Performance Evaluation of 5G mmWave Networks with Physical-Layer and Capacity-Limited Blocking

Jingjin Wu *†, Meiqian Wang †, Yin-Chi Chan †, Eric W. M. Wong †, and Taejoon Kim ‡
* Department of Statistics, BNU-HKBU United International College, Zhuhai, Guangdong, P. R. China
† Department of Electrical Engineering, City University of Hong Kong, Hong Kong SAR, P. R. China
‡ Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA
Email: jj.wu@ieee.org, meiqiwang2@cityu.edu.hk, ycchan26@cityu.edu.hk, eeewong@cityu.edu.hk, taejoonkim@ku.edu

***Abstract*—We propose a versatile cross-layer framework to analyze performance metrics for mobile traffic in fifth-generation (5G) millimeter wave (mmWave) networks. Our proposed framework is based on stochastic geometry, teletraffic models, and the classical Erlang Fixed Point Approximation method, with the objective of evaluating blocking probability, mean service time of user requests, and utilization rate of base stations by taking into account practical concerns in mmWave networks including blockages encountered by mmWaves in the physical layer, capacity constraints in the network layer, and the stochastic nature of mobile traffic. We demonstrate by numerical results that our analytical method is accurate and computationally-efficient.**

***Index Terms*—Millimeter wave (mmWave) networks, cross-layer analytical framework, performance analysis, teletraffic model, fixed-point approximation.**

I. Introduction

Academia and industry have been turning to the under-utilized millimeter wave (mmWave) bands (30–300 GHz) to address the current scarcity of wireless broadband spectrum [1]. Broadband transmission in the mmWave bands enables fifth-generation (5G) wireless networks to meet stringent throughput, reliability and latency requirements [2]. However, it also poses new challenges, as the weak penetration and diffraction of mmWave radio propagation makes it susceptible to physical blockages [3]. This problem is exacerbated in densely populated urban areas where pedestrians, vehicles or buildings would frequently block the line-of-sight (LoS) or low-order reflection paths between the base station (BS) and mobile users (MUs). Because of the weak propagation and penetration abilities of mmWave, mmWave BSs will be deployed to form small cells [4]. However, due to the increasing amount of network traffic, limited capacity at each small cell makes the system vulnerable to capacity-limited blocking. Both physical blockages and capacity limits may lead to connection failures between BSs and MUs [1]. The connectivity issue adversely affects user experience, which can be quantitatively measured by Quality of Service (QoS) metrics such as blocking probability, which is defined as the ratio of user requests that cannot be served by any BS due to either physical layer blockage (PLB) or capacity-limited blocking (CLB).

978-1-7281-4846-5/20/$31.00 ©2020 IEEE

Intuitively, deploying more BSs reduces both PLB and CLB. However, over-deployment of BSs leads to unnecessary deployment cost and low utilization rate of deployed BSs. The latter is more significant in 5G use cases, such as Augmented Reality (AR), machine type communications or industrial automation, where the traffic requiring reliable transmissions are usually sporadic. As 5G covers a wide range of such new use cases that require reliable transmissions, accurate evaluations of the blocking probability, mean service time of MU requests, and utilization rate are necessary for mobile network operators to make planning, design, dimensioning and operational decisions, such as BS deployment [5], BS sleeping [6] and resource reservation [7], to improve network performance or reduce power consumption. As a large number of evaluations of different configurations are required in such applications, computationally-efficient analytical methods are usually preferred over straightforward but time-consuming computer simulations, especially when the scale of the considered network is large.

PLB in terms of path blockage probability in urban mmWave networks has been analytically characterized for the LoS [8] and first-order reflection [9] paths based on stochastic geometry. On the other hand, teletraffic models can be applied to capture the stochastic dynamics of network traffic to address CLB and other network-layer performance metrics in mobile wireless networks. Particularly, appropriate queue size thresholds were derived in [10] for load sharing in an integrated wireless network by modeling a wireless channel as a finite-buffer M/G/1/k queue. In [11], closed-form expressions were obtained for evaluating the mean energy consumption and mean request service time in a BS for different BS sleeping schemes based on a single vacation queue model.

The stringent QoS requirement of 5G applications makes BS-cooperative association and scheduling necessary in the case of PLB or CLB [2]. For example, if physical obstacles such as pedestrians and vehicles block the transmission path between an MU and a BS for a relatively long time compared to the latency requirement of an MU request, it is impossible for the MU to wait for the path to clear in order to transmit to the BS. Instead, such requests should be directed to other nearby BSs for service. This creates interaction of traffic in different BSs and invalidates single-server queuing models as in [10] and [11] for the purpose of performance evaluation.

