Achieving Delay Rate-function Optimality in OFDM Downlink with Time-correlated Channels

Zhenzhi Qian*, Bo Ji†, Kannan Srinivasan*, Ness B. Shroff*‡
*Department of Computer Science and Engineering, The Ohio State University, Columbus 43210, OH
†Department of Computer and Information Sciences, Temple University, Philadelphia 19122, PA
‡Department of Electrical and Computer Engineering, The Ohio State University, Columbus 43210, OH

***Abstract*—There have been recent attempts to develop scheduling schemes for downlink transmission in a single cell of a multi-channel (e.g., OFDM-based) cellular network. These works have been quite promising in that they have developed low-complexity index scheduling policies that are delay-optimal (in a large deviation rate-function sense). However, these policies require that the channel is ON or OFF in each time-slot with a fixed probability (i.e., there is no memory in the system), while the reality is that due to channel fading and doppler shift, channels are often time-correlated in these cellular systems. Thus, an important open question is whether one can find simple index scheduling policies that are delay-optimal even when the channels are time-correlated. In this paper, we attempt to answer this question for time-correlated ON/OFF channels. In particular, we show that the class of *oldest packets first* (OPF) policies that give a higher priority to packets with a large delay is delay rate-function optimal under two conditions: 1) The channel is *non-negatively correlated*, and 2) The distribution of the OFF period is *geometric*. We use simulations to further elucidate the theoretical results.**

I. Introduction

Orthogonal frequency division multiplexing (OFDM) is a digital multi-carrier modulation method that has been widely used in wideband digital communications. A practical and important application is the downlink phase of a single cell of OFDM-based cellular networks, where the wideband can be divided into a large number of orthogonal sub-carriers, which can be used to carry data for different users. In this system, the Base Station (BS) maintains a separate queue to store data packets requested by each user. When the sub-carrier seen by a user is in good channel condition, the sub-carrier can successfully transmit a packet to the user from its designated queue. We will focus on the setting of a single-hop multi-user multi-channel system and study the delay performance of this system from a large-deviations perspective.

In wireless networks, a key problem that has been extensively studied is the design of high-performance scheduling policies. It is well known from the seminal work [1] that the MaxWeight policy is throughput-optimal, in the sense that it can stabilize the system under any feasible arrival rates. However, it has been shown in [2] that the MaxWeight policy sacrifices the delay performance (and may lead to very large queue lengths) for better throughput. This fact has motivated researchers to look for policies that can improve the delay performance measured by a queue-length-based metric. In [3], the authors showed that the maximum-throughput and load-balancing (MTLB) policy can achieve delay optimality for two special cases of ON/OFF channels with a two-user system or a system that allows fractional server allocation. However, this problem becomes much harder in general cases. On the other hand, in cellular networks, minimizing average delay may cause a large delay for certain users that have stringent delay requirements.

Another line of works focus on designing scheduling policies that maximize the rate-function of the steady-state probability that the largest queue length exceeds a given threshold when the number of channels and users both go to infinity. In [4] and [5] the authors showed that their proposed policy can achieve both throughput optimality and queue length rate-function optimality. However, simulations in [6] - [8] show that good queue length performance does not necessarily imply good delay performance. In fact, queue-length-based policies usually suffer from the so called "last packet" problem, which occurs in the situation where a certain queue has a very small number of packets. Hence, this queue is rarely scheduled by the queue-length-based policies, resulting in large packet delays.

To that end, a delay-based metric has been investigated in recent works in [7], [10] and [11]. The authors developed several policies that achieve both throughput optimality and delay rate-function optimality (or near-optimality). Although the results hold for general arrivals (e.g., time-correlated arrivals are allowed), the channels are assumed to be *i.i.d.* over time. In practice, the current channel condition could depend on past channel conditions. Therefore, the following important question remains: *How do we design a low-complexity scheduling policy that achieves provably good throughput and delay performance in the OFDM downlink system with time-correlated channels?*

While it is relatively straightforward to develop throughput optimal policies even for time-correlated channels, developing policies that are delay-optimal or delay-efficient for time-correlated channels remains an open problem.

To that end, we are motivated to consider the following question: *Can we find index scheduling policies that are delay-optimal even when the channels are time-correlated?* In this paper, we provide a positive answer in some cases. Specifically, we analyze the delay rate-function of the class of *oldest packets first* (OPF) policies which give a higher priority to packets with a large delay and present two conditions under which delay rate-function optimality can be achieved by any

OPF policy.

The key contributions of this paper are summarized as follows. We use an alternating renewal process to model a general ON/OFF time-correlated channel. We first prove an upper bound on the delay rate-function for any scheduling policy. Then, we analyze the delay rate-function of the class of OPF policies, which give a higher priority to older packets. We present two conditions and show that if both conditions are satisfied, delay rate-function optimality can be achieved by any OPF policy. The first condition requires that the channel condition is *non-negatively correlated* over time. This is often observed in practical time-correlated channels. The second condition requires that the "OFF" period distribution has the memoryless property, whereas the "ON" period distribution could be arbitrary.

