Heterogeneous Cellular Networks with Spatio-Temporal Traffic: Delay Analysis and Scheduling

Yi Zhong, Member, IEEE, Tony Q.S. Quek, Senior Member, IEEE and Xiaohu Ge, Senior Member, IEEE

Abstract—Emergence of new types of services has led to various traffic and diverse delay requirements in fifth generation (5G) wireless networks. Meeting diverse delay requirements is one of the most critical goals for the design of 5G wireless networks. Though the delay of point-to-point communications has been well investigated, the delay of multi-point to multi-point communications has not been thoroughly studied since it is a complicated function of all links in the network. In this work, we propose a novel tractable approach to analyze the delay in the heterogeneous cellular networks with spatio-temporal random arrival of traffic. Specifically, we propose the notion of delay outage and evaluated the effect of different scheduling policies on the delay performance. Our numerical analysis reveals that offloading policy based on cell range expansion greatly reduces the macrocell traffic while bringing a small amount of growth for the picocell traffic. Our results also show that the delay performance of round-robin scheduling outperforms first in first out scheduling for heavy traffic, and it is reversed for light traffic. In summary, this analytical framework provides an understanding and a rule-of-thumb for the practical deployment of 5G systems where delay requirement is increasingly becoming a key concern.

Index Terms—Delay, heterogeneous cellular networks, scheduling, spatio-temporal traffic, stochastic geometry

I. INTRODUCTION

A. Motivation

Evolution for high-speed data applications, such as high-quality wireless video streaming, social networking, and machine-to-machine communications, has led to the explosive growth of data traffic demand for 5G wireless networks. It is envisioned that the average capacity requirement of 5G wireless networks might reach 25 Gbps/km², which is 100 times higher compared to the current system capacity. Meanwhile, the delay requirements for different types of traffic become more diversified due to the emergence of new types of applications [1], such as latency-critical applications like command-and-control of drones, advanced manufacturing, and tactile Internet [2], [3]. As wireless networks evolve, the coupling between traffic and network services becomes increasingly strong. The tremendous traffic and its dynamic variations play increasingly crucial roles in affecting the delay performance of 5G wireless networks. Therefore, theoretical analysis of delay in wireless networks is imperative to guide us in our design and deployment.

Though the delay of point-to-point communication has been well investigated using classical queueing theory, the delay of multi-point to multi-point communications, such as multi-party video conference or multi-player online gaming, has not been well studied. This is mainly due to the fact that delay is a complicated function of all links in the network and is affected by a variety of factors such as the traffic, the medium access protocol, the path loss, and so on. In order to effectively analyze traffic and delay in heterogeneous cellular networks [4]–[9], a reasonable model to characterize both the spatial distribution and the temporal variation of traffic should first be established (see Fig. 1). The spatial distribution of traffic could be described by the locations of users. If the mobility of users is not considered, the temporal variation of traffic includes two parts, i.e., the random arrival process of packets for each user and the dynamic serving process of arrived packets.

B. Related Works

Most of the previous works only model one aspect of traffic, i.e., either only model the spatial distribution of users or only model the temporal arrival process of packets. In [10], the
distribution proposed in [11] was used to approximate the probability density function (pdf) of the area for Voronoi cell formed by single-tier macrocell base stations (BSs). With the obtained pdf, by modeling the locations of users as a homogeneous Poisson point process (PPP) [12]–[15], the probability distribution of the number of users in each Voronoi cell was derived. The works [16], [17] extended such pdfs to the case of multi-tier heterogeneous networks. On the other hand, the analysis of temporal variation of traffic were generally based on queueing theory and model the traffic as the function of random arrival of packets. For example, in [18]–[22], a discrete-time slotted ALOHA system with multiple terminals was considered, and each terminal maintains a buffer of infinite capacity to store incoming packets. The statuses (idle or busy) of terminals change over time due to the random arrival and serving of the traffic, and the serving rates also rely on the statuses of terminals since the idle terminals will not cause any interference.

Existing works that simultaneously model the spatio-temporal arrival of traffic, such as [23]–[25], considered the traffic generated at random spatial regions, other than modeling the flow at each independent user. The analysis in these works was based on the granularity of total traffic in each cell, i.e., the traffic generated in the coverage region of each BS is summed up as the traffic arrived at the corresponding BS. For example, the work in [23] modeled the spatio-temporal arrival of users as a homogeneous spatio-temporal PPP, then the traffic arrived at each BS was characterized by the amount of data all users demand in the coverage of associated BS. The main deficiencies of such a model are four aspects. First, since the users are not definitely characterized in these models, many network operations, such as scheduling of users in each cell and offloading traffic between cells cannot be well captured. Meanwhile, the metric for each independent user, such as throughput and delay, cannot be well defined. Second, due to the diversity of traffic, the quality of service (QoS) requirement and the arrival rate are different for different users, which cannot be appropriately modeled. Third, the interference and the interaction between the queues at different BSs are either ignored or only analyzed by approximations. The interacting queues problem [20] is notoriously difficult to cope with since the statuses of queues and the serving rates are highly coupled (see Fig. 2). Lastly, these models are not consistent with those used in the industrial simulations [27], in which the users are generated randomly and the packets arrive independently as a random process at each user.

Previous works using stochastic geometry to analyze the wireless networks focus on performance metrics such as success probability and average achievable rate by considering a snapshot of the network. However, these approaches are incapable to deal with long term metrics like delay, due to the difficulties such as the interacting queues problem, the spatial-temporal random arrival of traffic, and the static property of networks [28], [29]. The works [30]–[33] analyzed the transmission of traffic with delay constraint in heterogeneous cellular networks without considering the coupling between traffic and network services.

In this work, we model the spatio-temporal traffic by combining the tools from stochastic geometry and queueing theory. Specifically, we model the spatial distribution of users by PPP and model the temporal arrival of packets at individual users by independent Bernoulli processes. Then, the data flows for different users could be considered independently. Meanwhile, since both the traffic and the delay requirements are diversified for different users, we model the packet arrival process and the delay requirements for each individual user. The main contributions of this paper can be summarized as follows:

- Several tractable approaches, for example, considering the dominant system and the modified system, are proposed to analyze and bound the statistical distribution of the signal-to-interference ratio (SIR) and the delay in heterogeneous cellular networks based on the combination of stochastic geometry and queueing theory.
- Delay outage (DO) is proposed to characterize whether the delay requirement could be achieved. Furthermore, the effect of scheduling policies, i.e., the random scheduling, the first-input-first-output (FIFO) scheduling, and the round-robin scheduling, on the delay is evaluated, which to our knowledge has not been explored using stochastic geometry due to analytical difficulty.
- By numerical evaluations, statistics of traffic in heterogeneous cellular networks are investigated, and effect of spatial distribution and temporal variation of traffic on the delay is analyzed. Our proposed analytical framework provides a useful guideline for the system design and deployment of 5G wireless systems where the delay requirement becomes the key concern.

Our results reveal that the delay performance of round-robin scheduling outperforms FIFO scheduling for heavy traffic, and it is reversed for light traffic. Moreover, it shows that the offloading policy based on cell range expansion (CRE) greatly reduces the macrocell traffic while bringing a small amount of growth for the picocell traffic.