In this paper, we propose a cross-layer analytical framework that accounts for both PLB in the physical layer and CLB in the network layer to evaluate the blocking probability, mean service time of MU requests and utilization rate based on stochastic geometry, teletraffic theory, and the classical and computationally-efficient Erlang Fixed Point Approximation (EFPA) method [12]–[14], providing further insight into research involving these measurements in such networks. With the integration of these tools, our method can capture the dynamics of stochastic network traffic, as well as important features of 5G mmWave networks including PLB and CLB, and a dynamic user association mechanism that determines the BS to serve an MU request based on the states of the BSs upon the arrival of the request. The fundamental queuing model used in this paper is the M/G/k/k queue, which is an appropriate model for wireless networks with Orthogonal Frequency Division Multiplexing (OFDM) [15], a classical modulation method that is expected to be applied in 5G networks as well [16].

To the best of our knowledge, there are no existing analytical methods that can accurately and efficiently address the blocking constraints in both the physical and network layers, thus capturing both PLB and CLB in a mmWave network with dynamic user association mechanisms, while performance evaluation by simulation is too slow for practical large-scale 5G networks. Therefore, the main contribution of our proposed cross-layer analytical framework is to provide a comprehensive, computationally-efficient and realistic evaluation of the utilization rate and blocking probability in 5G mmWave networks, which may in turns lead to useful insights in network planning, dimensioning, and optimization for such networks.

II. System Model

We consider a 2D multi-BS network and denote the set of BSs as $\mathcal{U} = \{1, 2, ..., U\}$. Each BS $i \in \mathcal{U}$ is deployed at a known location with coordinates (x_i, y_i), and has a finite capacity of c_i. By "capacity", we mean that the radio resources in BS i are allowed to serve at most c_i MU requests concurrently. This guarantees that sufficient radio resources can be allocated to each admitted MU request such that their reliability and latency requirements can be satisfied, since a large value of c_i means more requests may be admitted simultaneously, leading to less radio resources for each request and therefore higher average latency. We focus on downlink delay-sensitive traffic generated by the applications mentioned in Section I, such that a finite capacity is necessary to guarantee that sufficient amount of radio resources can be allocated to each admitted request. We also consider that such delay-sensitive traffic is always given preemptive priority over delay-tolerant traffic for transmission, such that the transmission of delay-sensitive traffic is never affected by the existence of delay-tolerant traffic. The utilization rate of a BS is defined as the time-average proportion of the occupied capacity.

We further assume that MU requests are generated according to a spatio-temporal Poisson Point Process (PPP). Each BS $i \in \mathcal{U}$ covers a circular area with a radius of D_i. As the delay-sensitive traffic is usually composed of small packets and thus the transmission time of a single request is relatively short [2], we can assume that, for the duration of an MU request, both the MU itself and the obstacles remain static. As a consequence, there is no disruption due to physical blockage once a connection between a BS and an MU is established. Finally, we consider that the physical blockages experienced by a link are independent [8], [9].

A. Physical Layer: Stochastic Geometry and PLB

As discussed in Section I, mmWave transmission is inherently susceptible to physical link blockages. Therefore, the *effective arrival rate* of MU requests at BS i depends on both the generation rate of requests, and physical layer parameters such as the spatial distribution of the BSs and obstacles.

As in [8], we assume that the obstacles are impenetrable rectangles with independently and identically distributed (i.i.d.) lengths and widths with means $\mathbb{E}[L]$ and $\mathbb{E}[W]$, respectively, and with center points generated by a homogeneous PPP with density λ_b. Therefore, the total number of obstacles blocking an LoS path with length d is Poisson distributed with mean $\beta d + p$, where $\beta = 2\lambda_b(\mathbb{E}[L] + \mathbb{E}[W])/\pi$ and $p = \lambda_b \mathbb{E}[L]\,\mathbb{E}[W]$ are two parameters related to the size and density of the obstacles. In this paper, we consider only LoS transmissions, thus the PLB probability between BS i and a location with coordinate $\mathbf{z} = (x_z, y_z)$ is equivalent to the probability that at least one obstacle is present across the LoS path, namely

$$B_B(i, \mathbf{z}) = \begin{cases} 1 - \exp\left[-\left(\beta d\left(i, \mathbf{z}\right) + p\right)\right] & \text{if } d\left(i, \mathbf{z}\right) \leq D_i; \\ 1 & \text{if } d\left(i, \mathbf{z}\right) > D_i, \end{cases} \tag{1}$$

where $d(i, \mathbf{z})$ is the Euclidean distance between BS i and $\mathbf{z}$ [8], and D_i is the radius of the BS's coverage region. Note that the PLB is dependent on the distance between the MU and BS alone and not on the absolute coordinates.

