at the SBS servers will suffer a longer delay. For a given type of tasks, that the average delay in Strategy I is always shorter than Strategy II. This is reasonable since the IAB setting in Strategy I enables computation load migration from SBSs to MBSs, which is beneficial to reduce the average response time.

Authorized licensed use limited to: Maynooth University Library. Downloaded on September 18,2023 at 15:07:55 UTC from IEEE Xplore. Restrictions apply.

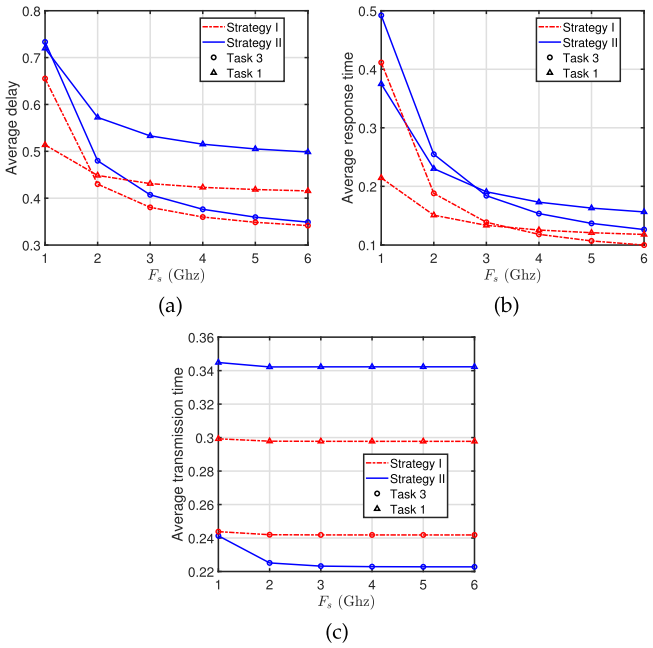


Fig. 8. Effect of SBS service rate.

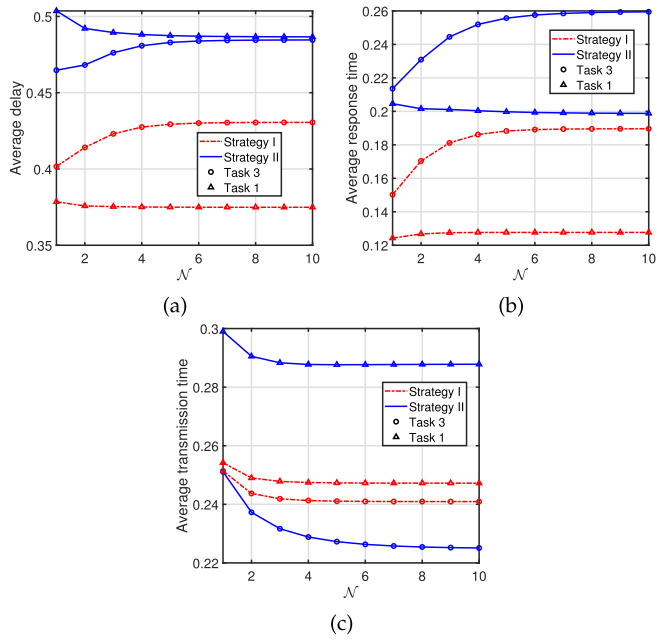


Fig. 9. Effect of SBS buffer size.

6.3 Effect of BS Density

Figs. 6 and 7 show the impacts of the SBS and the MBS density on the system performance, respectively. It is shown that the average delay is a monotonically decreasing function of MBS density in all considered cases. Compared with the SBS density, we can observe that increasing MBS density can achieve a more remarkable performance improvement. Due to the support of computation offloading from SBSs to MBSs, Figs. 7b and 6b indicate that the average response time in Strategy I are much smaller than Strategy II and less sensitive to the BS density. However, as shown in Figs. 7a and 7c, when λ_s/λ_u exceeds a certain level, the heavy-loaded backhaul link will degrade the average transmission time as well as the average delay for the MU with task 1, since the MU with task 1 has a higher probability to utilize the more powerful computation capacity at the MBS via the backhaul link, As for Strategy II, since there is no backhaul link bottleneck, the average delay is a monotonically decreasing function of λ_s/λ_u.

