Out-of-Band Millimeter Wave Beamforming and Communications to Achieve Low Latency and High Energy Efficiency in 5G Systems

Morteza Hashemi, *Student Member, IEEE,* C. Emre Koksal, *Senior Member, IEEE,* and Ness B. Shroff, *Fellow, IEEE*

***Abstract*—Communications in the millimeter wave (mmWave) band faces significant challenges due to variable channels, intermittent connectivity, and high energy usage. Moreover, speeds for electronic processing of data is of the same order as typical rates for mmWave interfaces, making the use of complex algorithms for tracking channel variations and adjusting resources impractical. In order to mitigate some of these challenges, we propose an architecture that integrates the sub-6 GHz and mmWave technologies. Our system exploits the spatial correlations between the sub-6 GHz and mmWave interfaces for *beamforming* and *data transfer*. Based on extensive experimentation in indoor and outdoor settings, we demonstrate that analog beamforming can be used in mmWave without incurring large overhead, thanks to the spatial correlations with sub-6 GHz. In addition, we incorporate the sub-6 GHz interface as a fallback (secondary) data transfer mechanism such that (i) the negative effects of highly intermittent mmWave connectivity is mitigated, and (ii) the abundant mmWave capacity is fully exploited. To achieve these goals, we consider the problem of scheduling the arrival traffic over the mmWave or sub-6 GHz in order to maximize the mmWave throughput while delay (due to mmWave outages) is guaranteed to be bounded. We prove using subadditivity analysis that the optimal scheduling policy is based on a single threshold that can be easily adopted despite high link variations. Numerical results demonstrate that our scheduler provides a bounded mmWave delay performance, while it achieves a similar throughput performance as the throughput-optimal policies (e.g., MaxWeight).**

***Index Terms*—Millimeter wave communication, 5G mobile systems, Out-of-band beamforming and communication**

I. Introduction

The annual data traffic generated by mobile devices is expected to surpass 130 exabits by 2020 [1]. This deluge of traffic will significantly exacerbate the spectrum crunch that cellular providers are already experiencing. To address this issue, it is envisioned that in 5G cellular systems certain portions of the mmWave band will be used, spanning the spectrum between 30 GHz to 300 GHz with the corresponding wavelengths between 1-10 mm [2]. This will substantially increase the spectrum available to cellular providers, which is currently between 700 MHz and 2.6 GHz with only 780 MHz of bandwidth allocation for all current cellular technologies.

M. Hashemi and C. E. Koksal are with the Department of Electrical and Computer Engineering, Ohio State University, Columbus, OH, 43210 USA e-mail: hashemi.20@osu.edu, koksal.2@osu.edu

N. B. Shroff is with the Department of Electrical and Computer Engineering and Department of Computer Science and Engineering, Ohio State University, Columbus, OH, 43210 USA e-mail: shroff.11@osu.edu.

Before mmWave communications can become a reality, there are significant challenges that need to be overcome. Firstly, compared with sub-6 GHz, the propagation loss in mmWave is much higher due to atmospheric absorption and low penetration. Although large antenna arrays have the potential to make up for the mmWave losses, they cause several other issues such as high energy consumption by components (e.g., analog-to-digital converters (ADC)). For instance, at a sampling rate of 1.6 Gsamples/sec, an $8-$bit quantizer consumes ≈ 250mW of power. During active transmissions, this would constitute up to 50% of the overall power consumption for a typical smart phone. Moreover, in order to fully utilize the large directional antenna arrays, continuous beamforming and signal training at the receiver is needed [3]. Digital beamforming is highly efficient in delay, but there is a need for a separate ADC for each antenna, which may not be feasible for even a small to mid-sized antenna array due to high energy consumption. In contrast, analog beamforming requires only one ADC, but it can focus on one direction at a time, making the angular search costly in delay. There are also proposals on hybrid digital/analog beamforming, which strikes a balance between analog and digital beamforming, using a few ADCs rather than one per antenna [4].

In order to remedy the high beamforming overhead of mmWave, we exploit its spatial correlations with sub-6 GHz. In particular, due to high cost and energy consumption by ADCs in fully-digital beamforming as well as the delay in fully-analog beamforming, we investigate the feasibility of conducting a coarse angle of arrival (AoA) estimation on the sub-6 GHz channel and then utilizing the fully-analog beamforming for fine tuning and transmissions. To this end, we first experimentally verify the correlation between the sub-6 GHz and mmWave AoA, especially in the presence of line-of-sight (LOS). Our measurements taken jointly at different bands and for both *indoor and outdoor settings* show that under LOS conditions and in 94% of the measurements, the identified AoA of signal in the sub-6 GHz band is within $\pm 10^\circ$ accuracy for the AoA of the mmWave signal. Based on the estimated sub-6 GHz AoA, the angular range over which we scan for the mmWave transmitter reduces to no more than 20° on average, from 180° in stand-alone mmWave systems. Note that the authors in [5] also proposed a beamforming method based on out-of-band measurements for 60 GHz WiFi and under *static indoor* conditions.

In addition to beamforming overhead, mmWave channels

can be highly variable with intermittent connectivity since most objects lead to blocking and reflections as opposed to scattering and diffraction in typical sub-6 GHz frequencies. When the users and/or surrounding objects are mobile, different propagation paths become highly variable with intermittent on-off periods, which can potentially result in long outages and poor mmWave delay performance. To mitigate this issue, we envision an integrated sub-6 GHz/mmWave transceiver model, in which, in addition to beamforming, the sub-6 GHz interface is used as a fallback (secondary) data transfer mechanism. We note that the link speed of the mmWave interface (multi-Gbps) is comparable to the speed at which a typical processor in a smart device operates. This is different from classical wireless interfaces in which data rates are much smaller than the clock speeds of the processors. Thus, the mmWave interface cannot be assumed to operate at smaller time-scales and the algorithms run at the processor may not be able to respond to variations in real time and execute control decisions. *This necessitates the use of proactive queue-control solutions along with a reasonably large buffer at the mmWave interface.* For instance, if the queue size at the mmWave interface gets small, the risk of wasting the abundant capacity from mmWave increases. Conversely, if we keep the queue at the mmWave interface large, if the channel goes down, we incur a high delay.

To understand the tradeoff between full exploitation of the mmWave capacity and delay for the mmWave channel access, *we model the sub-6 GHz/mmWave transceiver as a communication network, and investigate an optimal scheduling policy using network optimization tools.* Specifically, in the equivalent network model, the sub-6 GHz and mmWave interfaces are represented by individual network nodes with dedicated queues. Hence, the optimal transmission policy across the sub-6 GHz and mmWave interfaces is transformed into an optimal scheduling policy across the sub-6 GHz and mmWave nodes. Our experimental results from the first part demonstrate that under the LOS conditions, there is about $10-15$ dB channel gain improvement due to the strongest eigenmode. Therefore, the state of the strongest eigenmode can be used to determine the availability of the mmWave link. Our experimental observations provide guideline to model the mmWave channel with a binary ON-OFF process to account for the outages of the mmWave link. Built upon this model, we formulate an optimal scheduling problem where the objective is to achieve maximum mmWave channel utilization with bounded delay performance. In order to determine "when" a data packet should be added to the sub-6 GHz or mmWave queues, we prove that the optimal policy is of the *threshold-type* such that the scheduler routes the arrival traffic to the mmWave queue if and only if its queue length is smaller than a threshold. Our numerical results show that the threshold-based scheduling policy efficiently captures the dynamics of the mmWave channel, and indeed maximizes the channel utilization.

In summary, our main contributions are as follows:

- We experimentally evaluate the spatial correlations between the channel gains for the mmWave (30 GHz) and sub-6 GHz (3 GHz) interfaces under various indoor and outdoor situations involving existence of LOS between the transmitter and receiver.
- We propose an integrated sub-6 GHz/mmWave system that exploits the cross-interface correlations for beamforming as well as data transfer. Our ADC follows the beamformer at the receiver and eliminates the need for a separate ADC for each element in the mmWave antenna-array.
- We propose a framework to model the integrated sub-6 GHz/mmWave transceiver as a network and jointly manage the transmission across the sub-6 GHz and mmWave interfaces. Our queue management formulation explicitly takes into account the mmWave channel dynamics, and our approach enables full utilization of the available mmWave channel capacity, despite the highly variable nature of the channel. We prove using subadditivity analysis that the optimal scheduling policy is a simple threshold-based one, which can be easily adopted despite the high link variations.

We should emphasize that the sub-6 GHz/mmWave correlation was studied in [6], and applied only for beamforming in [5]. However, a coherent design that fully integrates the sub-6 GHz and mmWave interfaces and optimally design the sub-6 GHz/mmWave transceivers is missing. Hence, we aim to develop an architecture for which the sub-6 GHz interface is utilized for both beamforming and data transfer. A preliminary version of our results was presented in [7].

We use the following notation throughout the paper. Bold uppercase and lowercase letters are used for matrices and vectors, respectively, while non-bold letters are used for scalers. In addition, $(.)^H$ denotes the conjugate transpose, $\mathrm{tr}(.)$ denotes the matrix trace operator, and $\mathbb{E}[.]$ denotes the expectation operator. The sub-6 GHz and mmWave variables are denoted by $(.)_{\text{sub-6}}$ and $(.)_{\text{mm}}$, respectively.

II. Related Work

We classify existing and related work across the following thrusts:

(I) Experimental studies and beamforming methods: Wireless channel fading is primarily studied under two disparate categories based on the impact and the time-scale of the associated variations: large-scale (due to shadowing, path loss, etc.) and small-scale (due to mobility combined with multi-path). There exist numerous measurement and experimentation efforts in order to understand mmWave propagation and the effect of slow scale and large scale fading in the mmWave band (see, for example, [2, 8]). The main objective has been to extend the existing far-field ray-tracing models to accurately represent various phenomena observed in mmWave. For example, in [9, 10], a model based on isolated clusters is argued to be more appropriate to capture the observed reflections in mmWave, as opposed to the uniform distribution across the delay taps. Extensive evaluations of mmWave propagation taken from hundreds of different locations and settings also exist, by the same group [2, 9, 10] as well as others [11]. Our goal is to neither replicate nor expand these observations. Instead, we are interested in the channel/propagation environment correlation

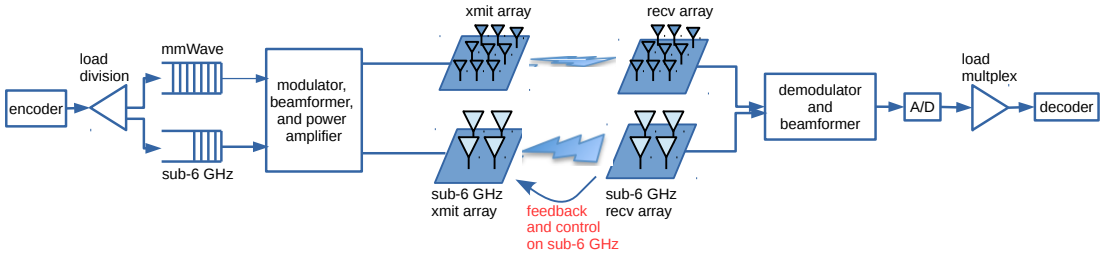


Fig. 1. Our integrated sub-6 GHz–mmWave architecture where the speed of mmWave interface necessitates the use of a separate queue.

across different interfaces under various conditions, including indoor and outdoor situations, with mobility, and LOS.