Let $\lambda(i, \mathbf{z})$ denote the intended arrival rate from coordinate $\mathbf{z}$ to BS i. The total intended arrival rate for BS i is $\lambda_i = \int_{\mathbf{R}_i} \lambda\left(i, \mathbf{z}\right) d\mathbf{z}$, where $\mathbf{R}_i = \{\mathbf{z} | d(i, \mathbf{z}) \leq D_i\}$ is the set of locations within the coverage region of BS i. However, MU requests encountering PLB will not be able to establish a connection with the BS and thus do not contribute to the traffic offered to the BS. The effective arrival rate for BS i, after encountering PLB, is $\lambda_i^* = \int_{\mathbf{R}_i} \lambda\left(i, \mathbf{z}\right)\left(1 - B_B\left(i, \mathbf{z}\right)\right) d\mathbf{z}$.

We define a *candidate BS* for a given MU request as any BS covering the location of the MU initiating the request. As the coverage of BSs may overlap, multiple candidate BSs are possible for a single request, and a request may be served by (associated to) any single candidate BS. We define *fresh traffic* as all requests making their first attempt to a candidate BS. If the attempt is not successful due to either PLB or CLB, the request will attempt (overflow to) another candidate BS that it has not attempted. This process continues until 1) the MU successfully connects to a BS and begins its service, or 2) the

request is rejected by all candidate BSs and leaves the network unserved. We further define *overflow traffic* as all requests that have attempted and been rejected from at least one BS. By definition, λ_i^* only includes *fresh traffic*.

Denoting the average transmission power of BS i to a single MU request as P_i, the received power of a request at coordinate $\mathbf{z}$ from BS i, when PLB does not occur, is

$$P_r(i,\mathbf{z}) = gd(i,\mathbf{z})^{-\alpha} P_i, \tag{2}$$

where g is a parameter reflecting the beamforming gain and small-scale fading effect, and α is the path loss exponent. The received power is 0 if PLB occurs.

For now, we ignore the reflection paths transmissions and dependencies among the paths as the focus of this paper is on the cross-layer modeling and performance evaluation of mmWave networks. The impact of reflection paths on overall network performance can be incorporated to our framework by replacing (1) with relevant analysis (e.g. [9]) to obtain the PLB, and consider a different loss-path exponent for reflection paths in (2) for calculating the received power.

B. Network Layer: Teletraffic Models and CLB

Assume that the number of data bits requested to transmit by each MU request in the network, Φ, is i.i.d. following a general distribution with mean $\mathbb{E}[\Phi]$, independent of time or location. Let $k_{\mathbf{z}}$ denote the set of candidate BSs for location $\mathbf{z}$, which depends on the location and coverage of each BS. We consider that each BS will cause interference to MU requests within its coverage but served by other BSs, and consider the upper bound on the level of interference for each MU request, where each potential beam from an interfering BS aims directly at the MU making the request, such that the interference is maximized. We compute the data rate for a request generated by an MU at location $\mathbf{z}$ served by BS i by

$$\bar{r}(i,\mathbf{z}) = N\log_2\left[1+\frac{P_r(i,\mathbf{z})}{N_0+\sum_{j\in k_{\mathbf{z}}}^{j\neq i} P_r(j,\mathbf{z})}\right], \tag{3}$$

where N_0 is the thermal noise, and N is the bandwidth equally allocated to each request. The last term in (3) is the average SINR experienced by such requests.

The mean service time (referred as mean holding time or delay in some other research, which is defined as the time elapsed from the moment when an MU initiates a request to the moment when the MU received the last required bit from the BS) of all requests connected to BS i is the ratio of the average number of data bits per request to the mean data rate, namely

$$\bar{t}_i = \frac{1}{\lambda_i^*}\int_{\mathbf{R}_i}\frac{\lambda(i,\mathbf{z})(1-B_B(i,\mathbf{z}))\mathbb{E}[\Phi]}{\bar{r}(i,\mathbf{z})}d\mathbf{z}. \tag{4}$$

In this sense, BS i can be modeled as an M/G/c_i/c_i queue with mean arrival rate of requests λ_i^* and mean service rate of requests $\bar{\mu}_i = 1/\bar{t}_i$, where the service time follows a general distribution [1]. We assume OFDM transmission such that intra-cell interference is eliminated. It was also mentioned in existing work such as [15] and [17] that loss systems like M/G/c_i/c_i queues are appropriate models for wireless networks with OFDM-FDMA and OFDM-TDMA transmissions. As discussed previously, by limiting the value of c_i, the network can provide delay guarantees.

When an MU initiates a request for transmission, it will begin by attempting a candidate BS. The order by which the request attempts its candidate BSs is determined by a user association strategy. If the request does not experience PLB during the attempt, it will examine whether the candidate BS has already reached the maximum capacity. If so, the incoming request will be rejected by the candidate BS due to CLB.