The rest of the paper is organized as follows. In Section II, we describe the system model and the performance metric. In Section III, we derive an upper bound on the rate-function for any possible policy, and in Section IV, we obtain an achievable rate-function of the class of OPF policies. Then in Section V, we propose two conditions that imply delay rate-function optimality of the class of OPF policies. We conduct simulations to validate our theoretical results in Section VI and make concluding remarks in Section VII.

II. System Model

We use a time-slotted multi-queue multi-server system to model the downlink phase of a single cell OFDM system. In particular, we assume that there are n servers which stand for frequency sub-carriers. Furthermore, we assume the number of users is equal to the number of channels for ease of presentation [10]. The Base Station maintains a queue/buffer to store packets requested by each user, hence there are also n queues in the queueing system. (We use terms "server" and "channel", "queue" and "user" interchangeably throughout this paper.) Next, we present several notations that will be used later in this paper. We use Q_i to denote the queue associated to the i-th user, and use S_j to denote the j-th server for $1 \leq i, j \leq n$. We use $Q_i(t)$ to denote the queue length of queue Q_i at the beginning of time-slot t immediately after new packet arrivals. All queues are assumed to have infinite buffer size. Further, we use $W_i(t)$ to denote the head-of-line (HOL) delay of queue Q_i at the beginning of time-slot t and use $W(t) = \max_{1 \leq i \leq n} W_i(t)$ to denote the largest packet delay in the system at the beginning of time-slot t. Finally, we use $\mathbb{1}_A$ to denote the indicator function that indicates whether event A occurs or not.

A. Arrival Process

The arrival process to each queue is assumed to be stationary and ergodic. We also assume the arrivals are *i.i.d.* across all users, but could be correlated over time. Let $A_i(t)$ denote the number of packet arrivals to queue Q_i in time-slot t. Let $A(t) = \sum_{i=1}^{n} A_i(t)$ denote the total packet arrivals coming into the system in time-slot t, and let $A(t_1, t_2) = \sum_{\tau=t_1}^{t_2} A(\tau)$ denote the cumulative packet arrivals to the system from time-slot t_1 to time-slot t_2.

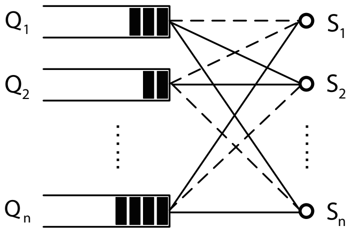

Fig. 1. A multi-queue multi-server system with stochastic connectivity. The connectivity between queue Q_i and server S_j is "ON" if they are connected by a solid line, and "OFF" otherwise (connected by a dashed line).

Next, we will introduce several assumptions on the arrival process for purpose of rate-function delay analysis.

Assumption 1: The number of arrivals are bounded, i.e., there exists a finite number L such that $A_i(t) \leq L$ for any i and t. Also, we assume $\mathbb{P}(A(s, s+t-1) = Lnt) > 0$ for any s, t and n.

Assumption 2: The arrival process are i.i.d across all users, and the mean arrival rate is p (we assume $p < 1$, otherwise the system could not be stable under any scheduling policy) for every user. Given any $\epsilon > 0$ and $\delta > 0$, there exists a positive function $I_B(\epsilon, \delta)$ independent of n and t such that

$$\mathbb{P}\Big(\frac{\sum_{\tau=1}^{t} \mathbb{1}_{\{|A(\tau)-pn|>\epsilon n\}}}{t} > \delta\Big) < \exp(-ntI_B(\epsilon, \delta)). \tag{1}$$

for all $t \geq T_B(\epsilon, \delta)$ and $n \geq N_B(\epsilon, \delta)$.

Assumptions 1 and 2 are mild. Packet arrivals per time-slot are typically bounded in practice. In addition, it has been shown in [7] that Assumption 2 is a general result of the statistical multiplexing effect of a large number of sources and holds for both *i.i.d.* arrivals and Markov chain driven arrivals.

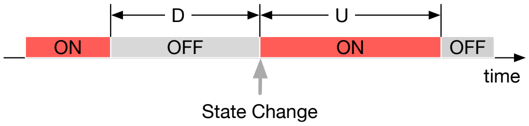

Fig. 2. Time-correlated channel model

B. Stochastic Connectivity

We assume that each channel has unit capacity and changes between "ON" state and "OFF" state from time to time. We use $C_{i,j}(t)$ to indicate the connectivity between queue Q_i and server S_j in time-slot t: $C_{i,j}(t) = 1$ when the channel is "ON" and $C_{i,j}(t) = 0$ when the channel is "OFF." We define "ON" period to be the number of time-slots between the last time the channel was "OFF" until the next time-slot it becomes "OFF" again. "OFF" period is defined in the similar way. From time to time, the channel state alternates between "ON" periods and "OFF" periods. We use an alternating renewal process to model the stochastic connectivity. In other words, the channel is initially "ON" for a time period U_1 and then "OFF" for a time period D_1, followed by another "ON" period U_2 and so on. In particular, the sequences of "ON" times $\{U_n : n \geq 1\}$ and "OFF" times $\{D_n : n \geq 1\}$ are independent sequences of *i.i.d.* positive random variables. Let U be a generic "ON" time and D be a generic "OFF" time. We use $F_U(\cdot)$ and $F_D(\cdot)$ to denote the CDF of random variable U and D, respectively.

