Delay-Optimal Scheduling for Integrated mmWave – Sub-6 GHz Systems with Markovian Blockage Model

Guidan Yao, Morteza Hashemi, Rahul Singh and Ness B. Shroff, *Fellow, IEEE*

Abstract—Millimeter wave (mmWave) communication has the potential to achieve very high data rates but is highly vulnerable to blockage. In this paper, we provision an integrated mmWavesub-6 GHz architecture to combat blockage and intermittent connectivity of the mmWave communications. To this end, we model the mmWave channel as a two-state Markov channel and investigate the problem of scheduling packets across the mmWave and sub-6 GHz interfaces such that the long-term average delay of system is minimized. We prove that the optimal policy is of a threshold-type with state-dependent thresholds, i.e., packets should always be routed to the mmWave interface as long as the number of packets in the system is smaller than the state-dependent threshold. Numerical results demonstrate that under heavy traffic, integrating sub-6 GHz with mmWave can reduce the average delay by over 70%. Moreover, considering the difficulty of tracking the mmWave channel state in practice, we develop heuristics of substituting a single fixed threshold (state-independent) for two state-dependent thresholds. Our simulation results indicate that the replacement only incurs a slight increase in average delay. Moreover, when system parameters are not known, we propose a certainty-equivalence threshold-based learning algorithm, and provide an upper bound on its regret.

Index Terms—Delay optimization, millimeter wave, Markov decision process, learning.

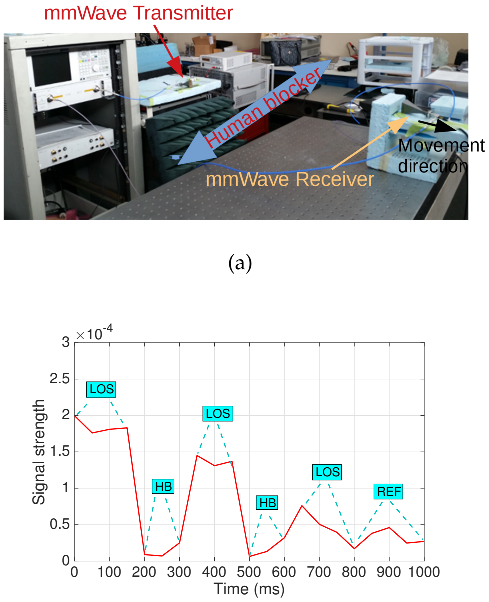
1 Introduction

The annual amount of mobile data traffic is projected to reach almost one zettabyte by 2022 [1]. The deluge of data traffic, especially the recent demands for data-intensive applications enabled by 5G and beyond, will only worsen the spectrum crunch that service providers are already experiencing. The bandwidth available in the millimeter wave (mmWave), ranging from 30 GHz to 300 GHz, has the potential to mitigate the spectrum scarcity in the sub-6 GHz band by enabling wireless communication at data rate of several Gbps.

One of the main hurdles to achieve reliable and robust mmWave communication is related to blockage. In particular, due to small wavelengths in the mmWave band, most objects such as concrete walls and human bodies cause blocking and reflections as opposed to scattering and diffraction in the sub-6 GHz frequencies. When the line-of-sight (LOS) is blocked, mmWave channel suffers from highly dynamic occasionally zero-throughput connectivity that further degrades upper layer performance [2], [3], [4]. Thus, the mmWave links may exhibit intermittent connectivity with ON/OFF (or available/unavailable) periods under blockage [2], [5]. For instance, the human body increases the mmWave path loss by more than 20 dB that can completely break the link and result in an almost zero data rate [2], [5], [6], [7]. To demonstrate the effect of human blockage on mmWave links, we have conducted a set of measurements with a stationary transmitter and a mobile receiver that moves away from the transmitter with the speed of 1 m/s. During the time intervals $200-300$ and $500-600$ ms, a human body blocks the line-of-sight (LOS) path between the transmitter and receiver. Figure 1a shows our basic experimental setup, and Fig. 1b depicts the strength of received signal at the mobile receiver over time [8]. From the results, we see that the received signal strength falls to almost zero under blockage, which can be modeled as an OFF or unavailable period. Therefore, the mmWave link exhibits an available/unavailable connectivity pattern under blockage scenarios such that during the unavailable periods, delivery rate and delay performance highly degrade. Besides, in this experiment, there are some metal reflectors placed around the transmitter and receiver. As such, when the mmWave receiver is moved away from the transmitter, the reflector provides an additional signal rays during specific time intervals, which in turn, increases the received signal strength.

In order to combat the blockage and achieve robust mmWave communication, some solutions have already been proposed, which will be further discussed in Section 1.1. Briefly speaking, one class of works study the ways to switch among mmWave paths to avoid blocked ones. However, it may happen that mmWave links (from a source or to a destination) in all directions are blocked. Thus, these methods cannot guarantee a robust connection, motivating the need to use a relatively reliable carrier to assist the mmWave. Another class of works suggest using sub-6 GHz

- *Guidan Yao is with the Department of ECE, The Ohio State University, Columbus, OH 43210 USA.
E-mail: yao.539@osu.edu*
- *Morteza Hashemi is with the Department of EECS, The University of Kansas, Lawrence, KS 66045 USA.
E-mail: mhashemi@ku.edu*
- *Rahul Singh is with the Department of ECE, Indian Institute of Science, Bangalore, Karnataka, 560012, India.
E-mail: rahulsiitk@gmail.com*
- *N. B. Shroff is with the Department of ECE and the Department of CSE, The Ohio State University, Columbus, OH 43210 USA.
E-mail: shroff.11@osu.edu*



(a)

(b)

Fig. 1: Effect of human blockage on mmWave channels. Figure (a) shows measurement setup and experiment scenario. Figure (b) depicts received mmWave signal strength under line of sight (LOS), human blocker (HB), and reflection (REF). The received signal power is measured and reported in terms of Watts. [8].

to assist the mmWave communication. But in these works on integrated mmWave and sub-6 GHz system, the delay optimization problem has not been extensively explored. Compared with these works, our paper aims to develop the optimal policies for the delay minimization problem in the integrated mmWave/sub-6 GHz system considering the rapid increase in delay-sensitive applications. In the preliminary version of this paper [9], we considered a memoryless available/unavailable mmWave channel *without* knowledge of the state of mmWave channel. To combat blockage, the sub-6 GHz interface is exploited as a fallback data transfer mechanism such that packets may be routed to the sub-6 GHz interface upon arriving to the system. Moreover, packets are allowed to be impatient in the sense that they can "renege" from the mmWave interface to the sub-6 GHz interface, when the waiting time of the head-of-line packet in the mmWave interface becomes large. Under this setting, we developed a threshold-type policy with a single fixed threshold and showed its optimality for minimizing both discounted and average delay in the integrated mmWave and sub-6 GHz system. Compared with the Bernoulli channel used in [9], this paper considers a more general channel model (Markov channel). In particular, we model the mmWave channel as a two-state Markov chain, which is experimentally shown to be able to characterize mmWave channel in the presence of blockage [10]. Although the model is simple, it roughly captures the intermittency of the mmWave channel. For mmWave channels with a larger state space i.e. the mmWave channel may have different average service rates (transmission rates) when they are not blocked, the optimal policy can be computed via dynamic programming but the computation cost is usually prohibitive. Thus, we believe the optimal policy obtained for the two-state Markov model is of interest since it allows us to gain deeper insight into how the channel parameters affect the scheduling choices.

To develop a theoretical framework for analyzing the delay performance, we first assume that the scheduler has access to the mmWave channel state. Under this assumption, we develop the optimal packet scheduling policy, which is expressed in terms of the mmWave channel state. In particular, the optimal policy is shown to be of a threshold-type with two state-dependent thresholds such that the optimal thresholds depend on the mmWave channel state. Next, we relax the assumption that the scheduler can access to the mmWave channel state in order to develop a scheduling policy that is independent of the mmWave channel state. This is especially important from practical point of view since the link speed of the mmWave interface (multi Gbps) is comparable to the speed at which a typical processor in a smart communication device operate, which makes tracking the channel state challenging. In particular, we develop heuristics of substituting a single fixed threshold (state-independent) for two state-dependent thresholds. Our numerical results indicate that such a replacement only incurs a slight increase in average delay over the optimal policy. In this case, the fixed threshold can be obtained using our model in [9] by replacing the "available" probability of the Bernoulli mmWave channel with the steady state probability that the Markovian mmWave channel is available. Moreover, we consider this scheduling problem in an unknown environment, where system parameters including arrival rates of packets and statistics of channel dynamics are not known a priori. This scenario can model the time-varying system. We employ bandit and threshold property to develop an efficient learning algorithm, which learns the parameters and thus minimize the delay. We also prove the upper bound for the regret of our learning algorithm.

1.1 Related works

There exist some works dealing with intermittent mmWave communications. One class of works investigates methods to switch from the unavailable mmWave path/link to the available mmWave path/link based on the state of mmWave links. In [11], the authors model the problem of cell selection in mmWave cellular networks as a Markov decision problem (MDP) to deal with intermittency of mmWave link. In [12], the authors develop a learning approach to access network selection in order to maximize throughput. The authors in [13], [14] consider multi-hop communication, in which relays are used to form an available path that goes around the blocked one. Another class of works suggest integrating a more reliable carrier, i.e., sub-6 GHz with mmWave. The authors in [8], [15] consider resource allocation and cooperative communication between the sub-6 GHz and mmWave to maximize the throughput of the system. In [16], the authors propose a novel dual-mode scheduling framework that jointly conducts user applications selection and scheduling over mmWave and sub-6 GHz bands in order to minimize the number of unsatisfied user applications. The paper [17] provides a roadmap towards realizing the integrated mmWave/sub-6 GHz networks, which may achieve joint mobile broadband and ultra-reliable low latency communication (URLLC), by introducing new designs for the radio interface and frame structure with flexible numerology and transmission time interval (TTI). However, delay optimization in the integrated mmWave and sub-6 GHz system, which is studied in this paper, has not been extensively studied in these works.

In the integrated mmWave and sub-6 GHz system, the mmWave interface acts as the fast but unreliable server whose service can be blocked completely, while the sub-6 GHz interface acts as a slow but reliable server. The most related class of works to our delay minimization problem is the *slow-server* problem, in which the goal is to obtain a delay optimal scheduling policy in a queuing system with heterogeneous (i.e., fast and slow) servers. The goal of this problem is to investigate the trade-off between waiting in queue and entering slow servers when fast servers are busy. The slow-server problem was first proposed in [18], where the authors presented a M/M/2 queuing system with two heterogeneous servers and conjectured that the optimal policy for minimizing the average delay and expected total discounted delay in system is of a threshold-type. The conjecture was then proven in [19] with policy iteration. Later, [20] and [21] showed the same result with coupling arguments and value iteration, respectively. Following these works, [22] extended the result to the system with multi-servers (i.e., more than two), and [23], [24] studied the delay minimization problem with different arrival and service processes. In [24], the authors took the failure of service into consideration and showed that the optimal policy to minimize the long-term average number of customers in system is also of a threshold type. Different from the parallel structure in the slow-server problem, our architecture is a mix of tandem and parallel queues (see Fig. 2). Moreover, we allow for a reneging action in the system which complicates the relationships among actions, i.e., we have to further consider the trade-off between waiting in the mmWave interface and reneging to the sub-6 GHz as well (details are discussed in section II).

1.2 Key contributions

We consider an integrated mmWave and sub-6 GHz system, and develop a delay-optimal scheduling policy for such a system. Our key contributions are as follows:

- We investigate the policy that minimizes the *expected total discounted delay* and through value iteration of Markov Decision Process (MDP), we obtain three rules that partially characterize the optimal policy. Based on the findings, we propose a threshold-type policy with state-dependent thresholds. Then, we collapse our system state space from five dimensions to four dimensions, and further demonstrate the optimality of the proposed policy. We further show that the proposed policy is also optimal for the *average delay problem*.
- We develop a technique for solving the delay minimization problem in settings consisting of tandem and parallel queues with heterogeneous servers. In particular, tandem queues exist in one branch of two parallel queues (see Fig. 2). This implies that our architecture is a mix of tandem and parallel queues, which is different from the parallel structure in the slow-server problem (introduced in Section 1.1).

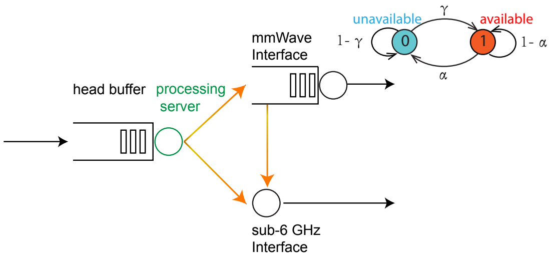


Fig. 2: Integrated sub-6 GHz and mmWave architecture. The system consists of a head buffer, a processing server, an mmWave interface and a sub-6 GHz interface. Packets that arrive at this system, wait in the head buffer to be processed and served by either mmWave or sub-6 GHz.

- We consider the case for which system parameters (service rates, arrival rate, and transition rates of channels) are not known. In this case, we propose a certainty equivalence-threshold based learning algorithm and provide an upper bound for the regret of the algorithm.
- We numerically verify that it is important to use the sub-6 GHz especially when the system is in heavy traffic. We develop heuristics of substituting a single fixed threshold for two state-dependent thresholds to make our policy implementable in practice. Numerical results indicate that the replacement only results in a slight increase in average delay over optimal policy.

To improve readability, we summarize primary notations of this paper in Table 1.

2 Problem Setup

In this section, we present the system model and formulate the delay minimization problem.

2.1 System Model

We consider an integrated communication architecture with dual sub-6 GHz and mmWave interfaces as shown in Fig. 2. The infinite *head buffer* is utilized to store all packets waiting to be processed and served by either mmWave or sub-6 GHz. The *processing server* is responsible for essential data processing before scheduling. In addition, the system includes two servers (mmWave and sub-6 GHz servers) with different service rates, e.g., mmWave spectrum can deliver theoretical speeds as high as 5Gb/s while sub-6 GHz, by far the most common, will usually deliver between 100 and 400 Mbit/s [25].

(i) Queueing Models: In our system model, we add a buffer to the mmWave server, which stores packets routed from the head buffer. The rationale behind our design (i.e., a separate queue for the mmWave interface) is described next. The service rate of the mmWave server is comparable to the processing server (i.e., processor speed). If we assume that there is no buffer for the mmWave server, then every packet would need to wait in the head queue until the mmWave server is available. In this case, the packet will experience the service time of both the processing and the mmWave

TABLE 1: List of Notations

Symbol	Description
γ	The transition rate from unavailable to available state.
α	The transition rate from available to unavailable state.
λ	Arrival rate.
μ_{mm}	Service rate of the mmWave interface.
μ_p	Service rate of the processing server.
$\mu_{\text{sub-6}}$	Service rate of the sub-6 GHz interface.
q_0	Queue length of the head buffer.
q_1	Queue length of the mmWave interface.
l_1	Busy/idle condition of the processing server.
l_2	Busy/idle condition of the sub-6 GHz interface.
s	State of the mmWave link.
$\mathcal{A}_0(\cdot)$	Arrival event.
$\mathcal{D}_1(\cdot)$	Departure of a packet from the mmWave interface.
$\mathcal{D}_2(\cdot)$	Departure of a packet from the sub-6 GHz interface.
$\mathcal{T}(\cdot)$	Processing completion.
$\mathcal{B}(\cdot)$	MmWave link becomes unavailable.
$\mathcal{G}(\cdot)$	MmWave link becomes available.
$K = \{A_h, A_1, A_2, A_b, A_r\}$	The set of allowed actions. A_h denotes holding action; A_1 dispatches a packet to the mmWave line; A_2 dispatches a packet to the sub-6 GHz server; A_b dispatches two packets to the two servers; A_r moves a packet from the mmWave line to the sub-6 GHz interface.
$K_{\mathbf{q}}$	The set of admissible actions in state $\mathbf{q}$.
$A_{r_{\text{p}}}$	The reneging action from the processing server.
$A_{r_{\text{mm}}}$	The reneging action from the mmWave interface.
$J_\beta(\mathbf{q})$	The optimal expected total discounted delay with initial state $\mathbf{q}$.
$\mathbf{m}^\star = (m_0^\star, m_1^\star)$	Optimal thresholds: $m_0^\star$, $m_1^\star$ denote optimal thresholds when mmWave channel is unavailable and available, respectively.
$\pi_{\mathbf{m}^\star} = \{D_{\mathbf{m}^\star}, D_{\mathbf{m}^\star}, \cdots\}$	Optimal threshold-type policy with decision rule $D_{\mathbf{m}^\star}$.
τ_k	The beginning of the k-th episode.
ϵ_k	The set of consecutive time slots that constitute the k-th episode.
$N_A(n)$	The number of arrivals until the n-th decision epoch.
$N_p(n)$	The number of service completions on processing server until the n-th decision epoch.
$N_{sub-6}(n)$	The number of service completions on sub-6 channel until the n-th decision epoch.
$N_{mm}(n)$	The number of service completions on mmWave channel until the n-th decision epoch.
$N_{i \to j}(n)$	The number of i to j channel state transitions until n-th decision epoch.
$I_A(\ell)$	The time between the arrival of ℓ-th and $\ell+1$-th packet.
$S_p(\ell)$	The time taken for completion of the ℓ-th service at the processing server.
$S_{sub-6}(\ell)$	The time taken for completion of the ℓ-th service at the sub-6 channel.
$S_{mm}(\ell)$	The time taken for completion of the ℓ-th service at the mmWave channel.

servers (almost double the service time of the mmWave) except waiting time in the head buffer. Then, the performance of mmWave is degraded by approximately half. In contrast, if the mmWave server has its own buffer for the processed packets, part of the waiting time in the head buffer can be utilized to process packets in advance, which reduces the experienced service time mentioned above. However, the sub-6 GHz link is much slower than the processing server. Therefore, the processing delay can be ignored compared to service time of the sub-6 GHz. In other words, it is not necessary for the sub-6 GHz server to have its own buffer. Thus, it is appropriate to assume that the *sub-6 GHz interface* acts as a server with a buffer size of one, while the *mmWave interface* consists of an infinite buffer and a server.

(ii) Two-state Markovian mmWave link; Available or Unavailable: As mentioned before, the mmWave link is highly variable with intermittent ON/OFF periods under the impact of blockage that can result in approximately zero throughput. As such, we model the mmWave channel as a Markov chain with two states: *available* state denoted by 1 and *unavailable* state denoted by 0 as in Fig. 2. The transition rate matrix governing the channel model is denoted as

$$R = \begin{bmatrix} 1-\alpha & \alpha \\ \gamma & 1-\gamma \end{bmatrix}, \tag{1}$$

where γ (or α) denotes the transition rate from unavailable (or available) to available (or unavailable) state. For the unavailable state, the mmWave channel is almost disconnected and thus we assume that the service (transmission) rate of the mmWave is 0. For the available state, we assume that the service (transmission) time of a packet is exponentially distributed with parameter μ_{mm}. Note that the Markov channel model is more general compared to the Bernoulli channel in the preliminary version of this paper [9].