In order to improve mmWave beamforming efficiency, there has been extensive amount of work on digital and analog beamforming (e.g., [3, 12]). There are also proposals on hybrid beamforming methods [4] in which the term "hybrid" refers to the mixture of analog/digital (different from our hybrid sub-6 GHz/mmWave system). The whole operation there is in sole mmWave domain. The authors in [13] propose a compressed beamforming method that is based on out-of-band spatial information obtained at the sub-6 GHz band. In this paper, we experimentally investigate the sub-6 GHz/mmWave spatial correlations under practical (e.g., with mobility) scenarios, and demonstrate that how mobility affects the channel conditions and cross-interface correlation.

(II) Out-of-Band communications and scheduling policies: In the second part of this paper, we use sub-6 GHz as an auxiliary data transfer interface. Similar to our model, the authors in [14] studied a dual-interface system to offload cellular data over WiFi network. In this work, the delay and efficiency are quantified in a simple interface selection strategy, where the WiFi interface is selected whenever it is available and the cellular interface is selected when a specified deadline, T_{out}, is expired for buffered packets. In order to enable dual-interface communications and data transfer using sub-6 GHz and mmWave, we include a load division component in our proposed architecture (see Fig. 1). The objective is to schedule the arrival traffic over the sub-6 GHz or mmWave interface such that the maximum mmWave throughput with a bounded delay performance is achieved.

In the context of wireless scheduling policies, backpressure algorithms [15] promise throughput-optimal performance, which leads to a problem, known as MaxWeight. Using this framework, the goal is to maximize the weighted sum of link rates, in which the weights are represented by backlog differences of queues. Although backpressure-type algorithms provide throughput-optimal performance, they suffer from high end-to-end delays. To mitigate this issue, several approaches have been proposed. For instance, [16] proposes backpressure with adaptive redundancy (BWAR) to improve the delay performance. [17] describes a backpressure-based per-packet randomized routing framework to improve the delay performance. The authors in [18] propose using delay information of packets in the queues instead of using queue differentials as weights of the MaxWeight problem. As a result, those packets that have experienced high delays are more likely to be scheduled in the next time slot.

Due to the high data rate of the mmWave interface, it may not be feasible to track the channel state in real-time. This makes the use of complex algorithms to track the channel variations impractical. In this paper, we devise a delay-constrained and throughput-optimal scheduling policy that is expressed in terms of the queue lengths. To the best of our knowledge, there is no previous work that considers an integrated sub-6 GHz/mmWave architecture for beamforming and optimal data scheduling across the sub-6 GHz and mmWave interfaces.

III. Integrated sub-6 GHz – mmWave Architecture

A. System Model

Figure 1 illustrates the basic components of our architecture. The proposed architecture exploits the cross-interface correlation to achieve the beamforming fully in the analog domain without incurring high delay overhead. Thus, the ADC follows the beamformer at the receiver, and eliminates the need for a separate one, for all elements in the mmWave antenna-array. In addition, we move all mmWave control signaling and channel state information (CSI) feedback to the sub-6 GHz interface, and thus we avoid the two-way beamforming and reverse channel transmission costs in mmWave.

In addition to high energy consumption by components, the mmWave channel is highly sensitive and outages can be long that can lead to unacceptably high delays for delay-sensitive applications. However, a conservative use of the mmWave link is not desirable either, since the upside of the mmWave channel can be enormous, especially in the presence of LOS that occurs intermittently. More importantly, the high data rate of the mmWave link necessitates the use of a reasonably large buffer at the mmWave interface along with proactive queue-control solutions. Therefore, we consider a load devision component in Fig. 1 along with the separate sub-6 GHz and mmWave queues. We derive an optimal scheduling policy to select which interface(s) to use and control the queue sizes of interfaces in order to achieve maximum mmWave throughput with constrained delay. We investigate the optimal interface scheduling in Section IV.

B. Sub-6 GHz System and Channel Model

For the sub-6 GHz system, we use digital beamforming as shown in Fig. 2. As a result, the received signal at the receiver can be written as:

$$\mathbf{y}_{\text{sub-6}} = \mathbf{H}_{\text{sub-6}} \cdot \mathbf{x}_{\text{sub-6}} + \mathbf{n}_{\text{sub-6}}, \tag{1}$$

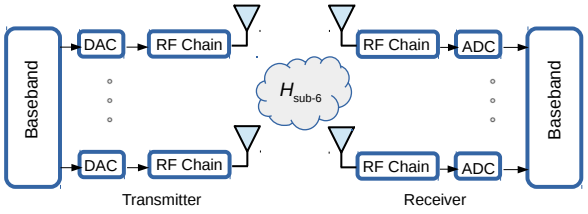


Fig. 2. Sub-6 GHz model based on digital beamforming.

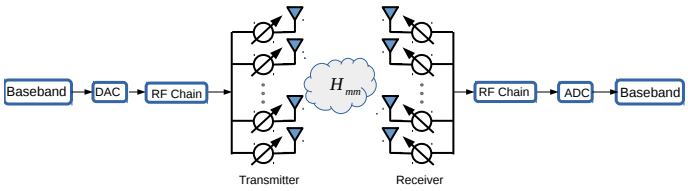


Fig. 3. MmWave system model based on analog beamforming.

where $\mathbf{H}_{\text{sub-6}} \in \mathcal{C}^{n_r \times n_t}$ is the matrix of complex channel gains from n_t transmit antennas to n_r receive antennas. In addition, $\mathbf{x}_{\text{sub-6}}$ is the transmitted signal vector in sub-6 GHz, and entries of circularly symmetric white Gaussian noise are denoted by $\mathbf{n}_{\text{sub-6}}$. The sub-6 GHz receiver uses the steering vector $\mathbf{w}_{\theta_{\text{sub-6}}}$ to align the received signals where the optimal steering direction $\theta^*_{\text{sub-6}}$ can be obtained based on maximizing the SNR, i.e.,:

$$\theta^*_{\text{sub-6}} = \arg\max_{\theta_{\text{sub-6}}} \frac{\mathbf{w}^H_{\theta_{\text{sub-6}}} \mathbf{H}_{\text{sub-6}} \mathbf{K}_{\mathbf{xx}} \mathbf{H}^H_{\text{sub-6}} \mathbf{w}_{\theta_{\text{sub-6}}}}{N_0}, \quad (2)$$

in which $\mathbf{K}_{\mathbf{xx}}$ is the covariance matrix of signal $\mathbf{x}$, and N_0 is the noise power.

C. MmWave System and Channel Model

The mmWave system model is shown in Fig. 3. Unlike sub-6 GHz, we use analog combining for mmWave via a single ADC to avoid high energy consumption. Consequently, the signal at the input of the decoder is a scalar, identical to a weighted combination of signal x_{mm} across all antennas. Thus, the received signal at the mmWave receiver can be written as:

$$y_{\text{mm}} = \mathbf{w}_r^H \mathbf{H}_{\text{mm}} \mathbf{w}_t \cdot x_{\text{mm}} + \tilde{n}_{\text{mm}}, \quad (3)$$

where $\mathbf{w}_r$ and $\mathbf{w}_t$ are the receive and transmit beamforming vectors, and $\tilde{n}_{\text{mm}}$ denotes the effective noise component after the combining step. In the mmWave domain, the channel matrix $\mathbf{H}_{\text{mm}}$ has a singular value decomposition $\mathbf{H}_{\text{mm}} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^H$, where $\mathbf{U} \in \mathcal{C}^{\bar{n}_r \times \bar{n}_r}$ and $\mathbf{V} \in \mathcal{C}^{\bar{n}_t \times \bar{n}_t}$ are rotation unitary matrices and $\mathbf{\Lambda} \in \mathcal{R}^{\bar{n}_r \times \bar{n}_t}$ is a diagonal matrix whose diagonal elements are nonnegative real numbers $\rho_1 \geq \rho_2 \geq ... \geq \rho_{\bar{n}_{\text{min}}}$, where $\bar{n}_{\text{min}} = \min(\bar{n}_r, \bar{n}_t)$, and $\bar{n}_t$ and $\bar{n}_r$ denote the number of mmWave transmit and receive antennas, respectively. The mmWave-channel matrix $\mathbf{H}_{\text{mm}}$ is low rank [19], and since the rank of $\mathbf{H}_{\text{mm}}$ is equal to the number of non-zero singular values, we restrict our attention to only the largest eigenvalue ρ_1 and assume that $\rho_1 \gg \rho_i$, and that $\rho_i \approx 0$ for $i \neq 1$. In fact, our experimental results show that under the LOS conditions, there is about $10 - 15$ dB gain improvement due to the strongest eigenmode, and thus we assume that the state of the mmWave link can be characterized based on the value of ρ_1. Next, we experimentally investigate the correlation between the sub-6 GHz and mmWave channels under various conditions.

D. Integrated sub-6 GHz – mmWave Experimentation

Experimental setup: We simultaneously observe the sub-6 GHz and mmWave channels via a dual transmitter-receiver pair in the same location. Our experimental setup is shown in Fig. 4. In the sub-6 GHz platform, we use an omni-directional antenna operating at 3 GHz as a transmitter and 4 omni-directional antennas as a receiver in order to observe the AoA for the incoming sub-6 GHz signal. The inter-element spacing of the sub-6 GHz array is about 10 cm, and the transmitter and receiver are located at a distance of 2 meters.

