

Delay-Aware Scheduling over mmWave/Sub-6 Dual Interfaces: A Reinforcement Learning Approach

Ying Cao, Bo Sun, Danny H.K. Tsang

Abstract—We consider a transmitter with mmWave/sub6 dual interfaces. Due to the intermittency of mmWave channel, the transmitter must schedule packets wisely across the interfaces to minimize the average delay by observing the system state. We use the well-known dynamic programming methods and Q-learning to find the optimal scheduling policy and investigate the influence of observing CSI on the optimal policy under different levels of knowledge of the environment. We find that only when the channel state transition model is not available, the instantaneous CSI can help in reducing system delay.

Index Terms—scheduling, millimeter wave, 5G, sub-6 GHz

I. INTRODUCTION

To enable emerging technologies such as augmented reality and connected autonomous vehicles, the fifth generation cellular wireless network will utilize the massive spectrum in millimeter wave bands (above 10 GHz), which can potentially boost the wireless capacity for eMBB services and reduce the transmission delay for low-latency applications. However, the mmWave band is inherently unstable for providing reliable connections, and thus the most promising solution is to integrate the stability of sub-6 GHz and the high capacity of mmWave networks [1]. Standardization bodies and industry partners have also recently emphasized the importance of *mmWave- μ W integrated technology* as a cost-effective solution to achieve high capacity, low latency and reliability for emerging wireless applications.

Nevertheless, the intermittency and fast-changing property of the mmWave channel still induce much difficulty in channel estimation and resource allocation. Since reliable mmWave transmission necessitates that both channel estimation and resource allocation keep up with the fast channel variations, it would be much easier if a certain statistical model exists in the channel variations. However, in most cases, we do not have a statistical model of the channel. To this end, reinforcement learning is deemed to be a viable tool, by utilizing observations from the past and making decisions based upon the gained knowledge.

A. Related Work

It is common to utilize channel knowledge to improve resource allocation in the wireless community. However, a specific probability distribution for the channel states is usually assumed. For an unknown channel model, [2] first provides an online implementation of the value iteration algorithm for optimal packet scheduling; however, this method only applies to slow-varying channels.

Recently, papers on learning a fast-changing wireless channel such as mmWave have been published. Some of these

formulate the problem into a multi-armed bandit (MAB) framework since the feedback of channels are usually bandit-formed. For example, [3] considers the optimal rate selection problem in rapidly-changing channels, where the user only has access to bandit feedback of a successful transmission over the channels. [4] studies the problem of learning channel statistics to efficiently schedule transmission in wireless networks under interference constraints. Our work differs in that we model the network as an Markov Decision Process (MDP) rather than a single state MAB problem, and thus the state transition process is more complicated than those.

Our work is inspired by [5], where the authors propose the system model that we build upon. Note that [5] added a processing server right after the head buffer. Since the delay incurred during this process only contains the delay of reading the packet from the head buffer and of writing it to the transmission buffer, and the data transfer rate of the latest variant of memory (DDR4) is around 25GB/s, which is an order of magnitude higher than the target peak download rate of 5G (i.e., 20Gb/s [6]), we assume the scheduling delay, i.e., the time from making the scheduling decision to the time the packet actually arrives at the corresponding server, is negligible. Due to this observation, we don't need a processing server after the head buffer. Moreover, we assume that the scheduler can access the mmWave server occupancy status with no delay, which is a common assumption in the literature [7] [8]. From the aforementioned research, intrinsically we ask the following questions:

- *When will channel state information (CSI) help scheduling?*
- *If the explicit channel model is not known, could the learning algorithm perform better than the queue-length-threshold policy?*

B. Our Contributions

Through extensive simulations, we find that under the full information of channel state transition kernel, the delay-optimal policy is still a threshold policy on queue length, which means the instantaneous CSI does not help. However, when the channel state transition kernel is inaccessible, the presence of CSI helps further improve the delay performance.

II. SYSTEM MODEL

We consider an integrated mmWave/sub-6 scheduler as shown in Fig. 1. The scheduler consists of one buffer and two servers, of which one is the mmWave interface and the other is the sub-6 GHz interface. Time is divided into equal-sized slots with length τ and is indexed by $t = 1, \dots, T$. We

consider the case of non-preemptive scheduling, i.e., the packet in the server cannot be interrupted during transmission, and the scheduler has access to the mmWave CSI only when it makes the scheduling decision but no knowledge of that when the packet is being transmitted. Meanwhile, the sub-6 GHz server is rather stable in terms of service time but the service rate is much slower compared to the average service rate at the mmWave interface. Confronted with two servers, one with high average service rate but highly dynamic nature and the other one with low but stable service rate, the objective of the scheduler is to wisely make the scheduling decision for the packet(s) at the beginning of the head queue at each time slot so as to minimize the average delay.