6.4 Effect of SBS Computing Capacity

Fig. 8 demonstrates the system performance versus the SBS service rate. The average delay can be regarded as a monotonically decreasing function of F_s in all considered cases. When F_s is small, the average response time for the type-3 MU is larger than that of the type-1 MU in both considered strategies, since the computation load at the type-1 SBS is smaller than the load at the type-3 SBS and P_{ofld} is smaller than 1. After certain values of F_s, the average response time of the type-3 MU starts to be smaller than type-1 MU. This is reasonable since a large F_s will lead to $P_{\text{ofld}} = 1$, which in turn causes the decrease of computation load at both SBSs and MBSs.

6.5 Effect of SBS Buffer Size

Fig. 9a shows the changes of average delay versus the buffer size at SBSs. In all considered strategies, before the average delay getting converged, the average delay for MUs with task 1 and task 3 is decreasing and increasing with the increasing of $\mathcal{N}$, respectively. Moreover, the average delay in Strategy I is smaller than than that of Strategy II for both types of MUs. With increasing $\mathcal{N}$, the typical MU will experience longer waiting time at its serving SBS. Meanwhile, P_{ofld} increases as $\mathcal{N}$ increases, which can alleviate the computing load at both MBS-tier and SBS-tier. Since the type-1 MU has a larger probability to associate with the MBS-tier and the response time at MBS-tier is smaller than type-1 SBS, the average response time decreases as $\mathcal{N}$ increases. On the contrary, the average response time for the type-3 MU increases with the increasing of $\mathcal{N}$. Since the computing load at the type-1 SBS in Strategy I is much lighter, $\mathcal{N}$ has a negligible impact on average response time for type-1 MUs in Strategy I.

7 Conclusion

In this paper, we formulated a two-tier MEC HetNet spatial model, which consisted of the multi-type MUs with the request of different service types and the two-tier MEC servers with different computing capacities. Due to the limited resource, SBS-tier MEC servers can only cache a specific type of service and their computing buffer is finite. We studied two strategies corresponding to two different settings: 1) Strategy I for an IAB-enabled MEC HetNet, and 2) Strategy II for traditional MEC HetNet. By using tools from stochastic geometry and queueing theory, we first derived the analytical expressions for coverage probability, and then analysed the average delay for both two considered strategies. Finally, we discussed and compared the performance of two proposed strategies by extensive simulations.

The current work and results presented in this paper can be extended in several future research directions. In this paper, we considered that each SBS only can cache one

Authorized licensed use limited to: Maynooth University Library. Downloaded on September 18,2023 at 15:07:55 UTC from IEEE Xplore. Restrictions apply.

specific service. However, considering multiple services caching makes the average delay analysis more challenging but is of practical relevance. In addition to delay, energy consumption is also a critical performance metric in MEC networks [13]. Hence, the design of energy-efficient offloading schemes aligned with heterogeneous services will be another promising direction.

REFERENCES

[1] U. Drolia *et al.*, "The case for mobile edge-clouds," in *Proc., IEEE Int. Conf. Ubiquitous Intell. Comput., IEEE 10th Int. Conf. Autonomic Trusted Comput.*, 2013, pp. 209–215.

[2] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint optimization of radio and computational resources for multicell mobile-edge computing," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 1, no. 2, pp. 89–103, Jun. 2015.

[3] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1628–1656, Third Quarter, 2017.

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

[5] H. A. Willebrand and B. S. Ghuman, "Fiber optics without fiber," *IEEE Spectr.*, vol. 38, no. 8, pp. 40–45, Aug. 2001.

[6] O. Teyeb, A. Muhammad, G. Mildh, E. Dahlman, F. Barac, and B. Makki, "Integrated access backhauled networks," in *Proc. IEEE 90th Veh. Technol. Conf.*, 2019, pp. 1–5.

[7] C. Dehos, J. L. Gonzalez, A. De Domenico, D. Ktenas, and L. Dussopt, "Millimeter-wave access and backhauling: The solution to the exponential data traffic increase in 5G mobile communications systems?," *IEEE Commun. Mag.*, vol. 52, no. 9, pp. 88–95, Sep. 2014.

[8] H. Guo, J. Liu, J. Zhang, W. Sun, and N. Kato, "Mobile-edge computation offloading for ultradense IoT networks," *IEEE Internet Things J.*, vol. 5, no. 6, pp. 4977–4988, Dec. 2018.

[9] S. Mukherjee and J. Lee, "Edge computing-enabled cell-free massive mimo systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2884–2899, Apr. 2020.

[10] Z. Xu, L. Zhou, S. C.-K. Chau, W. Liang, Q. Xia, and P. Zhou, "Collaborate or separate? Distributed service caching in mobile edge clouds," in *Proc. IEEE Conf. Comput. Commun.*, 2020, pp. 2066–2075.