Assumption 3: The sum of "ON" and "OFF" periods $U+D$ is aperiodic with $\mathbb{E}[U] < \infty$ and $\mathbb{E}[D] < \infty$.

It is well known (e.g. [9]) that under Assumption 3, we have:

$$\lim_{t\to\infty} \mathbb{P}(C_{i,j}(t) = 0) = \frac{\mathbb{E}[D]}{\mathbb{E}[U] + \mathbb{E}[D]} = \pi_0. \tag{2}$$

for any i, j.

Remark 1: If $U + D$ is periodic with period d, the above result is true if t is an integral multiple of d. For simplicity, we only focus on the aperiodic case.

Note that this is a general model that can capture the time-correlation of a channel. If U and D are *geometrically* distributed with parameters $1 - q$ and q, respectively, it degenerates to a static *i.i.d.* channel model with channel "ON" probability q. Similarly, if U and D have a *geometric* distribution with parameter p_{10} and p_{01}, respectively, it becomes the Markovian channel model with transition matrix $T = \begin{bmatrix} 1 - p_{01} & p_{01} \\ p_{10} & 1 - p_{10} \end{bmatrix}$.

In each time-slot, a scheduling policy allocates servers to serve packets from user queues. We further assume that a server can only serve one queue in a time-slot, however, a queue can get service from multiple servers simultaneously in one time-slot. In addition, one packet from queue Q_i can be served if an "ON" channel is allocated to queue Q_i.

C. Problem Formulation

In this paper, the metric we use to measure the delay performance is the large deviation rate-function of the steady-state probability that the largest packet delay exceeds a given threshold b. Assume the system starts at minus infinity, then $W(0)$ is the largest packet delay over all the queues in the steady-state. We define the rate-function $I(b)$ as the asymptotic decay-rate of the probability that $W(0) > b$ for a given threshold b:

$$I(b) \triangleq \lim_{n\to\infty} \frac{-1}{n} \log \mathbb{P}(W(0) > b). \tag{3}$$

Note that by the definition of rate-function $I(b)$, we can estimate the order of delay violation probability (i.e., $\mathbb{P}(W(0) > b)$ by $\exp(-nI(b))$. It is obvious that a larger rate-function implies a smaller delay violation probability and a better delay performance. In this paper, our objective is to maximize the rate-function $I(b)$. [1]

III. An Upper Bound On The rate-function

In this section, we derive an upper bound on the best achievable delay rate-function. Later, we will use this upper bound as a baseline to evaluate the delay performance of the OPF policies.

First, as in [6], [7], we define quantity $I_A(t, x)$ for any integer $t > 0$ and any real number $x \geq 0$:

$$I_A(t, x) \triangleq \sup_{\theta>0} [\theta(t + x) - \lambda_{A_i(-t+1,0)}(\theta)]. \tag{4}$$

where $\lambda_{A_i(-t+1,0)}(\theta) = \log \mathbb{E}[e^{\theta A_i(-t+1,0)}]$ is the cumulant-generating function of $A_i(-t + 1, 0) = \sum_{\tau=-t+1}^{0} A_i(\tau)$.

From Cramer's Theorem, $I_A(t, x)$ is equal to the asymptotic decay-rate of the probability that in any interval of t time-slots, the total number of packet arrivals to the system is no smaller than $n(t + x)$ as n tends to infinity, i.e.,

$$\lim_{n\to\infty} \frac{-1}{n} \log \mathbb{P}(A(-t + 1, 0) \geq n(t + x)) = I_A(t, x). \tag{5}$$

We define t_x for $L > 1$ and non-negative integer x:

$$t_x \triangleq \frac{x}{L - 1}. \tag{6}$$

Then we define an integer set $\Psi_b \triangleq \{c \in \{1, 2, \cdots, b\} | t_{b-c} \in \mathbb{Z}^+\}$. For any integer $b \geq 0$, let

$$\begin{aligned} I_U^*(b) \triangleq \min \Big\{ & \log \big(\max_{\tau\in\{0,1,\cdots\}} \frac{1 - F_D(\tau)}{1 - F_D(\tau + b)} \big) - \log \pi_0, \\ & \min \big\{ \inf_{t>t_b} I_A(t, b), \min_{1\leq c\leq b} \big\{ \inf_{t>t_{b-c}} I_A(t, b - c) - \log \pi_0 \\ & + \log \big(\max_{\tau\in\{0,1,\cdots\}} \frac{1 - F_D(\tau)}{1 - F_D(\tau + c - 1)} \big) \big\}, \\ & \min_{c\in\Psi_b} \{ I_A(t_{b-c}, b - c) - \log \pi_0 + \\ & \log \big(\max_{\tau\in\{0,1,\cdots\}} \frac{1 - F_D(\tau)}{1 - F_D(\tau + c)} \big) \big\} \Big\}. \end{aligned} \tag{7}$$