(a) Basic setup for indoor experiments

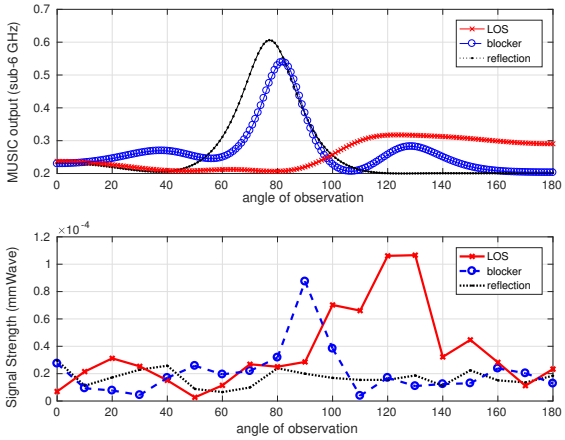
(b) Basic setup for outdoor experiments

Fig. 4. We use a joint mmWave and sub-6 GHz measurement setup to observe various properties of the propagation environment jointly across the two bands. The setup involves a network analyzer, horn antennas for mmWave, and omnidirectional 4-antenna array for sub-6 GHz measurements.

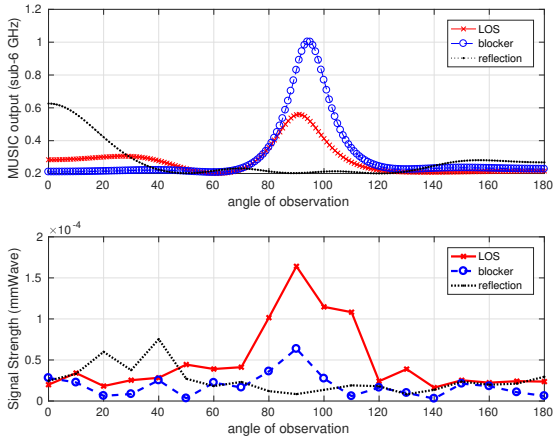
We use the Multiple Signal Classification (MUSIC) algorithm[1] [20] to evaluate the components of the signals across various angles. In principle, MUSIC algorithm decomposes the autocorrelation matrix of the received signal into signal subspace and noise subspace. Thereafter, the algorithm finds the peak value of the spectrum function calculated over the signal subspace. Finally, estimated AoA is set to the peak value of the spectrum function.

For mmWave, we use 30 GHz directional horn antennas to be able to align the beams. The half power beamwidth of the horn antennas is equal to 20 degrees, and we measure the channel across the 180° space with 10° step size. Based on a large set of measurements, we conclude that the propagation situations can be classified into three types as it pertains to summarizing the connection between the large-scale effects in sub-6 GHz and mmWave: line-of-sight (LOS), blocker, and non line-of-sight (NLOS). LOS implies that there is a strong line of sight path between the transmitter and the receiver; blocker indicates that the LOS path for the mmWave interface

[1]For the sake of clarity, we use MUSIC algorithm, but other estimators can be used as well.



(a) sub-6 GHz and mmWave activity vs. angle in *indoor* setting



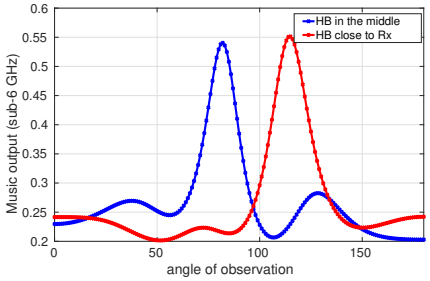
(b) sub-6 GHz and mmWave activity vs. angle in *outdoor* setting

Fig. 5. Indoor (5(a)) and outdoor (5(b)) associated with sub-6 GHz (top plots) and mmWave (bottom plots). In each case, we have tested three situations: LOS, blocker, and NLOS with reflector. The direction of strong signal is highly correlated between sub-6 GHz and mmWave if a LOS is present. The correlation is lost in part, if there is a blocker present and lost completely in the case of NLOS with reflections. MmWave signal strength is expressed in watts.

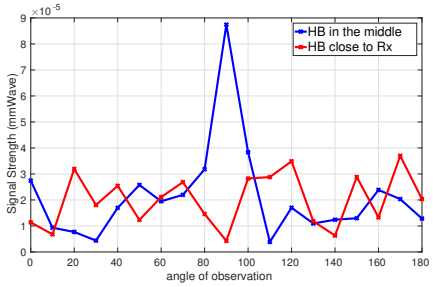
is being blocked by a non-stationary obstacle; and NLOS indicates the presence of a stationary obstacle, unlikely to change in time.

Experimental observations: Figure 5 provides our *indoor and outdoor* measurement results, taken simultaneously for sub-6 GHz and mmWave. The output of the MUSIC algorithm is given on the top plots, and the important thing to focus on is the correct AoA in each situation. Note that the AoA is different across different observations plotted. Once that AoA is identified, we compare it with the signal strength (bottom plots) we measured along that direction for the mmWave signal generated at the transmitter location of the sub-6 GHz. For the LOS situation, for both indoor and outdoor, there is a strong correlation in the angular composition and the strength of signal coming across all angles in sub-6 GHz and mmWave. This observation is in agreement with [5]. Indeed, in 94% of the LOS measurements, we have identified the AoA predicted by MUSIC within a $\pm 10^{\circ}$ accuracy for the AoA of the mmWave signal. As a result, based on sub-6 GHz measurements, the correct mmWave transmitter location can be almost perfectly identified under LOS. From Fig. 5, it is evident that as we lose the LOS, the sub-6 GHz/mmWave correlation is lost and the signal strength in mmWave starts to drop rapidly. However, depending on the size and the location of the blocker, AoA estimation accuracy varies. For instance, for a small/mid-size blocker in the middle, in 55% of the observations do the sub-6 GHz and mmWave signals have their strongest paths within $\pm 10^{\circ}$ of each other. In this context, Fig. 6 demonstrates the effect of human blocker located in the middle compared with when the blocker moves very close to the receiver. From the results, we note that as the blocker moves towards the receiver, the correlation decreases. Moreover, the mmWave signal strength drops when the human blocker is close to the receiver.

From the experimental results, our major observation is that in LOS situations, there is a high correlation between the observed sub-6 GHz and mmWave signals, both in signal strength and AoA. Therefore, LOS instances should be exploited in mmWave as much as possible, since there is an associated $10-15$ dB channel gain improvement as well. For instance,



(a) Sub-6 GHz activity vs. angle in indoor setting



(b) mmWave activity vs. angle in indoor setting

Fig. 6. Channel spatial behavior for human block (HB) in indoor environment for sub-6 GHz (top) and mmWave (bottom).

Fig. 7 illustrates the spatial variations of the mmWave channel gain in LOS and reflection situations. We observe that the LOS situation is quite robust with respect to slight movements, i.e., the large-scale effects lead to minor variations in the channel gain, if the presence of LOS is preserved. On the other hand, if the LOS is blocked and the connection depends on a strong reflector, channel gain becomes relatively unstable and slight movements can result in drastic changes in the channel. As a result, it is desirable to predict the loss of LOS and take the necessary precautions for a smoother transition in order to mitigate the negative effects of connection losses on the user experience. In order to detect the LOS situations, the authors in [5] use the ratio of the highest signal strength component to the average received signal energy (i.e., peak to average power ratio (PAPR)) as an indicator for LOS inference.

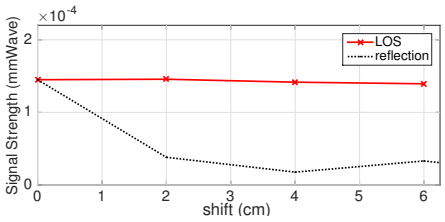


Fig. 7. Spatial variation in the channel gain. Small movements lead to significant variations without LOS.

E. Beamforming

In order to overcome the harsh nature of mmWave channels and compensate for large propagation losses, highly directional antenna arrays along with beamforming techniques are needed. However, deploying directional antenna arrays makes the cell discovery and access methods costly in terms of delay. This effect is more pronounced when mobility of users are taken into account where rapid handovers within small-scale cells are needed. Digital beamforming is highly efficient in delay where with the observations from all receive antennas, beamforming can be done by one-shot processing of the observed beacons. However, digital beamforming requires high energy consumption due to the need for a separate ADC per element in the antenna-array. Analog beamforming, on the other hand, is more energy efficient, but it can focus on one direction at a time, making the search process costly in delay.

Pertaining to the mmWave beamforming efficiency, the mmWave channel is often sparse in the angular domain, with a few scattering clusters, each with several rays, in addition to a dominant LOS path [19]. Thus, in order to find the optimal mmWave steering direction θ^*_{mm}, our proposed architecture exploits the correlation between the sub-6 GHz and mmWave AoA, and uses a coarse AoA estimation on the sub-6 GHz channel, followed by analog beamforming for fine tuning around the estimated AoA. The sub-6 GHz/mmWave AoA correlation reduces the angular search space and addresses the delay issue of fully-analog beamforming. The sub-6 GHz-assisted mmWave beamforming scheme is specified below, and is graphically illustrated in Fig. 8.

1) Start the system in the sub-6 GHz only mode.
2) Implement MUSIC algorithm in sub-6 GHz and estimate the angle of arrival $A_{\text{sub-6}}$ based on beacons.
3) Use analog beamforming to fine tune the mmWave beam in the range of $A_{\text{sub-6}} \pm 10°$:
   a) If the LOS is detected (e.g., by PAPR test), both interfaces operate jointly in the dual sub-6 GHz/mmWave mode in which resources and arrival traffic are allocated jointly.
   b) Otherwise, mmWave bemforming falls back to the conventional schemes.

Remark 1: As our experimental results show, the sub-6 GHz/mmWave correlation decreases as the LOS condition is lost. However, the sub-6 GHz assisted beamforming relies on the cross-interface correlation, and once the correlation is lost, it falls back to the conventional beamforming scheme.

Remark 2: The parameter of searching $\pm 10°$ around the estimated AoA is set based on our experimental setup. In general, it will be configured based on dynamics of the scenario and antenna beamwidth.

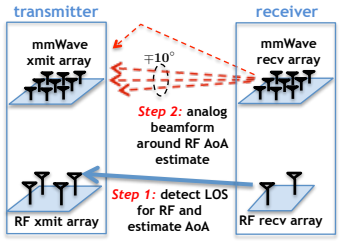


Fig. 8. Our beamformer works in two phases: (1) the presence of LOS is detected and AoA is estimated all in sub-6 GHz; (2) analog mmWave beamformer focuses on a small area around the estimated AoA.