[11] C. You and K. Huang, "Multiuser resource allocation for mobile-edge computation offloading," in *Proc. IEEE Global Commun. Conf.*, 2016, pp. 1–6.

[12] Q. Zhang, L. Gui, F. Hou, J. Chen, S. Zhu, and F. Tian, "Dynamic task offloading and resource allocation for mobile-edge computing in dense cloud RAN," *IEEE Internet Things J.*, vol. 7, no. 4, pp. 3282–3299, Apr. 2020.

[13] Y. Wu, B. Shi, L. P. Qian, F. Hou, J. Cai, and X. S. Shen, "Energy-efficient multi-task multi-access computation offloading via NOMA transmission for IoTs," *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4811–4822, Jul. 2020.

[14] K. Zhang, Y. Zhu, S. Leng, Y. He, S. Maharjan, and Y. Zhang, "Deep learning empowered task offloading for mobile edge computing in urban informatics," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 7635–7647, Oct. 2019.

[15] H. Guo, J. Liu, J. Ren, and Y. Zhang, "Intelligent task offloading in vehicular edge computing networks," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 126–132, Aug. 2020.

[16] H. Guo and J. Liu, "UAV-enhanced intelligent offloading for Internet of Things at the edge," *IEEE Trans. Ind. Informat.*, vol. 16, no. 4, pp. 2737–2746, Apr. 2020.

[17] Y. Mao, J. Zhang, and K. B. Letaief, "Dynamic computation offloading for mobile-edge computing with energy harvesting devices," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3590–3605, Dec. 2016.

[18] T. Zhao, I.-H. Hou, S. Wang, and K. Chan, "Red/LeD: An asymptotically optimal and scalable online algorithm for service caching at the edge," vol. 36, no. 8, pp. 1857–1870, Aug. 2018.

[19] Y. Liang, J. Ge, S. Zhang, J. Wu, Z. Tang, and B. Luo, "A utility-based optimization framework for edge service entity caching," *IEEE Trans. Parallel Distrib. Syst.*, vol. 30, no. 11, pp. 2384–2395, Nov. 2019.

[20] J. Xu, L. Chen, and P. Zhou, "Joint service caching and task offloading for mobile edge computing in dense networks," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, 2018, pp. 207–215.

[21] Q. Xie, Q. Wang, N. Yu, H. Huang, and X. Jia, "Dynamic service caching in mobile edge networks," in *Proc. IEEE 15th Int. Conf. Mobile Ad Hoc Sensor Syst.*, 2018, pp. 73–79.

[22] K. Poularakis, J. Llorca, A. M. Tulino, I. Taylor, and L. Tassiulas, "Joint service placement and request routing in multi-cell mobile edge computing networks," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, 2019, pp. 10–18.

[23] S. Bi, L. Huang, and Y.-J. A. Zhang, "Joint optimization of service caching placement and computation offloading in mobile edge computing systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4947–4963, Jul. 2020.

[24] Q. Fan, J. Lin, G. Feng, Z. Gao, H. Wang, and Y. Li, "Joint service caching and computation offloading to maximize system profits in mobile edge-cloud computing," in *Proc. Int. Conf. Mobility, Sens. Netw.*, 2020, pp. 244–251.

[25] W. Wen, Y. Cui, T. Q. S. Quek, F.-C. Zheng, and S. Jin, "Joint optimal software caching, computation offloading and communications resource allocation for mobile edge computing," *IEEE Trans. Veh. Technol.*, vol. 69, no. 7, pp. 7879–7894, Jul. 2020.

[26] X. He, R. Jin, and H. Dai, "Joint service placement and resource allocation for multi-UAV collaborative edge computing," in *IEEE Wireless Commun. Netw. Conf.*, 2021, pp. 1–6.

[27] C. Madapatha *et al.*, "Integrated access and backhaul networks: Current status and potentials," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 1374–1389, Sep. 2020.

[28] Y. Zhang, M. A. Kishk, and M.-S. Alouini, "A survey on integrated access and backhaul networks," *Front. Comms. Net.*, vol. 2, Jun. 2021.

[29] J. Y. Lai, W.-H. Wu, and Y. T. Su, "Resource allocation and node placement in multi-hop heterogeneous integrated-access-and-backhaul networks," *IEEE Access*, vol. 8, pp. 122 937–122 958, 2020.

[30] F. Gomez-Cuba and M. Zorzi, "Twice simulated annealing resource allocation for mmWave multi-hop networks with interference," in *Proc. IEEE Int. Conf. Commun.*, 2020, pp. 1–7.