In addition, we define:

$$I_U(b) = \begin{cases} -\log \pi_0 + \log \Big(\max_{\tau\in\{0,1,\cdots\}} \frac{1-F_D(\tau)}{1-F_D(\tau+b)} \Big), & \text{L=1} \\ I_U^*(b), & \text{L>1} \end{cases} \tag{8}$$

The following theorem shows that for any integer $b \geq 0$, $I_U(b)$ is an upper bound on the delay rate-function for any feasible scheduling policy.

Theorem 1: For any integer threshold $b \geq 0$ and any scheduling policy, we have:

$$\limsup_{n\to\infty} \frac{-1}{n} \log \mathbb{P}(W(0) > b) \leq I_U(b). \tag{9}$$

Proof: We consider two cases $L > 1$ and $L = 1$. For the case $L > 1$, we will consider three types of events: χ_1, χ_2^c and χ_3^c, which are subsets of the delay-violation event $\{W(0) > b\}$. Note that bursty arrivals and sluggish services both cause large packet delay in the system. In particular, χ_1 is the event with sluggish services while χ_2^c and χ_3^c are events with bursty arrivals and sluggish services. For detailed proof, please see our online technical report [14]. ■

[1] We mainly focus on the delay analysis, since the results for throughput performance can be easily generalized from [10].

IV. Achievable Rate-function of OPF Policies

In this section, we aim to derive a non-trivial achievable delay rate-function of the class of OPF policies. First, we state the definition of the class of OPF policies.

Definition 1: A scheduling policy **P** is said to be an OPF (*oldest packets first*) policy if in any time-slot, policy **P** can serve the k oldest packets in the system for the largest possible value of $k \in \{1, 2, \cdots, n\}$.

We want to show that the achievable rate-function of any OPF policy **P** is no smaller than $I_0(b)$, defined as:

$$I_0(b) = \begin{cases} -\log \pi_0 + b \cdot \log \frac{1}{1-\hat{q}}, & \text{L=1} \\ I_0^*(b), & \text{L>1} \end{cases} \tag{10}$$

where the parameter $\hat{q}$ is defined to be

$$\hat{q} \triangleq \min\{ \min_{k \in \{0,1,\cdots\}} \frac{\mathbb{P}(D = k+1)}{1 - F_D(k)}, \min_{k \in \{0,1,\cdots\}} \frac{1 - F_U(k+1)}{1 - F_U(k)} \}. \tag{11}$$

and

$$\begin{aligned} I_0^*(b) \triangleq \min \Big\{ & b \cdot \log \frac{1}{1-\hat{q}} - \log \pi_0, \\ & \min\{ \inf_{t > t_b} I_A(t, b), \min_{1 \le c \le b} \{ \inf_{t > t_{b-c}} I_A(t, b-c) \\ & \quad - \log \pi_0 + (c-1) \cdot \log \frac{1}{1-\hat{q}} \}\} \\ & \min_{c \in \Psi_b} \{ I_A(t_{b-c}, b-c) - \log \pi_0 + c \cdot \log \frac{1}{1-\hat{q}} \} \Big\}. \end{aligned} \tag{12}$$

The analysis of delay rate-function follows a similar line of argument as in the case of *i.i.d.* channels. Specifically, we analyze the rate-function of the Frame Based Scheduling (FBS) policy and the perfect-matching policy and exploit the dominance property of the OPF policies over both of them. However, in the case of time-correlated channels, it becomes more challenging to derive a good lower bound on the achievable rate-function. Since the channel has different behaviors (distributions) for state-change and state-keeping. To address this key challenge, we prove two important properties of the FBS policy and the perfect-matching policy (Section IV.A), which will play a key role in the proof. We start by briefly describing the operations of the FBS policy and the perfect-matching policy.

Under the FBS policy, packets are served in unit of frames. Each frame is constructed according to a given operating parameter h, such that: 1) the difference of the arrival times of any two packets within a frame must be no greater than h; and 2) the total number of packets in each frame is no greater than $n_0 = n - Lh$. In each time-slot, the packets arrived at the beginning of this time-slot are filled into the last frame until any of the above two conditions are violated, in which case a new frame will be opened. In each time-slot, the HOL frame can be served only if there exists a matching that can serve all the packets in the HOL frame. Otherwise, no packet will be served. In any time-slot, the FBS policy serves the HOL frame that contains the oldest (up to n_0) packets with high probability for a large n. Under the perfect-matching policy, if a perfect matching can be found, i.e., every queue can be matched with a different server that is connected to this queue, the HOL packet of every queue will be served by the respective server determined by the perfect matching. Otherwise, none of the packets will be served. It has been shown in [10] that any OPF policy dominates the FBS policy and the perfect-matching policy, i.e., given the same packet arrivals and channel realization, any OPF policy will serve every packet that the FBS policy has served up to time t; and the same for the perfect-matching policy. Therefore, the FBS and perfect-matching policy will provide lower bounds on the delay rate-function that any OPF policy can achieve.