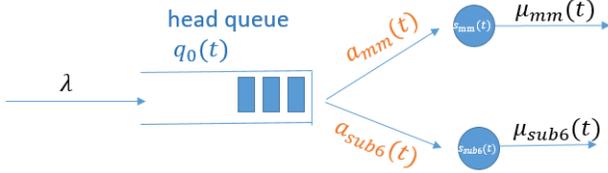


Fig. 1: System model: mmWave sub-6 GHz scheduler.

A. Two-Layer Markov mmWave Channel Model

It is well-known that the mmWave link is highly dynamic and easily blocked. The more widely-used mmWave channel model in the mmWave networking community is the 3-state Markov chain with line of sight (LoS), non-line of sight (NLoS) and outage states [9]. A more accurate finite state Markov chain (FSMC) model with 2 layers of variations is proposed in [10]. Here, we summarize the general workings of the model and show the diagram in Fig. 2.

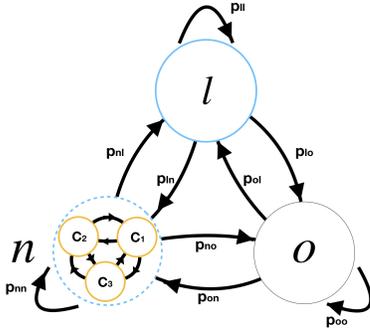


Fig. 2: Two-layer mmWave channel model

1) *Long-Term Link State Model*: The random process describing the transitions between macro-scale shadowing states is modeled as a Markov chain with states $st = \{l, n, o\}$, denoting LoS, NLoS and outage, respectively. We let $st(t)$ denote the link state at time t . The state transition kernel $\{p_{ij}\}_{i,j \in \{l,n,o\}}$ and the steady state probability $\mathbb{P}(st)$ defines the long-term link state model completely, where $p_{ij} = \mathbb{P}(st(t+1) = j | st(t) = i)$.

2) *Small-Scale Capacity Model*: The model characterizes the small-scale fading effect when the long-term link state is fixed. The capacity is calculated as

$$C = W \log\left(\frac{P_{TX} G_{mm}}{N_0 W}\right), \quad (1)$$

where G_{mm} is the squared magnitude of the mmWave channel gain. Here we do not consider the channel matrix but only focus on the magnitude of the channel gain, since in many cases, for example, admission control, the SNR is sufficient for making decision [11].

The small-scale channel capacity model is another Markov chain within state C . For NLoS state, the channel capacity is quantized into N levels, and the channel capacity in the i th level is denoted by c_i^n , $i \in \{1, \dots, N\}$. For LoS state, the channel capacity is quantized into L levels and the channel capacity in i th level is c_i^l , $i \in \{1, \dots, L\}$. For outage state, the channel capacity $c^o = 0$. Let $C(t)$ denote the link capacity at time t . For the NLoS state, the small-scale FSMC is defined by the state transition model $\{q_{ij}^n\}_{i,j \in \{1, \dots, N\}}$ and $\mathbb{P}^{(n)}(C)$, where $q_{ij}^n = \mathbb{P}(C(t+1) = c_j^n | C(t) = c_i^n)$ and $\mathbb{P}^{(n)}(c_j^n)$ is the steady state probability of having capacity c_j^n in link state n . Likewise, for the LoS state, the subscript changes from n to l . For the sake of brevity, Fig. 2 shows the case for $N = 3$, $L = 1$ and omits the self-transition loop for the small-scale capacity model due to the space limit.

Denote the combined channel state as $ch(t) = (st(t), C(t))$. The general two-layer state transition model can be derived as

$$\mathbb{P}(ch(t+1) = (j, c_k^j) | ch(t) = (i, c_m^i)) = \begin{cases} p_{ii} q_{mk}^i, & j = i \\ p_{ij} \mathbb{P}^{(j)}(c_k^j), & j \neq i. \end{cases}$$

Compared with the two-layer mmWave channel, the sub-6 GHz channel is much more stable, and thus we assume that it is a single-state channel with simple small-scale capacity level dynamics.