If we define $A_i = \lambda_i^*/\bar{\mu}_i$ as the effective offered traffic in Erlangs to BS i, the CLB probability is equivalent to the probability that there are c_i requests served in the BS, i.e., when the BS is "full" and has to reject a newly-initiated request for admission. By the Erlang B formula, this can be calculated as

$$B_C(A_i,c_i) = \frac{\frac{A_i^{c_i}}{c_i!}}{\sum_{y=0}^{c_i}\frac{A_i^y}{y!}}. \tag{5}$$

Note that in an M/G/c_i/c_i queue, the mean service time and blocking probability are insensitive to the distribution of Φ [18]. Additionally, since the queue has a finite buffer, the system is stable even when the arrival rate is greater than the service rate.

The utilization rate for BS i is given by [18]

$$\hat{U}_i = \frac{(1-B_C(A_i,c_i))A_i}{c_i}. \tag{6}$$

Without overflow traffic among candidate BSs, we can calculate blocking probability, mean service time and utilization rate for each BS individually by the above equations. However, overflow traffic is an intrinsic feature of wireless networks with dynamic user association mechanisms, which is essential for mmWave networks as discussed in Section I. Therefore, capturing overflow traffic is crucial for the accurate evaluation of blocking probability, mean service time and utilization rate in such networks. Such evaluation requires analysis of traffic interaction among a network of BSs, which is provided in the next section.

III. Performance Evaluation

In this section, we apply EFPA to address the overflow traffic. By integrating EFPA with our analysis in the previous section on PLB and CLB, we propose a novel analytical framework to evaluate the blocking probability of MU requests and utilization rate of BSs in urban 5G mmWave networks.

EFPA was initially proposed for approximating the blocking probability (CLB in our case) in overflow loss systems where mutual overflow traffic exists among non-hierarchical subsystems. The key idea of EFPA is to assume that both fresh and overflow traffic are Poisson, and decompose the entire system into independent Erlang B subsystems, with each

[1]The M/D/k/k and M/M/k/k queues discussed in [2] for ultra-reliable low-latency traffic are special cases of the M/G/c_i/c_i queue described in this paper, with deterministic and exponential distributions of service times respectively.

server group as a subsystem. Kelly [12] suggested that EFPA can be applied in cellular networks with channel borrowing capabilities, where the entire network is a system while each BS is an independent subsystem. The concept of channel borrowing is still applicable in modern cellular networks, including 5G mmWave networks, with dynamic user association mechanisms [14]. It is noted that there is some room for improvement on the accuracy of EFPA when the blocking probability is lower than 1% [13], [14], [19]. However, as the PLB in a typical mmWave network is around 10% [8], which constitutes the lower bound of the overall blocking probability, the original EFPA is sufficient in estimating the overall blocking probability (PLB and CLB combined).

For simplicity, in this paper we will use the Shortest Distance First (SDF, or equivalently the Smallest Path-Loss First) user association mechanism, in which a request attempts candidate BSs in an increasing order of distance. This mechanism favors a BS that can offer the highest signal power to the MU that initiates the request, and is considered as a state-of-art benchmark in literature [20]. In the rest of the section, we apply EFPA to evaluate the blocking probability and mean service time under the SDF mechanism. However, the proposed analytical framework is general and can apply to other user association mechanisms as well.

We start by defining the following:

- Δ: the sequence of attempted candidate BSs of an MU request. $|\Delta|$ is the number of candidates that the request has attempted;
- $\mathbf{S}_{\mathbf{z},n} = \{s_{\mathbf{z},1}, ..., s_{\mathbf{z},n}\}$: the sequence of the n closest BSs to location $\mathbf{z}$, with $s_{\mathbf{z},1}, ..., s_{\mathbf{z},n} \in \mathcal{U}$ and $d(s_{\mathbf{z},1}, \mathbf{z}) \leq d(s_{\mathbf{z},2}, \mathbf{z}) \leq ... \leq d(s_{\mathbf{z},n}, \mathbf{z})$;
- $a'_{\mathbf{z},n}$: Intended offered traffic for BS $s_{\mathbf{z},n}$ from location $\mathbf{z}$ with $\Delta = \mathbf{S}_{\mathbf{z},n-1}$ and $|\Delta| = n-1$;
- $a_{\mathbf{z},n} = a'_{\mathbf{z},n}\left(1 - B_B\left(s_{\mathbf{z},n}, \mathbf{z}\right)\right)$: Effective offered traffic, after accounting for PLB, to BS $s_{\mathbf{z},n}$ from location $\mathbf{z}$ with $\Delta = \mathbf{S}_{\mathbf{z},n-1}$ and $|\Delta| = n-1$;
- $a_{i,n} = \int_{\mathbf{R}_i} a_{\mathbf{z},n} d\mathbf{z}$: Effective offered traffic to BS i with $|\Delta| = n-1$, summing over all possible locations $\mathbf{z}$ with $\mathbf{S}_{\mathbf{z},n} = i$;
- $\lambda_{i,n}$: Arrival rate to BS i with $|\Delta| \leq n$. By definition, $\lambda_{i,0} = \lambda_i^*$ is the arrival rate of the fresh traffic offered to BS i;
- $A_{i,n} = \sum_{j=0}^{n} a_{i,j}$: Effective offered traffic to BS i with $|\Delta| \leq n$;
- $v_{\mathbf{z},n}$: Overflow traffic from BS $s_{\mathbf{z},n}$, offered from location $\mathbf{z}$, with $\Delta = \mathbf{S}_{\mathbf{z},n}$ and $|\Delta| = n$;