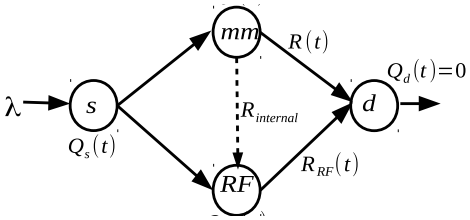


Fig. 9. An equivalent network model for the integrated sub-6 GHz/mmWave transceiver in which the mmWave (denoted by m) and sub-6 GHz (denoted by r) interfaces are viewed as individual nodes of the network.

IV. Out-of-Band mmWave Communications with Optimal Scheduling

In the previous section, we exploited the spatial correlations between the sub-6 GHz and mmWave interfaces in order to mitigate the overhead of analog beamforming in standalone mmWave systems. Next, we envision the use of the sub-6 GHz interface as the secondary data transfer mechanism. In particular, in the proposed architecture, once the dual sub-6 GHz/mmWave mode is activated, the load division component (in Fig. 1) schedules the arrival traffic over the sub-6 GHz and mmWave interfaces. The objective is to achieve maximum mmWave throughput with bounded delay performance.

In order to obtain the optimal scheduling policy, we model our integrated sub-6 GHz/mmWave transceiver as a *diamond network* (see Fig. 9) in which each of the sub-6 GHz and mmWave interfaces are represented as a network node. Moreover, a virtual destination (i.e., receiver) node d has been added, and since all data packets are destined for node d, its queue length, $Q_d(t)$, is set to 0 for all t. We further assume that the equivalent network model evolves in discrete (slotted) time $t \in \{0, 1, 2, ...\}$, and there is an exogenous packet arrival with rate λ. Our experimental results in Section III-D demonstrate that under the LOS conditions, there is about $10 - 15$ dB channel gain improvement due to the strongest eigenmode. Therefore, the state of the strongest eigenmode can be used to determine the availability of the mmWave link. To quantify the behavior of the mmWave link using the strongest eigenmode (i.e., corresponding to ρ_1), a two-state model (outage and non-outage) can be used such that the probability of being in each state can be characterized

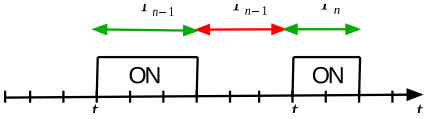


Fig. 10. ON-OFF periods of the mmWave link availability

as a function of distance and statistical models [2]. Hence, we use the binary process $\{L(t)\}_{t=1}^{\infty}$ to account for mmWave outage and non-outage situations such that $L(t) := 1$ implies the availability of the mmWave link (i.e., ON state) during time slot t and $L(t) := 0$ otherwise (i.e., OFF state). As we also experimentally show in Section V (see Fig. 13), $L(t) = 1$ corresponds to LOS situations, while $L(t) = 0$ can be mapped to the NLOS situations like human blockers or when there are no strong reflectors. We further assume that T_n^{on} and T_n^{off} (with general random variables T_{on} and T_{off}) denote the n-th ON and OFF periods respectively, as shown in Fig. 10. The sequence of ON times $\{T_n^{\text{on}} : n \geq 1\}$ and OFF times $\{T_n^{\text{off}} : n \geq 1\}$ are independent sequences of i.i.d positive random variables. Unlike mmWave, the sub-6 GHz link is much less sensitive to blockage due to diffraction. Thus, for the sake of simplicity, we assume that the sub-6 GHz link is available during all time slots even when $L(t)$ takes on the value of 0 due to blockers.

State variables and scheduling policy: The dynamics of the mmWave link during time slot t is denoted by $\mathbf{x}(t) = \big(Q(t), D(t)\big)$ in which $Q(t)$ (with a general random variable Q) is the mmWave queue length, and $D(t)$ is the waiting time of the head-of-line packet[2]. The state space is denoted by $\mathcal{S}$, and a scheduling policy $\pi \in \Pi$ determines the assignment of packets to the mmWave or sub-6 GHz queue, i.e., $\pi : Q \to \{0, 1\}$ in which Π denotes the class of *feasible causal* policies in a sense that scheduling decisions are made based on current state. The decision variable $\pi(Q) = 1$ (or, in short, $\pi = 1$) implies that the packet is routed to the mmWave queue, and $\pi(Q) = 0$ (or $\pi = 0$) otherwise. The scheduler node s queries the state information of its neighbor nodes, and assigns packets accordingly. Due to the high data rate of the mmWave interface, real time tracking of the channel state may not be feasible, and thus, it is desirable to obtain a scheduling policy that is not directly expressed in terms of the CSI. This is in contrast with the classical MaxWeight scheduling policies (e.g., Backpressure) [15] that require CSI information.

The number of packets added to the mmWave queue at time t is denoted by $\beta^{\pi}(t)$. To avoid a large waiting time in the mmWave queue due to intermittent connectivity, we require the packets to be *impatient* in the sense that if the waiting time of the head-of-line packet in the mmWave queue exceeds a timeout T_{out} (i.e., if $D(t) \geq T_{\text{out}}$ holds), the packet "reneges" (is moved to) to the sub-6 GHz queue. To account for packets reneging, we consider a virtual link between the mmWave and sub-6 GHz queues with a rate equal to the internal read/write speed of processor, as shown in Fig. 9. In this case, $\gamma^{\pi}(t, T_{\text{out}})$ denotes the number of reneged packets and $\alpha^{\pi}(t)$ is the number of packets that are completely served by the mmWave queue. Therefore, the mmWave queue evolves as $Q^{\pi}(t) = \max[0, Q^{\pi}(t-1) + \beta^{\pi}(t) - \alpha^{\pi}(t) - \gamma^{\pi}(t, T_{\text{out}})]$.

[2]For the sake of notations, we drop the subscript $(.)_{\text{mm}}$ from the mmWave variables.

A. Problem Formulation

Definition 1 Given that $\beta^{\pi}(t)$ packets are added to the mmWave queue; $\alpha^{\pi}(t)$ packets are completely served by the mmWave queue; and $\gamma^{\pi}(t, T_{\text{out}})$ packets renege at time t, their corresponding average quantities are respectively defined as:

$$\bar{\beta}(\pi) = \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T} \beta^{\pi}(t)\right], \tag{4a}$$

$$\bar{\alpha}(\pi) = \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T} \alpha^{\pi}(t)\right], \tag{4b}$$

$$\bar{\gamma}(\pi, T_{\text{out}}) = \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T} \gamma^{\pi}(t, T_{\text{out}})\right]. \tag{4c}$$

In order for the above expectations to exist, we assume that $\beta^{\pi}(t)$, $\alpha^{\pi}(t)$, and $\gamma^{\pi}(t, T_{\text{out}})$ are stationary ergodic. In this model, imposing the service deadline T_{out} ensures that the average waiting time of packets in the mmWave queue is smaller than or equal to T_{out}. Hence, the reneging mechanism explicitly dictates a constraint on the mmWave waiting time. Therefore, our goal is to derive a throughput-optimal policy with bounded reneging rate.

Problem 1 (Constrained Throughput Optimization) *Given that there is a timeout T_{out} for packets in the mmWave queue, for a given $\epsilon < \lambda$, we define Problem 1 as follows:*

$$\begin{aligned} &\max_{\pi\in\Pi} \quad \bar{\alpha}(\pi) \\ &\textit{s.t.} \quad \bar{\gamma}(\pi, T_{out}) \leq \epsilon \quad \textit{and} \quad \bar{\beta}(\pi) \leq \lambda, \end{aligned} \tag{5}$$

In (5), the objective function and the first constraint can be relaxed as: $\max_{\pi\in\Pi} \quad \bar{\alpha}(\pi) - b\big(\bar{\gamma}(\pi, T_{\text{out}}) - \epsilon\big)$, where b is a positive Lagrange multiplier. For any particular fixed value of b, it is straightforward to show that there is no loss of optimality in the relaxed problem. To see this, let π^* be the optimal policy for the original problem, and π_R^* be the optimal policy for the relaxed problem. We have: $\bar{\alpha}(\pi^*) \leq \bar{\alpha}(\pi^*) - b\big(\bar{\gamma}(\pi^*, T_{\text{out}}) - \epsilon\big) \leq \bar{\alpha}(\pi_R^*) - b\big(\bar{\gamma}(\pi_R^*, T_{\text{out}}) - \epsilon\big)$. The first inequality holds since π^* is feasible in the original problem, and the second inequality holds because π_R^* is the optimal solution for the relaxed problem. Thus, there is no loss of optimality in the relaxed problem. We note that the relaxed formulation can be interpreted as an optimization over obtained *rewards* and paid *costs*. In particular, each packet that receives service from the mmWave link, results in r units of reward (i.e., in terms of mmWave throughput), while a packet reneging incurs a cost of c (i.e., in terms of wasted waiting time in the mmWave queue). This leads to the following problem.

Problem 2 (Total Reward Optimization) *We consider the maximization problem over total rewards obtained as a result of serving packets, and costs due to packets reneging, i.e.,:*

$$\begin{aligned} &\max_{\pi\in\Pi} \ \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T} r\alpha^{\pi}(t) - c\gamma^{\pi}(t, T_{out})\right] \\ &\textit{s.t.} \quad \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T} \beta^{\pi}(t)\right] \leq \lambda, \end{aligned} \tag{6}$$

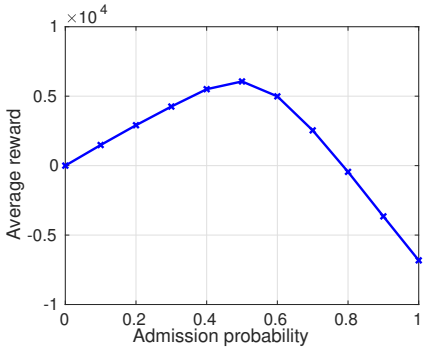


Fig. 11. Probabilistic admission policy where the objective value of (6) first increases by admitting more packets into the mmWave queue, and thereafter it decreases due to dominant reneging cost.

where the constraint $\bar{\alpha}(\pi) \leq \bar{\beta}(\pi)$ is implicit.