The remaining part of the paper is organized as follows. Section II describes the spatial distribution model, the arrival process, and the scheduling policies. Sections III and IV evaluate the statistic of traffic, success probability, and mean delay. Section V numerically analyzes the effect of the spatial-temporal traffic on the statistical distribution of mean delay and the delay outage. Finally, Section VI concludes the paper.
II. System Model

We consider the downlink of a multi-tier heterogeneous cellular network with $K$ tiers of BSs. The locations of the $k$th tier ($k = 1, \ldots, K$) BSs are modeled as a homogeneous PPP $\Phi_k$ with intensity $\lambda_k$ BSs per km$^2$, and $\Phi_1, \Phi_2, \ldots, \Phi_K$ are assumed to be independent. Let $\Phi_b$ be the superposition of $\Phi_1, \Phi_2, \ldots, \Phi_K$ as follows:

$$\Phi_b \triangleq \Phi_1 \cup \Phi_2 \cup \ldots \cup \Phi_K,$$

which denotes the locations of all tiers of BSs in the network. The transmit power of the BSs of the $k$th tier is fixed as $P_k$. We assume that the time is slotted into discrete time slots.

The traffic model in the network consists of two parts: the spatial distribution model of the users and the temporal arrival model of the packets at each user. The mobility of the users is not modeled in the following, i.e. the temporal variation of the locations of the users are not considered. However, a variety of mobility models, such as random walk model, random waypoint model, high mobility random walk model [34], could be applied to extend this work. Therefore, our network model considered here is static since the locations of the BSs and the users remain static during all time after they are deployed [35–37]. We assume that the users are distributed as a homogeneous PPP $\Phi_u = \{x_i\}$ with intensity $\lambda_u$ users per m$^2$. As for the temporal arrival of the downlink traffic at BSs, we assume that the packets arrival process for each user $x_i \in \Phi_u$ is an independent Bernoulli process of arrival rate $\xi_i$ (packets per time slot), i.e., $\xi_i$ is the probability of an arrival at user $x_i$ in any given time slot. Without loss of generality, we assume that the size of each packet is fixed and it requires exactly one time slot to transmit a packet. Each BS will maintain an independent queue of infinite size for each user within the coverage of associated BS to save the generated packets; thus the number of queues at a BS equals to the number of users served by the associated BS. Due to diverse latency requirements in future wireless network, we assume that the mean delay requirement for user $x_i$ is $\beta_i$ (in number of time slots), i.e., the mean delay of user $x_i$ should be less than $\beta_i$ time slots. With these notations, a user can be described by a triple $(x_i, \xi_i, \beta_i)$ with $x_i$ being the location, $\xi_i$ being the arrival rate of packets, and $\beta_i$ being the mean delay requirement. In the following, we assume that $\{\xi_i\}$ and $\{\beta_i\}$ are independent identically distributed (i.i.d.) uniform random variables within $[\xi_{\text{min}}, \xi_{\text{max}}]$ and $[\beta_{\text{min}}, \beta_{\text{max}}]$, respectively. 

Therefore, the traffic in the network can be modeled by a marked Poisson process $\tilde{\Phi}_u$ defined as follows:

$$\tilde{\Phi}_u \triangleq \{(x_i, \xi_i, \beta_i)\}.$$  

The propagation loss is modeled by two parts, the path loss and the fading. A deterministic path loss function $l(r) = Ar^{-\alpha}$ is used to model the path loss, where $A > 0$ is the reference path loss and $\alpha > 2$ is the path loss exponent. The fading is assumed to be Rayleigh block fading with unit mean, i.e., the power fading coefficients in different time slots are i.i.d.

Note that our analysis can be extended to the case of general probability distributions for $\xi_i$ and $\beta_i$. are constant during one time slot. We consider the interference-limited regime and ignore the thermal noise. If the signal-to-interference ratio (SIR) at a user is larger than a certain threshold $\theta$, a packet will be successfully decoded. Otherwise, if the acceptance of a packet fails, the packet will be added into the head of the queue and wait to be rescheduled again. This model is based on the Shannon’s theorem declaring that it is possible to communicate digital information nearly error-free up to certain rate determined by the SIR. We summarized several crucial system parameters in Table I.

The random arrival of traffic may greatly affect the performance of the network. In order to improve the overall performance with random arrival of traffic, we consider several mechanisms to manage the traffic. We first consider the CRE in which a bias factor for each tier of BSs is determined for offloading [38]. By setting different bias factors for different tiers of BSs, the coverage region of the small cells can be extended to serve more users. Meanwhile, the number of users served by the macro BSs can then be reduced. Another mechanism is to improve the SIR at the users by mitigating the interference through randomly muting the transmission of the BSs in each time slot. In practical wireless network, in the muted time slots, only control channels and cell-specific reference signals are transmitted with reduced power, and no user data is transmitted. During these time slots, the interference can be greatly reduced. The last mechanism is to introduce the scheduling of users in each cell, i.e., the BS in each cell selects an appropriate user to serve. Several scheduling mechanisms have been considered, such as the random scheduling in which each user is selected to be served randomly with equal probability, the FIFO scheduling in which the packets are served in the manner of first come first serve, and the round-robin scheduling in which the users are scheduled one by one.

As for the user association strategy, we assume that the CRE is employed, where the association is based on the maximum average biased received power. Let $B_k$ be the association bias for the $k$th tier of BSs. Thus, a user located at $x$ will associate the BS $X(x)$ denoted by

$$X(x) = \arg \max_{k \in \{1,2,\ldots,K\}, y \in \Phi_k} B_k |y - x|^{-\alpha}.$$  

By setting $B_k = 1$, $\forall k$, the association strategy is exactly based on the maximum received power. For the BSs from $k$th tier with small transmit power $P_k$, the association bias $B_k$ is set to be larger so that users can be offloaded from the overload macro cells to the light loaded small cells, and the overall interference in the network can also be reduced. With these settings, the spatial tessellation of cells construes a weighted

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K$</td>
<td>Number of tiers</td>
<td>$\Phi_b$</td>
<td>Locations of $k$th tier BSs</td>
</tr>
<tr>
<td>$\lambda_u$</td>
<td>Density of users</td>
<td>$A$</td>
<td>Reference path loss</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>Path loss exponent</td>
<td>$\beta_i$</td>
<td>Delay requirement for $x_i$</td>
</tr>
<tr>
<td>$\theta$</td>
<td>SIR threshold</td>
<td>$\Phi_u$</td>
<td>Locations of the users</td>
</tr>
<tr>
<td>$\xi_i$</td>
<td>Arrival rate for $x_i$</td>
<td>$P_k$</td>
<td>Power of $k$th tier BSs</td>
</tr>
</tbody>
</table>
Voronoi tessellation. Mathematically, the association region of a BS from the $k$th tier located at $X_k$ is given by

$$
\mathcal{C}(X_k) = \left\{ y \in \mathbb{R}^2 : |y - X_k| \leq \left( \frac{B_k P_k}{B_i P_i} \right)^{1/\alpha} |y - z|, \quad \forall i \in \{1, 2, \ldots, K\}, \forall z \in \Phi_t \right\}.
$$

(4)

As for the random muting, we assume that the blank time slots appear randomly at each BS with probability $1 - p$ ($0 \leq p \leq 1$) and independently among different BSs. With these assumptions, in each time slot, each BS will suppress its transmission with probability $1 - p$, thus postponing its own transmission but reducing its interference to the network.

Since there are multiple users served by each BS and only one user in a cell can be scheduled for transmission in each time slot, we consider three scheduling strategies as follows:

1) **Random Scheduling:** In random scheduling, each active BS will randomly choose one user within the coverage region of that BS to serve in each time slot. Since there are multiple queues at each BS corresponding to multiple users served by that BS, each queue will be scheduled with the same probability in each time slot. Due to the retransmission mechanism, if a packet fails for transmission in the scheduled time slot, the packet will be added into the head of the queue of the corresponding user and wait to be rescheduled again.

2) **FIFO Scheduling:** In FIFO scheduling, the packets arrived at each BS are served by the manner of FIFO, i.e., a packet arrived first will be scheduled first. The result of FIFO scheduling only relies on the time when the packets arrive at a BS and is irrelevant to the particular user a packet belonging to. Therefore, all queues at a BS can be considered as a large queue, a packet arrived first will be served first.