A. Properties of FBS and Perfect Matching Policy

In this subsection, we derive the following properties of FBS and perfect matching policy, which will later be used for the rate-function analysis. For ease of presentation, we define function $X_F(t)$ as:

$$X_F(t) = \begin{cases} 1 & \text{if a frame can be served in time-slot } t \\ & \quad \text{under FBS policy,} \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

We have the following lemma that gives a lower bound on the probability that $X_F(t) = 1$.

Lemma 1: Consider an $n \times n$ bipartite graph G, where the time-varying connectivity has the general time-correlation property described in Section II. Then, there exists an $N_F > 0$, such that for all $n \ge N_F$ the conditional probability that $X_F(t) = 1$ is bounded by:

$$\begin{aligned} & \mathbb{P}(X_F(t) = 1 | \mathcal{S}(t_1), \cdots, \mathcal{S}(t_d), \mathcal{S}(t-1)) \\ & \ge 1 - \big(\frac{n}{1-\hat{q}}\big)^{7H} e^{-n \log \frac{1}{1-\hat{q}}}. \end{aligned} \tag{14}$$

for any positive integer d, $t_1 < t_2 < \cdots < t_d < t-1$, and any $\mathcal{S}(t_1), \cdots, \mathcal{S}(t_d)$ and all $n > N_F$, where $\mathcal{S}(\cdot)$ is the connectivity in the corresponding time-slot.

Proof: We provide the proof in APPENDIX A. ■

Lemma 1 shows that given the past channel state information, a frame can be successfully served with high probability. We are interested in finding an upper bound on the probability that during the time interval $[-t-b, -1]$, exactly $t+a$ frames can be successfully served by the FBS scheduling policy. We have the following lemma:

Lemma 2: For all $a \le b - 1$, we have:

$$\begin{aligned} & \mathbb{P}(\sum_{\tau=-t-b}^{-1} X_F(\tau) = t + a) \\ & \le 2^{t+b} \big(\frac{n}{\pi_0}\big)^{7H} \big(\frac{n}{1-\hat{q}}\big)^{7bH} e^{-n\{-\log \pi_0 + (b-a-1) \log \frac{1}{1-\hat{q}}\}}. \end{aligned} \tag{15}$$

Proof: We provide the proof in APPENDIX B. ■

Likewise, we define X_{PM} as:

$$X_{PM}(t) = \begin{cases} 1 & \text{if } G \text{ has a perfect matching at time-slot } t, \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

Similarly, we have the following lemma:

Lemma 3: Consider an $n \times n$ bipartite graph G, where the time-varying connectivity has general time-correlation property. There exists an $N_{PM} > 0$, for all $n \geq N_{PM}$ the probability that G has no perfect matching can be bounded as:

$$\begin{aligned}&\mathbb{P}(X_{PM}(t) = 0|\mathcal{S}(t_1), \cdots, \mathcal{S}(t_d), \mathcal{S}(t-1)) \\ &\quad \leq 3ne^{-n \log \frac{1}{1-\hat{q}}}. \end{aligned} \tag{17}$$

Proof: We omit the proof here, as the same technique used in the proof of Lemma 1 can be applied. ■

Similarly, it can be shown that for all $a \leq b-1$:

$$\begin{aligned}&\mathbb{P}(\sum_{\tau=-t-b}^{-1} X_{PM}(\tau) = t+a) \\ &\quad \leq 2^{t+3b} n^b e^{-n\{-\log \pi_0 + (b-a-1) \log \frac{1}{1-\hat{q}}\}} \end{aligned} \tag{18}$$

The above inequality holds for sufficiently large $n \geq N_{PM}$. Note that the R. H. S. of inequalities (15) and (18) are both monotonically increasing with respect to a.

B. Achievable Rate-function

We first consider the case where $L > 1$. We need to pick an appropriate choice for the value of parameter h for FBS based on the statistics of the arrival process. We fix $\delta < \frac{2}{3}$ and $\epsilon < p/2$. Then, from Assumption 2, there exists a positive function $I_B(\epsilon, \delta)$ such that for all $n \geq N_B(\epsilon, \delta)$ and $t \geq T_B(\epsilon, \delta)$, we have

$$\mathbb{P}\Big(\frac{\sum_{\tau=l+1}^{l+t} \mathbb{1}_{\{|A(\tau)-pn|>\epsilon n\}}}{t} > \delta\Big) < \exp(-ntI_B(\epsilon, \delta)). \tag{19}$$

where l is any arbitrary integer. Choose parameter h to be:

$$h = \max\Big\{T_B(\epsilon, \delta), \Big\lceil \frac{1}{(p-\epsilon)(1-\frac{3\delta}{2})} \Big\rceil, \Big\lceil \frac{2I_0(b)}{I_B(\epsilon, \delta)} \Big\rceil\Big\} + 1. \tag{20}$$

and define $H = Lh$.