Given $A_{i,U-1}$, the blocking probability at BS i is $b_i = B_C\left(A_{i,U-1}, c_i\right)$ can be obtained by the Erlang B formula as

$$b_i = B_C\left(A_{i,U-1}, c_i\right) = \frac{\frac{A_{i,U-1}^{c_i}}{c_i!}}{\sum_{y=0}^{c_i} \frac{A_{i,U-1}^{y}}{y!}}, \tag{7}$$

as any single request is allowed to overflow at most $U-1$ times. A request will be blocked and leave the network without being served when it is rejected by the U-th candidate BS.

The overflow traffic with $|\Delta| = n$ from BS $s_{\mathbf{z},n}$ is

$$v_{\mathbf{z},n} = a_{\mathbf{z},n} b_{s_{\mathbf{z}}} + a'_{\mathbf{z},n} B_B(s_{\mathbf{z},n}, \mathbf{z}) \tag{8}$$

for $n < U$. In (8), the first term represents the overflow traffic when an incoming request with $|\Delta| = n-1$ finds BS $s_{\mathbf{z},n}$ fully occupied with other requests in service. The second term represents the overflow traffic due to PLB, is unique to the mmWave network model in this work. The overflow traffic $v_{\mathbf{z},n}$ adds $s_{\mathbf{z},n}$ in the attempted sequence and thus increase $|\Delta|$ by 1, and will be offered to the next closest unattempted candidate BS $s_{\mathbf{z},n+1}$ as $a'_{\mathbf{z},n+1}$. After accounting for different data rates offered by $s_{\mathbf{z},n+1}$ and $s_{\mathbf{z},n}$, we have

$$a'_{\mathbf{z},n+1} \bar{r}(s_{\mathbf{z},n+1}, \mathbf{z}) = v_{\mathbf{z},n} \bar{r}(s_{\mathbf{z},n}, \mathbf{z}), \tag{9}$$

for $n < U-1$. For any $n \geq U-1$, $a'_{\mathbf{z},n+1} = 0$. By (7) to (9), we can iteratively calculate the CLB probability for requests with $|\Delta| = n$ and the amount of overflow traffic with $|\Delta| = n+1$ at each BS, starting with the initial values $a_{i,0} = A_{i,0} = \lambda_{i,0}/\bar{\mu}_i$. Due to the circular dependencies between $a_{\mathbf{z},n}$, $v_{\mathbf{z},n}$ and b_i as in (7) to (9), a fixed-point iteration is needed to determine their values [12]–[14], [19].

After the fixed-point solutions for $A_{i,U-1}$ and b_i are obtained for every $i \in \mathcal{U}$, the overall blocking probability of the network can be computed as

$$\hat{B} = 1 - \frac{\sum_{i \in \mathcal{U}} A_{i,U-1}\left(1 - b_i\right)}{\sum_{i \in \mathcal{U}} \left(\lambda_i / \bar{\mu}_i\right)}, \tag{10}$$

where $\sum_{i \in \mathcal{U}} \left(\lambda_i / \bar{\mu}_i\right)$ is the total intended offered traffic for all BSs in the network, and $C_i = A_{i,U-1}\left(1 - b_i\right)$ is the carried traffic of BS i (which includes all the traffic with smaller $|\Delta|$ by definition). The network throughput, or total carried traffic, is $\sum_{i \in \mathcal{U}} C_i$. For an individual BS i, the total effective offered traffic $A_{i,U-1}$ and the CLB probability b_i can also be obtained.