3) **Round-robin Scheduling:** In round-robin scheduling, the users served by each BS are scheduled one by one. For example, if there are $N$ users served by a BS, $N$ time slots will form a complete cycle in that cell, and the scheduling of each user will occupy one of the $N$ time slots. Similarly, retransmission mechanism is introduced for the failed packets. Note that each BS may suppress its transmission in a time slot with probability $1 - p$ due to the random muting, the transmission of the scheduled user in that time slot will be suppressed.

Though several simple assumptions are used to make the analysis feasible, the analytical framework proposed in this work could be easily extended to model the particular scenarios and detailed technologies in wireless networks.

### III. Traffic Statistic

In this section, we evaluate the traffic statistic to characterize the spatio-temporal traffic, which is characterized by the spatial distribution, the temporal arrival rate, and the delay requirements for the users. In order to characterize the traffic, we first introduce the following lemma.

**Lemma 1.** The probability that the typical user associates to the BSs of the $k$th tier is given by

$$
P_k = \frac{\lambda_k (P_k B_k)^{2/\alpha}}{\sum_{i=1}^{K} \lambda_i (P_i B_i)^{2/\alpha}}.
$$

(5)

**Proof:** The proof is similar to Lemma 1 in [10], and we present a simplified version here for completeness. Let $Z_t$ be the distance between the typical user at the origin and the nearest BS in the $i$th tier of BSs. The pdf of $Z_t$, given by

$$
f_{Z_t}(z) = 2\pi \lambda_z \exp(-\pi \lambda_z z^2),
$$

(6)

can be obtained by the void probability of the PPP in a ball. The probability that the typical user associates to the BSs of the $k$th tier is

$$
P_k = \mathbb{P} \left( \bigcap_{i=1, i \neq k}^{K} \{P_k B_k Z_k^{-\alpha} > P_i B_i Z_i^{-\alpha}\} \right)
$$

(7)

where $(a)$ follows from the independence of $\{\Phi_k\}$. Plugging in the pdf of $Z_t$, we obtain $(5)$. With the association probability given by Lemma 1, we obtain the following lemma.

**Lemma 2.** The pdf of the coverage area $S_k$ of a BS of the $k$th tier is approximated as

$$
f_{S_k}(x) \simeq \frac{343}{15} \sqrt{\frac{3.5}{\pi}} x^{5/3} \exp \left( -\frac{3.5 x \lambda_k}{P_k} \right) \frac{\lambda_k}{P_k}.
$$

(8)

**Proof:** The pdf of the area $S$ of a Voronoi cell for a PPP with intensity $\lambda$ is [10, 11]

$$
f_S(x) \simeq \frac{343}{15} \sqrt{\frac{3.5}{\pi}} x^{5/3} \exp \left( -\frac{3.5 x \lambda}{\lambda} \right).
$$

(9)

Due to the ergodicity of the PPP, since $P_k$ is the probability that a user associates to the BSs of the $k$th tier, the average fraction of total area covered by the BSs of the $k$th tier is $P_k$. Therefore, the average coverage area of a BS of the $k$th tier is $P_k / \lambda_k$. By replacing $\lambda$ with $\lambda_k / P_k$ in (9), we get the approximated result (8).

The following lemma gives the statistics for different kinds of users in each cell.

**Lemma 3.** Let $N_{k,\xi,\beta}$ be the number of users served by a BS of the $k$th tier with arrival rates less than $\xi \in [\xi_{\min}, \xi_{\max}]$ and with delay requirements less than $\beta \in [\beta_{\min}, \beta_{\max}]$. The probability generating functions (PGFs) of $N_{k,\xi,\beta}$ are given by

$$
G_{N_{k,\xi,\beta}}(z) \simeq (1 + C_0 - C_0 z)^{-\frac{1}{2}}.
$$

(10)

The probability mass functions (PMF) of $N_{k,\xi,\beta}$, denoted by $\mathbb{P}(N_{k,\xi,\beta} = n) = \frac{1}{n!} G_{N_{k,\xi,\beta}}^{(n)}(0)$, is given by

$$
\mathbb{P}(N_{k,\xi,\beta} = n) = \frac{(2n + 5)!!}{15 \cdot n!} \left( \frac{C_0}{2} \right)^n (C_0 + 1)^{-n-\frac{1}{2}}.
$$

(11)

where $n!!$ is the double factorial and

$$
C_0 = \frac{(\beta - \beta_{\min})(\xi - \xi_{\min})}{(\beta_{\max} - \beta_{\min})(\xi_{\max} - \xi_{\min})} \frac{P_k}{\lambda_k}.
$$

(12)

**Proof:** Since the spatial distribution of users construes a PPP with intensity $\lambda_u$, given the area $S_k$ of a cell, the number of users within the cell with arrival rate less than $\xi$ and with delay requirement less than $\beta$ is a Poisson random variable.
while that served by each picocell does not increase much, the number of users served by each macrocell is greatly reduced than the picocells because they provide larger coverage. Figure 1 also indicates that in the case where some users are offloaded (with mean $\xi = \xi_{\text{max}}$) the macrocells will serve more users as the bias factor for macrocells is increased and the picocells are set as $B_1 = 1$ and $B_2 = 0$. The parameters are set as $\alpha = 2.5$ and $\lambda_u = 0.0001m^{-2}$. Therefore, we get the PMF as follows

$$G_{N_k,\xi,\beta}(z) = \mathbb{E}[z^{N_k,\xi,\beta}],$$

with mean $\mathbb{E}[N_k,\xi,\beta]$. Thus, the PGF of $N_k,\xi,\beta$, $G_{N_k,\xi,\beta}(z) = \mathbb{E}[z^{N_k,\xi,\beta}]$, is

$$G_{N_k,\xi,\beta}(z) = \mathbb{E}[z^{N_k,\xi,\beta}] = \int_0^\infty \frac{\lambda_k}{1 - \frac{\lambda_k}{\beta_{\text{max}} - \beta_{\text{min}}}} \left( \frac{\lambda_k}{\beta_{\text{max}} - \beta_{\text{min}}} \right) x^{2.5} \exp \left( -\frac{3.5\lambda_k}{\beta_{\text{max}} - \beta_{\text{min}}} x - \frac{(\beta_{\text{max}} - \beta_{\text{min}})(\xi - \xi_{\text{min}})\lambda_u (z - 1)x}{\beta_{\text{max}} - \beta_{\text{min}}(\xi_{\text{max}} - \xi_{\text{min}})} \right) dx$$

$$= \left( 1 - \frac{(\beta_{\text{max}} - \beta_{\text{min}})(\xi - \xi_{\text{min}})\lambda_u (z - 1)x}{\beta_{\text{max}} - \beta_{\text{min}}(\xi_{\text{max}} - \xi_{\text{min}})} \right)^{-3.5}. \tag{13}$$

The $n$th derivative of $G_{N_k,\xi,\beta}(z)$ is

$$G^{(n)}_{N_k,\xi,\beta}(z) = \frac{C_n}{2^n} \frac{(2n + 5)!}{15} (-C_0 z + C_0 + 1)^{-n - \frac{7}{2}}, \tag{14}$$

where $C_0$ is given by $\binom{12}{n}$ and $n!!$ is the double factorial. Therefore, we get the PMF by

$$\mathbb{P}(N_k,\xi,\beta = n) = \frac{1}{n!} G^{(n)}_{N_k,\xi,\beta}(0) = \left( \frac{2n + 5}{15} \right) \left( \frac{C_0}{2} \right)^n (C_0 + 1)^{-n - \frac{7}{2}}. \tag{15}$$

Figure 3 plots the PMFs of the number of users associating to different tiers of BSs in the two tiered heterogeneous networks formed by macrocells and picocells. When the bias factors for macrocells and picocells are the same ($B_1 = B_2 = 1$), i.e., no bias exists, the macrocells will serve more users than the picocells because they provide larger coverage. Figure 3 also indicates that in the case where some users are offloaded from the macrocells to the picocells ($B_1 = 1, B_2 = 10$), the number of users served by each macrocell is greatly reduced while that served by each picocell does not increase much, implying that the offloading policy is potential for better utilization of the idle resources of the small cells.