We further assume that arrivals to the system form a Poisson process with rate λ, and that the service times of the processing server and the sub-6 GHz interface are exponentially distributed with means $\frac{1}{\mu_{\text{p}}}$ and $\frac{1}{\mu_{\text{sub-6}}}$, respectively. Given that the mmWave service rate is of the same order as the clock speed of the processor (i.e., multi-GHz), we assume that μ_{p} is much faster than $\mu_{\text{sub-6}}$ but on the same order as μ_{mm} (i.e., $\mu_{\text{p}} > \mu_{\text{sub-6}}$ and $\mu_{\text{mm}} > \mu_{\text{sub-6}}$). Since the delay of the processing server becomes negligible compared with the sub-6 GHz interface, we consider the equivalent model depicted in Fig. 3 where we call the processing server and mmWave interface as *mmWave line.*

Within this content, we further clarify the difference of our problem from previous work, which has been briefly discussed in Section 1.1. In Fig. 3, packets that are scheduled to the mmWave line have to go through a processing server first. This makes our system a mix of the tandem and parallel queues, which implies that our problem is more complex than the classic slow-server problem.

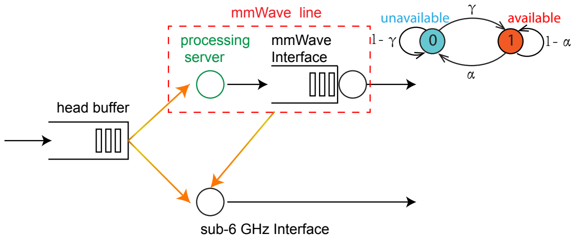


Fig. 3: Equivalent system model. We omit the processing server on the way where a packet is dispatched from the head buffer to the sub-6 GHz since the delay of the processing server becomes negligible compared with the sub-6 GHz interface.

To avoid a large waiting time in the mmWave queue due to the intermittent channel (e.g., due to blockage), 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 becomes large, the packet "reneges" (is moved to) from the mmWave line or "routes" from the head buffer (is dispatched to) to the sub-6 GHz interface. Note that the packet in the sub-6 GHz server cannot be sent back to the mmWave line or the head buffer. Adding the reneging concept introduces new challenges such as: *should the packets be moved from the head buffer or the mmWave queue to the sub-6 GHz server?* Therefore, in addition to the trade-off between waiting in the head queue and entering the slow server, which is investigated in the slow-server problem, we investigate: (i) the trade-off between waiting in the mmWave line and entering the slow server, and (ii) trade-off between dispatching packets from the head buffer and the mmWave line.

2.2 System Dynamics

(i) States: Let q_0, q_1 $\in \mathbb{N}$ denote the queue length of the head buffer and mmWave interface, respectively. Let l_1, l_2 $\in \{0,1\}$ denote the busy/idle condition of the processing server and sub-6 GHz interface, respectively. In this case, $l_1 = 1$ ($l_2 = 1$) implies that the processing server (sub-6 GHz interface) is busy. Moreover, $s \in \{0,1\}$ denotes the state of the mmWave link, where $s = 1$ ($s = 0$) corresponds to available (unavailable) state. Therefore, the system state can be expressed by a five-dimensional vector $\mathbf{q} \triangleq (q_0, l_1, q_1, l_2, s)$ with the state space of $\mathcal{X} \triangleq \mathbb{N} \times \{0,1\} \times \mathbb{N} \times \{0,1\} \times \{0,1\}$.

(ii) Events: Six different events that happen in the system are defined as follows:

(1) Arrival of a packet to the head buffer: $\mathcal{A}_0(\mathbf{q}) \triangleq (q_0 + 1, l_1, q_1, l_2, s)$.

(2) Departure of a packet from the mmWave interface: This happens only when the mmWave link is available ($s = 1$) and changes the system state as: $\mathcal{D}_1(\mathbf{q}) \triangleq \left(q_0, l_1, (q_1 - 1)^+, l_2, 1\right)$, where $(\cdot)^+ = \max(\cdot, 0)$.

(3) Departure of a packet from the sub-6 GHz interface: $\mathcal{D}_2(\mathbf{q}) \triangleq \left(q_0, l_1, q_1, (l_2 - 1)^+, s\right)$.

(4) Processing completion: If the processing server delivers a packet to the mmWave queue, the system state changes as: $\mathcal{T}(\mathbf{q}) \triangleq \left(q_0, (l_1 - 1)^+, l_1 + q_1, l_2, s\right)$.

(5) mmWave link becomes unavailable: $\mathcal{B}(\mathbf{q}) \triangleq (q_0, l_1, q_1, l_2, 0)$.

(6) mmWave link becomes available: $\mathcal{G}(\mathbf{q}) \triangleq (q_0, l_1, q_1, l_2, 1)$.

Note that we introduce "dummy" packets for events in (2)-(4) when $q_1 = 0$, $l_2 = 0$ and $l_1 = 0$, respectively. This is further elaborated in Section 2.3.

(iii) Actions: $K = \{A_h, A_1, A_2, A_b, A_r\}$ is the set of allowed controls or actions. $K_{\mathbf{q}} \subseteq K$ denotes the set of admissible actions in state $\mathbf{q}$. Each action in set K is defined as follows:

(1) Holding: A_h keeps the system state unchanged, i.e., $A_h(\mathbf{q}) \triangleq (q_0, l_1, q_1, l_2, s)$, $\mathbf{q} \in \mathcal{X}$.

(2) Scheduling-on-mmWave: A packet can be routed to the mmWave line if the processing server is idle, i.e., $A_1(\mathbf{q}) \triangleq (q_0 - 1, 1, q_1, l_2, s)$, $\mathbf{q} \in \{\mathbf{q} : q_0 \geq 1, l_1 = 0\}$.

(3) Scheduling-on-sub-6: A packet can be routed to the sub-6 GHz interface if the sub-6 GHz server is idle, i.e., $A_2(\mathbf{q}) \triangleq (q_0 - 1, l_1, q_1, 1, s)$, $\mathbf{q} \in \{\mathbf{q} : q_0 \geq 1, l_2 = 0\}$.

(4) Scheduling-on-both: If both the sub-6 GHz and processing servers are available, two packets can be dispatched to the two servers simultaneously, i.e.,

$$A_b(\mathbf{q}) \triangleq (q_0 - 2, 1, q_1, 1, s), \quad \mathbf{q} \in \{\mathbf{q} : q_0 \geq 2, l_1 = l_2 = 0\}.$$

(5) Reneging: Action A_r moves a packet from the mmWave line to the sub-6 GHz interface, and it is defined on the set $\{\mathbf{q} : q_1 + l_1 \geq 1, l_2 = 0\}$. Let $A_{r_{\mathrm{p}}}$ and $A_{r_{\mathrm{mm}}}$ denote the reneging actions from the processing server and mmWave interface, respectively. Therefore, we have:

$$\begin{aligned} &A_{r_{\mathrm{p}}}(\mathbf{q}) \triangleq (q_0, 0, q_1, 1, s), && \mathbf{q} \in \{\mathbf{q} : l_1 = 1,\ l_2 = 0\}; \\ &A_{r_{\mathrm{mm}}}(\mathbf{q}) \triangleq (q_0, l_1, q_1 - 1, 1, s), && \mathbf{q} \in \{\mathbf{q} : q_1 \geq 1,\ l_2 = 0\}. \end{aligned}$$

Then, A_r is expressed as follows:

$$A_r(\mathbf{q}) \triangleq \begin{cases} A_{r_{\mathrm{p}}}(\mathbf{q}) & \text{if } l_1 = 1,\ q_1 = 0 \\ A_{r_{\mathrm{mm}}}(\mathbf{q}) & \text{if } l_1 = 0,\ q_1 \geq 1 \\ \min\limits_{a \in \{r_{\mathrm{p}}, r_{\mathrm{mm}}\}} v(A_a(\mathbf{q})) & \text{otherwise,} \end{cases}$$

where $v(\cdot)$ denotes the delay cost. Note that if $A_{r_{\mathrm{p}}}$ and $A_{r_{\mathrm{mm}}}$ are admissible, we select an action that results in a smaller cost. In Section 3, we show that $A_r = A_{r_{\mathrm{p}}}$ for the discounted delay problem when both $A_{r_{\mathrm{p}}}$ and $A_{r_{\mathrm{mm}}}$ are admissible.

2.3 Problem Formulation

A scheduling policy π specifies the action selection rule for each decision epoch when an event happens. We use Π to denote the set of all admissible scheduling policies, where admissible means that each action selected in a certain state is admissible in that state. Further, whenever an event happens, we select a control variable from the set $U \triangleq \{(u_0, u_1, u_2, u_3, u_4, u_5) : u_0, u_1, u_2, u_3, u_4, u_5 \in K\}$, where $u_0, u_1, u_2, u_3, u_4, u_5$ are selected actions corresponding to events $\mathcal{A}_0, \mathcal{T}, \mathcal{D}_1, \mathcal{D}_2, \mathcal{B}, \mathcal{G}$, respectively. In particular, if the system occupies state $\mathbf{q}$ and $\mathbf{u} = (u_0, u_1, u_2, u_3, u_4, u_5)$ is selected at a certain decision epoch, then we know that if an arrival occurs at the next decision epoch, we would take action u_0. Similar explanation applies to u_1, u_2, u_3, u_4, u_5.

Average Delay Problem: Our objective is to find an admissible scheduling policy π that minimizes the average delay in the system. By Little's Law, minimizing the delay is equivalent to minimizing the total number of packets in the system. Thus, the problem is expressed as follows:

$$\min_{\pi \in \Pi} \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}^{\pi} \left[\int_{t=1}^{T} c(\mathbf{q}(t))\, dt \right], \tag{2}$$

where $\mathbb{E}^{\pi}$ denotes the conditional expectation given policy π, $\mathbf{q}(t) \in \mathcal{X}$ is the system state at time t, and $c\left(\mathbf{q}\right)$ is a function of the system state that computes the total number of packets in the system, i.e.,

$$c\left(\mathbf{q}\right) \triangleq q_0 + q_1 + l_1 + l_2. \tag{3}$$

The problem is a continuous-time MDP, and for simplicity, we convert the continuous-time MDP problem into an equivalent discrete-time MDP problem using uniformization [26]. In particular, we assume that all servers will serve "dummy" packets whenever they are idle. Then, we separate continuous time into time slots with sequences when either a packet arrival or a packet (real or dummy) departure (from the processing server or interfaces) or change of the state of the mmWave link occurs.

The technical analysis for uniformization can be found in [26]. The heuristic explanation is as follows. To ensure that the transformed discrete-time system is equivalent to the continuous-time system, the sampling rate across the continuous time should be the same. If we only consider real packets, then the sampling rate can be different. For example, for the case that the mmWave channel is in good state, the sampling rate is $\lambda + \mu_{\text{mm}} + \alpha$ when only the mmWave interface is busy, while it is $\lambda + \mu_{\text{sub-6}} + \alpha$ when only the sub-6 GHz interface is busy.

Let $n \in \mathbb{N}$ denote the beginning of the n-th time slot (decision epoch) when a certain event happens. Furthermore, without loss of generality, we scale time and assume that $\lambda + \mu_{\text{p}} + \mu_{\text{mm}} + \mu_{\text{sub-6}} + \alpha + \gamma = 1$.

Let $\mathbf{q}$ be the system state just before an event happens. After an event happens, certain actions will be selected. Let $\mathbf{q}'$ denote the system state after the action is taken. Then, the transition probability that goes from one state to another under certain action is expressed as:

$$\mathbb{P}\left(\mathbf{q}'|\mathbf{q},\mathbf{u}\right) = \begin{cases} \lambda & \text{if } \mathbf{q}' = u_0\left(\mathcal{A}_0\left(\mathbf{q}\right)\right) \\ \mu_{\text{p}} & \text{if } \mathbf{q}' = u_1\left(\mathcal{T}\left(\mathbf{q}\right)\right) \\ \mu_{\text{mm}} & \text{if } \mathbf{q}' = u_2\left(\mathcal{D}_1\left(\mathbf{q}\right)\right) \text{ and } s_{\mathbf{q}} = 1 \\ \mu_{\text{sub-6}} & \text{if } \mathbf{q}' = u_3\left(\mathcal{D}_2\left(\mathbf{q}\right)\right) \\ \alpha & \text{if } \mathbf{q}' = u_4\left(\mathcal{B}\left(\mathbf{q}\right)\right) \text{ and } s_{\mathbf{q}} = 1 \\ \gamma & \text{if } \mathbf{q}' = u_5\left(\mathcal{G}\left(\mathbf{q}\right)\right) \text{ and } s_{\mathbf{q}} = 0, \end{cases} \tag{4}$$

where $s_{\mathbf{q}}$ denotes the state of the mmWave link in system state $\mathbf{q}$.

Then, with the discrete-time MDP, the uniformized problem is formulated as:

$$\min_{\pi \in \Pi} \limsup_{N \to \infty} \frac{1}{N} \mathbb{E}^{\pi} \left[\sum_{n=1}^{N} c(\mathbf{q}(n)) \right], \tag{5}$$

where $\mathbf{q}(n)$ denotes the system state at the n-th time slot.

A policy is called a *stationary deterministic* policy if it is time independent and can be expressed as $\pi = \{D, D, \cdots\}$, where D is a deterministic function that maps states to actions. Stationary deterministic policies are the easiest to be implemented and evaluated. However, there may not exist a stationary deterministic policy that is optimal for average delay problem [27]. In Section 3, we show that there exists an optimal policy for average delay problem.

Discounted Delay Problem: A method for studying average cost MDPs is to relate them to discounted cost MDPs. Specifically, average cost optimal policy can be regarded as a limit of a sequence of discounted cost optimal policies. Thus, we begin with discounted delay problems and extend our results to the average delay problem in the end. The discounted delay problem in the equivalent discrete-time MDP is expressed as follows:

$$\min_{\pi \in \Pi} \limsup_{N \to \infty} \mathbb{E}^{\pi} \left[\sum_{n=0}^{N-1} (\beta)^n c(\mathbf{q}(n)) \right], \tag{6}$$

where $\mathbf{q}(n)$ denotes the system state at the n-th time slot, and β is a discount factor such that $0 \le \beta < 1$.

3 Delay Optimal Policy

3.1 Discounted Delay Problem

Note that one-step cost $c(\mathbf{q})$ is unbounded. Therefore, there may not exist a stationary deterministic policy for the discounted delay problem. In Proposition 1, we will show that there do exist optimal stationary deterministic policies, and we provide a method to study the structure of the optimal policies. Before that, we provide some notations and definitions which will be used in the proposition and following content.

Let w be a positive real-valued function on $\mathcal{X}$ defined by $w(\mathbf{q}) = \max(c(\mathbf{q}), 1)$. Define the weighted supremum norm $||\cdot||_w$ for real-valued functions v on $\mathcal{X}$ by

$$||v||_w = \sup_{\mathbf{q} \in \mathcal{X}} \frac{v(\mathbf{q})}{w(\mathbf{q})}.$$

Let V be the space of real-valued functions v on $\mathcal{X}$ that satisfies $||v||_w < \infty$. We define an operator $B : V \to V$ as:

$$Bv(\mathbf{q}) \triangleq \min_{u \in K_{\mathbf{q}}} v\left(u\left(\mathbf{q}\right)\right). \tag{7}$$

Then, we define operator $\mathcal{L} : V \to V$ as:

$$\begin{aligned} &\mathcal{L}v\left(\mathbf{q}\right) \\ &\triangleq c(\mathbf{q}) + \beta \Big\{ \lambda Bv\left(\mathcal{A}_0\left(\mathbf{q}\right)\right) + \mu_{\text{mm}} Bv\left(\mathcal{D}_1\left(\mathbf{q}\right)\right) \cdot \mathbb{1}_{\{s_{\mathbf{q}}=1\}} \\ &\quad + \mu_{\text{sub-6}} Bv\left(\mathcal{D}_2\left(\mathbf{q}\right)\right) + \mu_{\text{p}} Bv\left(\mathcal{T}\left(\mathbf{q}\right)\right) \\ &\quad + \alpha Bv\left(\mathcal{B}\left(\mathbf{q}\right)\right) \cdot \mathbb{1}_{\{s_{\mathbf{q}}=1\}} + \gamma Bv\left(\mathcal{G}\left(\mathbf{q}\right)\right) \cdot \mathbb{1}_{\{s_{\mathbf{q}}=0\}} \\ &\quad + p\left(\mathbf{q}\right) \cdot Bv\left(\mathbf{q}\right) \Big\}, \end{aligned} \tag{8}$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator function, $v\left(\cdot\right) \in V$ and $p\left(\mathbf{q}\right)$ denotes the total probability of impossible events in state $\mathbf{q}$. For example, when $s = 0$, a departure from the mmWave will not happen. The expression of $p\left(\mathbf{q}\right)$ is

$$p(\mathbf{q}) = 1 - \lambda - \mu_{\text{p}} - \mu_{\text{sub-6}} - (\mu_{\text{mm}} + \alpha) \cdot \mathbb{1}_{\{s_{\mathbf{q}}=1\}} - \gamma \cdot \mathbb{1}_{\{s_{\mathbf{q}}=0\}}. \tag{9}$$

Let $J_{\beta}\left(\mathbf{q}\right)$ denote the optimal expected total discounted delay function of initial state $\mathbf{q}$.

Proposition 1. *(a) The optimal expected total discounted delay J_{β} satisfies the following optimality equation:*

$$J_{\beta}\left(\mathbf{q}\right) = \mathcal{L} J_{\beta}\left(\mathbf{q}\right). \tag{10}$$

(b) There exists a stationary deterministic policy for the discounted delay problem and it is determined by the right-hand-side of (10).

(c) For any function $v \in V$, we have $\lim_{n\to\infty} \mathcal{L}^{(n)} v = J_\beta$

Proof. Please see Appendix A. □

Next, in Theorem 1, we will show some rules that the optimal policy must satisfy via value iteration provided in Proposition 1. In particular, we define Θ as a set of all $v \in V$ that satisfy the following properties from (11) to (15), and show that $J_\beta \in \Theta$.

$$v(A_1(\mathbf{q})) \le v(A_h(\mathbf{q})) \quad \text{if } q_0 \ge 1,\ l_1 = 0; \tag{11}$$

$$v(A_2(\mathbf{q})) \le v(A_r(\mathbf{q})) \quad \text{if } q_0 \ge 1,\ l_1 + q_1 \ge 1, \text{ and } l_2 = 0; \tag{12}$$

$$v(\mathcal{T}(\mathbf{q})) \le v(\mathbf{q}) \quad \text{if } l_1 = 1; \tag{13}$$

$$v(A_1(\mathbf{q})) \le v(A_2(\mathbf{q})) \quad \text{if } \mathbf{q} = (q_0, 0, 0, 0, s) \text{ and } q_0 \ge 1; \tag{14}$$

$$v(\mathbf{q}_1) \le v(\mathbf{q}_2) \quad \text{if } \mathbf{q}_1, \mathbf{q}_2 \in \mathcal{X}, \ \mathbf{q}_2 - \mathbf{q}_1 \in \{(1,0,0,0,0), (0,1,0,0,0), (0,0,1,0,0), (0,0,0,1,0)\}. \tag{15}$$

Note that the function set Θ is not empty since any constant function belongs to Θ.