Taking overflow traffic into consideration, the utilization rate of BS i is calculated by replacing A_i with $A_{i,U-1}$ and $B_C\left(A_i, c_i\right)$ with b_i in (6), that is,

$$\hat{U'}_i = \frac{\left(1 - b_i\right) A_{i,U-1}}{c_i}. \tag{11}$$

The utilization rate of the network is the weighted average of the utilization rate of each BS, namely,

$$\hat{U'} = \frac{\sum\limits_{i \in \mathcal{U}} c_i \hat{U'}_i}{\sum\limits_{i \in \mathcal{U}} c_i} \tag{12}$$

The mean service time of requests is also affected by the overflow traffic, as requests may be served by further BSs and thus experience a lower mean data rate. For an individual BS i, the mean service time can be calculated from the total effective offered traffic $A_{i,U-1}$ and the CLB probability b_i obtained by the above iterative process. Specifically, the mean service time of requests served by each BS is obtained by Little's Law as $\mathbb{E}\left[t_i\right] = \mathbb{E}\left[Q_i\right] / \lambda_{i,U-1}$, where $\mathbb{E}[Q_i] = \sum_{x=1}^{c_i} x(A_{i,U-1}^x / x!) / \sum_{y=0}^{c_i} (A_{i,U-1}^y / y!)$ is

the average number of requests (i.e., the mean queue size of an M/G/c_i/c_i queue with offered traffic $A_{i,U-1}$) at BS i [18]. For the network, the mean service time is

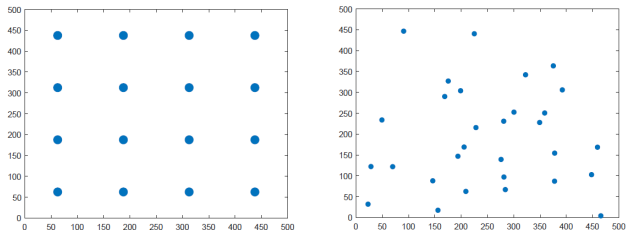
$$\mathbb{E}[t] = \frac{\sum\limits_{i\in\mathcal{U}} \lambda_{i,U-1}(1-b_i)\mathbb{E}[t_i]}{\sum\limits_{i\in\mathcal{U}} \lambda_{i,U-1}(1-b_i)}. \tag{13}$$

EFPA-based methods have been demonstrated to be computationally efficient [13], [14], [19]. We will report the improvement in running time of EFPA over simulation along with the accuracy of EFPA in the next section.

IV. Numerical Results

We consider two different BS layouts for a network with dimensions $500\,\mathrm{m} \times 500\,\mathrm{m}$. The first layout contains 16 BSs deployed in a rectangular grid and the other contains 30 BSs with locations randomly generated, as shown in Fig. 1. For the ease of demonstration, we assume, without loss of generality, that the network is discretized into $1\,\mathrm{m} \times 1\,\mathrm{m}$ grids so that x_z and y_z are all integers. We set $g = 1$, $\alpha = 3$, $N = 1\,\mathrm{GHz}$, $N_0 = -90\,\mathrm{dBm}$, and let Φ follows an exponential distribution with $\mathbb{E}[\Phi] = 500\,\mathrm{Mbits}$. We generate a set of random obstacles in each layout, with $\mathbb{E}[W] = 1\,\mathrm{m}$, $\mathbb{E}[L] = 4\,\mathrm{m}$, and $\lambda_b = 0.003$. We further assume homogeneous offered traffic in each grid, denoted as $\lambda(\mathbf{z}) = \lambda/(500 \times 500)$, where λ is the generation rate of requests per second in the whole network. Additionally, we set the radius of coverage the same for all BSs in the network, denoted by $D = D_i = 100\,\mathrm{m}$ for all $i \in \mathcal{U}$. The maximum number of requests allowed is $c_i = c = 50$, and the average transmit power is $P_i = P = 30\,\mathrm{dBm}$.

We perform discrete-event simulations [21] by MATLAB to obtain simulation results. The locations of BSs are fixed in all simulation runs. The 95% confidence intervals for all simulation results presented in this section, based on Student's t-distribution, are within 1% of the observed mean.



(a) (b)

Figure 1. Network layout with (a) 16 BSs deployed in a rectangular grid; (b) 30 BSs with randomly generated locations.

Simulation and analytical results for blocking probability, mean service time and utilization rate derived from Section III are presented in Fig. 2 for the rectangular BS layout and Fig. 3 for the random BS layout.

As shown in both Figs. 2a and 3a, when the arrival rate is low, the probability of CLB is very small. The network is considered to be within the PLB-constrained regime where the blocking probability is approximately equal to the probability of PLB (around 20% in both cases and not dependent on the arrival rate). This is the only case considered in most existing studies on physical layer blockage (e.g. [8], [9]). However, as the arrival rate increases, the network will move to the mixed regime where both PLB and CLB are significant, due to the finite capacities of BSs. In this case, blocking probability evaluations only based on PLB will underestimate the overall blocking probability of the network. As 5G networks are expected to accommodate large volume of traffic, this scenario cannot be ignored. Such underestimations may lead to inaccurate formulation and solution of relevant problems.