[31] M. Polese, M. Giordani, A. Roy, S. Goyal, D. Castor, and M. Zorzi, "End-to-end simulation of integrated access and backhaul at mmWaves," in *Proc. IEEE 23rd Int. Workshop Comput. Aided Model. Des. Commun. Links Netw.*, 2018, pp. 1–7.

[32] M. Hashemi, M. Coldrey, M. Johansson, and S. Petersson, "Integrated access and backhaul in fixed wireless access systems," in *Proc. IEEE 86th Veh. Technol. Conf.*, 2017, pp. 1–5.

[33] 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.

[34] 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.

[35] C. Saha, M. Afshang, and H. S. Dhillon, "Bandwidth partitioning and downlink analysis in millimeter wave integrated access and backhaul for 5G," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 8195–8210, Dec. 2018.

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

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

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

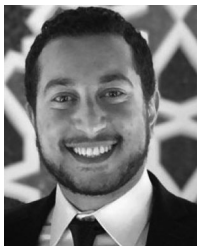
[39] M. Liu, F. R. Yu, Y. Teng, V. C. Leung, and M. Song, "Computation offloading and content caching in wireless blockchain networks with mobile edge computing," *IEEE Trans. Veh. Technol.*, vol. 67, no. 11, pp. 11 008–11 021, Nov. 2018.

[40] S.-W. Ko, K. Han, and K. Huang, "Wireless networks for mobile edge computing: Spatial modeling and latency analysis," *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 5225–5240, Aug. 2018.

Authorized licensed use limited to: Maynooth University Library. Downloaded on September 18,2023 at 15:07:55 UTC from IEEE Xplore. Restrictions apply.

[41] Y. Gu, C. Li, B. Xia, D. Xu, and Z. Chen, "Modeling and performance analysis of stochastic mobile edge computing wireless networks," in *Proc. IEEE 89th Veh. Technol. Conf.*, 2019, pp. 1–5.
[42] C. Park and J. Lee, "Mobile edge computing-enabled heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1038–1051, Feb. 2021.
[43] E. Gilbert, "Random subdivisions of space into crystals," *Annals Math. Statist.*, vol. 33, no. 3, pp. 958–972, 1962.
[44] H.-S. Jo, Y. J. Sang, P. Xia, and J. G. Andrews, "Heterogeneous cellular networks with flexible cell association: A comprehensive downlink SINR analysis," *IEEE Trans. Wireless Commun.*, vol. 11, no. 10, pp. 3484–3495, Oct. 2012.
[45] 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.
[46] J. G. Andrews, F. Baccelli, and R. K. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE Trans. Commun.*, vol. 59, no. 11, pp. 3122–3134, Nov. 2011.
[47] L. Kleinrock, *Queueing Systems, Volume 1: Theory*. New York: NY, USA: Wiley, 1975.
[48] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products*. Oxford, U.K.: Academic, 2007.

Yongqiang Zhang (Graduate Student Member, IEEE) received the BSc degree in communication engineering from the Southwest University, Chongqing, China, in 2019 and the MS degree in electrical and computer engineering in 2021 from the King Abdullah University of Science and Technology, Saudi Arabia where, he is currently working toward the PhD degree. His research interests include performance analysis and optimization of the integrated access and backhaul (IAB) networks.



Mustafa A. Kishk (Member, IEEE) received the BSc and MSc degree from Cairo University in 2013 and 2015, respectively, and the PhD degree from Virginia Tech in 2018. He is currently a postdoctoral research fellow with the Communication Theory Lab, King Abdullah University of Science and Technology. His research interests include stochastic geometry, energy harvesting wireless networks, UAV-enabled communication systems, and satellite communications.

Mohamed-Slim Alouini (Fellow, IEEE) was born in Tunis, Tunisia. He received the PhD degree in electrical engineering from the California Institute of Technology (Caltech), Pasadena, CA, USA, in 1998. He was a faculty member with the University of Minnesota, Minneapolis, MN, USA, then with the Texas A&M University at Qatar, Education City, Doha, Qatar. Before joining King Abdullah University of Science and Technology (KAUST), Thuwal, Makkah Province, Saudi Arabia, he was a professor of electrical engineering in 2009. His research interests include modeling, design, and performance analysis of wireless communication systems.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/csdl.

Authorized licensed use limited to: Maynooth University Library. Downloaded on September 18,2023 at 15:07:55 UTC from IEEE Xplore. Restrictions apply.