Except that the mmWave channel is extremely intermittent, the average service rate of the mmWave is much higher than the sub-6 GHz (e.g., two orders of magnitude). Besides, the service rate of the mmWave and processing server are in the same order. Hence, it is reasonable to assume that the expected time for a packet to go through empty mmWave line is less than empty sub-6 GHz interface, i.e., $\frac{1}{\mu_\text{p}} + \frac{\gamma+\alpha}{\gamma} \cdot \frac{1}{\mu_\text{mm}} < \frac{1}{\mu_\text{sub-6}}$. With this assumption, we have the following theorem

Theorem 1. *Given that $\frac{1}{\mu_p} + \frac{\gamma+\alpha}{\gamma} \cdot \frac{1}{\mu_{mm}} < \frac{1}{\mu_{sub\text{-}6}}$, we have $J_\beta \in \Theta$.*

Proof. Please see Appendix B. □

Remark: Note that in the following, we say that action A_i has a higher priority than action A_j if action A_i incurs no more costs than action A_j, where $i, j \in \{1, 2, r, b, h\}$.

By Theorem 1, we obtain three rules that partially characterize the optimal policy:

- **Rule 1: Holding is not preferable as long as the processing server is available:** Property (11) implies that A_1 has priority over A_h.
- **Rule 2: Keeping the mmWave line busy:** Properties (11) and (14) imply that a packet should be scheduled on the mmWave line whenever the mmWave line is empty and the head buffer (see Fig. 3) is not empty.
- **Rule 3: Head buffer is the first choice for the sub-6 GHz interface:** That is to say, moving a packet from the head buffer to sub-6 GHz interface incurs no more cost than reneging a packet from the mmWave line does. In particular, property (12) says that A_2 has priority over A_r. In addition, if both A_{r_p} and A_{r_mm} are admissible. then $J_\beta(A_{r_\text{p}}(\mathbf{q})) = J_\beta(\mathcal{T}(A_{r_\text{mm}}(\mathbf{q})))$. By property (13), we have $J_\beta(A_{r_\text{p}}(\mathbf{q})) \le J_\beta(A_{r_\text{mm}}(\mathbf{q}))$, which implies that $A_r = A_{r_\text{p}}$ when A_{r_p} and A_{r_mm} are admissible.

Optimal Policy: Based on these rules, we propose a threshold-type policy $\pi_{\mathbf{m}^\star} \triangleq \{D_{\mathbf{m}^\star}, D_{\mathbf{m}^\star}, \cdots\}$ with state-dependent thresholds $\mathbf{m}^\star \triangleq (m_0^\star, m_1^\star)$, where $m_0^\star$ and $m_1^\star$ denote optimal thresholds when mmWave channel is unavailable and available, respectively, and $D_{\mathbf{m}^\star}$ is defined as follows:

$$D_{\mathbf{m}^\star}(\mathbf{q}) = \begin{cases} A_1 & \text{if } \mathbf{q} = (q_0, 0, q_1, 1, s),\ q_0 \ge 1, \\ & \text{or } \mathbf{q} = (q_0, 0, q_1, 0, s),\ q_0 \ge 1,\ q_0 + q_1 \le m_s^\star, \\ A_2 & \text{if } \mathbf{q} = (q_0, 1, q_1, 0, s),\ q_0 \ge 1,\ q_0 + q_1 + 1 > m_s^\star, \\ & \text{or } \mathbf{q} = (1, 0, q_1, 0, s),\ q_1 \ge m_s^\star, \\ A_r & \text{if } \mathbf{q} = (0, l_1, q_1, 0, s),\ l_1 + q_1 > m_s^\star, \\ A_b & \text{if } \mathbf{q} = (q_0, 0, q_1, 0, s),\ q_0 + q_1 > m_s^\star, q_0 \ge 2, \\ A_h & \text{otherwise.} \end{cases}$$

Note that $\pi_{\mathbf{m}^\star}$ follows all above rules. Next, we show the optimality of $\pi_{\mathbf{m}^\star}$ for the discounted delay problem. To this end, we name the action sets $\{A_1, A_h\}$ and $\{A_2, A_r\}$ as *"not-adding-to-sub-6"* and exclusively *"adding-to-sub-6"*, respectively. We already know the priority between A_1 and A_h and the priority between A_2 and A_r. Thus, it remains to determine the priority between the sets not-adding-to-sub-6 and adding-to-sub-6. To show this, we dub the path consisting of the head buffer, the processing server, and the mmWave queue as *"FastLane"*. We claim that in the discounted delay optimal policy, adding-to-sub-6 obtains priority over not-adding-to-sub-6 when the queue length of FastLane exceeds certain state-dependent threshold $m_s^\star$, i.e., a threshold-type policy as expressed by $\pi_{\mathbf{m}^\star}$. We will show this via value iteration. For simplicity, we re-express the system state $\mathbf{q}$ in the form of (x, q_1, l_2, s) where x denotes the number of packets in the head buffer and processing server. Note that if $x > 0$, then the processing server should be busy by Rule 1. For the sake of exposition in the following proof, we define a term in Definition 1.

Definition 1. *Let $J_\beta^n(x, q_1, l_2, s)$ denote the optimal expected total discounted delay over the next n time slots with initial state (x, q_1, l_2, s). Then, $J_\beta^{n+1}(x, q_1, l_2, s)$ is written as:*

$$\begin{aligned} & J_\beta^{n+1}(x, q_1, l_2, s) \\ = & (x + q_1 + l_2) + \beta \Big(\lambda B J_\beta^n(x+1, q_1, l_2, s) \\ & + \mu_\text{p} B J_\beta^n\left((x-1)^+, x + q_1 - (x-1)^+, l_2, s\right) \\ & + \mu_\text{mm} B J_\beta^n\left(x, (q_1 - 1)^+, l_2, s\right) \cdot \mathbb{1}_{\{s=1\}} \\ & + \mu_\text{sub-6} B J_\beta^n(x, q_1, 0, s) + \gamma B J_\beta^n(x, q_1, l_2, 1-s) \cdot \mathbb{1}_{\{s=0\}} \\ & + \alpha B J_\beta^n(x, q_1, l_2, 1-s) \cdot \mathbb{1}_{\{s=1\}} + p(\mathbf{q}) B J_\beta^n(x, q_1, l_2, s) \Big). \end{aligned} \tag{16}$$

Moreover, $J_\beta^0(x, q_1, l_2, s) = x + q_1 + l_2$.

Next, we define a class of functions with threshold, supermodular and monotonicity properties in Definition 2 and Lemma 1 proves that J_β^n has these properties.

Definition 2. *We define a class of functions $\mathscr{F}$ that satisfy the following properties where $f \in \mathscr{F}$, $l_2, s \in \{0,1\}$, and $x, q_1 \in \mathbb{N}$:*

$$\begin{aligned} f(x{+}1, q_1, 0, s) + f(x{+}1, q_1, 1, s) & \\ \le f(x, q_1, 1, s) + f(x{+}2, q_1, 0, s) & \quad (17) \\ f(x{+}1, q_1, 0, s) + f(x, q_1{+}1, 1, s) & \\ \le f(x, q_1, 1, s) + f(x{+}1, q_1{+}1, 0, s) & \quad (18) \\ f(0, q_1{+}1, 0, s) + f(0, q_1{+}1, 1, s) & \\ \le f(0, q_1, 1, s) + f(0, q_1{+}2, 0, s) & \quad (19) \\ f(x, q_1{+}1, l_2, s) \le f(x{+}1, q_1, l_2, s) & \quad (20) \end{aligned}$$

together with supermodularity:

$$\begin{aligned} f(x, q_1, 1, s) + f(x{+}1, q_1, 0, s) & \\ \le f(x, q_1, 0, s) + f(x{+}1, q_1, 1, s) & \quad (21) \\ f(x, q_1, 1, s) + f(x, q_1{+}1, 0, s) & \\ \le f(x, q_1, 0, s) + f(x, q_1{+}1, 1, s) & \quad (22) \end{aligned}$$

and monotonicity:

$$\begin{aligned} f(x, q_1, l_2, s) \le f(x{+}1, q_1, l_2, s) & \quad (23) \\ f(x, q_1, l_2, s) \le f(x, q_1{+}1, l_2, s) & \quad (24) \\ f(x, q_1, 0, s) \le f(x, q_1, 1, s) & \quad (25) \end{aligned}$$

Eq. (17) to (19) describe the threshold property.

Lemma 1. *J_β^n satisfies all properties in Definition 2, i.e., $J_\beta^n \in \mathscr{F}$ for each $n \in \mathbb{N}$.*

Proof. Please see Appendix C. □

Now we are ready to provide our main result that the optimal policy is of the threshold-type for both "available" and "unavailable" states of the mmWave link.

Theorem 2. *If $s = 0$ $(s = 1)$, then there exists an $m_0^\star$ $(m_1^\star) \in \mathbb{N}$ such that if the number of packets in the system is larger than $m_0^\star$ $(m_1^\star)$, i.e., $c(\mathbf{q}) > m_0^\star$ $(c(\mathbf{q}) > m_1^\star)$, then it is optimal to add a packet to the sub-6 interface.*

Proof. Please see Appendix D. □

Thus far, we have proven that the optimal policy for the discounted delay problem is of a threshold type. However, note that we allow controllers to take actions when fictitious events happen, where the fictitious events refer to the events that do not alter the system state. Actually, the fictitious events make changes to the "dummy" packets. For example, if the current system state is $(10, 1, 0, 0, 1)$ and the event departure from the mmWave interface occurs, the mmWave interface finishes serving a "dummy" packet rather than a "real" packet, in which case the system state does not change. However, it is impossible to track these fictitious events in practice. In fact, our current policy will preserve its optimality when actions are limited to real events. This is because when the fictitious events happen, the optimal action does not change the system state as long as the system starts from a proper state. We will prove this via the following theorem. Moreover, we call a state to be *proper* if no optimal actions can be taken in this state. For example, state $(5, 0, 0, 1, 1)$ is not proper since it is optimal to do A_1 by property (11); however, state $(0, 0, 0, 0, 0)$ is proper since no actions can be taken.

Theorem 3. *If the system begins with a proper state and the optimal policy is followed, then the system state will not be changed when fictitious events happen.*

Proof. Please see Appendix E. □

Optimal Threshold: From Theorem 2, we already know that for both cases that the mmWave link is either unavailable $s = 0$ or available $s = 1$, the optimal policy is of the threshold type. The following theorem will provide the relationship between the threshold values $m_0^\star$ and $m_1^\star$.

Theorem 4. *The optimal threshold values for the sub-6 GHz interface satisfy $m_0^\star \le m_1^\star$, where $m_0^\star$ and $m_1^\star$ denote the optimal threshold values for cases that the mmWave link is unavailable and available, respectively.*

Proof. Please see Appendix F. □

3.2 Average Delay Problem

The following theorem extends our results to the average delay problem.

Theorem 5. *There exists a threshold-type optimal stationary deterministic policy that minimizes the average delay in our system.*

Proof. According to [28], $\lim_{\beta_n \to 1} (1 - \beta_n) J_{\beta_n}^{\pi_{\beta_n}^\star}(\mathbf{q}) = J^{\pi^\star}(\mathbf{q})$, $\forall \mathbf{q} \in \mathcal{X}$, where $J_{\beta_n}^{\pi_{\beta_n}^\star}(\mathbf{q})$ denotes optimal expected total discounted delay under optimal policy $\pi_{\beta_n}^\star$ associated with discount factor β_n and $J^{\pi^\star}(\mathbf{q})$ denotes optimal average delay under optimal policy $\pi^\star$. Since our action set is finite, by [28], there exists an optimal stationary policy for the average delay problem such that $\pi_{\beta_n}^\star \to \pi^\star$, which implies the optimal policy is of a threshold-type. □

In Fig. 4, we provide a flowchart on how decision is made based on a state. Briefly speaking, when the sub-6 GHz interface is idle, we decide whether to use it based on checking $q_0 + q_1 + l_1 > m_s^\star$. As long as the condition is satisfied, the sub-6 GHz interface is used either by dispatching a packet from the head queue or reneging a packet from the mmWave line.

4 Learning to Minimize Delay in Unknown Environment

While deriving scheduling policies, we assumed that the following parameters were given:

- Packet arrival rate λ,
- Service rate of the processing server μ_{p} ,
- Service rate of the sub-6 GHz channel $\mu_{\text{sub-6}}$,
- Service rate of the mm Wave channel μ_{mm},
- Parameters α, γ that describe the dynamics of the mmWave channel.

However, in practice these may not be known a priori. In this section we design "learning" algorithms that simultaneously optimize delays while "learning" these unknown parameters. This allows us to derive delay-minimizing schemes for the case when these parameters are not known to the scheduler. We denote these parameters collectively by the vector θ, i.e., $\theta := (1/\lambda, 1/\mu_{\text{p}}, 1/\mu_{\text{sub-6}}, 1/\mu_{\text{mm}}, 1/\alpha, 1/\gamma)$.

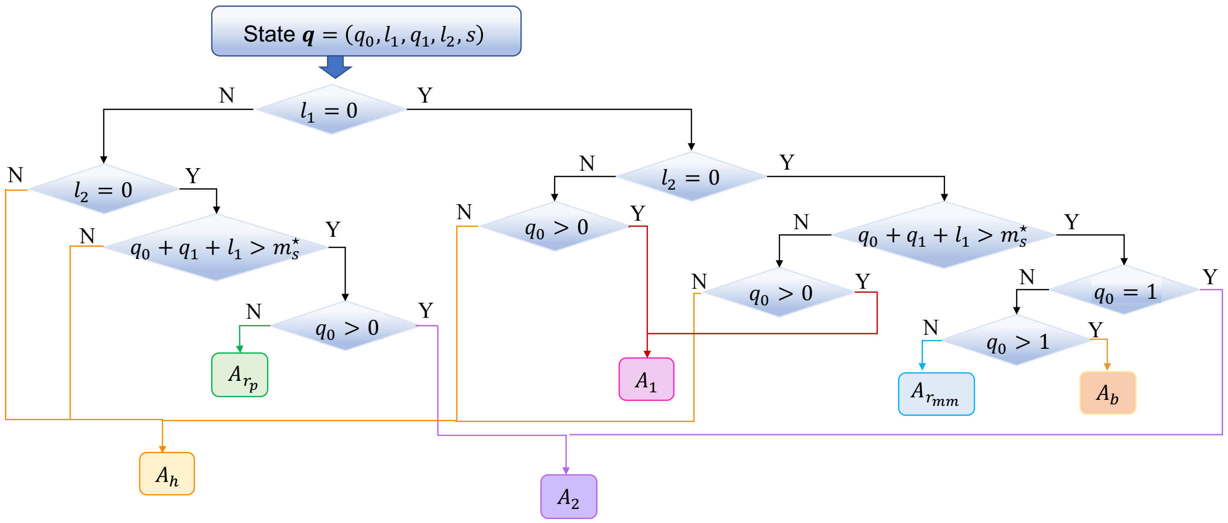


Fig. 4: Decision making. The Y and N in the figure denote Yes and No, respectively. The $m_s^\star$ is state-dependent threshold.

The reason why we use the reciprocals of the arrival rates, service rates and channel transition rates rather than the rates themselves to parameterize the system is that the empirical estimates for the former are unbiased. In what follows, we will work exclusively with the discrete-time system that has been obtained by sampling the original continuous-time system at those time instants when either of the following events occur: (i) service completion, (ii) packet arrival, and (iii) the channel state changes.

A learning rule ϕ is a collection of maps that at each decision epoch n maps the operational history of the system until n to a scheduling decision. We will derive learning policies under the following assumption.

Assumption 1 (Episodic Setup)**.** *The state of the network is "reset" at times $t = kH$ where $k = 1, 2, \ldots$ as follows: both the queues are emptied, and the channel state is reset to 1. This setup resonates with the episodic reinforcement learning (RL) setup [29] in which the system state is "reset" to a designated "start state" at the end of each "episode".*

We note that the above assumption is not very restrictive since we can always reset the system state at any desired time as follows. We simply stop the arrivals by not letting packets enter the queueing system until the queues are "drained". Thereafter, we wait for the channel state to become equal to 1. Thus, we can always perform a system reset after every H time steps, in order to start a new episode.

Thus, the k-th episode begins at time $\tau_k := kH$, and lasts for a duration of H time-steps. We denote $\mathcal{E}_k := \{\tau_k, \tau_k + 1, \ldots, \tau_k + H - 1\}$ as the set of consecutive time-slots that constitute the k-th episode.

Let $\pi^\star$ be an optimal stationary policy for the network that minimizes the average delay when the parameter θ is known. Let $\bar{c}(\pi^\star)$ be the optimal average cost (delay) under the policy $\pi^\star$. In order to measure the efficiency of the proposed learning algorithm, we will quantify its learning regret [30]. The regret $R(\phi, T)$ of a learning algorithm ϕ is given as follows,

$$R(\phi, T) := \sum_{k=1}^{K} \left[\sum_{n \in \mathcal{E}_k} c(\mathbf{q}(n)) - H\bar{c}(\pi^\star) \right], \qquad (26)$$

We are interested in designing learning rules ϕ for which the expected value of regret $\mathbb{E}R(\phi, T)$ is low, where the expectation is taken with respect to the probability measure induced by ϕ.

4.1 Certainty Equivalence Learning Rule

We begin with some notation. Let $N_A(n)$ denote the number of arrivals, and $N_p(n), N_{sub-6}(n), N_{mm}(n)$ the number of service completions on processing server, sub-6 channel and mmWave channel respectively until the n-th decision epoch. Let $N_{i \to j}(n)$ be the number of i to j channel state transitions until n-th decision epoch. Let $S_{i \to j}(\ell)$ denote the holding time for the ℓ-th $i \to j$ transition. Let $I_A(\ell)$ denote the time between the arrival of ℓ-th and $\ell + 1$-th packet, while $S_p(\ell), S_{sub-6}(\ell), S_{mm}(\ell)$ denote the time taken for completion of the ℓ-th service at the processing server, sub-6 channel and mmWave channel respectively. The empirical estimates of these rates are obtained as follows:

$$\left[\widehat{\frac{1}{\lambda}}\right](n) := \frac{\sum_{l=1}^{N_A(n)} I_A(l)}{N_A(n)}, \qquad (27)$$

$$\left[\widehat{\frac{1}{\mu_p}}\right](n) := \frac{\sum_{l=1}^{N_p(n)} S_p(l)}{N_p(n)}, \qquad (28)$$

$$\left[\widehat{\frac{1}{\mu_{sub-6}}}\right](n) := \frac{\sum_{l=1}^{N_{sub-6}(n)} S_{sub-6}(l)}{N_{sub-6}(n)}, \qquad (29)$$

$$\left[\widehat{\frac{1}{\mu_{mm}}}\right](n) := \frac{\sum_{l=1}^{N_{mm}(n)} S_{mm}(l)}{N_{mm}(n)}. \qquad (30)$$

We now obtain estimates of the parameters α, γ. We observe the following: the holding time in state 0 (1) is an exponential random variable with rate α (γ). Thus, the empirical estimates for the parameters $1/\alpha$, $1/\gamma$ are given as follows:

$$\left[\hat{\frac{1}{\alpha}}\right](n) := \frac{\sum_{\ell=1}^{N_{0\to1}(n)} S_{0\to1}(n)}{N_{0\to1}(n)}, \tag{31}$$

$$\left[\hat{\frac{1}{\gamma}}\right](n) := \frac{\sum_{\ell=1}^{N_{1\to0}(n)} S_{1\to0}(n)}{N_{1\to0}(n)}, \tag{32}$$

We let $\hat{\theta}(n)$ be the vector that contains these estimates (27)-(32). Let $\pi^\star(\theta)$ be the scheduling policy that is optimal when the true system has parameters θ. It was shown in Theorem 5 that the optimal policy for the delay-minimization problem (6) is of a threshold-type, with two state-dependent thresholds $m_0^\star, m_1^\star$. Recall that we denote this policy as $\pi_{\mathbf{m}^\star}$. We will make the following assumption while designing the learning algorithm.