The reason for choosing this value of h will later become clearer. Note that in Assumption 2, the maximum number of arrivals in a time-slot is L.

Let $L(-b)$ be the last time before time-slot $-b$, when the backlog is empty, i.e., all the queues have a queue-length of zero. Also, let $\mathcal{E}_t$ be the set of sample paths such that $L(-b) = -t-b-1$ and $W(0) > b$ under policy **P**. Then, we have

$$\mathbb{P}(W(0) > b) = \sum_{t=1}^{\infty} \mathbb{P}(\mathcal{E}_t). \tag{21}$$

Let $\mathcal{E}_t^F$ and $\mathcal{E}_t^{PM}$ be the set of sample paths such that given $L(-b) = -t-b-1$, the event $W(0) > b$ occurs under the FBS policy and the perfect-matching policy, respectively. Recall that policy **P** dominates both the FBS policy and the perfect-matching policy. Since each packet not served by the OPF policy is also not served by the FBS policy or perfect matching policy, then for any $t > 0$ we have

$$\mathcal{E}_t \subseteq \mathcal{E}_t^F \cap \mathcal{E}_t^{PM}. \tag{22}$$

Recall that p is the mean arrival rate to a queue. Now, we choose any fixed real number $\hat{p} \in (p, 1)$, and fix a finite time t^* as

$$t^* \triangleq \max\Big\{T_1, \Big\lceil \frac{I_0(b)}{I_{BX}} \Big\rceil, \max\{t_{b-c}|c \in \Psi_b\}\Big\}, \tag{23}$$

where T_1 and I_{BX} are constants determined by $\hat{p}$.

Hence, if we let

$$P_1 \triangleq \sum_{t=1}^{t^*} \mathbb{P}(\mathcal{E}_t^F \cap \mathcal{E}_t^{PM}), \tag{24}$$

and

$$P_2 \triangleq \sum_{t=t^*}^{\infty} \mathbb{P}(\mathcal{E}_t^F \cap \mathcal{E}_t^{PM}). \tag{25}$$

From the relation in (22), we can bound $\mathbb{P}(\mathcal{E}_t)$ as:

$$\mathbb{P}(\mathcal{E}_t) \leq P_1 + P_2. \tag{26}$$

Hence, we can divide the rate-function analysis into two parts. In part 1, we show that there exists a finite $N_1 > 0$ such that for all $n \geq N_1$, we have

$$P_1 \leq C_1 n^{7(b+1)H} e^{-nI_0(b)}. \tag{27}$$

Then, in part 2, we show that there exists a finite $N_2 > 0$ such that for all $n \geq N_2$,

$$P_2 \leq 4e^{-nI_0(b)}. \tag{28}$$

By combining part 1 and part 2, there exists a finite $N \triangleq \max\{N_1, N_2\}$, such that for all $n \geq N$,

$$\mathbb{P}(W(0) > b) \leq \Big(C_1 n^{7(b+1)H} + 4\Big)e^{-nI_0(b)}. \tag{29}$$

If we take logarithm and limit as n goes to infinity, we obtain $\liminf_{n\to\infty} \frac{-1}{n} \log \mathbb{P}(W(0) > b) \geq I_0(b)$, which is the desired result.

Using the properties we derived in Section IV.A, we can prove part 1 and part 2 following a similar argument as in the proof of [10]. The detailed proof is provided in our online technical report [14] for completeness.

V. The Relationship Between $I_U(b)$ and $I_0(b)$

We have already shown that $I_U(b)$ is an upper bound on the delay rate-function under any possible scheduling policies. Also, we show that the delay rate-function that can be achieved by any OPF policy is no smaller than $I_0(b)$. In this section, we investigate the relationship between the values of these two rate-functions. We show that if the channel is *non-negatively correlated* (Condition A) and the distribution of the OFF period is memoryless (Condition B), any OPF policy can achieve the optimal delay rate-function, i.e., $I_U(b) = I_0(b)$ for *any* fixed integer $b \geq 0$.

1) Condition A: Any vector of finite channel states satisfies non-negative correlation condition:
In statistics, two random variables X, Y are *non-negatively correlated* if $cov(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \geq 0$. The following definition from [13] is a reasonable generalization of non-negative correlation to a set of random variables.

Definition 2: (**Non-negative Correlation Condition**) Let $\mathbf{X} = (X_1, \cdots, X_n)$ be a vector of random variables. Then the random vector $\mathbf{X}$ satisfies non-negative correlation condition if the conditional expectation $\mathbb{E}[X_i, i \in \mathbb{I} | X_j = t_j, \text{for } \forall j \in \mathbb{J}]$ is non-decreasing in each t_j, $j \in \mathbb{J}$ for any disjoint index set $\mathbb{I}, \mathbb{J} \subseteq [n]$.