B. System Dynamics

1) *MDP Formulation*: Let $A(t)$ denote the random new packet arrivals at the end of the t -th scheduling slot, which is assumed to be i.i.d. Poisson distributed over scheduling slots with mean $\mathbb{E}[A] = \lambda$. $X(t)$ denotes the packet size of the packet in the front of the head queue at the beginning of the t -th time slot and is assumed to be i.i.d. exponential distributed over scheduling slots with mean \bar{X} . The departure rate is state-dependent.

- *State Space*: $\mathbf{s}(t) = (q(t), ch_{mm}(t)) \in \mathcal{S}$, where $q(t) = (q_0(t), s_{mm}(t), s_{sub6}(t))$ and $ch_{mm}(t) = (st_{mm}(t), C_{mm}(t))$. Note that we introduce the subscript for mmWave channel in order to differentiate it from the sub6-6 GHz channel. Let $q_0(t) \in \{0, 1, \dots\}$ denote the head queue length and $s_i(t) \in \{0, 1\}$, $i \in \{mm, sub6\}$ denote the occupancy of the mmWave server and the sub-6 GHz server, respectively. Given the state $\mathbf{s}(t)$, the average departure rate at the mmWave interface is given by $\mu_{mm}(\mathbf{s}(t)) = \mathbb{E}\left[\frac{R(\mathbf{s}(t))}{X(t)} | \mathbf{s}(t)\right] = \frac{R(\mathbf{s}(t))}{\bar{X}}$ and the

probability of a packet departure there at the t -th slot can be approximated by $\mu_{mm}(s(t))\tau$ ¹.

- **Action Space:** $\mathbf{a}(t) = (a_{mm}(t), a_{sub6}(t))$. The action space is state-dependent:
 - $\mathbf{q} = (0, s_{mm}, s_{sub6})$: $\mathcal{A} = \{(0, 0)\}$
 - $\mathbf{q} = (1, s_{mm}, s_{sub6})$:
 $\mathcal{A} = \{(a_{mm}, a_{sub6}) | a_{mm} + s_{mm} < 2, a_{sub6} + s_{sub6} < 2, a_{mm} + a_{sub6} < 2\}$
 - $\mathbf{q} = (q_0 \geq 2, s_{mm}, s_{sub6})$:
 $\mathcal{A} = \{(a_{mm}, a_{sub6}) | a_{mm} + s_{mm} < 2, a_{sub6} + s_{sub6} < 2\}$

A decision rule, $\pi : \mathcal{S} \rightarrow \mathcal{A}(s)$, is a function mapping from the state space \mathcal{S} to the action space $\mathcal{A}(s)$.

- **Transition Kernel:** $\mathbb{P}(s(t+1)|s(t), \mathbf{a}(t))$. The generic expression of this kernel is:

$$\begin{cases} \lambda\tau\mathbb{P}(ch_{mm}(t+1)|ch_{mm}(t)), C_1 \\ \mu_{mm}\tau\mathbb{P}(ch_{mm}(t+1)|ch_{mm}(t)), C_2 \\ \mu_{sub6}\tau\mathbb{P}(ch_{mm}(t+1)|ch_{mm}(t)), C_3 \\ 1 - (\lambda + \mu_{mm} + \mu_{sub6})\tau\mathbb{P}(ch_{mm}(t+1)|ch_{mm}(t)), C_4, \end{cases}$$

where $\mu_{mm} = \mu_{mm}(s(t))$ and

$$\begin{aligned} C_1 &= \begin{cases} q_0(t+1) = q_0(t) - a_{mm}(t) - a_{sub6}(t) + 1 \\ s_1(t+1) = \min(s_1(t) + a_{mm}(t), 1) \\ s_2(t+1) = \min(s_2(t) + a_{sub6}(t), 1), \end{cases} \\ C_2 &= \begin{cases} q_0(t+1) = q_0(t) - a_{mm}(t) - a_{sub6}(t) \\ s_1(t+1) = \max(\min(s_1(t) + a_{mm}(t) - 1, 1), 0) \\ s_2(t+1) = \min(s_2(t) + a_{sub6}(t), 1), \end{cases} \\ C_3 &= \begin{cases} q_0(t+1) = q_0(t) - a_{mm}(t) - a_{sub6}(t) \\ s_1(t+1) = \min(s_1(t) + a_{mm}(t), 1) \\ s_2(t+1) = \max(\min(s_2(t) + a_{sub6}(t) - 1, 1), 0), \end{cases} \\ C_4 &= \begin{cases} q_0(t+1) = q_0(t) - a_{mm}(t) - a_{sub6}(t) \\ s_1(t+1) = \min(s_1(t) + a_{mm}(t), 1) \\ s_2(t+1) = \min(s_2(t) + a_{sub6}(t) - 1, 1). \end{cases} \end{aligned}$$

Note that C_1 , C_2 and C_3 are the queue dynamics for different events, which correspond to the arrival, the departure from the mmWave server and the departure from the sub6 GHz server, respectively. C_4 means nothing happens.