**Remark 1.** From Lemma 3 we get the mean number of users served by a BS of the $k$th tier with arrival rate less than $\xi$ and with delay requirement less than $\beta$ as

$$\mathbb{E}[N_k,\xi,\beta] = \frac{(\beta_{\text{max}} - \beta_{\text{min}})(\xi - \xi_{\text{min}})}{\beta_{\text{max}} - \beta_{\text{min}}(\xi_{\text{max}} - \xi_{\text{min}})}. \tag{16}$$

By setting $\xi = \xi_{\text{max}}$ and $\beta = \beta_{\text{max}}$, we get the total number of users served by a BS of the $k$th tier as $N_k = N_{k,\xi_{\text{max}},\beta_{\text{max}}}$ whose PGF, PMF and mean value are

$$G_{N_k}(z) = \left( 1 - \frac{2p_k}{\lambda_k} \frac{\lambda_u (z - 1)}{\xi_{\text{max}} - \xi_{\text{min}}} \right)^{-3.5}, \tag{17}$$

$$\mathbb{P}(N_k = n) = \frac{1}{n!} G_{N_k}^{(n)}(0), \tag{18}$$

$$\mathbb{E}[N_k] = \frac{p_k}{\lambda_k} \lambda_u. \tag{19}$$

Letting $\xi_{k,\text{total}}$ be the total arrival rate of the packets at a BS of the $k$th tier, we get

$$\xi_{k,\text{total}} = \mathbb{E}[\xi_{k,\text{total}}] = \frac{\xi_{\text{max}} + \xi_{\text{min}}}{2} \frac{p_k}{\lambda_k} \lambda_u. \tag{20}$$

Figure 4 plots the total arrival rate at different BSs in the two tiered heterogeneous networks formed by the macrocells and the picocells. It is observed that when fixing the bias factor for macrocells and increasing the bias factor for picocells, the total arrival rate at each macrocell decreases dramatically while the total arrival rate at each picocell increases slowly, indicating that the traffic for macrocells is effectively offloaded with a little cost of increasing the traffic for each picocell. Meanwhile, Figure 4 also reveals that the propagation environment may greatly affect the offloading of the traffic. For example, the coverage region of the macrocells may shrink...
when increasing the path loss exponents $\alpha$, resulting in less traffic for the macrocells.

IV. SUCCESS PROBABILITY AND DELAY

In the following, we analyze the statistics of success probability and delay in the heterogeneous networks. The success probability is the probability that a scheduled packet can be successfully delivered. The delay consists of two parts, one is the queuing delay, which measures the delay between the time a packet arrives at the queue and the time when it starts to be served; the other is the service time, which is the time to transmit the given packet. We assume that the delay mentioned here is evaluated in terms of number of time slots. Since the considered network is static, i.e., the locations of the BSs are deployed at first and keep unchanged during the following time, the success probabilities for different links are different, resulting in different mean delays for different queues. Thus, a suitable metric to characterize the delay performance of the large system is the statistical cumulative distribution function (cdf) of the mean delay of all users in the network. From the ergodicity of the PPP, the ensemble averages obtained by averaging over the point process equal the spatial averages obtained by averaging over an arbitrary realization of the PPP over a large region. Therefore, in order to obtain the statistical cdf of the mean delay obtained from the spatial statistics of the mean delay samples of a large number of transmissions in the network, we could analytically evaluate the cdf of the mean delay at the typical transmission for different realizations of the PPP.

Without loss of generality, we consider a typical user located at the origin, denoted by $(x_0, \xi_0, \beta_0)$ with $x_0 = 0$, and a typical BS $y_0$ associated with the typical user. For each realization of the spatial locations of BSs $\Phi_b$, we obtain a mean delay for the typical user. By traversing all the realizations of $\Phi_b$, we obtain a statistical distribution of the mean delay (for example, the cdf) for the typical user at the origin. Thus, the cdf of the mean delay for the typical user equals the statistical cdf of the mean delays of all users in the network. To avoid confusion, we denote the mean delay of the typical user at the origin conditioning on the realization of $\Phi_b$ as the conditional mean delay, denoted by $D_{\Phi_b}$, which is a random variable uniquely defined by $\Phi_b$ [39 Def. 1]. In the following discussions, we derive the pdf of the conditional mean delay $D_{\Phi_b}$, which is suitable to characterize the delay performance of the large network since it also gives the statistical cdf of the mean delays of all users in the network.

The typical user will associate the BS from the $k$th tier with probability $p_k$. Let $Z_k$ be the distance between the typical user at the origin and the nearest BS in the $k$th tier of BSs. Let $y_0 \in \Phi_b$ be the serving BS; then $Z_k = |y_0|$ holds when a user associates to BS from the $k$th tier. The following lemma gives the pdf of the link distance $|y_0|$ conditioned on the user associating to BS from the $k$th tier.

Lemma 4. Conditioned on the typical user associating to BS $y_0$ from the $k$th tier, the pdf of the distance $Z_k = |y_0|$ between the typical user at the origin and its serving BS $y_0$ is

$$f_{Z_k|y_0|(\Phi_k)}(r) = \frac{2\pi r \lambda_k}{\text{P}_k} \exp\left(-\frac{\pi r^2 \lambda_k}{\text{P}_k}\right).$$

Proof: Conditioned on the typical user associating to BS from the $k$th tier, the cdf of the distance between the typical user and its serving BS $y_0$ is

$$P\left(|y_0| \leq r \mid y_0 \in \Phi_k\right) = \frac{P\left(|y_0| \leq r \mid y_0 \in \Phi_k\right)}{P(y_0 \in \Phi_k)} = \frac{1}{\text{P}_k} \Phi_k \left\{Z_k \leq r\right\} \bigg\{ \bigcap_{i=1,i\neq k}^K \left\{P_k B_k Z_k^{-\alpha} > P_i B_i Z_i^{-\alpha}\right\} \bigg\}\bigg\} = \frac{1}{\text{P}_k} \int_0^r \prod_{i=1,i\neq k}^K P(\lambda_i P_i B_i Z_i^{-\alpha} > \lambda_k P_k B_k z^{-\alpha}) f_{Z_k}(z)dz = 1 - \exp\left(-\pi r^2 \sum_{i=1}^K \lambda_i \left(\frac{P_i B_i}{P_k B_k}\right)^{2/\alpha}\right),$$

where the last equation follows by plugging in the pdf of $Z_i$ given by (22). By differentiating with respect to $r$, we get the result in the lemma.

Remark 2. The form of the conditional pdf given by (21) is very similar to the contact distance distribution of the PPP given by (30). The only difference lies in that the intensity of the PPP $\lambda_i$ is replaced by $\frac{P_i}{\lambda_k}$. Intuitively, it can be interpreted as that, since $\text{P}_k$ is the average fraction of total area covered by BSs from the $k$th tier, the spatial distribution of the BSs from the $k$th tier could be considered as a PPP of intensity $\frac{P_k}{\lambda_k}$.