Assumption 2. *The scheduler knows an upper-bound on the optimal thresholds $m_0^\star, m_1^\star$ of Theorem 5, i.e., it knows the value of M such that $m_0^\star, m_1^\star \le M$. Thus, it knows that the optimal policy belongs to the following set of policies*

$$\Pi_{opt} := \left\{\pi_{(m_0,m_1)} : m_0, m_1 \in \{0,1,\ldots,M\}\right\}. \tag{33}$$

Note that the above assumption allows the scheduler to obtain some "partial knowledge" about the underlying system.

Remark 1. *The above assumption is justified since it suffices to choose the parameter M in* (33) *to be sufficiently large. For example, it could be taken to be greater than a known upper-bound on the optimal value of average delay* (6). *Any crude upper-bound on the optimal delay would be sufficient for our purpose. To see that why would such a technique work, we note that a policy with both the thresholds m_0, m_1 set equal to M does not transmit unless the queue lengths exceed M, and hence necessarily has average delay greater than M. Thus, when M is chosen so as to satisfy this condition, Π_{opt} would contain an optimal threshold policy.*

Certainty Equivalence Learning Rule: The proposed algorithm proceeds in episodes. Let $\pi^\star_{\Pi_{opt}}(\theta')$ denote the policy from the set Π_{opt} that achieves the smallest average delay when the true system parameter is equal to θ'. At the beginning of each episode k, i.e. at time τ_k, the learning rule derives the empirical estimates $\hat{\theta}(\tau_k)$ as in (27)-(32). It then solves for the policy $\pi^\star_{\Pi_{opt}}(\hat{\theta}(\tau_k))$ that is optimal when the true value of the system parameters is equal to $\hat{\theta}(\tau_k)$. It then implements the policy $\pi^\star_{\Pi_{opt}}(\hat{\theta}(\tau_k))$ during $\mathcal{E}_k$. Thus, the learning algorithm implements a "certainty equivalent" (CE) controller [31], [32], [33] that during each episode makes scheduling decisions that are optimal when the true value of the system parameters are equal to the current estimates. Though the CE based learning algorithm is very simple to implement, it is well-known that such a CE rule needs not always yield the optimal performance [31], [32], [33], [34]. In general, while optimizing the performance of an unknown system, one also needs to take into account the estimation errors while making sequential decisions. For example, the Upper Confidence Bound (UCB) [34], [35], [36] rule maintains a high-probability confidence ball around the empirical estimates, and implements a policy that is optimal for the "optimistic estimate" from within this ball. Reward Biased Maximum Likelihood Estimate (RBMLE) [32] derives an optimistic estimate of the parameter by maximizing an objective function that is sum of log-likelihoods and a "bias" term that gives more weightage to those system parameters that yield better performance. Even though these learning algorithms are known to yield state-of-the-art performance in RL tasks, they entail higher computational complexity than the CE rule, and are also harder to analyze. However, as our analysis shows, for the task of minimizing the average delays in the wireless networks, the time-average learning regret of the CE rule is asymptotically 0. Our proof relies on the key result Theorem 5 which shows that while searching for an optimal policy, one can safely restrict to the class of threshold policies.

Algorithm 1: Certainty Equivalence Learning Rule

```
Initialize the empirical estimate $\hat{\theta}(1)$.
for $n = 1, 2, \ldots$ do
    if $n = \tau_k$ then
        Calculate $\hat{\theta}(n)$ as in (27)-(32).
        Calculate the policy from the set $\Pi_{opt}$ (33) that has
        the smallest cost
    end if
    Implement this policy within the current episode.
end for
```

4.2 Preliminary Results

We will analyze regret of Algorithm 1 under the following assumption.

Assumption 3. *Under each threshold policy from the set Π_{opt}* (33), *the average cost* (6) *is finite, i.e., $\bar{c}(\pi) < \infty$ for all $\pi \in \Pi_{opt}$.*

This amounts to the assumption that under each policy from Π_{opt}, the system is "stable", i.e., has a finite average cost. In order to verify this assumption, we could utilize Lyapunov functions [37] in order to verify whether for each policy the network remains stable as the system parameter values range over the set of possible values that could be assumed by them. We now derive some results that are useful while analyzing the learning regret of the CE rule.

Lemma 2. *Consider a sequence of parameters $\theta_n, n \in \mathbb{N}$ that satisfies $\theta_n \to \theta$. Let Assumption 2 hold true. We then have $\pi^\star_{\Pi_{opt}}(\theta_n) \to \pi^\star_{\Pi_{opt}}(\theta)$[1]. Consequently, there is an $\epsilon_p > 0$ such that whenever[2] $\|\theta' - \theta\| < \epsilon_p$, then $\pi^\star_{\Pi_{opt}}(\theta') = \pi^\star_{\Pi_{opt}}(\theta)$.*

Proof. Please see Appendix G. □

Lemma 3. *There exists a constant $\eta > 0$ and natural number n_0 such that for all $n \ge n_0, \pi \in \Pi_{opt}$,*

$$\mathbb{E}_\pi(N_A(n)), \mathbb{E}_\pi(N_p(n)), \mathbb{E}_\pi(N_{sub-6}(n)), \mathbb{E}_\pi(N_{mm}(n)) \ge \eta n, \tag{34}$$

and also, for $\pi \in \Pi_{opt}, i, j \in \{0, 1\}$,

$$\mathbb{E}_\pi N_{i\to j}(n) \ge \eta n, \forall n \ge n_0.$$

[1] Recall that $\pi^\star_{\Pi_{opt}}(\theta')$ denotes the optimal policy from within the set Π_{opt}, when the true parameter is θ'.

[2] For a vector x, we let $\|x\|$ denote its Euclidean norm.

Proof. Please see Appendix H. □

Lemma 4. *(Sufficient sampling) Define,*

$$N_\theta(n) := \min\left\{N_A(n), N_p(n), N_{sub-6}(n),\right.$$
$$\left. N_{mm}(n), N_{0\to1}(n), N_{1\to0}(n)\right\}. \quad (35)$$

Define the following event,

$$\mathcal{G}_1 := \left\{\omega : N_\theta(n) \geq \eta_1 n, \quad \forall n \in \{\tau_k\}_{k>k_0}\right\}, \quad (36)$$

where $\eta_1 \in (0,\eta)$, η is as in Lemma 3, and k_0 is the smallest episode that satisfies

$$\tau_{k_0} \geq \frac{2}{\eta L}\log T, \quad (37)$$

where T is the operating time horizon. We then have that

$$\mathbb{P}\left(\mathcal{G}_1^c\right) \leq \sum_{k>k_0} \exp\left(-\delta\tau_k\right) \leq \frac{\log T}{T^2}.$$

Proof. Please see Appendix I. □

Lemma 5. *(Concentration Result) Fix a $\xi > 3$. Define the confidence intervals for empirical estimates $\hat{\theta}(n)$ (27)-(32) as follows:*

$$\mathcal{C}_\lambda(n) := \left\{z > 0 : \left|z - \left[\hat{\frac{1}{\lambda}}\right](n)\right| \leq \sqrt{\frac{\xi\log n}{N_A(n)}}\right\},$$

$$\mathcal{C}_{\mu_p}(n) := \left\{z > 0 : \left|z - \left[\hat{\frac{1}{\mu_p}}\right](n)\right| \leq \sqrt{\frac{\xi\log n}{N_p(n)}}\right\}$$

$$\mathcal{C}_{\mu_{sub\text{-}6}}(n) := \left\{z > 0 : \left|z - \left[\hat{\frac{1}{\mu_{sub-6}}}\right](n)\right| \leq \sqrt{\frac{\xi\log n}{N_{sub\text{-}6}(n)}}\right\},$$

$$\mathcal{C}_{\mu_{mm}}(n) := \left\{z > 0 : \left|z - \left[\hat{\frac{1}{\mu_{mm}}}\right](n)\right| \leq \sqrt{\frac{\xi\log n}{N_{mm}(n)}}\right\}$$

$$\mathcal{C}_\alpha(n) := \left\{z > 0 : \left|z - \left[\hat{\frac{1}{\alpha}}\right](n)\right| \leq \sqrt{\frac{\xi\log n}{N_\alpha(n)}}\right\},$$

$$\mathcal{C}_\gamma(n) := \left\{z > 0 : \left|z - \left[\hat{\frac{1}{\gamma}}\right](n)\right| \leq \sqrt{\frac{\xi\log n}{N_\gamma(n)}}\right\}.$$

Let $\mathcal{C}_\theta(n)$ be the set of those possible values of the system parameters $\theta' := \left(1/\lambda', 1/\mu'_p, 1/\mu'_{sub\text{-}6}, 1/\mu'_{mm}, 1/\alpha', 1/\gamma'\right)$ for which the individual elements belong to the corresponding confidence intervals. Define

$$\mathcal{G}_2(n) := \left\{\omega : \theta \in \mathcal{C}_\theta(n)\right\}, \text{ and } \mathcal{G}_2 := \cap_{n>n_0}\mathcal{G}_2(n), \quad (38)$$

where $n_0 \in \mathbb{N}$. We have

$$\mathbb{P}\left(\mathcal{G}_2\right) \geq 1 - \sum_{n>n_0}\frac{6}{n^\xi}. \quad (39)$$

Proof. Please see Appendix J. □

Corollary 1. *Consider the operation of the CE learning (Algorithm 1), and let the events $\mathcal{G}_1, \mathcal{G}_2$ be as in* (36)*, and* (38) *respectively. We then have that*

$$\mathbb{P}\left(\mathcal{G}_1 \cap \mathcal{G}_2\right) \geq 1 - \left(\sum_{n>n_0}\frac{6}{n^\xi} + \frac{\log T}{T^2}\right).$$

4.3 Regret Analysis of CE Learning Rule

We will analyze the regret on the sets $\mathcal{G}_1 \cap \mathcal{G}_2$ and $(\mathcal{G}_1 \cap \mathcal{G}_2)^c$ separately.

Regret on $\mathcal{G}_1 \cap \mathcal{G}_2$: In this case, the confidence interval $\mathcal{C}_\theta(n)$ holds true, and moreover its radius is smaller than $\sqrt{\frac{\xi\log n}{\eta_1 n}}$. Let n_1 be the smallest integer that satisfies $\sqrt{\frac{\xi\log n}{\eta_1 n}} \leq \epsilon_p$ (ϵ_p is as in Lemma 2). Since the radius of the confidence ball is less than ϵ_p, in this scenario the policy produced by the CE rule is optimal after time n_1. Thus, the regret after n_1 is 0. n_1 is clearly upper-bounded by $\frac{\xi\log T}{\eta_1\epsilon_p^2}$, where T is the operating time-horizon. Thus, the regret is upperbounded by $\frac{\xi\log T}{\eta_1\epsilon_p^2}$.

Regret on $(\mathcal{G}_1 \cap \mathcal{G}_2)^c$: It follows from Lemma 4 and Lemma 5 that the probability of the event that either $\mathcal{C}(\tau_k)$ fails, or the number of samples $N_\theta(\tau_k)$ is less than $\eta_1\tau_k$, is upper-bounded by $\frac{1}{\tau_k^\xi} + \exp(-\eta_1\tau_k)$. Under Assumption 3, the regret within such a "bad" episode can be trivially upper-bounded by a constant C_1 (the maximum cost incurred by a policy from the set Π_{opt}) times its duration H. Hence, the reget during such an episode can be upper-bounded by

$$C_1\left[\frac{1}{\tau_k^\xi} + \exp(-\eta_1\tau_k)\right]\tau_k = \frac{C_1}{\tau_k^{\xi-1}} + C_1\tau_k\exp(-\eta_1\tau_k).$$

Summing up the above expression over episodes k, we conclude that this regret is bounded by

$$\sum_k \frac{1}{\tau_k^{\xi-1}} + \sum_k \tau_k\exp(-\eta_1\tau_k)$$
$$= \sum_k \frac{1}{H^{\xi-1}k^{\xi-1}} + H\sum_k k\exp(-H\eta_1 k).$$

Since $\xi > 3$, this sumation is bounded.

We summarize our result below.

Theorem 6. *Consider the CE based learning rule (Algorithm 1) that is used for making scheduling decisions for the network described in Section 2. The expected value of its regret* (26) *$R(T)$ until time T can be bounded as follows:*

$$\mathbb{E}R(T) \leq \frac{\xi\log T}{\eta_1\epsilon_p^2} + C_1\sum_{k=1}^{\infty}\frac{1}{H^{\xi-1}k^{\xi-1}} + HC_1\sum_{k=1}^{\infty}\exp(-H\eta_1 k),$$

where ϵ_p is a problem-dependent constant as in Lemma 2, the constant ξ can be taken to be 3, and the constant C_1 is the maximum cost incurred by a policy from the set Π_{opt}.

5 Simulation Results

In this section, we numerically investigate the performance of our proposed policy. To this end, we first investigate the relationship between the arrival rate and the optimal thresholds. Next, we demonstrate the benefits of utilizing the sub-6 GHz paired with our threshold-type policy especially in heavy traffic scenarios. Finally, we compare the proposed policy with other policies. In particular, our results show that replacing the state-dependent optimal thresholds with one fixed threshold incurs slight cost increase over the optimal policy.

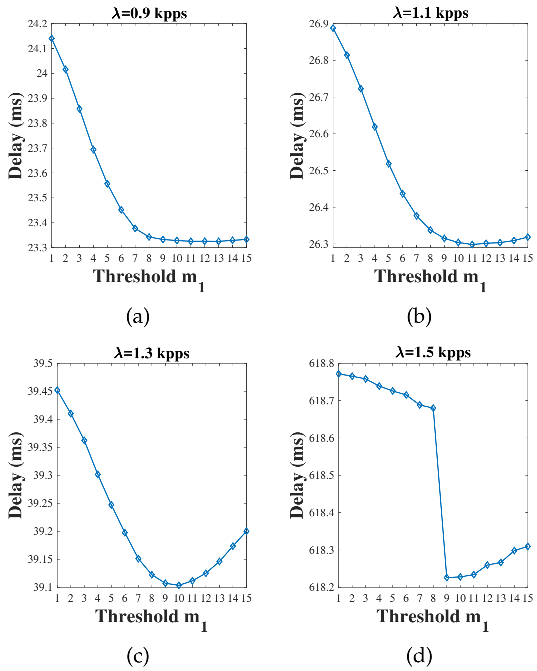
According to 802.11ad, with different Modulation and Coding Scheme (MCS), the data rate of the mmWave ranges from 27.5 Mbps to 6756.75 Mbps while according to 802.11n, the data rate of the 20MHz channel ranges from 6.5 Mbps to 288.9 Mbps. Based on the statistics, we assume that the length of the packet is 64kB [1] and set $\mu_{\text{mm}} = \mu_{\text{p}} = 2000$ pps (packets per second) and $\mu_{\text{sub-6}} = 40$ pps. In addition, [10] conducted a field measurement to study the impact of blockage caused by typical pedestrian traffic on the mmWave link and formulated the mmWave link with a two-state Markov channel according to their measurement. We use the transition rates of the continuous Markov mmWave channel obtained in this paper, i.e. we set $\alpha = 0.18$ and $\gamma = 3.85$.

5.1 Optimal Threshold

As in [9], we use simulations to obtain optimal thresholds. In particular, for each fixed arrival rate, we simulate the delay performance as the thresholds m_1 and m_0 vary. From our results, for all arrival rates, the optimal threshold when mmWave is unavailable ($m_0^\star$) is 1. This is reasonable since the expected transition time from unavailable mmWave to available mmWave in the simulation is $1/0.18$ s, which is is large enough to serve at least one packet (average service time is $1/40$ s). In addition, as shown in Fig. 5, for fixed arrival rate and the threshold for unavailable state m_0, the average delay first decreases and then increases as the threshold for available state m_1 increases, and the smallest average delay corresponds to optimal threshold $m_1^\star$. Then, it is easy to observe that the optimal threshold for available state $m_1^\star$ decreases with arrival rate. The unit kpps in the figure is short for kilo packets per second. The abrupt change in Fig. 5d, is due to the fact that the arrival rate approaches the service capability of the system. As the system is in heavy load, utilizing the sub-6 GHz interface at the right time is extremely important. Before the transition point, lots of packets which could be served by the mmWave are served by the sub-6 GHz since the sub-6 GHz is overused. After using the optimal threshold, the sub-6 GHz is utilized at the right time. Because of the large load, using the sub-6 GHz appropriately will allow the mmWave to serve significantly increased number of packets over using a smaller threshold. This is why the change is abrupt.

5.2 Benefits from the Sub-6 GHz with Threshold-Type Policy

In this section, we demonstrate benefits of the sub-6 GHz interface to combat the effects of blockage and intermittent connectivity, especially under heavy traffic scenarios. Considering the extremely different service rates of the mmWave and the sub-6 GHz interfaces, we raise the following question: *How much does the sub-6 GHz interface improve the average delay of the system?* To answer this question, we compare delay performance in systems with and without the sub-6 GHz. For the system with the sub-6 GHz (our integrated system), the proposed threshold-type policy is utilized. For the system without the sub-6 GHz server, no scheduling policy applies since only mmWave interface exists in the system. To provide a more clear exhibition of our simulation results, we define relative delay improvement as follows:



(a)

(b)

(c)

(d)

Fig. 5: Delay vs threshold m_1 (when $m_0 = 1$). Given arrival rate, the smallest delay in each figure corresponds to the optimal threshold $m_1^\star$. As arrival rate increases (from (a) to (d)), the optimal threshold $m_1^\star$ decreases.

$$\hat{W} = \frac{\bar{W}(\text{no sub-6}) - \bar{W}(\text{with sub-6})}{\bar{W}(\text{no sub-6})},$$

where $\bar{W}(\text{with sub-6})$ and $\bar{W}(\text{no sub-6})$ denote the average delay in the integrated system and that in the system without the sub-6 GHz server, respectively. As shown in Fig. 6, we study how the relative delay improvement changes as arrival rate varies. We note that the relative delay improvement increases with arrival rate and reach more than 70% in heavy traffic scenarios.