2) *Two levels of CSI knowledge:* The queue state information (QSI) can be accessed in real time, which is a common assumption in the literature [14] [15]. Regarding the CSI, although it is known that the complete instantaneous mmWave CSI $ch_{mm}(t)$ is usually unavailable, knowledge of the channel transition kernel and the large-scale CSI $st_{mm}(t)$ can be acquired easily in some cases, where dynamic programming methods can be used to obtain the optimal policy. For example, when the environment is rather static, the channel transition matrix can be estimated offline; Advanced channel tracking technologies [13] can be utilized to estimate $st_{mm}(t)$. In other

cases, when we do not know the channel transition matrix in advance, traditional dynamic programming methods will fail and learning methods will surge. Based on this divergence, we divide the possible channel state information into 2 levels.

- **Known $\mathbb{P}\{ch_{mm}(t+1)|ch_{mm}(t)\}$, known $ch_{mm}(t)$:** Full information of the stationary channel state transition model is known. Additionally, we assume the scheduler can observe the complete instantaneous CSI, in order to investigate how CSI can help scheduling.
- **Unknown $\mathbb{P}\{ch_{mm}(t+1)|ch_{mm}(t)\}$, known $st_{mm}(t)$:** The scheduler has no information about the channel state transition kernel except the instantaneous large-scale CSI.

C. Problem Formulation

We are concerned with minimizing the average delay incurred in the system. Based on Little's law, the average delay experienced by one packet is proportional to the average number of packets in the system, thus the problem is

$$\min_{\pi} \limsup_{T \rightarrow \infty} \frac{1}{T} \mathbb{E}^{\pi} \sum_{t=1}^T q_t,$$

where $q_t = q_0(t) + s_{mm}(t) + s_{sub6}(t)$.

III. TECHNIQUES

A. Dynamic Programming

For the known channel transition matrix case, the optimal policy can be calculated by exact dynamic programming methods. Both relative value iteration (RVI) and linear programming (LP) are implemented in this work. The key update step in RVI for infinite horizon average cost problem is:

$$h_{i+1}(s) = (Th_i)(s) - (Th_i)(s_0),$$

where $(Th_i)(s) = \min_{\mathbf{a} \in \mathcal{A}(s)} [C(s) + \sum_{s' \in \mathcal{S}} P(s'|s, \mathbf{a})h_i(s')]$ and s_0 is a fixed state. $C(s)$ is the cost function, which is set to $q_0 + s_{mm} + s_{sub6}$ in our problem. As $i \rightarrow \infty$, $h_i(s)$ will converge to the differential cost $h^*(s)$ for state s w.r.t. the optimal average cost for all states $(Th^*)(s_0)$. The optimal policy $\pi(s) = \operatorname{argmin}_{\mathbf{a} \in \mathcal{A}(s)} [C(s) + \sum_{s' \in \mathcal{S}} P(s'|s, \mathbf{a})h_i(s')]$.

For LP, we solve the optimization problem as follows:

$$\begin{aligned} \min_{q(s, \mathbf{a})} & \sum_{s \in \mathcal{S}} \sum_{\mathbf{a} \in \mathcal{A}(s)} q(s, \mathbf{a})C(s) \\ \text{s.t.} & \sum_{\mathbf{a} \in \mathcal{A}(s)} q(s, \mathbf{a}) = \sum_{s' \in \mathcal{S}} \sum_{\mathbf{a} \in \mathcal{A}(s')} q(s', \mathbf{a})P(s|s', \mathbf{a}), \forall s \\ & \sum_{\mathbf{a} \in \mathcal{A}(s)} q(s, \mathbf{a}) = 1, \forall s \\ & q(s, \mathbf{a}) \geq 0, \forall s, \mathbf{a}, \end{aligned}$$

where $q^*(s, \mathbf{a})$ denotes the state-action probability under the optimal policy. It is direct to see that the policy output by linear programming can be randomized due to the probability distribution, while value iteration is guaranteed to output a deterministic optimal policy since the optimal value function is unique.

¹ To show this, we need another assumption: $\mu_{mm}(s(t))\tau \ll 1$, the detailed proof is given in [12].