In the mixed regime where the CLB is significant, the mean service time also increases as the arrival rate increases in both layouts as shown in Figs. 2b and 3b. This is because that MU requests are more likely to overflow to further BSs for service under heavy traffic condition, resulting in a lower data rate and thus a longer mean service time.

For the network with rectangular BS layout as shown in Fig. 2, the relative difference in blocking probability between EFPA and simulation are within 1.5% for all values of arrival rates in the considered range. The results indicate that EFPA provides fairly accurate estimates of blocking probability, mean service time and utilization rate for the rectangular BS layout. For the network with random BS layout as shown in Fig. 3, EFPA gives relatively conservative (higher) estimations of blocking probability and mean service time compared to simulation results, which is preferable and often adopted for design and optimization problems.

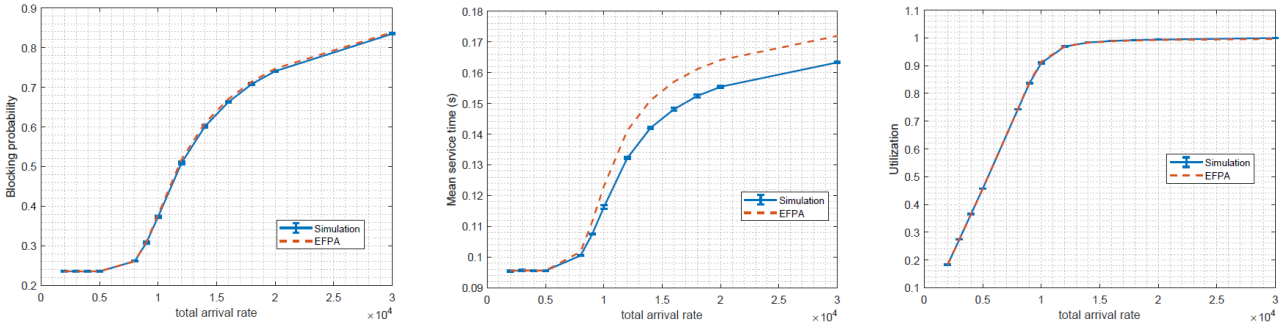
The running time of EFPA is, on average, shorter than 10% that of simulation. This makes EFPA well-suited for network planning and optimization scenarios where a large number of network configurations must be compared.

V. Conclusion

We propose an accurate and computationally-efficient cross-layer analytical framework to compute the blocking probability and mean service time in 5G mmWave networks. We demonstrate that our results provide reasonable estimates for the Shortest-Distance-First user association mechanism. The results obtained by our analytical model based on EFPA will provide useful insights in network planning, dimensioning, and optimization. A potential future extension is to integrate power control and opportunistic user scheduling in our model to further improve the network performance. Another possible extension is to apply new approximation methods which are based on EFPA but have incorporated improvements to address certain features of specific models (e.g., [13], [14], [19]), to further increase the accuracy of evaluations.

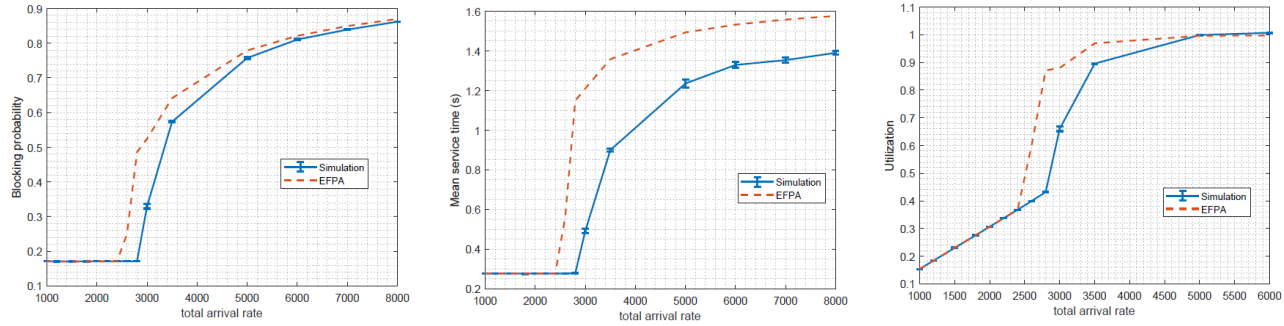
Acknowledgement

The work described in this paper was partly supported by College Research Grant from BNU-HKBU United International College [UIC-R201811] and a grant from the Innovation and Technology Funding (ITF) of the Hong Kong Special Administrative Region, China [ITS/359/17].