It is straightforward to show that an optimal solution π_R^* for the relaxed formulation of Problem 1 is the optimal solution for (6) and vice versa. To see this, assume that the set of feasible solutions for (5) is denoted by Π. For all $\pi \in \Pi$, we have: $\bar{\alpha}(\pi) - b\big(\bar{\gamma}(\pi, T_{\text{out}})\big) \leq \bar{\alpha}(\pi_R^*) - b\big(\bar{\gamma}(\pi_R^*, T_{\text{out}})\big)$. Multiplying both sides by $r \geq 0$, and setting $c := rb$, we conclude that π_R^* is the optimal solution for (6) as well since their feasible sets are identical. For $r = 1$ and $c = b$, two formulations will be identical. In general, the values of r and c are set based on the application and performance requirements. For instance, a large value of r ensures high throughput, while a large value of c prioritizes low-latency performance (i.e., a *conservative policy*). Therefore, (6) captures the tradeoff between full exploitation of the mmWave capacity and the delay for mmWave channel access through the control knob $\beta^\pi(t)$: if $\beta^\pi(t)$ is set to a very small value for all time slots t (i.e., a conservative policy) then $\alpha^\pi(t)$ would be small as well, and the objective function reduces due to the first term. On the other hand, if $\beta^\pi(t)$ is set to a large value (e.g., matched to the arrival rate λ for all time slots t) and the link state fluctuates according to the process $\{L(t)\}_{t=1}^{\infty}$, then the objective function could decrease due to the reneging cost that is captured by the second term. Therefore, there is an optimal value of $\beta^\pi(t)$ within these two extreme cases that results in the maximum return rate. The following example illustrates this point clearly.

Example (Probabilistic Admission Policy): Using a probabilistic admission policy π, the input arrival rate λ is mapped to the admitted rate $\beta^\pi(t)$ such that a packet is admitted to the mmWave queue with an admission probability p. We assume that the arrival process is a batch arrival such that there is a batch of size randomly distributed with normal distribution with mean 20 and standard deviation 1. The probability of a batch arrival within a time slot is set to 0.9. The length of ON and OFF periods of the mmWave channel is set according to the log-normal distributions with mean 20 and 3.5, respectively. Fig. 11 demonstrates behavior of the objective function in (6) as p increases. We observe that the objective value increases by admitting more packets into the queue up to a certain threshold, and thereafter the objective value decreases due to the dominant reneging cost. Next, we tackle the problem of deriving an optimal admission policy.

B. Optimal Scheduling Policy

From (6) and using the Lagrangian relaxation, we define:

$$\begin{aligned} g(W) &= \max_{\pi\in\Pi} \Big[r\bar{\alpha}(\pi) - c\big(\bar{\beta}(\pi) - \bar{\alpha}(\pi)\big) + W\big(\lambda - \bar{\beta}(\pi)\big) \Big] \\ &= \max_{\pi\in\Pi} \Big[(r+c)\bar{\alpha}(\pi) + (W+c)\big(\lambda - \bar{\beta}(\pi)\big) \Big] - c\lambda, \end{aligned} \tag{7}$$

in which the Lagrange multiplier W is positive, and it can be interpreted as a *subsidy* for taking the *passive action*. In our problem, passive action is defined as adding packets to the sub-6 GHz queue, while active action corresponds to admitting packets into the mmWave queue. Hence, the objective is to maximize the long-term expected reward by balancing the reward for serving and the subsidy for passivity.

The solution of (7) partitions the state space $\mathcal{S}$ into three sets, $\mathcal{S}_0$, $\mathcal{S}_1$ and $\mathcal{S}_{01}$, where, respectively, the optimal action is $\pi(\mathbf{x}) = 0$ for $\mathbf{x} \in \mathcal{S}_0$, $\pi(\mathbf{x}) = 1$ for $\mathbf{x} \in \mathcal{S}_1$, or some randomization between $\pi(\mathbf{x}) = 0$ and $\pi(\mathbf{x}) = 1$ for $\mathbf{x} \in \mathcal{S}_{01}$. From [21], it is known that in a Markov Decision Process if the state space contains a finite number of states, then set $\mathcal{S}_{01}$ does not contain more than one state. This case holds in our model since the mmWave queue length is upper-bounded, and the waiting time in the queue can be at most T_{out}. Thus, set $\mathcal{S}_{01}$ does not contain more than one state. The following theorem states that Problems 1 and 2 are solved by a *monotone* policy, where a class of policies Π has monotone structure if for $\pi \in \Pi$, there exists $h^* \in \{1, 2, ..\}$ such that: $\pi(Q) = 0 \Longleftrightarrow Q \geq h^*$, where Q denotes the mmWave queue length. In other words, *the optimal policy routes a packet to the mmWave queue if and only if the mmWave queue length is smaller than an optimal threshold h^*.*

Theorem 1. *(Optimality of monotone policy) The solution for the reward optimization in* (7) *has a monotone structure.*

Proof. Let us denote by $v(Q, D, \pi)$, the value function corresponding to Problem 2 when mmWave is at state (Q, D) and action $\pi \in \{0, 1\}$ is taken. In addition, let us define $V(Q, \pi) = \sum_{1\leq D\leq T_{\text{out}}} v(Q, D, \pi)$. From the Bellman equation [21], we have:

$$\begin{aligned} f(Q, 0) &= r + W + (\sigma + \mu)V\left((Q-1)^+, 0\right); \\ f(Q, 1) &= r + \lambda V(Q+1, 1) + (\sigma + \mu)V\left((Q-1)^+, 1\right), \end{aligned}$$

in which, σ is the reneging rate, μ is the mmWave service rat, and $(.)^+ = \max(., 0)$. We prove that if passive action is optimal in Q then passive action is optimal in $Q' \geq Q$. Similar to [22], let us define $\varphi(Q) = \arg\max_{\pi\in\{0,1\}} f(Q, \pi)$. It then suffices to show that $\varphi(Q') \leq \varphi(Q)$ for $Q' \geq Q$. We have $\pi \leq \varphi(Q')$, and by definition $f(Q', \varphi(Q')) - f(Q', \pi) \geq 0$. Let us now prove that $V(Q, \pi)$ has the *subadditivity* property.

Definition 2 (Subadditive Function) Let X and Y be partially ordered sets and $u(x, y)$ a real-valued function on $X \times Y$. We say that u is subadditive if for $x^+ \geq x^-$ in X and $y^+ \geq y^-$ in Y we have: $u(x^+, y^+) + u(x^-, y^-) \leq u(x^+, y^-) + u(x^-, y^+)$.

To prove that $V(Q,\pi)$ is a subadditive function, it suffices to show that for all $Q' \geq Q$ and $\pi \in \{0,1\}$, the inequality $f(Q',\varphi(Q')) + f(Q,\pi) \leq f(Q',\pi) + f(Q,\varphi(Q'))$ holds. If $\varphi(Q') = \pi = 0$ or $\varphi(Q') = \pi = 1$, then the inequality is satisfied. If $\varphi(Q') = 1$ and $\pi = 0$, then we show that $f(Q,0) - f(Q,1) \leq f(Q,1) - f(Q',1)$. By replacing the corresponding terms, we need to show:

$$\begin{aligned}(\sigma+\mu)\left[V\left((Q-1)^+,0\right) - V\left((Q'-1)^+,0\right)\right] \leq \\ \lambda\left[V\left(Q+1,1\right) - V\left(Q'+1,1\right)\right] \\ +(\sigma+\mu)\left[V\left((Q-1)^+,1\right) - V\left((Q'-1)^+,1\right)\right]. \quad (8)\end{aligned}$$

In order to show this inequality, we note that $V(Q,1)$ is non-increasing and $V(Q,0)$ is non-decreasing. The reason is that when the action $\pi = 1$ is chosen, all packets will be added to the mmWave queue. The likelihood that an admitted packet reneges before receiving service increases with the number of queued packets, and thus the incurred reneging cost increases. Therefore, the value function $V(Q,1)$ is a non-increasing function of the queue length. A similar argument holds for $V(Q,0)$. Therefore, the inequality $f(Q,0) - f(Q,1) \leq f(Q,1) - f(Q',1)$ holds, and the theorem statement follows. □

Intuitively, for a first-in-first-out (FIFO) queue, the likelihood that an admitted packet reneges before receiving service increases with the number of queued packets. Therefore, given that the reneging and moving packets from the mmWave queue to the sub-6 GHz queue incurs a delay cost, it is in the scheduler interest to exercise admission control and deny entry to packets when the mmWave queue grows and becomes larger than a threshold. Next, we characterize the optimal threshold.

C. Optimal Threshold

Optimal policy π^* imposes a threshold $h^* \in \{0,1,2,..\}$ such that $\pi^*(Q) = 1$ if and only if $Q < h^*$. Under the ergodicity assumption, we rewrite Problem 2 as:

$$\max_{h\in\{0,1,2,..\}} \Big(\big(r+c\big)\mathbb{E}[\alpha_h] - \big(W+c\big)\mathbb{E}[\beta_h] \Big). \quad (9)$$

Lemma 1. *Given an admission threshold h, if*

$$\psi(h) := \frac{\mathbb{E}[\beta_h] - \mathbb{E}[\beta_{h-1}]}{\mathbb{E}[\alpha_h] - \mathbb{E}[\alpha_{h-1}]}, \quad (10)$$

then $\psi(h)$ is non-decreasing in h.

Proof. In order to prove that $\psi(h)$ is non-decreasing, we note that both $\mathbb{E}[\alpha_h]$ and $\mathbb{E}[\beta_h]$ are non-decreasing in h since a larger threshold h results in admitting more packets (i.e., a larger $\mathbb{E}[\beta_h]$) and potentially a higher throughput $\mathbb{E}[\alpha_h]$. Moreover, $\mathbb{E}[\beta_h]$ is assumed to be an affine function of h. In order to prove that $\psi(h)$ is non-decreasing in h, we need to show that $\psi(h+1) - \psi(h) \geq 0$ for $h \geq 0$. Therefore, we have:

$$\begin{aligned}\psi(h+1) - \psi(h) = \frac{\left(\mathbb{E}[\alpha_h] - \mathbb{E}[\alpha_{h-1}]\right)\left(\mathbb{E}[\beta_{h+1}] - \mathbb{E}[\beta_h]\right)}{\left(\mathbb{E}[\alpha_{h+1}] - \mathbb{E}[\alpha_h]\right)\left(\mathbb{E}[\alpha_h] - \mathbb{E}[\alpha_{h-1}]\right)} \\ - \frac{\left(\mathbb{E}[\beta_h] - \mathbb{E}[\beta_{h-1}]\right)\left(\mathbb{E}[\alpha_{h+1}] - \mathbb{E}[\alpha_h]\right)}{\left(\mathbb{E}[\alpha_{h+1}] - \mathbb{E}[\alpha_h]\right)\left(\mathbb{E}[\alpha_h] - \mathbb{E}[\alpha_{h-1}]\right)}\end{aligned}$$

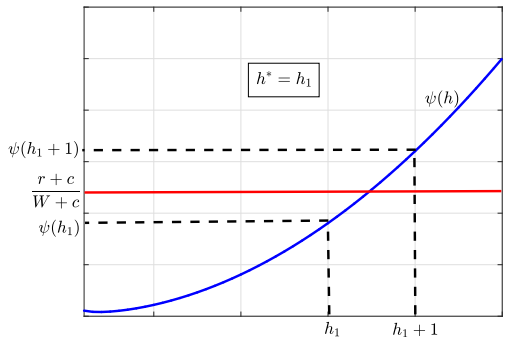


Fig. 12. A sample path of $\psi(h)$ function and finding the optimal admission threshold.