B. Q-learning

Q-learning is a classic model-free reinforcement learning (RL) method proposed originally in [16]. The essence is to approximate the optimal action value function using learned action value function from sampled rewards:

$$Q_{t+1}(\mathbf{s}_t, \mathbf{a}_t) = (1 - \alpha)Q_t(\mathbf{s}_t, \mathbf{a}_t) + \alpha(R(\mathbf{s}_t) + \gamma \max_{\mathbf{a} \in \mathcal{A}(\mathbf{s}_{t+1})} Q_t(\mathbf{s}_{t+1}, \mathbf{a})),$$

where $\alpha \in (0, 1)$ is the learning rate and $\gamma \in (0, 1)$ is the discounted ratio. It can be proven that when $\gamma \rightarrow 1$, the system converges to the average reward case. $R(\mathbf{s})$ is the reward observed in state \mathbf{s} , which is set to $\frac{1}{q_0 + s_{mm} + s_{sub6}}$. In this work, since we have an explicit expression on the queueing dynamics, the queue size will not increase significantly as long as the system satisfies the queue stability condition, the state space can be handled only by tabular Q-learning. However, note that if considering a wireless network, the state space may be too large for tabular methods, and thus one can seek the help of function approximation. This is beyond the scope of this paper.

IV. SIMULATION: SETTINGS AND RESULTS

Since the techniques we use are different under the two information levels, we divide the simulation section into two parts.

A. Known $\mathbb{P}(ch_{mm}(t+1)|ch_{mm}(t))$, known $ch_{mm}(t)$

Under this channel information level, we implemented the value iteration and linear programming methods to find the optimal policy. The case without considering channel state information is studied in [5]. Since our objective is to investigate the effect of CSI on the scheduling policy, for comparison, we use a similar system model as in [5], which adds a buffer before the mmWave server. Nevertheless, it is further verified by the output optimal policy that, if the instantaneous CSI is known and there is no scheduling delay, the buffer before the mmWave server is not needed.

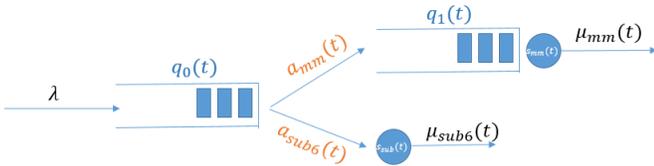


Fig. 3: System model: modified mmWave-sub 6GHz scheduler.

1) Simulation Parameters:

- **State Space:** $q_0 \in \{0, 1, \dots, 6\}$, $q_1 \in \{0, 1, \dots, 5\}$, $s_{mm}, s_{sub6} \in \{0, 1\}$, $st_{mm} \in \{l, n, o\}$. We follow the parameters shown in Fig. 2 and $C_{mm} \in \{c^l, c_1^n, c_2^n, c_3^n, c^o\}$. Here we normalize the maximum channel capacity to 1 and set $(c^l, c_1^n, c_2^n, c_3^n, c^o) = (1, 0.05, 0.004, 0.002, 0)$. The state space size is 840. We simulate the mmWave channel state transition kernel based on the sub-6 GHz channel. Assume the sub-6 GHz channel has two states

denoted by $st_{sub6} \in \{\text{bad}, \text{good}\}$. The detailed parameters of the mmWave/sub6 CSI model are as follows:

- sub-6: $\mathbb{P}(st_{sub6} = \text{bad}) = 0.2$, $\mathbb{P}(st_{sub6} = \text{good}) = 0.8$.
- mmWave: see Table I.

TABLE I: Conditional distribution of C_{mm}

$\mathbb{P}(C_{mm} st_{sub6})$	st_{sub6}	bad	good
C_{mm}			
c^l		0.1	0.7
c_1^n		0.15	0.15
c_2^n		0.15	0.05
c_3^n		0.15	0.05
c^o		0.45	0.05

- **Action Space:** $a_{mm} \in \{0, 1\}$, $a_{sub6} \in \{0, 1\}$, and the action space size is 4.
- As shown in Table II, the small-scale mmWave channel gain follows Gaussian processes with different means and variances in different channel states.

TABLE II: Small-scale mmWave channel gain model

LoS	NLoS(<i>i</i> th level)	outage
$N(c^l, 0.001)$	$N(c_i^n, 0.1)$	0

- The arrival rate λ is normalized to 1 pkt/s, and the departure rate is shown in Tables III and IV. We can see that under undesirable mmWave channel conditions, the departure rate is slower than the sub-6 GHz interface given the same power. The mean packet size \bar{X} is set to 500 kbits. In order to avoid introducing irrelevant variables, the transmission power at both interfaces, denoted by P_{mm} and P_{sub6} are set to 30 W.