(a) (b) (c)

Figure 2. (a) Blocking probability; (b) mean service time; and (c) utilization rate for the network with rectangular BS layout.



(a) (b) (c)

Figure 3. (a) Blocking probability; (b) mean service time; and (c) utilization rate for the network with random BS layout.

REFERENCES

[1] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tut.*, vol. 18, no. 3, pp. 1617–1655, third quarter 2016.

[2] Chih-Ping Li, Jing Jiang, W. Chen, Tingfang Ji, and J. Smee, "5G ultra-reliable and low-latency systems design," in *2017 European Conference on Networks and Communications (EuCNC)*, June 2017, pp. 1–5.

[3] S. Hur, T. Kim, D. J. Love, J. V. Krogmeier, T. A. Thomas, and A. Ghosh, "Millimeter wave beamforming for wireless backhaul and access in small cell networks," *IEEE Trans. Commun.*, vol. 61, no. 10, pp. 4391–4403, October 2013.

[4] M. Marcus and B. Pattan, "Millimeter wave propagation: spectrum management implications," *IEEE Microwave Magazine*, vol. 6, no. 2, pp. 54–62, June 2005.

[5] M. Dong, T. Kim, J. Wu, and E. W. M. Wong, "Cost-efficient millimeter wave base station deployment in Manhattan-type geometry," *IEEE Access*, pp. 1–1, 2019.

[6] J. Wu, E. W. M. Wong, Y. C. Chan, and M. Zukerman, "Energy efficiency-QoS tradeoff in cellular networks with base-station sleeping," in *IEEE GLOBECOM 2017*, Dec 2017, pp. 1–7.

[7] M. Wang, S. Li, E. W. M. Wong, and M. Zukerman, "Blocking probability analysis of circuit-switched networks with long-lived and short-lived connections," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 5, no. 6, pp. 621–640, June 2013.

[8] T. Bai, R. Vaze, and R. W. Heath, "Analysis of blockage effects on urban cellular networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 9, pp. 5070–5083, Sept 2014.

[9] M. Dong and T. Kim, "Interference analysis for millimeter-wave networks with geometry-dependent first-order reflections," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 12 404–12 409, Dec 2018.

[10] W. Song and W. Zhuang, "Multi-service load sharing for resource management in the cellular/WLAN integrated network," *IEEE Trans. Wireless Commun.*, vol. 8, no. 2, pp. 725–735, Feb. 2009.

[11] X. Guo, Z. Niu, S. Zhou, and P. R. Kumar, "Delay-constrained energy-optimal base station sleeping control," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1073–1085, May 2016.

[12] F. Kelly, "Blocking probabilities in large circuit-switched networks," *Advances in Applied Probability*, vol. 18, pp. 473–505, 1986.

[13] E. W. M. Wong, J. Guo, B. Moran, and M. Zukerman, "Information exchange surrogates for approximation of blocking probabilities in overflow loss systems," in *Proc. The 25th International Teletraffic Congress (ITC)*, Sep. 2013.

[14] J. Wu, E. W. M. Wong, J. Guo, and M. Zukerman, "Performance analysis of green cellular networks with selective base-station sleeping," *Perform. Eval.*, vol. 111, pp. 17–36, May 2017.

[15] Jui-Chi Chen and Wen-Shyen E Chen, "Call blocking probability and bandwidth utilization of OFDM subcarrier allocation in next-generation wireless networks," *IEEE Commun. Lett.*, vol. 10, no. 2, pp. 82–84, Feb 2006.

[16] P. Guan, D. Wu, T. Tian, J. Zhou, X. Zhang, L. Gu, A. Benjebbour, M. Iwabuchi, and Y. Kishiyama, "5G field trials: OFDM-based waveforms and mixed numerologies," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 6, pp. 1234–1243, June 2017.

[17] D. Niyato and E. Hossain, "Queueing analysis of OFDM/TDMA systems," in *IEEE GLOBECOM 2005*, vol. 6, Nov 2005.

[18] M. Zukerman, "Introduction to queueing theory and stochastic teletraffic models." [Online]. Available: http://www.ee.cityu.edu.hk/~zukerman/classnotes.pdf

[19] Y. C. Chan and E. W. M. Wong, "Blocking probability evaluation for non-hierarchical overflow loss systems," *IEEE Trans. Commun.*, vol. 66, no. 5, pp. 2022–2036, May 2018.

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

[21] M. Mezzavilla, M. Zhang, M. Polese, R. Ford, S. Dutta, S. Rangan, and M. Zorzi, "End-to-end simulation of 5G mmWave networks," *IEEE Commun. Surveys Tut.*, vol. 20, no. 3, pp. 2237–2263, thirdquarter 2018.