In the following, we first condition on the realization of $\Phi_b = \bigcup_{i=1}^K \Phi_i$ and derive the conditional mean delay $D_{\Phi_b}$. The statuses of BSs in the heterogeneous network switch between idle and busy according to the random muting, the statistics of the queues, and the scheduling results at the BSs. In each time slot, a BS will suppress its transmission and not cause interference if the BS is muted or the queue of the scheduled user of the BS is empty. Let $\zeta_{y,t} \in \{0, 1\}$ be the indicator showing whether the BS located at $y \in \Phi_b$ is idle ($\zeta_{y,t} = 0$) or busy ($\zeta_{y,t} = 1$) in the time slot $t$. Then, if the typical user located at the origin associates to a BS $y_0$ from the $k$th tier, the SIR at the user in the time slot $t$ is

$$\text{SIR}_t = \frac{P_k h_{y_0} |y_0|^{-\alpha}}{\sum_{i=1}^K \sum_{y \in \Phi_b \backslash \{B(o, z_{ik})\}} |y_{i,t} P_i h_{y}|^{-\alpha}},$$

where $h_y$ is the power fading coefficient between the typical user and the BS located at $y$ with $h_y \sim \text{Exp}(1)$, and $B(o, z_{ik})$ is a ball centered at the origin with radius $z_{ik} = \left(\frac{P_i B_i}{\text{P}_k B_k}\right)^{1/\alpha} |y_0|$.

The difficulty of the delay analysis lies in the fact that the SIR relies on the statuses of the queues at all BSs, and on the other hand the SIR at the users also affects the statuses of the queues. Decoupling the SIR analysis and the statuses of the queues at all BSs is difficult; thus, we propose to bound the SIR as follows:
Lower bound of SIR: Consider a dominant system in which the typical user and its serving BS behave exactly the same as that in the original system; however, for other BSs in the dominant system, we assume that when the queues of the scheduled users at those BSs become empty, those BSs continue to transmit “dummy” packets, thus continuing to cause interference to other users no matter whether the scheduled queues are empty or not. The queue size at each BS in the dominant system will never be smaller than that in the original system, resulting in smaller SIR and larger delay. Therefore, the obtained SIR under these assumptions will be a lower bound for the SIR in the original system.

Upper bound of SIR: Consider a modified system as follows: the typical user and its serving BS behave exactly the same as that in the original system; however, if an arrived packet at an interfering BS is not scheduled or failed for the transmission in current time slot, it will be dropped rather than being retransmitted. In this way, since the interference in the modified system is always smaller than that in the original system and the packets will not accumulate at the interfering BSs, the obtained SIR will be an upper bound for the SIR in the original system.

A. Statistic of Success Probability

By using the aforementioned bounding approaches, \{\zeta_{y,t}\} become independent random variables and are not related to the time slot index \(t\). In these cases, the SIRs at the typical user in different time slots, denoted by \{SIR_t\}, are independent random variables and also not related to the time slot index \(t\). For simplicity and without ambiguity, we omit the subscript \(t\) and use \(\zeta_y\) and SIR when deriving these bounds. Therefore, the success probability conditioned on \(\Phi_b\) when the typical user is served by a BS \(y_0\) from the \(k\)th tier with link distance \(Z_k = |y_0|\) is given by

\[
\mathbb{P}(\text{SIR} > \theta \mid \Phi_b, y_0 \in \Phi_k, Z_k = |y_0|) = \mathbb{P} \left( \frac{P_k h_{y_0}|y_0|^{-\alpha}}{\sum_{i=1}^{K} \sum_{y \in \Phi_1 \setminus B_{(\alpha, z_{ik})}} \zeta_y P_i h_y |y|^{-\alpha}} > \theta \mid \Phi_b, y_0 \in \Phi_k, Z_k = |y_0| \right).
\]

(24)

The probability of \(\zeta_y = 1\) relies on the bounding approaches and the scheduling strategies. We first assume

\[
\mathbb{P}(\zeta_y = 1) = q, \tag{25}
\]

and analyze the success probability; then we determine the values of \(q\) for different bounding approaches and scheduling strategies.

Lemma 5. If each interfering BS is active independently with probability \(q\), i.e., \(\mathbb{P}(\zeta_y = 1) = q\), the cdf of success probability for the typical user conditioned on \(\Phi_b\) when it associates to BS from \(k\)th tier, denoted by \(\mathbb{P}(\text{SIR} > \theta \mid \Phi_b, y_0 \in \Phi_k)\), is

\[
\mathbb{P}(\mathbb{P}(\text{SIR} > \theta \mid \Phi_b, y_0 \in \Phi_k) \leq u) = \frac{1}{2} \left( 1 - \frac{1}{\pi \rho_k} \int_0^\infty \frac{1}{\omega} \text{Im}\left\{ e^{-j\omega \kappa^{-1}(\omega, k, q)} \right\} d\omega \right).
\]

(26)

where \(\kappa(\omega, k, q)\) is defined as

\[
\kappa(\omega, k, q) = -2 \sum_{i=1}^{K} \frac{\lambda_i}{\lambda_k} \left( \frac{P_i B_i}{P_k B_k} \right)^{\delta} \times \int_0^1 \left( 1 - \left( \frac{q}{1 + \theta y^{\alpha} B_i^{\alpha}} + 1 - q \right)^{j\omega} \right)^{y^{-3}} dy - \frac{1}{f_k}.
\]

(27)

Proof: See Appendix [A].

Figure [5] plots the cdf of success probabilities for different tiers of BSs in the two tiered heterogeneous networks formed by macrocells and picocells. If there is no bias (\(B_1 = B_2 = 1\)), the cdfs of the success probabilities for macrocell users and picocell users coincide, implying that the statistical success probability of users associating to different tiers of BSs is the same for the maximum average received power association. The cdfs given by Figure [5] shows that when increasing the bias factor for picocells (\(B_1 = 1, B_2 = 4\)), the proportion of picocell users with small success probability increases since the offloaded users are distributed at the edge of the picocells and may experience poor SIR performance. For the case of large path loss exponent, the electromagnetic signal attenuates fast, thus suppressing the inter-cell interference and increasing the success probability for both macrocell and picocell users.
This observation indicates that offloading is more effective when the signal attenuates fast.

Define the success probability of the typical user conditioned on \( \Phi_b \) as

\[
P_{\Phi_b} \triangleq \mathbb{P}(\text{SIR} > \theta \mid \Phi_b).
\]  

Then, \( P_{\Phi_b} \) is a random variable uniquely determined by the realization of \( \Phi_b \). By traversing all the realizations of \( \Phi_b \), we get the cdf of \( P_{\Phi_b} \). Due to the ergodicity of the PPP, the cdf of \( P_{\Phi_b} \) also gives the statistical cdf of the success probabilities of all users within a large region. Therefore, we get the following theorem.

**Lemma 6.** If each interfering BS is active independently with probability \( q \), i.e., \( \mathbb{P}(\zeta_u = 1) = q \), the statistical cdf of success probabilities of all users in the network equals the cdf of the success probability of the typical user conditioned on \( \Phi_b \), which is

\[
\mathbb{P}(P_{\Phi_b} \leq u) = F(u, q) \triangleq \frac{1}{2} + \frac{1}{\pi} \sum_{k=1}^{K} \int_{0}^{\infty} \frac{1}{\omega} \text{Im}\left\{u^{-j\omega}K^{-1}(\omega, k, q)\right\}d\omega.
\]  

**Proof:** Combining Lemma 5 and using the total probability formula

\[
P_{\Phi_b} = \sum_{k=1}^{K} P_{k} \mathbb{P}(\text{SIR} > \theta \mid \Phi_k, y_0) \leq u),
\]

we obtain the theorem. \( \square \)

**Remark 3.** Particularly, in single-tier cellular network, i.e., \( K = 1 \), the statistical cdf of success probabilities of all users in the network is

\[
\mathbb{P}(P_{\Phi_b} \leq u) = \frac{1}{2} + \frac{1}{\pi} \int_{0}^{\infty} \frac{1}{\omega} \text{Im}\left\{u^{-j\omega}\right\}d\omega.
\]

**B. Statistic of Delay**

In the following, we discuss the statistics of the delay for different scheduling policies.