Lemma 4: If condition A holds, then the class of OPF policies can achieve a delay rate-function of $I_0(b)$ with parameter $\hat{q}$ replaced by $\tilde{q}$, which is given by:

$$\tilde{q} = \min_{k \in \{0,1,\cdots\}} \frac{\mathbb{P}(D = k+1)}{1 - F_D(k)}. \tag{30}$$

Proof: The proof follows a similar argument as in the proof of Lemma 1. We provide the proof in our online technical report [14]. ■

2) Condition B: Distribution D has a memoryless property:
When the distribution D has a memoryless property, namely D is *geometrically* distributed, we have:

$$\begin{aligned} \frac{1 - F_D(k+n-1)}{1 - F_D(k+n)} &= \frac{1 - F_D(k+n-2)}{1 - F_D(k+n-1)} \\ &= \cdots = \frac{1 - F_D(k)}{1 - F_D(k+1)}. \end{aligned} \tag{31}$$

for any $k \geq 1$ and $1 \leq n \leq b$. Multiplying all these n fractions, we can obtain the following equation:

$$\left(\frac{1 - F_D(k)}{1 - F_D(k+1)}\right)^n = \frac{1 - F_D(k)}{1 - F_D(k+n)}. \tag{32}$$

Finally, if the above two conditions are both satisfied, we have the following theorem:

Theorem 2: The class of OPF policies achieve optimal delay rate-function performance under the general time-correlated channel model if conditions A and B both hold:

Proof: Condition A ensures that $\tilde{q}$ is related to the distribution of random variable D and does not depend on U. If we substitute the value of $\tilde{q}$ into $I_0(b)$, it is easy to see that the expression for $I_0(b)$ is very similar to $I_U(b)$, except for the terms related to the CDF of D. Applying condition B, we can obtain $I_0(b) \geq I_U(b)$ directly. Since $I_0(b)$ is an lower bound on the delay rate-function that can be achieved by any OPF policy and $I_U(b)$ is an upper bound on the delay rate-function under any possible scheduling policies, we can conclude that the class of OPF policies achieve delay rate-function optimality in general correlated channel model. ■

In fact, *i.i.d.* channel and *non-negatively correlated* Markovian channel are two special cases, in which both conditions A and B are satisfied, and thus, the optimal rate-function is achieved.

Remark 2: Under *i.i.d.* channel model with channel "ON" probability q, conditions A and B always hold. In this case, U has a *geometric* distribution with parameter $1-q$, and D has a *geometric* distribution with parameter q.

$$\mathbb{E}[C_{i,j}(t) | C_{i,j}(t_1) = c_1, \cdots, C_{i,j}(t_k) = c_k] = \mathbb{E}[C_{i,j}(t)] = q.$$

Since the conditional expectations remain the same for any c, the non-negative correlation condition (condition A) holds. On the other hand, since random variable D is *geometrically* distributed, D has a memoryless property, i.e., condition B holds.

Remark 3: Under Markovian channel model with transition matrix T, condition A is equivalent to the standard notion of non-negative correlation for a two-state Markov chain. In this case, U has a *geometric* distribution with parameter p_{10}, and D has a *geometric* distribution with parameter p_{01}. Substituting the PMF of the *geometric* distribution, we have

$$\begin{aligned} &\mathbb{E}[C_{i,j}(t) | C_{i,j}(t_1) = c_1, \cdots, C_{i,j}(t-1) = 1] \\ &= \mathbb{P}(C_{i,j}(t) = 1 | C_{i,j}(t-1) = 1) = 1 - p_{10}. \end{aligned} \tag{33}$$

and

$$\begin{aligned} &\mathbb{E}[C_{i,j}(t) | C_{i,j}(t_1) = c_1, \cdots, C_{i,j}(t-1) = 0] \\ &= \mathbb{P}(C_{i,j}(t) = 1 | C_{i,j}(t-1) = 0) = p_{01}. \end{aligned} \tag{34}$$

Hence, condition A is equivalent to:

$$1 - p_{10} \geq p_{01} \iff p_{01} + p_{10} \leq 1. \tag{35}$$

which is the condition for non-negative correlation in a two-state Markov chain. Similarly, condition B is satisfied because D is also *geometrically* distributed.

Theorem 3: Under *negatively correlated* Markovian channel model, i.e., $p_{01} + p_{10} > 1$, the class of OPF policies can achieve a delay rate-function that is no smaller than $\frac{\log p_{10}}{\log(1-p_{01})}$-fraction of the optimal value, where p_{01} and p_{10} come from the transition probability.