Due to the fact that $\mathbb{E}[\alpha_h]$ is non-decreasing and concave, and $\mathbb{E}[\beta_h]$ is increasing and affine in terms of h, we conclude that (**??**) is non-negative, and thus $\psi(h+1) - \psi(h) \geq 0$ for $h \geq 0$. □

Theorem 2. *Given an admission threshold h, we define*

$$\phi(h) := (W+c)\psi(h). \quad (11)$$

If $\phi(h) < r + c \leq \phi(h+1)$, then $h^* = h$.

Proof. From Lemma 1, we conclude that $\phi(h)$ is non-decreasing in h as well, i.e., $\phi(h+1) \geq \phi(h), \forall h \geq 0$. For a given threshold h that satisfies $r + c \leq \phi(h+1)$, we conclude that:

$$(r+c)\mathbb{E}[\alpha_{h+1}] - (W+c)\mathbb{E}[\beta_{h+1}] \leq (r+c)\mathbb{E}[\alpha_h] - (W+c)\mathbb{E}[\beta_h].$$

Therefore, h achieves a higher objective value than $h+1$. Now in order to establish this result for $h+2$, we can show that:

$$r + c \leq \phi(h+1) \leq \phi(h+2) \leq (W+c)\frac{\mathbb{E}[\beta_{h+2}] - \mathbb{E}[\beta_h]}{\mathbb{E}[\alpha_{h+2}] - \mathbb{E}[\alpha_h]},$$

from which we conclude that h is optimal with respect to $h+2$ as well. By induction, we extend this result for all $h' > h$. Similarly, based on the constraint $\phi(h) < r + c$ we prove that h is optimal with respect to all $h' < h$ as well. Thus, h is optimal in general, and we have $h^* = h$. Note that $\mathbb{E}[\beta_h] = \lambda \sum_{Q<h,D} \xi_{(Q,D)}$ and $\mathbb{E}[\alpha_h] = \mathbb{E}[\beta_h] - \sigma \sum_{Q,D=T_{\text{out}}} \xi_{(Q,D)}$, where $\xi_{(Q,D)}$ denotes the limiting probability of the state $\mathbf{x} = (Q,D)$. Calculation of the limiting distribution is presented in Appendix A. □

Theorem 3. *The optimal threshold h^* is an increasing function of r and a decreasing function of c.*

Proof. We note that $\frac{r+c}{W+c}$ is increasing in r and decreasing in c due to the fact that $W \leq r$. The condition $W \leq r$ is necessary in order to avoid the trivial scenario where the subsidy is larger than the reward of successful transmission. The trivial scenario leads to always choosing the passive action, and thus we pose the constraint $W \leq r$ to avoid the trivial condition. From Lemma (1), we note that $\phi(h)$ is non-decreasing in h. From Theorem 2 and because $\frac{r+c}{W+c}$ is an increasing function of r and a decreasing function of c, we conclude that the optimal threshold h^* increases in r and decreases in c as well. □

Algorithm 1 Online Threshold-based Scheduling Policy

1: $t \leftarrow 1$ // Set the time to 1
2: $h^*(t) \leftarrow K$ // Set h^* equal to mmWave buffer size
3: $Q(t) \leftarrow 0$ // $Q(t)$: mmWave queue length at time 0
4: $Q_{\text{sub-6}}(t) \leftarrow 0$ // $Q_{\text{sub-6}}(t)$: sub-6 GHz queue length at time 0
5: **while** $Q_s(t) \neq 0$ **do** // Continue until there is no packet
6: **if** $Q(t) \leq h^*(t)$ **then**
7: Set $\pi = 1$ // Add the packet to the mmWave queue
8: **else**
9: Set $\pi = 0$ // Add the packet to the sub-6 GHz queue
10: **end if**
11: Update $Q_s(t), Q_{\text{sub-6}}(t), Q(t)$, $\bar{\alpha}(t)$ and $\bar{\beta}(t)$
12: $h^*(t+1)$ = UPDATE-THRESHOLD $\big(\bar{\alpha}(t), \bar{\beta}(t), \bar{\alpha}(t-1), \bar{\beta}(t-1), h^*(t)\big)$
13: **end while**
14:
15: **function** UPDATE-THRESHOLD($\alpha(t), \beta(t), \alpha(t-1)$, $\beta(t-1)$, $h(t)$)
16: Calculate $\psi(t) = \frac{\beta(t)-\beta(t-1)}{\alpha(t)-\alpha(t-1)}$
17: **if** $\psi(t) \geq \frac{r+c}{W+c}$ **then**
18: $h(t+1) \leftarrow h(t) - 1$
19: **end if**
20: **return** $h(t+1)$
21: **end function**

The above theorem shows that if the value of r increases, throughput performance will have a higher priority than delay, and thus optimal threshold increases, as expected. On the other hand, by increasing the value of c, the optimal threshold decreases to avoid high reneging costs. As a result, based on the performance requirements, the tradeoff between full exploitation of the mmWave capacity and the delay for mmWave channel access is adjusted through the use of parameters r and c.

D. Online Scheduling Policy

In the previous section, we derived the optimal scheduling policy along with the optimal admission threshold. In practice, the mmWave link is highly dynamic such that the data rate can vary over two orders of magnitude, and thus it is desirable to be able to adjust the admission threshold on-the-fly and accommodate the dynamics of the mmWave channel. In what follows, we provide an online scheduling policy that preserve the form of optimal policy, while adjusts the admission threshold on-the-fly. In order to obtain the online algorithm of Theorem 2, we note that the optimal threshold h^* is a function of the ratio $\frac{r+c}{W+c}$ with $W \leq r$. Moreover, the optimal threshold is expressed in terms of function $\psi(h)$ that is non-decreasing with respect to h. As an example, Fig. 12 demonstrates a sample path of the $\psi(h)$ function introduced in Lemma 1. In order to calculate the optimal threshold h^* at time t, we consider the value of function $\psi(h)$ up to time t and adjust the optimal threshold h^* accordingly (as shown in Fig. 12). Algorithm 1 provides an online scheme to calculate the optimal threshold, and works as follows. In line 2, the admission threshold h^* is set to its maximum value, the mmWave buffer size. Next, the threshold-type scheduler assigns the packets to the mmWave (line 7) or sub-6 GHz queue (line 9). Thereafter, based on the outcome of the transmissions, queue lengths and throughput value $\bar{\alpha}(t)$ are updated (line 11). Hence, at each time slot the algorithm takes into account the state of the mmWave channel and updates the corresponding parameters. Finally, in the `Update-Threshold` subroutine, the value of admission threshold $h^*(t)$ is updated.

V. NUMERICAL RESULTS

In this section, we investigate the performance of our proposed scheduling policy. To this end, we use the experimental traces to model the ON-OFF mmWave link. In our experiment, a mobile receiver moves with the speed of 1 m/s over a path characterized by sudden link transitions due to human blockers (HB) and reflectors (REF). Figure 13 illustrates the received signal strength as the mobile moves away from the transmitter. We assume a signal reception cutoff threshold δ (determined based on the hardware used and environment) such that if the signal strength is below δ, the channel is in the OFF state. Moreover, in order to adequately capture the dynamics of the mmWave channels, the timeout value T_{out} is set on-the-fly such that at time t, we set $T_{\text{out}}(t) = \bar{Z}_{\text{sub-6}}(t)$ with $\bar{Z}_{\text{sub-6}}(t)$ to be the sub-6 GHz average waiting time. Thus, on average, packets would not get stuck in the mmWave queue longer than if they would have joined the sub-6 GHz queue.

A. Optimality Results

We first investigate the tradeoff between the mmWave throughput (or, conversely, *link wastage*) and the average waiting time. Link wastage is defined as the fraction of time slots that there are packets in the system, but the mmWave queue is empty and the mmWave link is available (i.e., $L(t) = 1$). The tradeoff between link wastage and the average waiting time is shown in Fig. 14(a). From the results, we observe that if there

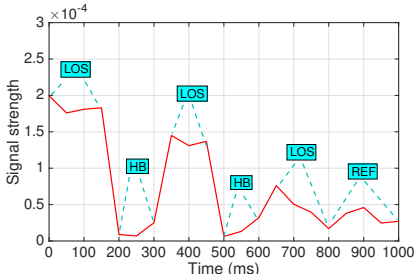
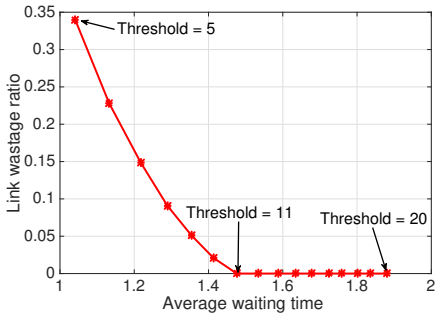
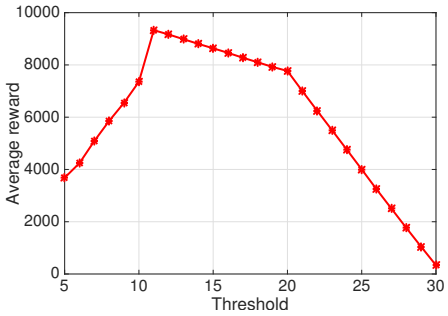


Fig. 13. MmWave channel temporal correlation under line of sight (LOS), human block (HB), and reflection (REF)

are so many packets added to the mmWave queue and if the mmWave link becomes unavailable, high average delay incurs. On the other hand, a conservative policy is not desirable either such that due to lack of packets in the mmWave queue, the link wastage increases. Figure 14(b) illustrates the total reward obtained as a function of the admission threshold where the maximum reward is obtained for threshold 11, which is the same threshold value with zero link wastage and the smallest average waiting in Fig. 14(a).