TABLE III: Departure rate for sub-6 GHz interface

st_{sub6}	$\mu_{sub6}(\text{pkts/s})$
bad	0.99
good	1.45

TABLE IV: Departure rate for mmWave interface

C_{mm}	$\mu_{mm}(\text{pkts/s})$
c^l	49.54
c_1^n	13.22
c_2^n	1.64
c_3^n	0.84
c^o	0

Although the values are not from real data, the construction is based on verified facts:

- 1) When the mmWave is in the LoS state, the correlation between mmWave and sub-6 GHz is strong [17].
- 2) The mmWave channel in the LoS state is dozens of times the data rate of sub-6 GHz.

2) Simulation Results:

- 1) State Space Reduction: Using LP, we can calculate the occurrence probability for all state-action pairs, as shown in Fig. 4. After projecting it to the state space, we can find the recurrent states, which are the states of the state-action pairs with non-zero probability.

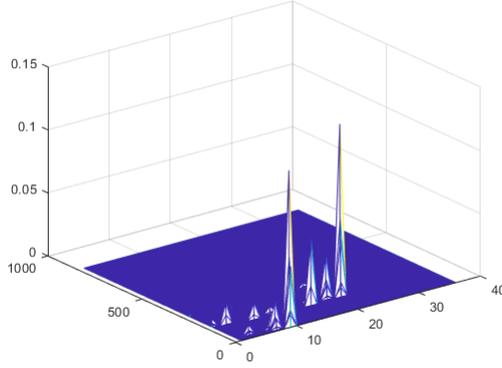


Fig. 4: Occurrence probability of state-action pairs

After carefully examining the recurrent states, it is found that the states with $q_1 > 1$ never appear, which confirms that the buffer before the mmWave server is not needed. To be specific, in terms of queue state (q_0, q_1, s) , the recurrent states are $(n, 0, 0)$, $(n, 0, 1)$, $(n, 1, 0)$, $n \in \{0, 1, \dots, 5\}$. The reason why states with $q_0 = 6$ do not appear is that the arrival rate is limited. To show this, we increase the buffer size, and the recurrent states do not change.

2) Optimal Policy Structure:

Table V shows the optimal policy. Note that when different states correspond to the same optimal action, we use "x" to represent all the possible and unspecified state components for that action. From Table V, we can see that the scheduling decision depends on the departure rate of each interface. When $(q_0, q_1, s) = (1, 0, 0)$, the policy schedules packets to the mmWave interface when the mmWave channel offers a higher rate than sub-6 GHz. When $(q_0, q_1, s) = (n, 0, 0)$, $n \geq 2$, if the departure rates at the mmWave and sub-6 GHz are comparable, the policy schedules packets to both interfaces, but when the mmWave can offer a much higher departure rate, sub-6 GHz is put aside.

When $(q_0, q_1, s) = (n, 1, 0)$, $n \geq 1$, the policy operates in a similar way as previously: it will send packets to sub-6 GHz only when the rate of the mmWave interface degenerates to the same level as that of sub-6 GHz.

Interestingly, we find that even if we have knowledge of the instantaneous CSI, the optimal policy is still the threshold type w.r.t. the queue length. In other words, the policy is not improved by the observed CSI.

TABLE V: Optimal policy

Queue State (q_0, s_{mm}, s_{sub6})	CSI (C_{mm}, st_{sub6})	Action (a_{mm}, a_{sub6})
$(0, x, x)$	(x, x)	$(0, 0)$
$(1, 0, 0)$	(c^o, x)	$(0, 1)$
	(c_3^n, bad)	$(1, 0)$
	(c_3^n, good)	$(0, 1)$
	(c_2^n, x)	$(1, 0)$
	(c_1^n, x)	
(c^l, x)		
$(n, 0, 0)$ $2 \leq n \leq 6$	(c^o, x)	$(0, 1)$
	(c_3^n, x)	$(1, 1)$
	(c_2^n, x)	$(1, 0)$
	(c_1^n, x)	
	(c^l, x)	
$(n, 0, 0)$ $n \geq 7$	(c^o, x)	$(0, 1)$
	(c_3^n, x)	$(1, 1)$
	(c_2^n, x)	
	(c_1^n, x)	
	(c^l, x)	
$(n, 1, 0)$ $1 \leq n \leq 5$	(c^o, x)	$(0, 1)$
	(c_3^n, x)	
	(c_2^n, x)	$(0, 0)$
	(c_1^n, x)	
	(c^l, x)	
$(n, 1, 0)$ $n \geq 6$	(c^o, x)	$(0, 1)$
	(c_3^n, x)	
	(c_2^n, x)	$(0, 1)$
	(c_1^n, x)	
	(c^l, x)	
$(n, 0, 1)$ $n \geq 1$	(c^o, x)	$(0, 0)$
	(x, x)	$(1, 0)$
$(n, 1, 1), n \geq 1$	(x, x)	$(0, 0)$