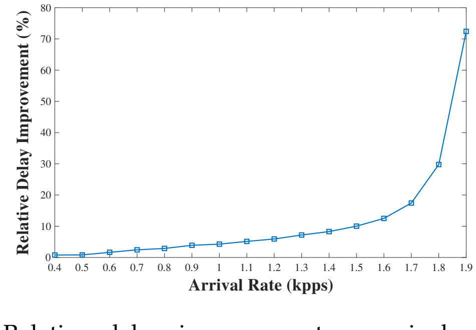


Fig. 6: Relative delay improvement vs arrival rate. Delay improvement from using sub-6 GHz and our threshold-type policy increases with arrival rate.

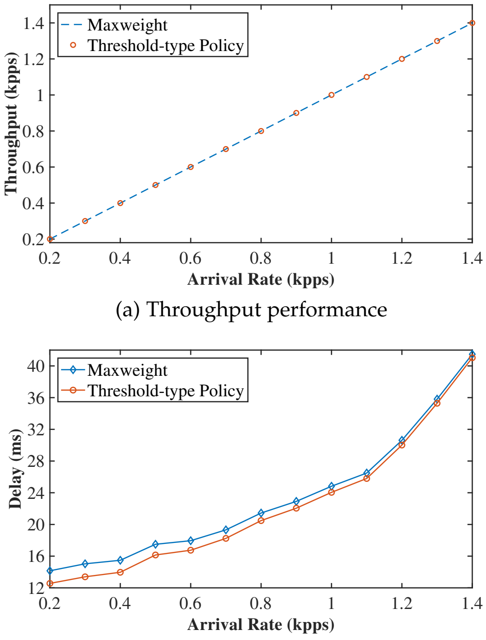
5.3 Comparison with other Policies

5.3.1 Throughput and Delay Comparison with MaxWeight

In this section, we investigate the performance of threshold-type policy compared with the MaxWeight policy. From

[1] Considering that the maximum size of the data frame in MAC layer of 802.11n is 65kB and we focus on scheduling policy in this layer, we assume that our packet size is 64kB.

the results shown in Fig. 7, we note that the threshold-type policy has better delay performance while achieving almost the same throughput performance compared with MaxWeight. We also observe that when arrival rate exceeds 1.1 kpps, the gap between the two policies is very small. This is because that the system is in a very heavy load scenario. That is, the chance that the number of packets in the head queue and mmWave line is less than the threshold, is small. Thus, by our threshold-type policy, the sub-6 GHz will be utilized almost every time when it becomes idle. This is similar to what Maxweight does in the scenario. When Maxweight is utilized, the sub-6 GHz is utilized whenever it becomes idle and the head queue has more than one packet, which is true when arrival rate is high. Thus, the gap is very small in this case.



(a) Throughput performance

(b) Delay performance

Fig. 7: Comparison with MaxWeight. (a) Our threshold-type policy achieves almost the same throughput performance compared with MaxWeight. (b) Our threshold-type policy outperforms MaxWeight in delay performance.

5.3.2 *Policy with Fixed Threshold*

Recall that the optimal thresholds are state-dependent. However, the link speed of the mmWave interface (multi-Gbps) is comparable to the speed at which a typical processor in a smart device operates. Thus, from a practical perspective it will be challenging to track and respond to channel variations in real time. Within this context, it is desirable to devise a more practical policy that can achieve a similar delay performance as the optimal policy. To this end, we substitute the two state-dependent thresholds with one fixed threshold that does not depend on the mmWave channel state. Further, the only difference in flowchart of decision making (Fig. 4) is that $m_s^\star$ is fixed rather than state-dependent. In order to find the fixed threshold, we deploy our method in section 5.1. But at this time, no action will be taken when the mmWave channel state changes. Moreover, we use the steady state probability of the available state as the “available" probability in the policy proposed in our previous paper [9] and obtain a threshold-type policy with a fixed threshold. For the sake of exposition, we refer to this policy as $\bar{\mu}$ policy. Table 2 illustrates the delay increase of $\bar{\mu}$ over the optimal policy vs arrival rate. The results in the table demonstrate that the benefits of varying thresholds as a function of the mmWave channel state are slight. This is because when the system is steady, the queue length in the system does not approach the threshold very frequently.

TABLE 2: Cost increase over optimal policy

Arrival Rate (kpps)	0.7	0.9	1.1	1.3	1.5	1.7
Cost Increase over Optimal Policy (%)	0.0272	0.002	0.0101	0.0133	0.0036	0

5.4 Performance of learning algorithm

Fig. 8 illustrates the delay performance of our CE-Threshold based learning (Algorithm 1) over time. In particular, we set $H = 20$ and $\lambda = 0.5$kpps. We can observe that the delay obtained from the CE-Threshold based learning algorithm converges to the delay obtained from the optimal policy which has a priori knowledge of system parameters. The delay increase of the learning algorithm over the optimal policy is less than 10% after 10000 iterations.

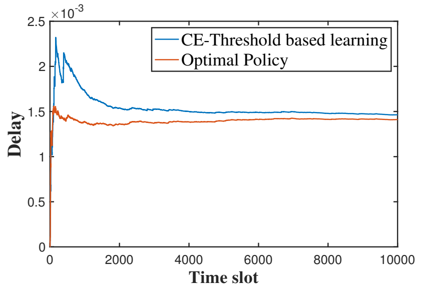


Fig. 8: Comparison between CE-Threshold based learning (Algorithm 1) for $H = 20$ and optimal policy

5.5 Performance of the threshold-type policy with real-world traffic

To show whether our threshold-type policy is still useful for a real-world traffic, we apply the threshold-type policy and Q-learning method to the scenario with inter-arrival distribution replaced by Pareto distribution. Fig. 9 compares the delay performance of the threshold-type policy and model-free Q-learning. In the simulation, we keep scale parameter of Pareto distribution as 2000 and simulate delay performance with different arrival rates by varying shape parameter. We can observe that the threshold-type policy does better than Q-learning in delay.

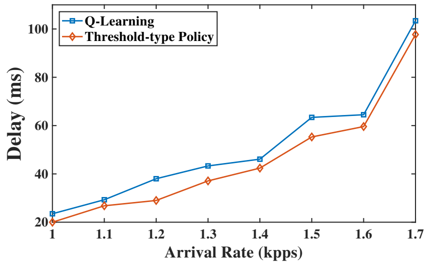
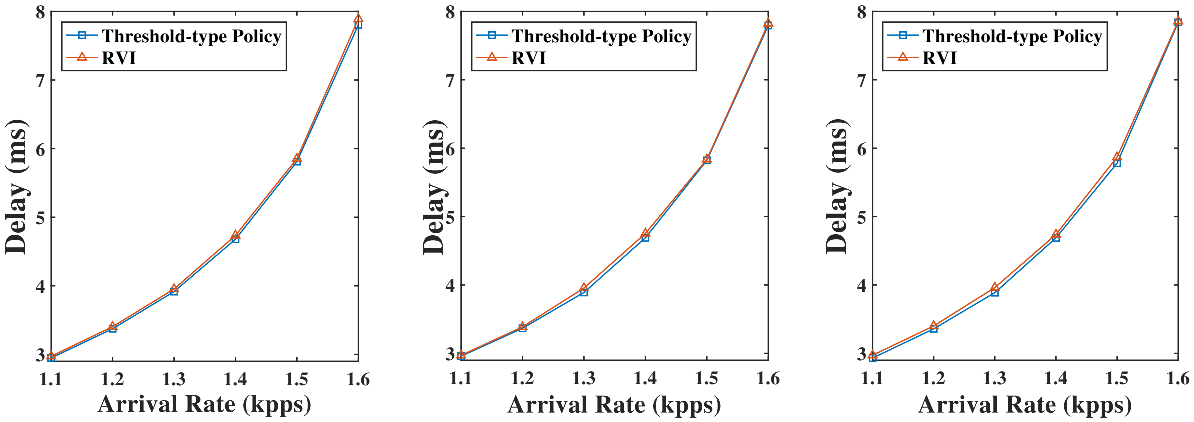


Fig. 9: Delay with Pareto distributed inter-arrival time



(a) Scenario 1

(b) Scenario 2

(c) Scenario 3

Fig. 10: Delay performance with four-state mmWave channel. Scenario 1 has parameter $(p_{\text{decay}}, p_{\text{shad}}, p_{\text{unshad}}, p_{\text{rise}}) = (0.21, 10.49, 5.48, 9.79)$; scenario 2 has parameter $(p_{\text{decay}}, p_{\text{shad}}, p_{\text{unshad}}, p_{\text{rise}}) = (0.18, 11.3, 6.88, 10.36)$; scenario 3 has parameter $(p_{\text{decay}}, p_{\text{shad}}, p_{\text{unshad}}, p_{\text{rise}}) = (0.21, 7.88, 7.67, 7.7)$.

5.6 Performance of the threshold-type policy with multi-state mmWave channel

In the paper, we theoretically prove that the proposed threshold-type policy is optimal for the setting with a two-state mmWave channel. Using [10], the mmWave can be also modeled by a four-state Markov chain. In addition to shadowed and unshadowed periods, the four-state model considers a decaying signal level region from unshadowed to shadowed, and a rising signal level region from shadowed to unshadowed. In this section, we simulate the performance of the proposed threshold-type policy in the setting with a four-state mmWave channel and compare it with relative value iteration (RVI) to see whether our policy works for more complex scenario. Since the state in the problem is unbounded, we make truncation to the state before applying RVI. In the simulation, we use bound $N = 1000$ for truncation and test delay performance in three scenarios using parameters in [10], i.e., $(p_{\text{decay}}, p_{\text{shad}}, p_{\text{unshad}}, p_{\text{rise}}) \in \{(0.21, 10.49, 5.48, 9.79), (0.18, 11.3, 6.88, 10.36), (0.21, 7.88, 7.67, 7.7)\}$, where p_{decay}, p_{shad}, p_{unshad} and p_{rise} denote transition rate from unshadowed to decay state, from decay to shadowed state, from rising to unshadowed state and from shadowed to rising state, respectively. Further, we set service rate in decay and rising state the half of the service rate in unshadowed state.

As you can see in Fig. 10, the delay performance gap between the threshold-type policy and RVI is quite small. It is known that RVI is a classical method to find optimal solution for MDP problem. Although the truncation may introduce some offset from optimality, RVI can be treated as a standard to test performance of other policies. Hence, the threshold-type policy has near-optimality performance in four-state scenario while it has reduced complexity compared to RVI.

6 Conclusion

In this paper, we proposed utilizing the sub-6 GHz interface as a fallback data transfer mechanism to mitigate blockage in the mmWave bands. In this case, packets can be transmitted through the mmWave or sub-6 GHz interface or both. We investigated the optimal scheduling policy such that the expected total discounted delay and the average delay are minimized when system parameters (service rates, arrival rates and transition rates of the channel) are given. Using value iteration, we proved that the optimal policy is of a threshold-type with two state-dependent thresholds. Based on this, we propose a certainty equivalence-threshold based learning algorithm for the case that system parameters are unknown. Also, an upper bound of its regret is provided. Numerical results verified that utilization of sub-6 GHz paired with our threshold-type policy can highly improve delay performance, especially under heavy traffic. Moreover, our results demonstrated that the threshold-type policy outperforms the MaxWeight policy in terms of average delay. We also showed that the delay increase incurred by replacing the state-dependent thresholds in the optimal policy with a single fixed threshold is less than 0.023%, which implies the feasibility of using a single fixed threshold as an alternative in practice.

Acknowledgments

This work was funded in part through NSF grants: CNS-2112471, CNS-1955535, CNS-2106932, CNS-2106933, CNS-1901057, CNS- 2007231, CNS-1618520, and CNS-1409336, an Office of Naval Research under Grant: N00014-17-1-241, and Army Research Office grant: W911NF-21-1-0244.

References

[1] Global Mobile Data Traffic Forecast. Cisco visual networking index: Global mobile data traffic forecast update, 2017–2022. *Update*, 2017:2022, 2019.

[2] Sylvain Collonge, Gheorghe Zaharia, and Ghais El Zein. Influence of the human activity on wide-band characteristics of the 60 ghz indoor radio channel. 2004.

[3] Su Khiong Yong and Chia-Chin Chong. An overview of multigigabit wireless through millimeter wave technology: Potentials and technical challenges. *EURASIP journal on wireless communications and networking*, 2007:1–10, 2006.

[4] Sumit Singh, Federico Ziliotto, Upamanyu Madhow, E Belding, and Mark Rodwell. Blockage and directivity in 60 ghz wireless personal area networks: From cross-layer model to multihop mac design. *IEEE Journal on Selected Areas in Communications*, 27(8):1400–1413, 2009.

[5] Sanjib Sur, Xinyu Zhang, Parmesh Ramanathan, and Ranveer Chandra. Beamspy: enabling robust 60 ghz links under blockage. In *13th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 16)*, pages 193–206, 2016.

[6] Sanjib Sur, Vignesh Venkateswaran, Xinyu Zhang, and Parmesh Ramanathan. 60 ghz indoor networking through flexible beams: A link-level profiling. In *ACM SIGMETRICS Performance Evaluation Review*, volume 43, pages 71–84. ACM, 2015.

[7] Christopher Slezak, Vasilii Semkin, Sergey Andreev, Yevgeni Koucheryavy, and Sundeep Rangan. Empirical effects of dynamic human-body blockage in 60 ghz communications. *arXiv preprint arXiv:1811.06139*, 2018.

[8] Morteza Hashemi, C Emre Koksal, and Ness B Shroff. Out-of-band millimeter wave beamforming and communications to achieve low latency and high energy efficiency in 5G systems. *IEEE Transactions on Communications*, 66(2):875–888, 2018.

[9] Guidan Yao, Morteza Hashemi, and Ness B Shroff. Integrating sub-6 ghz and millimeter wave to combat blockage: delay-optimal scheduling. In *2019 International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOPT)*, pages 1–8. IEEE, 2019.

[10] George R MacCartney, Theodore S Rappaport, and Sundeep Rangan. Rapid fading due to human blockage in pedestrian crowds at 5g millimeter-wave frequencies. In *GLOBECOM 2017-2017 IEEE Global Communications Conference*, pages 1–7. IEEE, 2017.

[11] Marco Mezzavilla, Sanjay Goyal, Shivendra Panwar, Sundeep Rangan, and Michele Zorzi. An mdp model for optimal handover decisions in mmwave cellular networks. In *2016 European conference on networks and communications (EuCNC)*, pages 100–105. IEEE, 2016.

[12] Michael Wang, Aveek Dutta, Swapna Buccapatnam, and Mung Chiang. Smart exploration in hetnets: Minimizing total regret with mmwave. In *Proc. IEEE Int. Conf. Sens., Commun. Netw.*, pages 1–10, 2016.

[13] Sumit Singh, Federico Ziliotto, Upamanyu Madhow, Elizabeth M Belding, Mark JW Rodwell, et al. Millimeter wave wpan: Cross-layer modeling and multi-hop architecture. In *INFOCOM*, pages 2336–2340, 2007.

[14] Mei Sun, Yue Ping Zhang, YX Guo, KM Chua, and LL Wai. Integration of grid array antenna in chip package for highly integrated 60-ghz radios. *IEEE Antennas and Wireless Propagation Letters*, 8:1364–1366, 2009.

[15] Morteza Hashemi, C Emre Koksal, and Ness B Shroff. Energy-efficient power and bandwidth allocation in an integrated sub-6 GHz–millimeter wave system. *arXiv preprint arXiv:1710.00980*, 2017.

[16] Omid Semiari, Walid Saad, and Mehdi Bennis. Joint millimeter wave and microwave resources allocation in cellular networks with dual-mode base stations. *IEEE Transactions on Wireless Communications*, 16(7):4802–4816, 2017.

[17] Omid Semiari, Walid Saad, Mehdi Bennis, and Merouane Debbah. Integrated millimeter wave and sub-6 ghz wireless networks: A roadmap for joint mobile broadband and ultra-reliable low-latency communications. *IEEE Wireless Communications*, 26(2):109–115, 2019.

[18] Ronald L. Larsen and Ashok K. Agrawala. Control of a heterogeneous two-server exponential queueing system. *IEEE Transactions on Software Engineering*, (4):522–526, 1983.

[19] Woei Lin and P Kumar. Optimal control of a queueing system with two heterogeneous servers. *IEEE Transactions on Automatic control*, 29(8):696–703, 1984.

[20] P WALRAND. A note on'optimal control of a queueing system with two heterogeneous serves. *Systems and Control Letters*, 4:131–134, 1984.

[21] Ger Koole. A simple proof of the optimality of a threshold policy in a two-server queueing system. *Systems & Control Letters*, 26(5):301–303, 1995.

[22] VV Rykov. Monotone control of queueing systems with heterogeneous servers. *Queueing systems*, 37(4):391–403, 2001.

[23] Ioannis Viniotis and Anthony Ephremides. Extension of the optimality of the threshold policy in heterogeneous multiserver queueing systems. *IEEE Transactions on Automatic Control*, 33(1):104–109, 1988.

[24] Erhun Özkan and Jeffrey P Kharoufeh. Optimal control of a two-server queueing system with failures. *Probability in the Engineering and Informational Sciences*, 28(4):489–527, 2014.

[25] Jessica Dolcourt. We tested 5g speeds across the globe. *CNET*, Retrieved January 2, 2020.

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

[27] Dimitri P Bertsekas, Dimitri P Bertsekas, Dimitri P Bertsekas, and Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA, 1995.

[28] Steven A Lippman. Semi-markov decision processes with unbounded rewards. *Management Science*, 19(7):717–731, 1973.

[29] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[30] Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[31] Vivek Borkar and Pravin Varaiya. Adaptive control of markov chains. In *Stochastic Control Theory and Stochastic Differential Systems*, pages 294–296. Springer, 1979.

[32] Akshay Mete, Rahul Singh, and PR Kumar. Reward biased maximum likelihood estimation for reinforcement learning. *Learning for Dynamics and COntrol (L4DC*, 2021.

[33] Henk Van de Water and J Willems. The certainty equivalence property in stochastic control theory. *IEEE Transactions on Automatic Control*, 26(5):1080–1087, 1981.

[34] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[35] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[36] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In *Advances in neural information processing systems*, pages 89–96, 2009.

[37] Harold J Kushner. Stochastic stability and control. Technical report, Brown Univ Providence RI, 1967.

[38] Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

[39] Alan F. Karr. Weak convergence of a sequence of markov chains. *Zeitschrift f ″u r Probability Theory and Related Areas*, 33(1):41–48, 1975.

[40] Maxim Raginsky and Igal Sason. Concentration of measure inequalities in information theory, communications and coding. *arXiv preprint arXiv:1212.4663*, 2012.

[41] Terence Tao, Van Vu, et al. Random matrices: universality of local spectral statistics of non-hermitian matrices. *Annals of probability*, 43(2):782–874, 2015.