(a) Tradeoff between delay and link wastage ratio

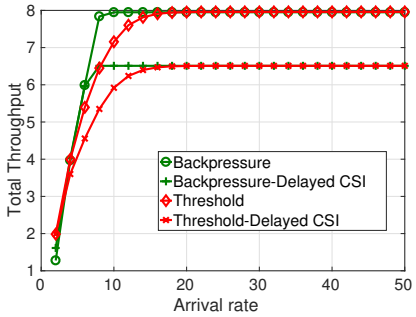


(b) Total reward obtained as a function of threshold

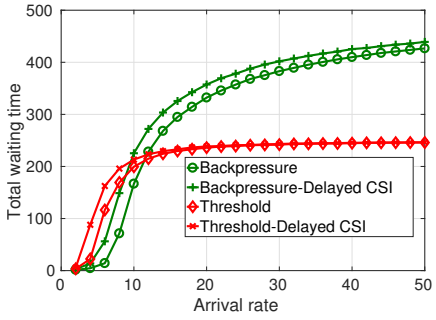
Fig. 14. (a) Trade-off between the average waiting time and link wastage. The control knob is the admission threshold (b) Performance of our proposed framework that maximizes total reward. The optimal threshold results in zero wastage and the lowest delay in Fig. (a).

B. Comparison with Backpressure

In order to optimally design the sub-6 GHz/mmWave transceiver, we represented the transceiver node as a communication network where the sub-6 GHz and mmWave interfaces are modeled as individual network nodes. The objective is to fully exploit the abundant mmWave capacity, while the waiting time is guaranteed to be bounded. In the context of throughput-optimal scheduling, it is well known that traditional network



(a) Throughput performance



(b) Delay performance

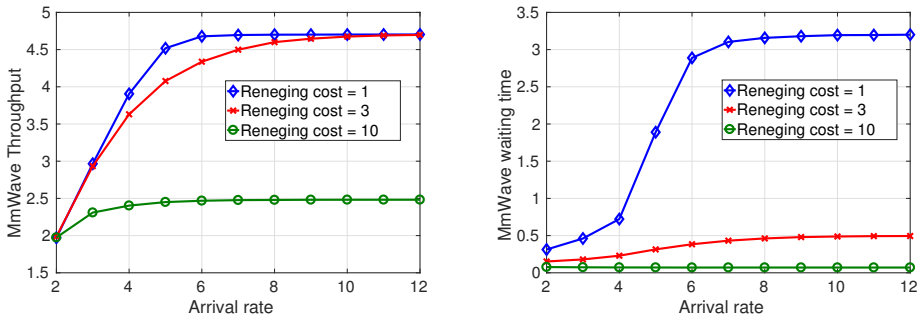
Fig. 15. Throughput and delay performance of our proposed threshold-based policy compared with the Backpressure policy under delayed CSI conditions.

utility optimization, such as Backpressure policy, promises optimal throughput performance for a wide range of networking problems [15]. However, Backpressure policy does not provide any guarantee on the delay performance. Moreover, Backpressure policy requires knowledge of channel state (i.e., link rate), while due to the high data rate of the mmWave interface, real-time tracking of the link state may not be feasible. Therefore, the scheduler node s (in Fig. 9) may obtain information of the data rate of interface $a \in \{\text{mm}, \text{sub-6}\}$ with a delay of τ_a. Under the assumption of delayed network state information, the authors in [23] have shown that the following link selection policy achieves optimal throughput, i.e.,:

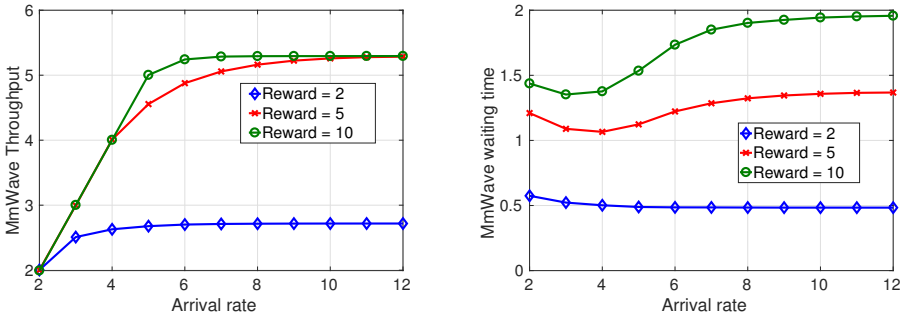
$$a^*(t) = \underset{a \in \{\text{mm, sub-6}\}}{\arg\max} \big[Q_s(t) - Q_a(t)\big]\mathbb{E}[R_a(t)|R_a(t-\tau_a)], \tag{12}$$

in which $a^*(t)$ is the optimal interface selected at time t, and $R_a(t)$ is the link rate of interface a at time t. Under the assumption of delayed CSI, the network stability region (and thus maximum achievable throughput) shrinks as the CSI delay increases [23].

Figure 15(a) and 15(b) demonstrate the throughput and delay performance of our threshold-based scheduler compared with the Backpressure algorithm applied for the network model in Fig. 9. We investigate the performance under both real-time CSI and delayed CSI scenarios. From the results, we observe that the threshold-based scheme achieves a similar throughput performance to Backpressure. However, delay of Backpressure increases with the arrival traffic, while threshold-based policy provides a bounded delay performance. We note that delayed CSI degrades the performance of both policies, however, the threshold-based policy is more robust (in terms



(a) Throughput performance (b) Delay performance



(c) Throughput performance (d) Delay performance

Fig. 16. Throughput and delay performance of our proposed threshold-based policy compared with different values of reneging cost.

of delay performance) towards CSI delay since it is not *directly* expressed in terms of CSI for scheduling, while the Backpressure policy requires CSI for scheduling.

C. Throughput and Delay Tradeoff

We showed that the optimal threshold increases by the reward value r and decreases by the reneging cost c. As a result, depending on the application requirements (throughput vs. latency), the value of r and c are set, and the optimal threshold value is regulated accordingly. Figure 16(a) and 16(b) demonstrate the throughput and delay performance of the threshold-based policy as the reneging cost c increases. From the results, we note that by increasing the value of c, the optimal threshold decreases and thus less packets are admitted to the mmWave queue, as expected. As a result, mmWave waiting time decreases while due to the lack of backlogged packets, throughput performance degrades as well. On the other hand, the trade off between the mmWave throughput and waiting time can be balanced by adjusting the value of reward r. Figure 16(c) and 16(d) illustrate the throughput and delay performance as the reward value r increases. Similarly, as the reward value r increases, the optimal threshold increases and more packets are admitted into the mmWave queue, which results in a higher throughput at the cost of larger waiting time.

D. Temporal Performance

Next, we investigate the throughput performance of the mmWave interface and queue length of the sub-6 GHz interface as a function of time. Figure 17 demonstrates mmWave throughput over time. From the results, we observe that when

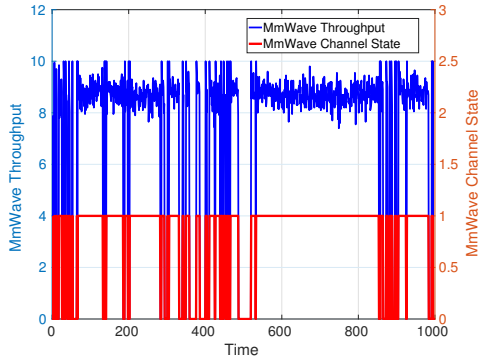


Fig. 17. MmWave throughput performance as a function of time.

the mmWave channel is available (ON state), there are enough packets in the mmWave queue to avoid wasting the abundant capacity available, while throughput drops to zero when the channel becomes unavailable.

Due to the offloading mechanism from mmWave to sub-6 GHz, it is desirable to consider the stability of the sub-6 GHz queue. Figure 18 presents the sub-6 GHz queue length as a function of time where we observe that when the mmWave interface becomes unavailable, the sub-6 GHz queue grows due to the reneging. On the other hand, when the mmWave interface becomes available, the sub-6 GHz queue length shrinks and remains bounded.

VI. Conclusion

In this paper, we proposed an integrated sub-6 GHz/mmWave architecture for 5G cellular systems. Our proposed architecture includes an sub-6 GHz assisted beamforming that exploits the correlation between the

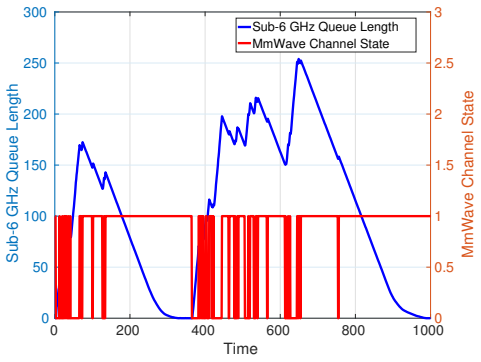


Fig. 18. Sub-6 GHz queue length as a function of time

sub-6 GHz and mmWave interfaces in order to enhance the energy efficiency of mmWave beamforming. In addition to beamforming, we utilized the sub-6 GHz interface for data transfer, and formulated an optimal scheduling policy in order to maximize the long-term throughput of the mmWave interface provided that the average delay is bounded. We cast the constrained throughput maximization as a reward optimization, and proved that the optimal scheduling policy has a simple monotone structure. As a result, using the sub-6 GHz interface as a secondary data transfer mechanism, the abundant yet intermittent mmWave bandwidth is fully utilized. Indeed, we believe that mmWave will most likely be deployed with an overlay of sub-6 GHz in 5G.

ACKNOWLEDGMENT

This work was supported in part by the following grants from the National Science Foundation CNS-1618566 and CNS-1421576.