B. Unknown $\mathbb{P}(ch_{mm}(t+1)|ch_{mm}(t))$, known $st_{mm}(t)$

Since we do not have the exact channel transition model in this case, the value iteration and linear programming methods cannot apply. It is known that model-free methods in RL are suitable for tackling problems without knowledge of the model. In this part, we implemented Q-learning to learn the unknown channel transition kernel under the system model in 1 and compare the performance with the queue-length-threshold policy.

1) Simulation Parameters:

- State Space: $q_0 \in \{0, 1, \dots, 10\}$, $s_{mm}, s_{sub6} \in \{0, 1\}$, $st_{mm} \in \{l, n, o\}$.
- Action Space: $a_{mm} = \{0, 1\}$, $a_{sub6} = \{0, 1\}$.
- To simulate the mmWave channel, we use the transition model in [18], which characterizes an urban scenario where the dominant link is NLoS:

$$\mathbb{P}(st_{mm}(t+1)|st_{mm}(t)) = \begin{bmatrix} 0.55 & 0.3 & 0.15 \\ 0.01 & 0.8 & 0.19 \\ 0.38 & 0.40 & 0.22 \end{bmatrix}$$

The average sojourn time in seconds for each state obeys $[t_l : t_n : t_o] = [5 : 25 : 3]$. The arrival rate $\lambda = 60$ pkts/s. The mean packet size \bar{X} is set to 500 kbits. The parameters for the small-scale channel model are the same as IV-A.

2) Simulation Results: Under different channel models, the average time needed to transmit a

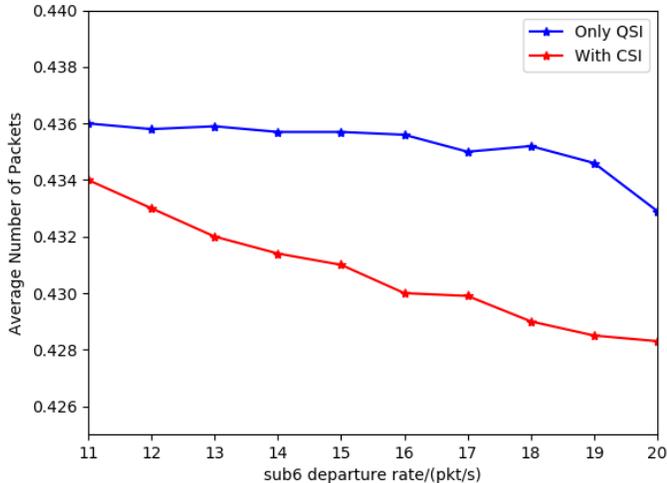


Fig. 5: Performance with different sub-6 GHz rate

packet at the mmWave interface is in Table VI.

TABLE VI: Average packet departure time at mmWave

st_{mm}	l	n	o	Average
$t(ms)$	3.2	20.4	32	19.79

In Fig. 5, we compare the performance of the queue-length-threshold policy that uses queue state information (QSI) only and the converged policy output by Q-learning that uses both CSI and QSI. The reason that the sub-6 GHz departure rate is set in $[11, 20]$ pkts/s is to ensure that the system is still in the stability region. It could be seen that the inclusion of CSI indeed helps reduce the average delay, which confirms our intuition. Compared with the results in IV-A, where the optimal policy is still queue length based, the impact of instantaneous CSI information on the scheduling policy is different. The reason could be that, the channel statistical model is higher-order information compared to the instantaneous CSI. However, since this high-order information is usually not available in real life, we show that under this case, it is of benefit to use instantaneous CSI. Due to the slow convergence of the Q-learning method, however, the algorithm may only apply to the static environment, where the channel state transition model is stationary. Thus, a possible extension of our work is to consider a more adaptive algorithm to learn the environment faster.