1) **Random Scheduling:** In the random scheduling, each active BS will randomly choose one user from all users within the coverage region of that BS to serve in each time slot. Let \( N \) be the number of users served by the BS. Considering the random muting mechanism that each BS is idle with probability \( 1 - p \) in each time slot, the average service rate (number of packets transmitted successfully per time slot) of the typical user conditioned on \( \Phi_b \) is

\[
\mu = \frac{P}{N} \mathbb{P}(\text{SIR} > \theta \mid \Phi_b) = \frac{P}{N} P_{\Phi_b}.
\]

The probability for a packet at the typical user being scheduled and successfully transmitted in a time slot is \( \mu \) given by \( (32) \). Since the transmissions of the typical user in different time slots are independent in the random scheduling, the probability for successfully transmitting a packet in any time slot is also \( \mu \). Therefore, the queueing system at the typical transmitter is equivalent to a \( \text{Geo/G/1} \) queue, or a discrete-time single server retrial queue \([41],[42]\). In the equivalent \( \text{Geo/G/1} \) queue, the time is slotted and the packets arrive according to a Bernoulli process with intensity \( \xi_0 \) packets per time slot. The arrival process is also called geometric arrival process since the probability that a packet arrives in a time slot is \( \xi_0 \), and the number of time slots between two adjacent arrivals is a geometric random variable. The success probability is \( \mu \) and the service times of packets are i.i.d. with geometric distribution. From \([41]\), we get the conditional mean delay \( D_{\Phi_b} \) in random scheduling as

\[
D_{\Phi_b} = \begin{cases} 
\frac{1-\xi_0}{\frac{\mu}{\xi_0}} & \text{if } \mu > \xi_0 \\
\infty & \text{if } \mu \leq \xi_0.
\end{cases}
\]

Therefore, we obtain the following theorem.

**Lemma 7.** In random scheduling, if each interfering BS is active independently with probability \( q \), i.e., \( \mathbb{P}(\zeta_u = 1) = q \), the statistical cdf of the mean delay of all users in the network equals the cdf of mean delay of the typical user conditioned on \( \Phi_b \), which is

\[
\mathbb{P}(D_{\Phi_b} \leq T) = H_1(T, q) \triangleq \frac{1}{2} - \frac{1}{\pi} \sum_{k=1}^{K} \int_{0}^{\infty} \frac{1}{\omega} \text{Im}\left\{\right\}d\omega.
\]

\[
\mathbb{P}(D_{\Phi_b} \leq T) = \mathbb{E}_{N_k \xi_0} \left[N_k j\omega \left(\frac{T-1}{T \xi_0}\right)^{-j\omega}\right]d\omega.
\]

**Proof:** See Appendix B \( \square \)

Particularly, by the previous mentioned bounding approaches, we obtain the following theorem which gives the bounds for the cdf of success probability and mean delay in random scheduling.

**Theorem 1.** In random scheduling, the statistical cdf of success probability and mean delay of all users in the network are bounded by

\[
F(u, p) \leq \mathbb{P}(P_{\Phi_b} \leq u) \leq F(u, \frac{\xi_{\text{max}} + \xi_{\text{min}}}{2} p),
\]  

\[
H_1(T, p) \leq \mathbb{P}(D_{\Phi_b} \leq T) \leq H_1(T, \frac{\xi_{\text{max}} + \xi_{\text{min}}}{2} p).
\]
5) On equals the cdf of the mean delay of the typical user conditioned on the statistical cdf of the mean delay of all users in the network.

2) FIFO Scheduling: In the FIFO scheduling, each active BS will schedule the earliest arriving packet at that BS in each time slot, and all queues at a BS could be considered as a single large queue. Notice that the delay of the typical user equals the delay of the packets arriving at the large queue. Therefore, the conditional mean delay of the typical user in round-robin scheduling is

\[ D_{\Phi_b} = \left\{ \begin{array}{ll} \frac{1 - \sum_{i=0}^{N-1} \xi_i}{\mu - \sum_{i=0}^{N-1} \xi_i} & \text{if } \mu > \sum_{i=0}^{N-1} \xi_i \\ \infty & \text{if } \mu \leq \sum_{i=0}^{N-1} \xi_i. \end{array} \right. \] (38)

Therefore, we obtain the following lemma and theorem.

Lemma 8. In FIFO scheduling, if each interfering BS is active independently with probability \( q \), i.e., \( \mathbb{P}(\zeta_i = 1) = q \), the statistical cdf of the mean delay of all users in the network equals the cdf of the mean delay of the typical user conditioned on \( \Phi_b \), which is

\[ \mathbb{P}(D_{\Phi_b} \leq T) = H_2(T, q) \triangleq \frac{1}{2} - \frac{1}{\pi} \sum_{k=1}^{K} \int_0^{\infty} \frac{1}{\omega} \text{Im} \left\{ \frac{p^{j\omega}}{\kappa(\omega, k, q)} \mathbb{E}_{N_k, \xi_i} \left[ \left( \frac{T}{T} + \frac{T - 1}{T} \sum_{i=0}^{N-1} \xi_i \right)^{\frac{-j\omega}{\omega}} \right] \right\} d\omega. \] (39)

Proof: The proof is similar to that of Lemma 7. \( \square \)

Theorem 2. In FIFO scheduling, the statistical cdf of success probability and mean delay of all users in the network are bounded by

\[ F(u, p) \leq \mathbb{P}(\Phi_b \leq u) \leq F\left( u, \min \left\{ 1, \frac{p}{2} \left( \xi_{\text{max}} + \xi_{\text{min}} \right) \min_{1 \leq k \leq K} N_k \right\} \right), \] (40)

\[ H_2(T, p) \leq \mathbb{P}(D_{\Phi_b} \leq T) \leq H_2\left( T, \min \left\{ 1, \frac{p}{2} \left( \xi_{\text{max}} + \xi_{\text{min}} \right) \min_{1 \leq k \leq K} N_k \right\} \right). \] (41)

Proof: The proof is similar to that of Theorem 1. \( \square \)

3) Round-robin Scheduling: In the round-robin scheduling, if there are \( N \) users within the typical cell, a complete cycle will occupy \( N \) time slots. Therefore, the duration between two scheduled time slots for the typical user is \( N \). If we only consider the time slots that the typical user is scheduled, the serving of the typical user could be considered as a new queueing system with the serving rate given by

\[ \mu = p \mathbb{P}(\text{SIR} > \theta | \Phi_b) = p \mathbb{P}_{\Phi_b}. \] (42)

The new queueing system is also a Geo/G/1 queueing system. Let \( D_{\text{RR}} \) be the mean delay of the new queueing system, which is the mean number of scheduled time slots required for the typical user to successfully transmit a packet. Then, we have

\[ D_{\text{RR}} = \begin{cases} \frac{1 - \xi_0}{\mu - \xi_0} & \text{if } \mu > \xi_0 \\ \infty & \text{if } \mu \leq \xi_0. \end{cases} \] (43)

For the typical user, note that \( N/2 \) time slots are required on average if a packet is successfully delivered in the first scheduled time slot; otherwise, the additional time slots required should be the additional number of scheduled time slots multiplied by \( N \). Therefore, the conditional mean delay \( D_{\Phi_b} \) in round-robin scheduling is

\[ D_{\Phi_b} = N(D_{\text{RR}} - 1) + N \leq \begin{cases} \frac{1 - \xi_0}{\mu - \xi_0} N - \frac{N}{2} & \text{if } \mu > \xi_0 \\ \infty & \text{if } \mu \leq \xi_0. \end{cases} \] (44)

The following lemma and theorem give the bounds for the cdf of the conditional mean delay \( D_{\Phi_b} \) in round-robin scheduling.