APPENDIX A
CALCULATING THE LIMITING DISTRIBUTION

In order to characterize the value of optimal threshold, we calculate the limiting distribution of the state of mmWave queue. To this end, the authors in [14] introduced an *embedding technique* such that an embedded process $\{\mathbf{x}_n\}_{n=1}^{\infty}$ is obtained by sampling the process $\{\mathbf{x}(t)\}_{t=1}^{\infty}$ at the beginning of each ON period (see [14] for details). We assume that the limiting distribution of the mmWave queue at state $\boldsymbol{i}$ is denoted by $\xi_{\text{off}}^{(\boldsymbol{i})}$ and $\xi_{\text{on}}^{(\boldsymbol{i})}$ for $L(t)=0$ and $L(t)=1$, respectively:

$$\begin{aligned}\xi_{\text{off}}^{(\boldsymbol{i})} &:= \lim_{t\to\infty}(\mathbf{x}(t)=\boldsymbol{i}, L(t)=0);\\ \xi_{\text{on}}^{(\boldsymbol{i})} &:= \lim_{t\to\infty}(\mathbf{x}(t)=\boldsymbol{i}, L(t)=1).\end{aligned} \tag{13}$$

As in [14], the limiting distribution of all states $\boldsymbol{i}\in\mathcal{S}$ under the OFF and ON link state is then obtained in a matrix form as follows:

$$\begin{aligned}\boldsymbol{\xi}_{\text{off}} &= \frac{\boldsymbol{\nu}\mathbb{E}\big[(\boldsymbol{M}_{\text{on}})^{T_{\text{on}}}\sum_{k=1}^{T_{\text{off}}}(\boldsymbol{M}_{\text{off}})^{k-1}\big]}{\mathbb{E}\big[T_{\text{on}}+T_{\text{off}}\big]};\\ \boldsymbol{\xi}_{\text{on}} &= \frac{\boldsymbol{\nu}\mathbb{E}\big[\sum_{k=1}^{T_{\text{on}}}(\boldsymbol{M}_{\text{on}})^{k-1}\big]}{\mathbb{E}\big[T_{\text{on}}+T_{\text{off}}\big]},\end{aligned} \tag{14}$$

where $\boldsymbol{\nu}$ is the vector of limiting distribution for the embedded process $\{\mathbf{x}_n\}_{n=1}^{\infty}$. Moreover, $\boldsymbol{M}_{\text{off}} = \big[P_{\text{off}}^{(\boldsymbol{i},\boldsymbol{j})}\big]$ and $\boldsymbol{M}_{\text{on}} = \big[P_{\text{on}}^{(\boldsymbol{i},\boldsymbol{j})}\big]$ such that:

$$\begin{aligned}P_{\text{off}}^{(\boldsymbol{i},\boldsymbol{j})} &:= P\big(\mathbf{x}(t+1)=\boldsymbol{j}|\mathbf{x}(t)=\boldsymbol{i}, L(t)=0\big),\\ P_{\text{on}}^{(\boldsymbol{i},\boldsymbol{j})} &:= P\big(\mathbf{x}(t+1)=\boldsymbol{j}|\mathbf{x}(t)=\boldsymbol{i}, L(t)=1\big).\end{aligned} \tag{15}$$

The proof is similar to [14]. Therefore, the limiting distribution vector of the state space $\mathcal{S}$ is obtained as: $\boldsymbol{\xi}=\boldsymbol{\xi}_{\text{off}}+\boldsymbol{\xi}_{\text{on}}$. A sufficient condition for existence of the limiting distribution is that the embedded process has finite state space, which holds in our model due to a bounded queue length and waiting time. Our model involves an admission policy that regulates the arrival process, and thus length of the mmWave queue does not exceed an optimal threshold h^*. To denote the limiting distribution at the state $\mathbf{x}=(Q,D)$, we use the notation $\xi_{(Q,D)}$.

Morteza Hashemi (S'10) is currently a postdoctoral researcher at the Ohio State University. He received his PhD and MSc degrees in Electrical and Computer Engineering from Boston University in 2015 and 2013, respectively. He received his BSc degree in Electrical Engineering from Sharif University of Technology, Tehran, Iran (2011). His research interests span the broadly defined areas of wireless communications, information systems, real-time data networking, networks performance evaluation, learning, and cyber-physical systems. He is the recipient of the Boston University Provost's Award in 2014 for excellence in research.

C. Emre Koksal (S'96-M'03-SM'13) received the B.S. degree in Electrical Engineering from the Middle East Technical University in 1996, and the S.M. and Ph.D .degrees from MIT in 1998 and 2002, respectively, in Electrical Engineering and Computer Science. He was a Postdoctoral Fellow at MIT until 2004, and a Senior Researcher at EPFL until 2006. Since then, he has been with the Electrical and Computer Engineering Department at Ohio State University, currently as an Associate Professor. His general areas of interest are wireless communication, communication networks, information theory, and cybersecurity.

He is the recipient of the NSF CAREER Award in 2011, a finalist of the Bell Labs Prize in 2015, the OSU College of Engineering Lumley Research Award in 2011 and 2017, Innovators Award in 2016, and the co-recipient of an HP Labs - Innovation Research Award in 2011. The paper he co-authored was a best student paper candidate in MOBICOM 2005. He has been an Associate Editor for IEEE Transactions on Wireless Communication between 2012-2016 and currently, he is an Associate Editor for IEEE Transactions on Information Theory and Elsevier Computer Networks.

Ness B. Shroff (S'91-M'93-SM'01-F'07) received his Ph.D. degree in Electrical Engineering from Columbia University in 1994. He joined Purdue university immediately thereafter as an Assistant Professor in the school of ECE. At Purdue, he became Full Professor of ECE in 2003 and director of CWSA in 2004, a university-wide center on wireless systems and applications. In July 2007, he joined The Ohio State University, where he holds the Ohio Eminent Scholar endowed chair in Networking and Communications, in the departments of ECE and CSE. He holds or has held visiting (chaired) professor positions at Tsinghua University, Beijing, China, Shanghai Jiaotong University, Shanghai, China, and the Indian Institute of Technology, Bombay, India. Dr. Shroff is currently an editor at large of IEEE/ACM Trans. on Networking, senior editor of IEEE Transactions on Control of Networked Systems, and technical editor for the IEEE Network Magazine. He has received numerous best paper awards for his research and listed Thomson Reuters Book on The World's Most Influential Scientific Minds as well as noted as a highly cited researcher by Thomson Reuters. He also received the IEEE INFOCOM achievement award for seminal contributions to scheduling and resource allocation in wireless networks.

REFERENCES

[1] F. Khan and Z. Pi, "mmWave mobile broadband (MMB): Unleashing the 3–300GHz spectrum," in *34th IEEE Sarnoff Symposium*, 2011.

[2] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *Access, IEEE*, vol. 1, pp. 335–349, 2013.

[3] W. Roh, J.-Y. Seol, J. Park, B. Lee, J. Lee, Y. Kim, J. Cho, K. Cheun, and F. Aryanfar, "Millimeter-wave beamforming as an enabling technology for 5G cellular communications: theoretical feasibility and prototype results," *IEEE Communications Magazine*, vol. 52, no. 2, 2014.

[4] J. Mo, A. Alkhateeb, S. Abu-Surra, and R. W. Heath Jr, "Hybrid architectures with few-bit ADC receivers: Achievable rates and energy-rate tradeoffs," *arXiv preprint arXiv:1605.00668*, 2016.

[5] T. Nitsche, A. B. Flores, E. W. Knightly, and J. Widmer, "Steering with eyes closed: mm-wave beam steering without in-band measurement," in *Computer Communications (INFOCOM), IEEE Conference on*. IEEE, 2015, pp. 2416–2424.

[6] A. Ali, N. Prelcic, and R. Heath, "Estimating millimeter wave channels using out-of-band measurements," *Information Theory and Applications Workshop (ITA)*, 2016.

[7] M. Hashemi, C. E. Koksal, and N. B. Shroff, "Hybrid RF-mmWave communications to achieve low latency and high energy efficiency in 5G cellular systems," in *Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), 15th International Symposium on*. IEEE, 2017.

[8] S. Collonge, G. Zaharia, and G. E. Zein, "Influence of the human activity on wide-band characteristics of the 60 GHz indoor radio channel," *Wireless Communications, IEEE Transactions on*, vol. 3, no. 6, pp. 2396–2406, 2004.

[9] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter-wave cellular wireless networks: Potentials and challenges," *Proceedings of the IEEE*, vol. 102, no. 3, pp. 366–385, 2014.

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

[11] V. Nurmela, A. Karttunen, A. Roivainen, L. Raschkowski, T. Imai, J. Jarvelainen, J. Medbo, J. Vihriala, J. Meinila, K. Haneda *et al.*, "METIS channel models," *Seventh FrameworN Programme ICT-317669*, 2015.

[12] A. Adhikary, E. Al Safadi, M. K. Samimi, R. Wang, G. Caire, T. S. Rappaport, and A. F. Molisch, "Joint spatial division and multiplexing for mm-wave channels," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1239–1255, 2014.

[13] A. Ali, N. González-Prelcic, and R. W. Heath Jr, "Millimeter wave beam-selection using out-of-band spatial information," *arXiv preprint arXiv:1702.08574*, 2017.

[14] Y. Kim, K. Lee, and N. B. Shroff, "An analytical framework to characterize the efficiency and delay in a mobile data offloading system," in *Proceedings of the 15th ACM international symposium on Mobile ad hoc networking and computing*. ACM, 2014, pp. 267–276.

[15] L. Georgiadis, M. J. Neely, and L. Tassiulas, *Resource allocation and cross-layer control in wireless networks*. Now Publishers Inc, 2006.

[16] M. Alresaini, M. Sathiamoorthy, B. Krishnamachari, and M. J. Neely, "Backpressure with adaptive redundancy (BWAR)," in *INFOCOM, 2012 Proceedings IEEE*. IEEE, 2012, pp. 2300–2308.

[17] X. L. Bui, E. Athanasopoulou, T. Ji, R. Srikant, and A. Stolyar, "Backpressure-based packet-by-packet adaptive routing in communication networks," 2012.

[18] B. Ji, C. Joo, and N. B. Shroff, "Delay-based back-pressure scheduling in multihop wireless networks," *IEEE/ACM Transactions on Networking (TON)*, vol. 21, no. 5, pp. 1539–1552, 2013.

[19] J. G. Andrews, T. Bai, M. Kulkarni, A. Alkhateeb, A. Gupta, and R. W. Heath Jr, "Modeling and analyzing millimeter wave cellular systems," *arXiv preprint arXiv:1605.04283*, 2016.

[20] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[21] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[22] M. Larrañaga, O. J. Boxma, R. Núñez-Queija, and M. S. Squillante, "Efficient content delivery in the presence of impatient jobs," in *Teletraffic Congress (ITC 27), 27th International*. IEEE, 2015, pp. 73–81.

[23] L. Ying and S. Shakkottai, "On throughput optimality with delayed network-state information," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 5116–5132, 2011.