V. CONCLUSION

In the forthcoming 5G era, there are emerging applications with more stringent delay and reliability requirements, and mmWave- μ W integrated technology is considered as a promising solution. This paper considers the dual-interface scheduling problem in a mmWave/sub6 integrated transmitter and investigates the role of different levels of CSI in the performance of the policy. It is found that the instantaneous CSI is only helpful to the policy when the statistical knowledge of the channel is not available, where the inclusion of CSI

indeed further reduces the average delay based on the delay-optimal policy with only QSI. Hopefully, the findings can help the system designers decide when it is necessary to take the instantaneous CSI into account for resource allocation.

REFERENCES

- [1] O. Semiari, W. Saad, M. Bennis, and M. 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*, vol. 26, no. 2, pp. 109115, Apr. 2019.
- [2] N. Salodkar, A. Bhorkar, A. Karandikar, and V. S. Borkar, An on-line learning algorithm for energy efficient delay constrained scheduling over a fading channel, *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 4, pp. 732742, May 2008.
- [3] H. Gupta, A. Eryilmaz, and R. Srikant, Low-Complexity, Low-Regret Link Rate Selection in Rapidly-Varying Wireless Channels, in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, 2018, pp. 540548.
- [4] T. Stahlbuhk, B. Shrader, and E. Modiano, Learning algorithms for scheduling in wireless networks with unknown channel statistics, *Ad Hoc Networks*, vol. 85, pp. 131144, Mar. 2019.
- [5] G. Yao, M. Hashemi, and N. B. Shroff, Integrating Sub-6 GHz and Millimeter Wave to Combat Blockage: Delay-Optimal Scheduling, arXiv:1901.00963 [cs, math], Jan. 2019.
- [6] *IMT Vision - "Framework and overall objectives of the future development of IMT for 2020 and beyond"*, Rec. ITU-R M.2083-0, International Telecommunications Union, Geneva, Switzerland, Sep. 2015.
- [7] A. Gopalan, C. Caramanis, and S. Shakkottai, On Wireless Scheduling With Partial Channel-State Information, *IEEE Trans. Inform. Theory*, vol. 58, no. 1, pp. 403420, Jan. 2012.
- [8] J. Liu, A. Eryilmaz, N. B. Shroff, and E. S. Bentley, Understanding the Impacts of Limited Channel State Information on Massive MIMO Cellular Network Optimization, *IEEE J. Select. Areas Commun.*, vol. 35, no. 8, pp. 17151727, Aug. 2017.
- [9] M. R. Akdeniz et al., Millimeter Wave Channel Modeling and Cellular Capacity Evaluation, *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 11641179, Jun. 2014.
- [10] R. Ford, S. Rangan, E. Mellios, D. Kong, and A. Nix, Markov Channel-Based Performance Analysis for Millimeter Wave Mobile Networks, in *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, 2017, pp. 16.
- [11] M. Li and Y. Wu, Performance Analysis of Adaptive Multicast Streaming Services in Wireless Cellular Networks, *IEEE Transactions on Mobile Computing*, pp. 11, 2018.
- [12] V. K. N. Lau and Y. Cui, Delay-optimal power and subcarrier allocation for OFDMA systems via stochastic approximation, *IEEE Transactions on Wireless Communications*, vol. 9, no. 1, pp. 227233, Jan. 2010.
- [13] X. Ma, H. Ye, and Y. Li, Learning Assisted Estimation for Time-Varying Channels, in *2018 15th International Symposium on Wireless Communication Systems (ISWCS)*, 2018, pp. 15.
- [14] Y. Cui, Q. Huang, and V. K. N. Lau, Queue-Aware Dynamic Clustering and Power Allocation for Network MIMO Systems via Distributed Stochastic Learning, *IEEE Transactions on Signal Processing*, vol. 59, no. 3, pp. 12291238, Mar. 2011.
- [15] S. Kittipiyakul and T. Javidi, Resource Allocation in OFDMA with Time-Varying Channel and Bursty Arrivals, *IEEE Communications Letters*, vol. 11, no. 9, pp. 708710, Sep. 2007.
- [16] C. J. C. H. Watkins and P. Dayan, Q-learning, *Mach Learn*, vol. 8, no. 3, pp. 279292, May 1992.
- [17] A. Ali, N. Gonzalez-Prelcic, and R. W. Heath, Estimating millimeter wave channels using out-of-band measurements, in *2016 Information Theory and Applications Workshop (ITA)*, 2016, pp. 16.
- [18] M. Mezzavilla, S. Goyal, S. Panwar, S. Rangan, and M. Zorzi, An MDP model for optimal handover decisions in mmWave cellular networks, in *2016 European Conference on Networks and Communications (EuCNC)*, 2016, pp. 100105.

$$\beta_j = \quad (2)$$