Lemma 9. In round-robin scheduling, if each interfering BS is active independently with probability \( q \), i.e., \( \mathbb{P}(\zeta_i = 1) = q \), the statistical cdf of the mean delay of all users in the network equals the cdf of the mean delay of the typical user conditioned on \( \Phi_b \), which is

\[ \mathbb{P}(D_{\Phi_b} \leq T) = H_3(T, q) \triangleq \frac{1}{2} - \frac{1}{\pi} \sum_{k=1}^{K} \int_0^{\infty} \frac{1}{\omega} \text{Im} \left\{ \frac{p^{j\omega}}{\kappa(\omega, k, q)} \mathbb{E}_{N_k, \xi_0} \left[ \left( \xi_0 + \frac{1 - \xi_0}{2 + T/N_k} \right)^{\frac{-j\omega}{\omega}} \right] \right\} d\omega. \] (45)

Proof: The proof is similar to that of Lemma 7. \( \square \)

Theorem 3. In round-robin scheduling, the statistical cdf of success probability and mean delay of all users in the network are bounded by

\[ F(u, p) \leq \mathbb{P}(\Phi_b \leq u) \leq F(u, \frac{\xi_{\text{max}} + \xi_{\text{min}}}{2}), \]

\[ H_3(T, p) \leq \mathbb{P}(D_{\Phi_b} \leq T) \leq H_3(T, \frac{\xi_{\text{max}} + \xi_{\text{min}}}{2}). \] (46)

Proof: The proof is similar to that of Theorem 1. \( \square \)

Other user scheduling mechanisms, such as FDMA, could also be analyzed similarly. In the case of FDMA, all users served by a same BS will be served simultaneously using orthogonal sub-channels. Since not all sub-channels are used at a BS, the set of BSs that cause interference to the typical
link will be only a subset of all BSs, i.e., the interfering BSs construct a thinning version of the original BSs. Therefore, the approaches used in our work can be also applied in the FDMA case, with the difference being that the interference is thinned, and the user scheduling is simplified.

C. Delay Outage

In the design of practical wireless network, an intuitive and meaningful metric is the proportion of users whose delay requirements cannot be achieved. The delay requirement of a user \(x_i\) cannot be achieved if the practical mean delay of the user is larger than the mean delay requirement \(\beta_i\) of the user. Therefore, we propose the notion of \textit{delay outage} which means that the delay requirement of a user cannot be achieved. The user \(x_i\) is delay outage if and only if its mean delay is larger than \(\beta_i\). We further evaluate the delay outage probability, which is the proportion of users that are delay outage in the wireless network, as

\[
\eta_{DO} \triangleq \mathbb{E}_{\beta_i} \left( \mathbb{P}(D_{\Phi_i} \geq \beta_i) \right) = \frac{1}{\beta_{\max} - \beta_{\min}} \int_{\beta_{\min}}^{\beta_{\max}} \mathbb{P}(D_{\Phi_i} \geq \beta_i) \, d\beta_i. \quad (47)
\]

The previous results could be plugged into (47) to bound the delay outage probability \(\eta_{DO}\).

V. NUMERICAL EVALUATION

To gain insight, we discuss the statistical cdf of mean delay and the delay outage numerically.

Figure 6 plots the upper and lower bounds for the cdfs of the mean delay of all queues in the two tiered heterogeneous networks formed by macrocells and picocells. The parameters are set as \(P_1 = 39\) dBm, \(P_2 = 24\) dBm, \(\lambda_1 = 0.00001\) m\(^{-2}\), \(\lambda_2 = 0.00005\) m\(^{-2}\), \(B_1 = B_2 = 1\), \(p = 0.5\) and \(\lambda_u = 0.00005\) m\(^{-2}\).

The curves for different arrival rates, we observe that the delay performance of random scheduling is the worst while the delay performance of round-robin scheduling outperforms FIFO scheduling when the traffic is heavy, and it is reversed for light traffic. This is because when the traffic is light, applying round-robin scheduling may waste some time slots to schedule the empty queues, and when the traffic is heavy, in the FIFO scheduling, the scheduling may be blocked at the users with poor link quality, resulting in large delay for the packets. Figure 7 shows that the cdf of the mean delay will not increase to one when increasing the value of the mean delay, due to the fact that some users may be affected by strong interference, and their queues increase to infinite resulting in infinite mean delay. We also observe that certain percentage of users in the network experience very low mean delay, i.e., the mean delay for those users approaches to zero, and the delay outage probability is the worst while the delay outage probability of random scheduling is the worst while the delay outage probability of round-robin scheduling reduces to zero, indicating that there are always certain percentage of users that experience a small interference and a low mean delay.

Figure 7 plots the upper and lower bounds for the delay outage probability as functions of bias factor \(B_2\) in the two tiered heterogeneous networks formed by macrocells and picocells. The parameters are set as \(P_1 = 39\) dBm, \(P_2 = 24\) dBm, \(\lambda_1 = 0.00001\) m\(^{-2}\), \(\lambda_2 = 0.00005\) m\(^{-2}\), \(\xi_{\text{min}} = 0.7\), \(\xi_{\text{max}} = 0.8\), \(\beta_{\min} = 18\), \(\beta_{\max} = 20\), \(B_1 = 1\), \(p = 1\) and \(\lambda_u = 0.00005\) m\(^{-2}\).
increasing the opportunity to schedule a packet is increased for large $p$. Utilized, and the delay outage probability will start to increase as the increasing of user density. This counterintuitive result can be interpreted as follows. When the user density $\lambda_u$ is much smaller than the density of BSs, many cells are empty and serve no user. Since we only consider the cells serving at least one user, slightly increasing $\lambda_u$ will increase the proportion between the number of cells serving one user and that of cells serving more than one users. Noting that the cells with only one user experience much smaller delay outage probability than that with more than one users, the overall delay outage probability will decrease when slightly increasing the user density $\lambda_u$ in the case where $\lambda_u$ is much smaller than the density of BSs. Figure 9 also reveals that deploying more tiers of BSs helps to decrease the delay outage probability only when the user density is large.

VI. Conclusion

In this paper, a tractable approach is proposed to analyze the delay in the heterogeneous cellular networks with spatio-temporal random arrival of traffic. Meanwhile, the statistic of spatio-temporal traffic with offloading policy is evaluated. The notion of delay outage is proposed to evaluate the effect of different scheduling policies, such as random scheduling, FIFO scheduling and round-robin scheduling. The obtained upper and lower bounds are useful to evaluate the delay performance of the large wireless networks. The numerical analysis shows that those bounds are tight especially for the case of heavy traffic and small active probability.

The results reveal that the offloading policy based on cell range expansion greatly reduces the macrocell traffic while bring a small amount of growth for the picocell traffic, implying that the offloading policy is potential for better utilization of the idle resources of the small cells. Our results also show that the delay performance of round-robin scheduling outperforms FIFO scheduling when the traffic is heavy, and it is reversed for light traffic since applying round-robin scheduling may waste some time slots to schedule the empty queues for light traffic, while in the FIFO scheduling, the scheduling may be blocked at the users with poor channel quality for heavy traffic, resulting in large delay.