Guidan Yao received the B.E. and M.E. degrees in electrical engineering from Tianjin University, Tianjin, China, in 2013 and 2016 respectively, She is currently working toward the Ph.D. degree in the Department of Electrical and Computer Engineering from The Ohio State University, OH, USA. Her research interests include wireless communication, resource allocation, information freshness, Markov decision process, and scheduling algorithms. She received China national scholarship for being one of the top 0.2% undergraduate students in China, for the year 2011 and 2012.

Morteza Heshemi is an Assistant Professor with the Department of Electrical Engineering and Computer Science at the University of Kansas, Lawrence, Kansas. He received his MSc and PhD degrees in Electrical Engineering from Boston University in 2013 and 2015, respectively. Before joining KU in 2019, he was a postdoctoral researcher and senior lecturer at the Ohio State University. His research interests span the areas of wireless communications, mmWave systems, real-time data networking, and networked cyber-physical systems.

Rahul Singh received the B. Tech. degree in Electrical Engineering from Indian Institute of Technology, Kanpur, India in 2009, M.S. degree in Electrical Engineering from the University of Notre Dame, South Bend, USA in 2011, and the Ph.D. degree from Texas A & M University, College Station, in 2015. Currently, he is an Assistant Professor at the Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore, India. He was a Postdoctoral Scholar at the Laboratory for Information and Decision Systems (LIDS), Massachusetts Institute of Technology, and the Ohio State University. His research interests include Networking, Stochastic Control and Machine Learning. His article was runner-up for the Best Paper Award of ACM MobiHoc 2020.

Ness B. Shroff (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Columbia University in 1994. Thereafter, he joined Purdue University as an Assistant Professor with the School of Electrical and Computer Engineering. At Purdue, he became a Full Professor of ECE and the Director of the University-Wide Center on wireless systems and applications in 2004. In 2007, he joined The Ohio State University, where he currently holds the Ohio Eminent Scholar Endowed Chair in networking and communications with the Department of Electronics and Communication Engineering and the Department of Computer Science and Engineering. He also holds or has held positions as a Visiting (chaired) Professor with Tsinghua University, Beijing, China, Shanghai Jiao Tong University, Shanghai, China, and the IIT Bombay, Mumbai, India. He has received numerous best paper awards for his research and is listed in the Thomson Reuter's on The World'Ăs Most Influential Scientific Minds, and is 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. He serves as the Steering Committee Chair for ACM Mobihoc and the Editor-in-Chief of the IEEE/ACM TRANSACTIONS ON NETWORKING.

Appendix A
Proof of Proposition 1

Note that our state space is countable and action set is finite. Thus, by Theorem 6.10.4 in [26], we only need to show the following two assumptions hold:
Assumption 1: There exists a constant $\mu < \infty$ such that

$$\sup_{\mathbf{q}\in\mathcal{X}} |c(\mathbf{q})| \leq \mu w(\mathbf{q}). \tag{40}$$

Assumption 2: (i) There exists a constant L, $0 \leq L < \infty$, such that

$$\sum_{\mathbf{q}'\in\mathcal{X}} \mathbb{P}(\mathbf{q}'|\mathbf{q},\mathbf{u})w(\mathbf{q}') \leq w(\mathbf{q})L, \tag{41}$$

for all $\mathbf{q}$ and $\mathbf{u}$. (ii) For each β, there exists an ζ, $0 \leq \zeta < 1$ and an integer I such that for all deterministic policies π,

$$\beta^I \sum_{\mathbf{q}'\in\mathcal{X}} \mathbb{P}^I_\pi(\mathbf{q}'|\mathbf{q})w(\mathbf{q}') \leq \zeta w(\mathbf{q}), \tag{42}$$

where $\mathbb{P}^I_\pi$ denotes the I-stage transition probability under policy π.

Assumption 1 holds with $\mu = 1$ obviously. Since

$$\sum_{\mathbf{q}'\in\mathcal{X}} \mathbb{P}(\mathbf{q}'|\mathbf{q},\mathbf{u})w(\mathbf{q}') \leq w(\mathcal{A}_0(\mathbf{q})) \leq 2w(\mathbf{q}), \tag{43}$$

Assumption 2(i) holds with $L = 2$. For any policy π, we have

$$\beta^I \sum_{\mathbf{q}'\in\mathcal{X}} \mathbb{P}^I_\pi(\mathbf{q}'|\mathbf{q})w(\mathbf{q}') \leq \beta^I w(\mathcal{A}^I_0(\mathbf{q})) \leq \beta^I(I+1)w(\mathbf{q}), \tag{44}$$

where $\mathcal{A}^I_0$ denotes the I arrivals to the system. Consequently, for I sufficiently large, $\beta^I(I+1) < 1$. Hence, assumption 2(ii) holds.

Appendix B
Proof of Theorem 1

Note that zero function (i.e., $v = 0$) belongs to the function set Θ. By Proposition 1, we have $\lim_{n\to\infty} \mathcal{L}^{(n)}v = J_\beta$. Thus, in order to show that $J_\beta \in \Theta$, we start with zero function and show that $\mathcal{L}v \in \Theta$ given that $v \in \Theta$. Before providing the proof, we first show that the operator B preserves the properties of functions in Θ in Lemma 6, which will be used in our main proof.

Lemma 6. *If $v \in \Theta$, then $Bv \in \Theta$, where B is defined by Eq.* (7).

Proof. First, let us consider the preservation of property (11). We want to show

$$\begin{aligned}&Bv(A_1(\mathbf{q})) - Bv(A_h(\mathbf{q}))\\ =&\min_{u\in K_{(A_1(\mathbf{q}))}} v\left(u\left(A_1\left(\mathbf{q}\right)\right)\right) - \min_{u\in K_{\mathbf{q}}} v\left(u\left(A_h\left(\mathbf{q}\right)\right)\right) \leq 0.\end{aligned}$$

It suffices to show that for each $u_2 \in K_{\mathbf{q}}$, there exists $u_1 \in K_{(A_1(\mathbf{q}))}$ such that $v(u_2(A_h(\mathbf{q}))) \geq v(u_1(A_1(\mathbf{q})))$. Same logic is also used in the following proof for other cases.

Generally, $\{A_h, A_1\} \subseteq K_{\mathbf{q}}$ and $A_h \in K_{(A_1(\mathbf{q}))}$. Then, we have $v\left(A_h\left(A_h\left(\mathbf{q}\right)\right)\right) \overset{(11)}{\geq} v\left(A_1\left(A_h\left(\mathbf{q}\right)\right)\right) = v\left(A_h\left(A_1\left(\mathbf{q}\right)\right)\right)$. If $l_2 = 0$, then $A_2 \in K_{\mathbf{q}}$ and $A_r \in K_{(A_1(\mathbf{q}))}$. But $v\left(A_2\left(A_h\left(\mathbf{q}\right)\right)\right) = (q_0-1, 0, q_1, 1, s) = v\left(A_r\left(A_1\left(\mathbf{q}\right)\right)\right)$. If $l_2 = 0$ and $q_1 \geq 1$, then $A_r \in K_{\mathbf{q}}$ and we have:

$$v\left(A_r\left(A_h\left(\mathbf{q}\right)\right)\right) \overset{(12)}{\geq} v\left(A_2\left(A_h\left(\mathbf{q}\right)\right)\right) = v\left(A_r\left(A_1\left(\mathbf{q}\right)\right)\right).$$

For property (12), generally, $A_h \in K_{(A_r(\mathbf{q}))}$ and $A_h \in K_{(A_2(\mathbf{q}))}$. Then, we have:

$$v\left(A_h\left(A_r\left(\mathbf{q}\right)\right)\right) = v\left(A_r\left(\mathbf{q}\right)\right) \overset{(12)}{\geq} v\left(A_2\left(\mathbf{q}\right)\right) = v\left(A_h\left(A_2\left(\mathbf{q}\right)\right)\right).$$

If $l_1 = 0$, then A_1 is admissible for both states. We obtain:

$$\begin{aligned}v\left(A_1\left(A_r\left(\mathbf{q}\right)\right)\right) =& (q_0-1, 1, q_1-1, 1, s)\\ &\overset{(13)}{\geq} (q_0-1, 0, q_1, 1, s) = v\left(A_h\left(A_2\left(\mathbf{q}\right)\right)\right).\end{aligned}$$

If $l_1 = 1$, then $A_1 \in K_{(A_r(\mathbf{q}))}$. By (13), it is better to renege a packet from processing server than from the mmWave interface and we have:

$$v\left(A_1\left(A_r\left(\mathbf{q}\right)\right)\right) = (q_0-1, 1, q_1, 1, s) = v\left(A_h\left(A_2\left(\mathbf{q}\right)\right)\right).$$

For property (13), if $l_1 = 0$, then the result holds obviously since $\mathcal{T}(\mathbf{q}) = \mathbf{q}$. If $l_1 = 1$, then generally A_h is admissible for both states and we have $v\left(A_h\left(\mathbf{q}\right)\right) \overset{(13)}{\geq} v\left(A_h\left(\mathcal{T}\left(\mathbf{q}\right)\right)\right)$. If $l_2 = 0$, then $A_r \in K_{(\mathcal{T}(\mathbf{q}))}$ and we have $v\left(A_r\left(\mathbf{q}\right)\right) = v\left(A_r\left(\mathcal{T}\left(\mathbf{q}\right)\right)\right)$. If $l_2 = 0$ and $q_0 > 0$, then $A_2 \in K_{(\mathcal{T}(\mathbf{q}))}$ and we have $v\left(A_2\left(\mathbf{q}\right)\right) \overset{(13)}{\geq} v\left(\mathcal{T}\left(A_2\left(\mathbf{q}\right)\right)\right) = v\left(A_2\left(\mathcal{T}\left(\mathbf{q}\right)\right)\right)$.

For property (14), generally A_h is admissible for both states and we have $v\left(A_h\left(A_2\left(\mathbf{q}\right)\right)\right) \overset{(14)}{\geq} v\left(A_h\left(A_1\left(\mathbf{q}\right)\right)\right)$. If $q_0 > 1$, then $A_1 \in K_{(A_2(\mathbf{q}))}$ and we obtain $v\left(A_1\left(A_2\left(\mathbf{q}\right)\right)\right) = v\left(A_2\left(A_1\left(\mathbf{q}\right)\right)\right)$.

For property (15), if $\mathbf{q}_1 = (q_0, l_1, q_1, l_2, s)$ and $\mathbf{q}_2 = (q_0+1, l_1, q_1, l_2, s)$, we have

$$\begin{aligned}Bv(\mathbf{q}_2) =& \min_{u\in K_{\mathbf{q}_2}} v(u(q_0+1, l_1, q_1, l_2, s))\\ &\overset{(15)}{\geq} v(A_h(q_0, l_1, q_1, l_2, s))\\ &\geq Bv(\mathbf{q}_1)\end{aligned}$$

Similarly, we can show that property (15) holds when $\mathbf{q}_2 - \mathbf{q}_1 \in \{(0,1,0,0,0), (0,0,1,0,0), (0,0,0,1,0)\}$. □

Next, we use this lemma to show our result.

Property* (11)*: If mmWave channel is unavailable $s = 0$, then we have:

$$\begin{aligned}&\mathcal{L}v\left(A_1(\mathbf{q})\right) - \mathcal{L}v\left(A_h(\mathbf{q})\right)\\ =&\beta\lambda\big(Bv\left(\mathcal{A}_0\left(A_1(\mathbf{q})\right)\right) - Bv\left(\mathcal{A}_0\left(A_h(\mathbf{q})\right)\right)\big)\\ &+ \beta\mu_{\text{p}}\big(Bv\left(\mathcal{T}\left(A_1(\mathbf{q})\right)\right) - Bv\left(\mathcal{T}\left(A_h(\mathbf{q})\right)\right)\big)\\ &+ \beta\mu_{\text{sub-6}}\big(Bv\left(\mathcal{D}_2\left(A_1(\mathbf{q})\right)\right) - Bv\left(\mathcal{D}_2\left(A_h(\mathbf{q})\right)\right)\big)\\ &+ \beta\gamma\big(Bv\left(\mathcal{G}\left(A_1(\mathbf{q})\right)\right) - Bv\left(\mathcal{G}\left(A_h(\mathbf{q})\right)\right)\big)\\ &+ \beta p(\mathbf{q})\big(Bv\left(A_1(\mathbf{q})\right) - Bv\left(A_h(\mathbf{q})\right)\big)\\ \overset{\text{(c1)}}{\leq}&\beta\lambda\big(Bv\left(A_1\left(\mathcal{A}_0(\mathbf{q})\right)\right) - Bv\left(A_h\left(\mathcal{A}_0(\mathbf{q})\right)\right)\big)\\ &+ \beta\mu_{\text{sub-6}}\big(Bv\left(A_1\left(\mathcal{D}_2(\mathbf{q})\right)\right) - Bv\left(A_h\left(\mathcal{D}_2(\mathbf{q})\right)\right)\big)\\ &+ \beta\mu_{\text{p}}\big(Bv\left(A_1(\mathbf{q})\right) - Bv\left(A_h(\mathbf{q})\right)\big)\\ &+ \beta\gamma\big(Bv\left(A_1\left(\mathcal{G}(\mathbf{q})\right)\right) - Bv\left(A_h\left(\mathcal{G}(\mathbf{q})\right)\right)\big)\\ &+ \beta p(\mathbf{q})\big(Bv\left(A_1(\mathbf{q})\right) - Bv\left(A_h(\mathbf{q})\right)\big)\\ \overset{\text{(c2)}}{\leq}&0\end{aligned}$$

where (c1) is due to $Bv\left(\mathcal{T}\left(A_1(\mathbf{q})\right)\right) \le Bv\left(A_1(\mathbf{q})\right)$ by Lamma 6 and $Bv\left(\mathcal{T}\left(A_h(\mathbf{q})\right)\right) = Bv\left(A_h(\mathbf{q})\right)$; (c2) is because that B preserves the property (11) by Lamma 6.

If mmWave channel is available $s = 1$, then we have:

$$\begin{aligned}
&\mathcal{L}v\left(A_1(\mathbf{q})\right) - \mathcal{L}v\left(A_h(\mathbf{q})\right)\\
=&\beta\lambda\big(Bv\left(\mathcal{A}_0\left(A_1(\mathbf{q})\right)\right) - Bv\left(\mathcal{A}_0\left(A_h(\mathbf{q})\right)\right)\big)\\
&+ \beta\mu_{\text{p}}\big(Bv\left(\mathcal{T}\left(A_1(\mathbf{q})\right)\right) - Bv\left(\mathcal{T}\left(A_h(\mathbf{q})\right)\right)\big)\\
&+ \beta\mu_{\text{mm}}\big(Bv\left(\mathcal{D}_1\left(A_1(\mathbf{q})\right)\right) - Bv\left(\mathcal{D}_1\left(A_h(\mathbf{q})\right)\right)\big)\\
&+ \beta\mu_{\text{sub-6}}\big(Bv\left(\mathcal{D}_2\left(A_1(\mathbf{q})\right)\right) - Bv\left(\mathcal{D}_2\left(A_h(\mathbf{q})\right)\right)\big)\\
&+ \beta\alpha\big(Bv\left(\mathcal{B}\left(A_1(\mathbf{q})\right)\right) - Bv\left(\mathcal{B}\left(A_h(\mathbf{q})\right)\right)\big)\\
&+ \beta p(\mathbf{q})\big(Bv\left(A_1(\mathbf{q})\right) - Bv\left(A_h(\mathbf{q})\right)\big)\\
\overset{\text{(c3)}}{\le}&\beta\lambda\big(Bv\left(A_1\left(\mathcal{A}_0(\mathbf{q})\right)\right) - Bv\left(A_h\left(\mathcal{A}_0(\mathbf{q})\right)\right)\big)\\
&+ \beta\mu_{\text{p}}\big(Bv\left(A_1(\mathbf{q})\right) - Bv\left(A_h(\mathbf{q})\right)\big)\\
&+ \beta\mu_{\text{mm}}\big(Bv\left(A_1\left(\mathcal{D}_1(\mathbf{q})\right)\right) - Bv\left(A_h\left(\mathcal{D}_1(\mathbf{q})\right)\right)\big)\\
&+ \beta\mu_{\text{sub-6}}\big(Bv\left(A_1\left(\mathcal{D}_2(\mathbf{q})\right)\right) - Bv\left(A_h\left(\mathcal{D}_2(\mathbf{q})\right)\right)\big)\\
&+ \beta\alpha\big(Bv\left(A_1\left(\mathcal{B}(\mathbf{q})\right)\right) - Bv\left(A_h\left(\mathcal{B}(\mathbf{q})\right)\right)\big)\\
&+ \beta p(\mathbf{q})\big(Bv\left(A_1(\mathbf{q})\right) - Bv\left(A_h(\mathbf{q})\right)\big)\\
\overset{\text{(c4)}}{\le}&0
\end{aligned}$$

where (c3) is due to $Bv\left(\mathcal{T}\left(A_1(\mathbf{q})\right)\right) \le Bv\left(A_1(\mathbf{q})\right)$ by Lemma 6 and $Bv\left(\mathcal{T}\left(A_h(\mathbf{q})\right)\right) = Bv\left(A_h(\mathbf{q})\right)$; (c4) is because that B preserves the property (11) by Lemma 6.