Appendix A

Proof of Lemma 5

Proof: The success probability conditioned on $\Phi_k$ when the typical user is served by a BS $y_0$ from the $k$th tier with...
link distance \( Z_k = |y_0| \) is

\[
P(\text{SIR} > \theta \mid \Phi_b, y_0 \in \Phi_k, Z_k = |y_0|) = \mathbb{P} \left( \sum_{i=1}^{K} \frac{\zeta_i P_i h_i}{1 + \theta |y_0|^a |y_j |^{-\alpha} P_i / P_k} + 1 - q \right) \bigg| \Phi_b, y_0 \in \Phi_k, Z_k = |y_0| \right)
\]

The moment generating function of \( Y \triangleq \ln \mathbb{P}(\text{SIR} > \theta \mid \Phi_b, y_0 \in \Phi_k, Z_k = |y_0|) \) is

\[
M_Y(s) = \mathbb{E} \left( \exp(sY) \right) = \mathbb{E} \left( \mathbb{P}(\text{SIR} > \theta \mid \Phi_b, y_0 \in \Phi_k, Z_k = |y_0|) \right) \]

\[
(48)
\]

The cdf of \( \mathbb{P}(\text{SIR} > \theta \mid \Phi_b, y_0 \in \Phi_k, Z_k = |y_0|) \) is evaluated as

\[
\mathbb{P}(\text{SIR} > \theta \mid \Phi_b, y_0 \in \Phi_k, Z_k = |y_0|) = \max \{\ln \mathbb{P}(\text{SIR} > \theta \mid \Phi_b, y_0 \in \Phi_k, Z_k = |y_0|), 0\} \]

\[
(51)
\]

The cdf of \( \mathbb{P}(\text{SIR} > \theta \mid \Phi_b, y_0 \in \Phi_k, Z_k = |y_0|) \) is obtained by the total probability formula as

\[
\mathbb{P}(\text{SIR} > \theta \mid \Phi_b, y_0 \in \Phi_k, Z_k = |y_0|) \leq u)
\]

\[
= \mathbb{P}(\text{SIR} > \theta \mid \Phi_b, y_0 \in \Phi_k, Z_k = |y_0|) \leq u)
\]

\[
(52)
\]

Plugging in \( z_{ik} = \left( \frac{P_i}{P_k} \right)^{1/\alpha |y_0|} \), we obtain

\[
\mathbb{P}(\text{SIR} > \theta \mid \Phi_b, y_0 \in \Phi_k, Z_k = |y_0|) \leq u)
\]

\[
= \mathbb{P}(\text{SIR} > \theta \mid \Phi_b, y_0 \in \Phi_k, Z_k = |y_0|) \leq u)
\]

\[
(53)
\]

Plugging in \( \alpha = 1 \) and calculating the integral with respective to \( r \), we obtain the theorem.

**APPENDIX B**

**PROOF OF LEMMA 7**

*Proof:* By the total probability formula, we obtain

\[
\mathbb{P}(D_{b_k} \leq T) = \sum_{k=1}^{K} \mathbb{P}_k \mathbb{P}(D_{b_k} \leq T \mid y_0 \in \Phi_k) \cdot \mathbb{P}(D_{b_k} \leq T) \cdot (55)
\]

Plugging in \( \text{26} \) and \( \text{33} \), we get the cdf as

\[
\mathbb{P}(D_{b_k} \leq T) = \sum_{k=1}^{K} \mathbb{P}_k \mathbb{P}(D_{b_k} \leq T \mid y_0 \in \Phi_k)
\]

\[
= \sum_{k=1}^{K} \mathbb{P}_k \mathbb{P}(D_{b_k} \leq T \mid y_0 \in \Phi_k)
\]

\[
(56)
\]
From Lemma\[5\] we have

\[
\mathbb{P}(D_{b^*} \leq T) = \frac{1}{2} - \frac{1}{\pi} \sum_{k=0}^{\infty} \left[ \int_{0}^{\infty} \frac{1}{\omega} \exp \left\{ -j\omega \left( \xi_0 + \frac{1}{T} \right) \right\} \kappa^{-1}(\omega, k, q) \omega d\omega \right].
\]

Therefore, we obtain the results. \(\blacksquare\)

REFERENCES


Yi Zhong received his B.S. and Ph.D. degree in Electronic Engineering from University of Science and Technology of China (USTC) in 2010 and 2015 respectively. From August to December 2012, he was a visiting student in Prof. Martin Haenggi’s group at University of Notre Dame. From July to October 2013, he was an research intern with Qualcomm Incorporated, Corporate Research and Development, Beijing. From July 2015 to December 2016, he was a Postdoctoral Research Fellow with the Singapore University of Technology and Design (SUTD) in the Wireless Networks and Decision Systems (WNDS) Group led by Prof. Tony Q.S. Quek. Now, he is an assistant professor with School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China. His main research interests include heterogeneous and femtocell-overlaid cellular networks, wireless ad hoc networks, stochastic geometry and point process theory.

Tony Q.S. Quek (S’98-M’08-SM’12) received the B.E. and M.E. degrees in Electrical and Electronics Engineering from Tokyo Institute of Technology, respectively. At MIT, he earned the Ph.D. in Electrical Engineering and Computer Science. Currently, he is a tenured Associate Professor with the Singapore University of Technology and Design (SUTD). He also serves as the Associate Head of ISTD Pillar and the Deputy Director of the SUTD-ZJU IDEA. His main research interests are the application of mathematical, optimization, and statistical theories to communication, networking, signal processing, and resource allocation problems. Specific current research topics include heterogeneous networks, green communications, wireless security, internet-of-things, and big data processing.

Dr. Quek has been actively involved in organizing and chairing sessions, and has served as a member of the Technical Program Committee as well as symposium chairs in a number of international conferences. He is serving as the Workshop Chair for IEEE Globecom in 2017, the Tutorial Chair for the IEEE ICC in 2017, and the Special Session Chair for IEEE SPAWC in 2017. He is currently an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS and an elected member of IEEE Signal Processing Society SPCOM Technical Committee. He was an Executive Editorial Committee Member for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and an Editor for the IEEE WIRELESS COMMUNICATIONS LETTERS. He is a co-author of the book “Small Cell Networks: Deployment, PHY Techniques, and Resource Allocation” published by Cambridge University Press in 2013 and the book “Cloud Radio Access Networks: Principles, Technologies, and Applications” by Cambridge University Press in 2017.

Dr. Quek was honored with the 2008 Philip Yeo Prize for Outstanding Achievement in Research, the IEEE Globecom 2010 Best Paper Award, the 2012 IEEE William R. Bennett Prize, the IEEE SPAWC 2013 Best Student Paper Award, the IEEE WCSP 2014 Best Paper Award, the 2015 SUTD Outstanding Education Awards – Excellence in Research, the 2016 Thomson Reuters Highly Cited Researcher, and the 2016 IEEE Signal Processing Society Young Author Best Paper Award.

Xiaohu Ge (M’09-SM’11) is currently a full Professor with the School of Electronic Information and Communications at Huazhong University of Science and Technology (HUST), China. He is an adjunct professor with with the Faculty of Engineering and Information Technology at University of Technolo gy Sydney (UTS), Australia. He is the director of China International Joint Research Center of Green Communications and Networking. He received his PhD degree in Communication and Information Engineering from HUST in 2003. He has worked at HUST since Nov. 2005. Prior to that, he worked as a researcher at Ajou University (Korea) and Politecnico Di Torino from Jan. 2004 to Oct. 2005. His research interests are in the area of mobile communications, traffic modeling in wireless networks, green communications, and interference modeling in wireless communications. He has published more than 130 papers in refereed journals and conference proceedings and has been granted about 15 patents in China. He received the Best Paper Awards from IEEE Globecom 2010.

Dr. Ge is a Senior Member of the China Institute of Communications and a member of the National Natural Science Foundation of China and the Chinese Ministry of Science and Technology Peer Review College. He has been actively involved in organizing more the ten international conferences since 2005. He served as the general Chair for the 2015 IEEE International Conference on Green Computing and Communications (IEEE GreenCom 2015). He serves as an Associate Editor for the IEEE Transaction on Green Communications and Networking, etc.