Property **(12)*:*** If mmWave channel is unavailable $s = 0$, then we have:

$$\begin{aligned}
&\mathcal{L}v\left(A_2(\mathbf{q})\right) - \mathcal{L}v\left(A_r(\mathbf{q})\right)\\
=&\beta\lambda\big(Bv\left(\mathcal{A}_0\left(A_2(\mathbf{q})\right)\right) - Bv\left(\mathcal{A}_0\left(A_r(\mathbf{q})\right)\right)\big)\\
&+ \beta\mu_{\text{sub-6}}\big(Bv\left(\mathcal{D}_2\left(A_2(\mathbf{q})\right)\right) - Bv\left(\mathcal{D}_2\left(A_r(\mathbf{q})\right)\right)\big)\\
&+ \beta\mu_{\text{p}}\big(Bv\left(\mathcal{T}\left(A_2(\mathbf{q})\right)\right) - Bv\left(\mathcal{T}\left(A_r(\mathbf{q})\right)\right)\big)\\
&+ \beta\gamma\big(Bv\left(\mathcal{G}\left(A_2(\mathbf{q})\right)\right) - Bv\left(\mathcal{G}\left(A_r(\mathbf{q})\right)\right)\big)\\
&+ \beta p(\mathbf{q})\big(Bv\left(A_2(\mathbf{q})\right) - Bv\left(A_r(\mathbf{q})\right)\big) \qquad (45)
\end{aligned}$$

Notice that by property (13), it is better to renege a packet from the processing server than from the mmWave interface if the processing server is not empty. Then, the operator $\mathcal{T}$ will not change the system state since the processing server must be empty after action A_r. Thus, $Bv\left(\mathcal{T}\left(A_r(\mathbf{q})\right)\right) = Bv\left(A_r(\mathbf{q})\right)$. We have:

$$\begin{aligned}
&Bv\left(\mathcal{T}\left(A_2(\mathbf{q})\right)\right) - Bv\left(\mathcal{T}\left(A_r(\mathbf{q})\right)\right)\\
\overset{(13)}{\le}&Bv\left(A_2(\mathbf{q})\right) - Bv\left(A_r(\mathbf{q})\right) \overset{(12)}{\le} 0
\end{aligned}$$

For the λ term, we have:

$$\begin{aligned}
&Bv\left(\mathcal{A}_0\left(A_2(\mathbf{q})\right)\right) - Bv\left(\mathcal{A}_0\left(A_r(\mathbf{q})\right)\right)\\
=&Bv\left(A_2\left(\mathcal{A}_0(\mathbf{q})\right)\right) - Bv\left(A_r\left(\mathcal{A}_0(\mathbf{q})\right)\right) \overset{(12)}{\le} 0
\end{aligned}$$

For the $\mu_{\text{sub-6}}$ term, if $l_1 = 1$, then we obtain:

$$\begin{aligned}
&Bv\left(\mathcal{D}_2\left(A_2(\mathbf{q})\right)\right) - Bv\left(\mathcal{D}_2\left(A_r(\mathbf{q})\right)\right)\\
=&Bv\left(q_0 - 1, 1, q_1, 0, 0\right) - Bv\left(q_0, 0, q_1, 0, 0\right)\\
=&Bv\left(A_1(q_0, 0, q_1, 0, 0)\right) - Bv\left(A_h(q_0, 0, q_1, 0, 0)\right) \overset{(11)}{\le} 0
\end{aligned}$$

If $l_1 = 0$, then we obtain:

$$\begin{aligned}
&Bv\left(\mathcal{D}_2\left(A_2(\mathbf{q})\right)\right) - Bv\left(\mathcal{D}_2\left(A_r(\mathbf{q})\right)\right)\\
=&Bv\left(q_0 - 1, 0, q_1, 0, 0\right) - Bv\left(q_0, 0, q_1 - 1, 0, 0\right)\\
\overset{(13)}{\le}&Bv\left(q_0 - 1, 1, q_1 - 1, 0, 0\right) - Bv\left(q_0, 0, q_1 - 1, 0, 0\right) \overset{(11)}{\le} 0
\end{aligned}$$

For the γ term, we have:

$$\begin{aligned}
&Bv\left(\mathcal{G}\left(A_2(\mathbf{q})\right)\right) - Bv\left(\mathcal{G}\left(A_r(\mathbf{q})\right)\right)\\
=&Bv\left(A_2\left(\mathcal{G}(\mathbf{q})\right)\right) - Bv\left(A_r\left(\mathcal{G}(\mathbf{q})\right)\right) \overset{(12)}{\le} 0
\end{aligned}$$

The last term is less than or equal to zero since B preserves property (12) by Lemma 6. Hence, (45) is less or equal to zero.

For the case that the mmWave channel is available $s = 1$, the proof of the terms λ, μ_{p}, $\mu_{\text{sub-6}}$ and last term ($p(\mathbf{q})$ term) are same with the case $s = 0$. It remains to show the μ_{mm} and α terms.

For the μ_{mm} term, if $l_1 = 1$ or $l_1 = 1, q_1 > 1$, then exchanging the order of operators $\mathcal{D}_1$ and A_r, operators $\mathcal{D}_1$ and A_2 does not affect results. Then, we have;

$$\begin{aligned}
&Bv\left(\mathcal{D}_1\left(A_2(\mathbf{q})\right)\right) - Bv\left(\mathcal{D}_1\left(A_r(\mathbf{q})\right)\right)\\
=&Bv\left(A_2\left(\mathcal{D}_1(\mathbf{q})\right)\right) - Bv\left(A_r\left(\mathcal{D}_1(\mathbf{q})\right)\right) \overset{(12)}{\le} 0
\end{aligned}$$

If $l_1 = 0$ and $q_1 = 1$, we have;

$$\begin{aligned}
&Bv\left(\mathcal{D}_1\left(A_2(\mathbf{q})\right)\right) - Bv\left(\mathcal{D}_1\left(A_r(\mathbf{q})\right)\right)\\
=&Bv\left(q_0 - 1, 0, 0, 1, 1\right) - Bv\left(q_0, 0, 0, 1, 1\right) \overset{(15)}{\le} 0
\end{aligned}$$

For the α term, we have:

$$\begin{aligned}
&Bv\left(\mathcal{B}\left(A_2(\mathbf{q})\right)\right) - Bv\left(\mathcal{B}\left(A_r(\mathbf{q})\right)\right)\\
=&Bv\left(A_2\left(\mathcal{B}(\mathbf{q})\right)\right) - Bv\left(A_r\left(\mathcal{B}(\mathbf{q})\right)\right) \overset{(12)}{\le} 0
\end{aligned}$$

Property **(13)*:*** This can be shown with similar argument in proof for property (11) and (12).

Property **(14)*:*** Note that

$$\begin{aligned}
&Bv\left(\mathcal{A}_0\left(A_1\left(\mathbf{q}\right)\right)\right) - Bv\left(\mathcal{A}_0\left(A_2\left(\mathbf{q}\right)\right)\right)\\
=&Bv\left(A_1\left(\mathcal{A}_0\left(\mathbf{q}\right)\right)\right) - Bv\left(A_2\left(\mathcal{A}_0\left(\mathbf{q}\right)\right)\right) \le 0.
\end{aligned}$$

With similar argument, we have $Bv\left(\mathcal{D}_1\left(A_1\left(\mathbf{q}\right)\right)\right) \le Bv\left(\mathcal{D}_1\left(A_2\left(\mathbf{q}\right)\right)\right)$, $Bv\left(\mathcal{B}\left(A_1\left(\mathbf{q}\right)\right)\right) \le Bv\left(\mathcal{B}\left(A_2\left(\mathbf{q}\right)\right)\right)$ and $Bv\left(\mathcal{G}\left(A_1\left(\mathbf{q}\right)\right)\right) \le Bv\left(\mathcal{G}\left(A_2\left(\mathbf{q}\right)\right)\right)$. By Eq. (8), it remains to show

$$\mu_{\text{p}}C_1 + \mu_{\text{sub-6}}C_2 \le \mu_{\text{p}}C_3 + \mu_{\text{sub-6}}C_4. \qquad (46)$$

where $C_1 \triangleq Bv\left(q_0 - 1, 0, 1, 0, s\right)$, $C_2 \triangleq Bv\left(q_0 - 1, 1, 0, 0, s\right)$, $C_3 \triangleq Bv\left(q_0 - 1, 0, 0, 1, s\right)$ and $C_4 \triangleq Bv\left(q_0 - 1, 0, 0, 0, s\right)$. Note that $C_4 \le C_1 \le C_2 \le C_3$ by Lemma 6.

First consider the difference between C_3 and C_1. For the best case, at current state the optimal action is to use the sub-6 GHz and the next event is processing completion. Then, $C_3 = v(q_0 - 2, 1, 0, 1, s)$ and $C_1 = v(q_0 - 2, 0, 1, 1, s)$. After processing completion, the states in two processes become the same. In the case, the difference between C_3 and C_1 is $\frac{1}{\mu_{\text{p}}}$. Thus, $C_3 - C_1 \ge \frac{1}{\mu_{\text{p}}}$.

Now consider the difference between C_2 and C_4. By Eq. (14), $C_4 = v(q_0 - 2, 1, 0, 0, s)$ if $q_0 \ge 2$. Then, $C_2 - C_4 \le$

$v(q_0-2,1,0,1,s)-v(q_0-2,1,0,0,s)$. For the worst case, in the process with initial state $(q_0-2,1,0,0,s)$, the sub-6 GHz is not used until in the process with initial state $(q_0-2,1,0,1,s)$ the sub-6 GHz completes its service. Thus, $C_2-C_4\le\frac{1}{\mu_{\text{sub-6}}}$. For case $q_0=1$, we can get same conclusion with similar argument. Thus, $\mu_p(C_3-C_1)\ge\mu_{\text{sub-6}}(C_2-C_4)$.

***Property** (15)**:** Now we check monotonicity:* If the mmWave is unavailable $s=0$, then:

$$\begin{aligned}
&\mathcal{L}v\left(\mathbf{q}_1\right)-\mathcal{L}v\left(\mathbf{q}_2\right)\\
=&\beta\lambda\big(Bv\left(\mathcal{A}_0\left(\mathbf{q}_1\right)\right)-Bv\left(\mathcal{A}_0\left(\mathbf{q}_2\right)\right)\big)\\
&+\beta\mu_{\text{p}}\big(Bv\left(\mathcal{T}\left(\mathbf{q}_1\right)\right)-Bv\left(\mathcal{T}\left(\mathbf{q}_2\right)\right)\big)\\
&+\beta\mu_{\text{sub-6}}\big(Bv\left(\mathcal{D}_2\left(\mathbf{q}_1\right)\right)-Bv\left(\mathcal{D}_2\left(\mathbf{q}_2\right)\right)\big)\\
&+\beta\gamma\big(Bv\left(\mathcal{G}\left(\mathbf{q}_1\right)\right)-Bv\left(\mathcal{G}\left(\mathbf{q}_2\right)\right)\big)\\
&+\beta p(\mathbf{q})\big(Bv\left(\mathbf{q}_1\right)-Bv\left(\mathbf{q}_2\right)\big)\\
\le&0
\end{aligned}$$

The inequality holds because for $\mathcal{E}\in\{\mathcal{A}_0,\mathcal{T},\mathcal{D}_2,\mathcal{G}\}$, we have $\mathcal{E}(\mathbf{q}_2)-\mathcal{E}(\mathbf{q}_1)\in\{(1,0,0,0,0),(0,1,0,0,0),(0,0,1,0,0),(0,0,0,1,0),(0,0,0,0,0)\}$, and the operator B preserves the property (15). With similar argument, we can prove the case that the mmWave is available $s=1$.

Appendix C
Proof of Lemma 1

Note that $J_\beta^0\left(x,q_1,l_2,s\right)=x+q_1+l_2$ and $J_\beta^0\in\mathscr{F}$ obviously. By Eq. (16), it remains to show that $BJ_\beta^n\in\mathscr{F}$ and then $J_\beta^{n+1}\in\mathscr{F}$ given $J_\beta^n\in\mathscr{F}$. Before our proof, we provide some properties extended from Definition 2, which will be used in the following proof.

Extended properties from Definition 2:

$$2f\left(x,q_1,1,s\right)\le f\left(x+1,q_1,1,s\right)+f\left(x-1,q_1,1,s\right)\quad(47)$$
$$2f\left(0,q_1,1,s\right)\le f\left(1,q_1,1,s\right)+f\left(0,q_1-1,1,s\right)\quad(48)$$
$$2f\left(0,q_1,1,s\right)\le f\left(0,q_1+1,1,s\right)+f\left(0,q_1-1,1,s\right)\quad(49)$$
$$2f\left(x+1,q_1,0,s\right)\le f\left(x+2,q_1,0,s\right)+f\left(x,q_1,0,s\right)\quad(50)$$
$$2f\left(0,q_1+1,0,s\right)\le f\left(0,q_1,0,s\right)+f\left(0,q_1+2,0,s\right)\quad(51)$$
$$\begin{aligned}&f\left(x,q_1,1,s\right)+f\left(x-1,q_1+1,1,s\right)\\&\qquad\le f\left(x,q_1+1,1,s\right)+f\left(x-1,q_1,1,s\right)\end{aligned}\quad(52)$$
$$\begin{aligned}&f\left(0,q_1+1,0,s\right)+f\left(0,q_1+1,1,s\right)\\&\qquad\le f\left(0,q_1,1,s\right)+f\left(1,q_1+1,0,s\right)\end{aligned}\quad(53)$$
$$\begin{aligned}&f\left(x+1,q_1,0,s\right)+f\left(x,q_1+1,0,s\right)\\&\qquad\le f\left(x+1,q_1+1,0,s\right)+f\left(x,q_1,0,s\right)\end{aligned}\quad(54)$$

These properties can be obtained from combinations of certain equations in Definition 2. We take Eq. (47) for example. It is obtained by adding Eq. (17) with x replaced by $x-1$ and Eq. (21).

With this, we first show that $BJ_\beta^n\in\mathscr{F}$ given $J_\beta^n\in\mathscr{F}$.

For Eq. (17)**:** Note that $BJ_\beta^n\left(x+2,q_1,0,s\right)=\min\{J_\beta^n\left(x+2,q_1,0,s\right),J_\beta^n\left(x+1,q_1,1,s\right)\}$ and $BJ_\beta^n\left(x,q_1,1,s\right)=J_\beta^n\left(x,q_1,1,s\right)$. Thus, we only need to consider two cases. Note that the proof for other equations also follow similar flow in which case we omit description of same analysis.

If $BJ_\beta^n\left(x+2,q_1,0,s\right)=J_\beta^n\left(x+2,q_1,0,s\right)$, then:

$$\begin{aligned}
&BJ_\beta^n\left(x+1,q_1,0,s\right)+BJ_\beta^n\left(x+1,q_1,1,s\right)\\
&\overset{(7)}{\le}J_\beta^n\left(x+1,q_1,0,s\right)+J_\beta^n\left(x+1,q_1,1,s\right)\\
&\overset{(17)}{\le}J_\beta^n\left(x,q_1,1,s\right)+J_\beta^n\left(x+2,q_1,0,s\right).
\end{aligned}$$

If $BJ_\beta^n\left(x+2,q_1,0,s\right)=J_\beta^n\left(x+1,q_1,1,s\right)$, then:

$$\begin{aligned}
&BJ_\beta^n\left(x+1,q_1,0,s\right)+BJ_\beta^n\left(x+1,q_1,1,s\right)\\
&\overset{(7)}{\le}J_\beta^n\left(x,q_1,1,s\right)+J_\beta^n\left(x+1,q_1,1,s\right).
\end{aligned}$$

Thus, (17) holds. Similarly, we can show that Eq. (18), (19) and (20) hold.

For Eq. (21) **:** If $BJ_\beta^n\left(x,q_1,0,s\right)=J_\beta^n\left(x,q_1,0,s\right)$, then:

$$\begin{aligned}
&BJ_\beta^n\left(x,q_1,1,s\right)+BJ_\beta^n\left(x+1,q_1,0,s\right)\\
&\overset{(7)}{\le}J_\beta^n\left(x,q_1,1,s\right)+J_\beta^n\left(x+1,q_1,0,s\right)\\
&\overset{(21)}{\le}J_\beta^n\left(x,q_1,0,s\right)+J_\beta^n\left(x+1,q_1,1,s\right).
\end{aligned}$$

If $x\ge1$ and $BJ_\beta^n\left(x,q_1,0,s\right)=J_\beta^n\left(x-1,q_1,1,s\right)$, then:

$$\begin{aligned}
&BJ_\beta^n\left(x,q_1,1,s\right)+BJ_\beta^n\left(x+1,q_1,0,s\right)\\
&\overset{(7)}{\le}2J_\beta^n\left(x,q_1,1,s\right)\\
&\overset{(47)}{\le}J_\beta^n\left(x-1,q_1,1,s\right)+J_\beta^n\left(x+1,q_1,1,s\right).
\end{aligned}$$

If $x=0$, $q_1\ge1$ and $BJ_\beta^n\left(0,q_1,0,s\right)=J_\beta^n\left(0,q_1-1,1,s\right)$, then:

$$\begin{aligned}
&BJ_\beta^n\left(0,q_1,1,s\right)+BJ_\beta^n\left(1,q_1,0,s\right)\\
&\overset{(7)}{\le}2J_\beta^n\left(0,q_1,1,s\right)\\
&\overset{(48)}{\le}J_\beta^n\left(0,q_1-1,1,s\right)+J_\beta^n\left(1,q_1,1,s\right).
\end{aligned}$$

For Eq. (22) **:** If $BJ_\beta^n\left(x,q_1,0,s\right)=J_\beta^n\left(x,q_1,0,s\right)$, then:

$$\begin{aligned}
&BJ_\beta^n\left(x,q_1,1,s\right)+BJ_\beta^n\left(x,q_1+1,0,s\right)\\
&\overset{(7)}{\le}J_\beta^n\left(x,q_1,1,s\right)+J_\beta^n\left(x,q_1+1,0,s\right)\\
&\overset{(22)}{\le}J_\beta^n\left(x,q_1,0,s\right)+J_\beta^n\left(x,q_1+1,1,s\right).
\end{aligned}$$

If $x\ge1$ and $BJ_\beta^n\left(x,q_1,0,s\right)=J_\beta^n\left(x-1,q_1,1,s\right)$, then:

$$\begin{aligned}
&BJ_\beta^n\left(x,q_1,1,s\right)+BJ_\beta^n\left(x,q_1+1,0,s\right)\\
&\overset{(7)}{\le}J_\beta^n\left(x,q_1,1,s\right)+J_\beta^n\left(x-1,q_1+1,1,s\right)\\
&\overset{(52)}{\le}J_\beta^n\left(x-1,q_1,1,s\right)+J_\beta^n\left(x,q_1+1,1,s\right).
\end{aligned}$$

If $x=0$, $q_1\ge1$ and $BJ_\beta^n\left(0,q_1,0,s\right)=J_\beta^n\left(0,q_1-1,1,s\right)$, then:

$$\begin{aligned}
&BJ_\beta^n\left(0,q_1,1,s\right)+BJ_\beta^n\left(0,q_1+1,0,s\right)\\
&\overset{(7)}{\le}2J_\beta^n\left(0,q_1,1,s\right)\\
&\overset{(49)}{\le}J_\beta^n\left(0,q_1-1,1,s\right)+J_\beta^n\left(0,q_1+1,1,s\right).
\end{aligned}$$

For Eq. (23) **:** If $BJ_\beta^n\left(x+1,q_1,l_2,s\right)=J_\beta^n\left(x+1,q_1,l_2,s\right)$, then:

$$BJ_\beta^n\left(x,q_1,l_2,s\right)\overset{(7)}{\le}J_\beta^n\left(x,q_1,l_2,s\right)\overset{(23)}{\le}J_\beta^n\left(x+1,q_1,l_2,s\right).$$

If $l_2 = 0$ and $BJ^n_\beta(x+1, q_1, 0, s) = J^n_\beta(x, q_1, 1, s)$, then:

$$BJ^n_\beta(x, q_1, 0, s) \overset{(7)}{\leq} J^n_\beta(x, q_1, 0, s) \overset{(25)}{\leq} J^n_\beta(x, q_1, 1, s).$$

Similarly, we obtain Eq. (24) and Eq. (25).

Next, we show that $J^{n+1}_\beta \in \mathscr{F}$. According to Eq. (16), we show seven terms, say λ, μ_{p}, μ_{mm}, $\mu_{\text{sub-6}}$, α, γ and $p(\mathbf{q})$ terms, satisfy properties in Definition 2, respectively. Note that $p(\mathbf{q})$ term holds obviously since $BJ^n_\beta \in \mathscr{F}$. Moreover, the α and γ terms hold since only mmWave changes status of availability (s turns into $1-s$), $\forall s \in \{0,1\}$, and $BJ^n_\beta \in \mathscr{F}$. Next, we focus on the remaining four terms.

For Eq. (17)**:** the difficulty falls in the μ_{p} term with $x = 0$ and $\mu_{\text{sub-6}}$ term. The former one can be proved with Eq. (53). and for the latter, Eq. (17) reduces to Eq. (50).

For Eq. (18)**:** For the μ_{p} term, when $x = 0$, it reduces to Eq. (19). For the μ_{mm} term, when $q_1 = 0$, Eq. (18) reduces to equality. For $\mu_{\text{sub-6}}$ term, Eq. (18) reduces to Eq. (54).

For Eq. (19)**:** the λ and μ_{p} terms obviously hold. For the μ_{mm} term, the difficulty falls in the case with $q_1 = 0$, where $BJ^n_\beta(0,0,0,s) \leq BJ^n_\beta(0,1,0,s)$. In fact, the inequality holds by Eq. (24). For the $\mu_{\text{sub-6}}$ term, Eq. (19) reduces to Eq. (51).

For Eq. (20)**:** the λ, μ_{p} with $x \geq 1$, μ_{mm} with $q_1 \geq 1$ and $\mu_{\text{sub-6}}$ terms hold obviously. As for the μ_{p} with $x = 0$ term, Eq. (20) reduces to an equation. As for the μ_{mm} with $q_1 = 0$ term, Eq. (20) reduces to Eq. (23) with $q_1 = 0$.

For Eq. (21)**:** the λ, μ_{p} with $x \geq 1$, and μ_{mm} terms hold obviously. Notice that for the $\mu_{\text{sub-6}}$ term, Eq. (21) reduces to an equality. For the μ_{p} term with $x = 0$, Eq. (21) reduces to Eq. (22).

For Eq. (22)**:** the difficulty falls in the $\mu_{\text{sub-6}}$ and μ_{mm} with $q_1 = 0$ terms. For both of the cases, Eq. (22) reduces to an equality.

For Eq. (23)**:** the difficulty falls in the μ_{p} term with $x = 0$. In the case, Eq. (23) reduces to Eq. (24) with $x = 0$.

For Eq. (24)**:** the difficulty falls in the μ_{mm} term with $q_1 = 0$. In the case, Eq. (24) reduces to an equality.

For Eq. (25)**:** the difficulty falls in the $\mu_{\text{sub-6}}$ term, in which case Eq. (25) reduces to an equality.

Appendix D
Proof of Theorem 2

If the mmWave channel is unavailable ($s = 0$), then according to Lemma 1, for each $n \in \mathbb{N}$, J^n_β satisfies properties (17), (18), (19) and (53). It implies that for either the case $x > 0$ or $x = 0$, $J^n_\beta(x+1, q_1, 0, 0) - J^n_\beta(x, q_1, 1, 0)$ or $J^n_\beta(0, q_1+1, 0, 0) - J^n_\beta(0, q_1, 1, 0)$ increases as the number of packets in the FastLane (i.e., $x + q_1 + 1$) increases (due to increase of x or q_1 or both). In other words, the difference between costs resulted from not-adding-to-sub-6 and adding-to-sub-6 increases as the number of packets in the FastLane increases. It is known that $J^n_\beta(0,1,0,0) \leq J^n_\beta(1,0,0,0) \leq J^n_\beta(0,0,1,0)$, which means that it's better to hold the packet in FastLane when only one packet is in the system. As $x + q_1$ increases, the difference becomes positive, which means that adding-to-sub-6 obtains priority.

To sum up, when $s = 0$, for each $n \in \mathbb{N}$, there exists a certain threshold for the queue length of FastLane (or the number of packets in the system) above which we should add a packet to the sub-6 GHz interface. In other words, for each round of value iteration, corresponding policy is of threshold-type. As $n \to \infty$, the corresponding policy is also of the threshold-type, and the policy is expected total discounted delay optimal policy. Thus, when $s = 0$, there exists some threshold m^*_0 such that when $c(\mathbf{q}) > m^*_0$, it is optimal to use the sub-6 GHz interface.

With similar argument, we can easily show that when $s = 1$, there exists a threshold m^*_1 such that when $c(\mathbf{q}) > m^*_1$, it is optimal to use the sub-6 GHz interface.

Appendix E
Proof of Theorem 3

Let us call an epoch fictitious transition epoch if fictitious events happen at the epoch. Consider the set of actions $\{A_h, A_1, A_2, A_r\}$. Action A_h cannot change system state. If we want to take action A_1 at a fictitious transition epoch, then the state just before the transition must satisfy the condition that the processing server is idle and the head queue is not empty. But this contradicts with the fact that action A_1 is always preferable to A_h by property (11). In other words, A_1 should have been taken in last epoch before this fictitious transition epoch. It remains to consider actions A_2 and A_r. We will show the result by contradiction. Suppose the first fictitious transition epoch that changes the system state is epoch t_0. Let t_1 be the last decision epoch before t_0.

If it is optimal to take A_2 at t_0. Then the state just before t_0 must be $(q_0, 1, q_1, 0, s)$ for $q_0 > 0$, $q_0 + q_1 + 1 > m^*_s$, $s \in \{0, 1\}$. According to what kind of events happen at t_1, there are different states for the time just before t_1. We will consider these possibilities separately and show that $(q_0, 1, q_1, 0, s)$ cannot be reachable if the system starts from a proper state.

(a) Packet arrival happens at t_1. In the case, the state just before epoch t_1 should be $(q_0 - 1, 1, q_1, 0, s)$. However, since $q_0 + q_1 + 1 > m^*_s$, one packet will be delivered to the sub-6 GHz interface upon packet arrival. And the state before t_0 should be $(q_0 - 1, 1, q_1, 1, s)$.

(b) Processing server completes serving a packet. In the case, the state just before epoch t_1 should be $(q_0 + 1, 1, q_1 - 1, 0, s)$. However, since $q_0 + q_1 + 1 > m^*_s$, one packet will be delivered to the sub-6 GHz interface upon service completion. And the state before t_0 should be $(q_0 - 1, 1, q_1, 1, s)$ rather than $(q_0, 1, q_1, 0, s)$.

(c) A packet departs from the sub-6 GHz interface. In the case, the state just before epoch t_1 should be $(q_0, 1, q_1, 1, s)$. However, since $q_0 + q_1 + 1 > m^*_s$, one packet will be delivered to the sub-6 GHz interface upon the departure. And the state will be $(q_0 - 1, 1, q_1, 1, s)$ rather than $(q_0, 1, q_1, 0, s)$ before t_0.

(d) MmWave link changes states. In the case, the state just before epoch t_1 should be $(q_0, 1, q_1, 0, 1 - s)$. When the mmWave link changes states from $1 - s$ to s at t_1, it is optimal to deliver a packet to the sub-6 GHz interface since $q_0 + q_1 + 1 > m^*_s$. Then, the state will be $(q_0 - 1, 1, q_1, 1, s)$ rather than $(q_0, 1, q_1, 0, s)$ before t_0.

(e) MmWave is available $s = 1$ and a packet departs the mmWave interface. If $s = 0$, then it is possible that a packet leaves the mmWave at t_1. In the case, the state just before t_1 should be

$(q_0, 1, q_1 + 1, 0, 1)$. But since $q_0 + q_1 + 1 > m_1^*$, one packet will be delivered to the sub-6 GHz interface. Then, the state will be $(q_0 - 1, 1, q_1, 1, s)$ rather than $(q_0, 1, q_1, 0, s)$ before t_0. Thus, the system cannot reach the state $(q_0, 1, q_1, 0, s)$ for $q_0 > 0$, $q_0 + q_1 + 1 > m_s^*$, $s \in \{0, 1\}$.

If it is optimal to take A_r at t_0. Then the state just before t_0 must be $(0, l_1, q_1, 0, s)$ for $l_1 + q_1 > m_s^*$, $l_1, s \in \{0, 1\}$. If $l_1 = 1$, we can show the result with similar argument in last case. Next we consider the case $l_1 = 0$. Notice that it is not possible for a packet to arrive at t_1 since the packet cannot be delivered to the mmWave interface within one time slot. Next, consider remaining possible states separately.
(a) Processing server completes serving a packet. In the case, the state just before epoch t_1 should be $(0, 1, q_1 - 1, 0, s)$. However, since $q_1 > m_s^*$, one packet will be delivered to the sub-6 GHz interface upon service completion. And the state before t_0 should be $(0, 0, q_1 - 1, 1, s)$.
(b) A packet departs from the sub-6 GHz interface. In the case, the state just before epoch t_1 should be $(0, 0, q_1, 1, s)$. However, since $q_1 > m_s^*$, one packet will be delivered to the sub-6 GHz interface upon the departure. And the state will be $(0, 0, q_1 - 1, 1, s)$ before t_0.
(c) MmWave link changes states. In the case, the state just before epoch t_1 should be $(0, 0, q_1, 0, 1 - s)$. When the mmWave link changes states from $1 - s$ to s at t_1, it is optimal to deliver a packet to the sub-6 GHz interface since $q_1 > m_s^*$. Then, the state will be $(0, 0, q_1 - 1, 1, s)$ rather than $(0, 0, q_1, 0, s)$ before t_0.
(d) MmWave is available $s = 1$ and a packet departs the mmWave interface. In the case, the state just before t_1 should be $(0, 0, q_1 + 1, 0, 1)$. But since $q_1 > m_1^*$, one packet will be delivered to the sub-6 GHz interface. Then, the state will be $(0, 0, q_1 - 1, 1, s)$ before t_0.

Therefore, the system cannot reach the state $(0, l_1, q_1, 0, s)$ for $l_1 + q_1 > m_s^*$, $l_1, s \in \{0, 1\}$. Thus, the fictitious transition epoch that can change system state does not exist if the system starts with a proper state and the optimal policy is followed at each epoch.

Appendix F
Proof of theorem 4

For ease of exposition, we re-express the system state as (y, l_2, s), where $y \in \mathbb{N}$ denotes the queue length of FastLane. Then, Theorem 2 is expressed as follows:

$$J_\beta(y + 1, 0, s) < J_\beta(y, 1, s), \text{ if } y + 1 \le m_s^*; \quad (55)$$

$$J_\beta(y + 1, 0, s) \ge J_\beta(y, 1, s), \text{ if } y + 1 > m_s^*, \quad (56)$$

We will show the result by contradiction. Suppose that $m_0^* > m_1^*$. Then, $m_0^* = m_1^* + k$, for $k \in \mathbb{N}^+$. Assume that at the beginning of the n-th time slot, the system state is $(m_1^* + k - 1, 0, 0)$ and a packet arrives at the system now. Since the mmWave is currently unavailable, by (55) the optimal action is to keep this packet in the FastLane, and the next system state will be $(m_1^* + k, 0, 0)$. However, if a suboptimal action is taken (the packet is delivered to the sub-6 GHz interface), then the new system state will be $(m_1^* + k - 1, 1, 0)$. Let us track these two processes, say P_1 and P_2. Assume that these two processes will experience same consecutive events, i.e., same arrivals, departures from same interfaces, same change of mmWave links. For process P_1, we will take optimal actions at each decision epoch. But for process P_2, we can choose both optimal and suboptimal actions. In the case, the values (delay) obtained at each epoch in P_1 should be less or equal to that in P_2. But in fact, it does not always hold.

Now consider the next decision epoch, the beginning of the $(n + 1)$-th time slot. There are four possible events that can happen, say arrival of a packet, service completion at processing server, departure from the sub-6 GHz interface and state change of the mmWave link. Let us examine the value functions for each case.
case 1: If a packet arrives at the system, in P_1 the optimal action is to deliver the packet to the sub-6 GHz interface by (56) and the new system state is $(m_1^* + k, 1, 0)$. In P_2, since the sub-6 GHz interface is occupied, the new state will be $(m_1^* + k, 1, 0)$, which is same with that in P_2.
case 2: If the processing server completes service, in both P_1 and P_2 no action can be taken and the system state will keep unchanged.
case 3: If a packet departs from the sub-6 GHz, in P_1 the optimal action is to keep the state $(m_1^* + k, 0, 0)$ by (55). In P_2, the new system state will be $(m_1^* + k - 1, 0, 0)$, which has smaller value than that in P_2 by monotonicity.
case 4: If the mmWave changes the state, in P_1 the optimal action is to deliver a packet to the sub-6 GHz interface by (56) and the new state is $(m_1^* + k - 1, 1, 1)$. In P_2, since the sub-6 GHz interface is occupied, the new state will be $(m_1^* + k - 1, 1, 1)$, which is same with that in P_2.

Appendix G
Proof of Lemma 2

Consider $\pi \in \Pi_{opt}$. It follows from Lemma 7 that $P_\pi^{(\infty)}(;\theta_n)$ converges weakly to $P_\pi^{(\infty)}(;\theta)$.

Lemma 7. *Let π be a stationary deterministic policy that makes scheduling decisions for the network described in Section 2. Consider a sequence of systems such that the corresponding sequence $\{\theta_n\}$ of parameters converges to θ, i.e. $\theta_n \to \theta$. We then have that the sequence of stationary probability distributions associated with the system that has parameter θ_n and operates under π, converges weakly to the stationary probability associated with the system that has parameter θ and operates π. (For a detailed discussion on weak convergence see [38]).*

Proof. We denote the controlled transition probability from state $\mathbf{q}$ to state $\mathbf{q}'$ under the control u by $P(\mathbf{q}, \mathbf{q}', u; \theta)$. Note that if θ is known, the transition probability is fixed (defined in (4)). Thus, the controlled transition probabilities for the n-th system are solely a function of θ_n, and moreover the probabilities $P(\mathbf{q}, \mathbf{q}', u; \theta_n) \to P(\mathbf{q}, \mathbf{q}', u; \theta)$ pointwise. Thus, the result follows from Theorem 1 of [39]. □

Thus, the estimates of the steady state delay resulting from π under θ_n also converge to the true value. Since there are only finitely many policies in Π_{opt}, we obtain $\pi^\star(\theta_n) = \pi^\star(\theta)$ if $\|\theta_n - \theta\|$ is sufficiently small. This completes the proof.

Appendix H
Proof of Lemma 3

Fix thresholds m_0, m_1. We will derive lower bounds on the stationary probability with which each parameter is sampled.

Sampling $\mu_{\text{sub-6}}$: The sub-6 channel is used only when the total packets are more than a certain threshold. We will derive a lower bound on the probability of the event that queue length exceeds this threshold. To do this, we consider a modified system, one in which the two server dynamic scheduling system is replaced by a single server that provides service at the rate $\mu_{\text{sub-6}} + \mu_{\text{mm}}$, and the two queues are replaced by a single queue. Clearly, the probability with which the queue length of this modified system exceeds $\max\{m_0, m_1\}$, serves as a lower bound on the probability of the original system. This probability for the original system is equal to $\left(\frac{\lambda}{\mu_{\text{sub-6}}+\mu_{\text{mm}}}\right)^{\max(m_0,m_1)}$.

Sampling μ_{mm}, μ_p: Consider a modified system, in which upon arrival, a packet is sent to the sub-6 GHz channel if it is available. Otherwise, the packet is "lost" from the system. The sub-6 GHz channel does not maintain any queue. This packet loss probability serves as a lower bound on the sampling frequency of μ_{mm}, μ_p.

Sampling λ, α, γ: A lower bound clearly follows since the associated events are not controlled by the scheduler, and these events (e.g. packet arrival, channel state value) occur with non-zero probability.

Appendix I
Proof of Lemma 4

We decompose the operation into consecutive "frames", each of duration L time-steps, where the parameter L is as in Lemma 3. Let $n_\theta(k)$ be the number of samples of θ obtained during the k-th such frame. Consider the following martingale difference sequence,

$$\Delta(k) := \mathbb{E}\left(n_\theta(k)|\mathcal{F}_k\right) - n_\theta(k).$$

Note that $n_\theta(k)$ is bounded by the frame duration L. It thus follows from Azuma's inequality [40] that (note that there are $\lfloor n/L \rfloor$ frames until n),

$$\mathbb{P}\left(|N_\theta(n) - \sum_{k=1}^{\lfloor n/L \rfloor} \Delta(k)| \geq z\right) \leq \exp\left(-\frac{z^2}{2\lfloor n/L \rfloor L^2}\right).$$

Since $\sum_{k=1}^{\lfloor n/L \rfloor} \Delta(k) \geq \eta n$, we have that $\{N_\theta(n) \leq \eta_1 n\} \subseteq \{|N_\theta(n) - \eta_1 n| \geq (\eta - \eta_1)n\}$, so that letting $z = (\eta - \eta_1)n$ in the above we obtain

$$\mathbb{P}\left(N_\theta(n) \leq \eta_1 n\right) \leq \exp\left(-L(\eta - \eta_1)n\right).$$

Letting $\eta_1 = \eta/2$, and using union bound on the episode starting times $\tau_k, k > k_0$, this probability can be upper-bounded by $\sum_{k>k_0} \exp\left(-\eta L \tau_k\right) \leq K \exp\left(-\eta L \tau_{k_0}\right) \leq \frac{K}{T^2}$, where K is the number of episodes until T. This completes the proof.

Appendix J
Proof of Lemma 5

Amongst these, we will only derive a (probabilistic) upper-bound on the error associated with $[\hat{\frac{1}{\lambda}}](n)$. Rest of the proof would involve using union bound to control the errors for estimates of all the parameters. Fix the number of samples used for estimating $1/\lambda$ at j, and let τ_h be a threshold. Note that $I_A(1), I_A(2), \ldots, I_A(j)$ are successive inter-arrival times. It follows from a modified version of Azuma-Hoeffding's inequality from [41] that if the probability that atleast one out of j service times exceeds τ_h is less than δ, then we have

$$\mathbb{P}\left(\left|\sum_{\ell=1}^{j} I_A(\ell) - \frac{j}{\lambda}\right| \geq z\right) \leq \exp\left(-\frac{z^2}{j\tau_h^2}\right) + \delta. \qquad (57)$$

Since the inter-arrival times are exponentially distributed with parameter λ, when we let $\delta = j\exp\left(-\lambda\tau_h\right)$ in (57) and threshold $\tau_h = \frac{(\xi+1)\log n}{\lambda}$, we obtain that the probability with which atleast one out of $I_A(\ell), \ell = 1, 2, \ldots, j$ exceeds the value $\frac{(\xi+1)\log n}{\lambda}$, is less than $\frac{1}{n^\xi}$. We then let $z = \sqrt{j\xi \log n}$ in the inequality (57) (and also use the fact that $j \leq n$) to obtain:

$$\mathbb{P}\left(\left|\frac{\sum_{\ell=1}^{j} I_A(\ell)}{j} - \frac{1}{\lambda}\right| \geq \sqrt{\frac{\xi \log n}{j}}\right) \leq \frac{1}{n^\xi} + \frac{1}{n^\xi}.$$

Concentration result for $[\hat{\frac{1}{\lambda}}](n)$ then follows by taking union bound with respect to the number of arrivals until n (this can assume values from 1 to n). The proof for concentration of $1/\mu_{\text{p}}, 1/\mu_{\text{sub-6}}, 1/\mu_{\text{mm}}, 1/\alpha, 1/\gamma$ is similar, and hence omitted. (39) then follows by using union bound in order to jointly control deviations of all the components of θ.