Proof: Since $p_{01} + p_{10} > 1$, the conditional probability $\mathbb{P}(C_{i,j}(t) = 1 | C_{i,j}(t-1))$ is lower bounded by $1 - p_{10}$. Thus, by using the same proof technique, we can show that the same results hold for $\log \frac{1}{1-p_{01}}$ replaced by $\log \frac{1}{p_{10}}$. Note that the upper bound still remains the same, therefore, it is easy to see that the delay rate-function achieved by the OPF policies is no smaller than $\frac{\log p_{10}}{\log(1-p_{01})}$-fraction of the optimal value. ■

VI. Numerical Results

In this section, we conduct simulations to compare scheduling performance under different channel settings. Among all the OPF policies such as delay weighted matching (DWM) [6], [7], DWM-n and hybrid policy [10], we choose DWM in our simulations as DWM has the best empirical performance in various scenarios [7]. The DWM policy considers at most n oldest packets from each queue, i.e., a total of at most n^2 packets and chooses the schedule that maximizes the sum of the delays in each time-slot. We consider 0-5 *i.i.d.* arrivals i.e.,

$$A_i(t) = \begin{cases} 5, & \text{with probability } \mu, \\ 0, & \text{with probability } 1 - \mu, \end{cases} \tag{36}$$

for all i. The arrival processes are assumed to be independent across all the queues. For the channel model, we assume that all the channels are homogeneous and consider the following seven channel settings, channel settings 1 and 2 are *i.i.d.* ON/OFF channels with "ON" probability $q_1 = 0.6$ and $q_2 = 0.5$, respectively, and channel settings 3, 4, 5, 6 and 7 are Markovian channels with transition matrix $T = \begin{bmatrix} 0.94 & 0.06 \\ 0.04 & 0.96 \end{bmatrix}$, $\begin{bmatrix} 0.85 & 0.15 \\ 0.1 & 0.9 \end{bmatrix}$, $\begin{bmatrix} 0.01 & 0.99 \\ 0.99 & 0.01 \end{bmatrix}$, $\begin{bmatrix} 0.1 & 0.9 \\ 0.9 & 0.1 \end{bmatrix}$ and $\begin{bmatrix} 0.25 & 0.75 \\ 0.75 & 0.25 \end{bmatrix}$, respectively. Note that channel settings 1, 2, 3, and 4 are *non-negatively correlated*, while channel settings 5, 6, and 7 are *negatively correlated*. In addition, we fix the channel/server number to 10, i.e., $n = 10$.

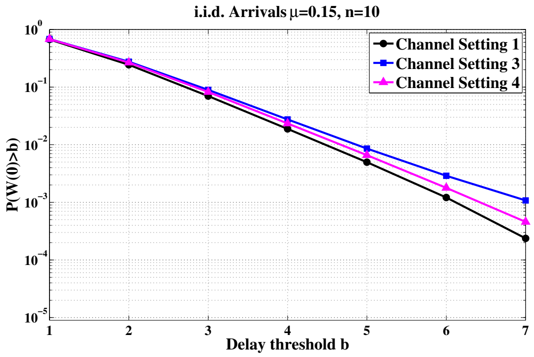

Fig. 3. Performance comparison under different channel settings with $\mu = 0.15$, $n = 10$. Channels are *i.i.d.* in channel setting 1 and from channel setting 4 to 3, channels become more *positively correlated*.

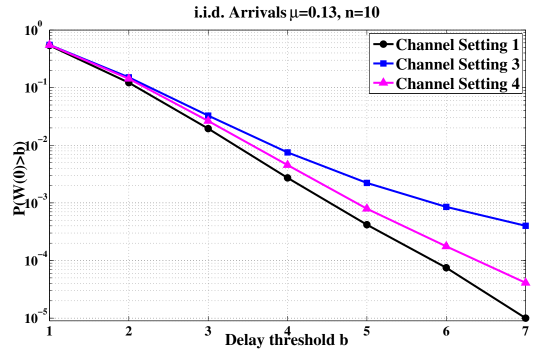

Fig. 4. Performance comparison under different channel settings, with $\mu = 0.13$, $n = 10$. Channels are *i.i.d.* in channel setting 1 and from channel setting 4 to 3, channels become more *positively correlated*.

First, we plot the delay violation probability against different delay thresholds b under channel settings 1, 3, 4 for $\mu = 0.15$ and $\mu = 0.13$, respectively. From Fig. 3 and Fig. 4, we can observe that the *positively correlated* Markovian channel settings have a larger delay than that in the *i.i.d.* channel setting. This result can also be seen through our theoretical results. The *i.i.d.* channel setting has a larger delay rate-function which implies good delay performance. Also, we can use a single-queue single-server system to mimic the multi-queue multi-server system here. As channels are more *positively correlated*, it is more likely to see longer "ON" and "OFF" periods. In this case, the sum of the total service rate could be very large (up to n) or very small with a non-trivial probability. However, in the *i.i.d.* case, according to the Chernoff bound, the sum of total service rate lies in a neighborhood of the mean value nq with high probability. Thus, the service variation under Markovian channels should be larger than the counterpart under the *i.i.d.* channels.

Given the same mean service rate, we know from basic queueing theory that the Markovian channel setting should have a larger delay. Moreover, if we further lower the arrival rate (e.g., decrease μ from 0.15 to 0.13), the simulation results show that as the channels become more *positively correlated* the delay gap increases further.

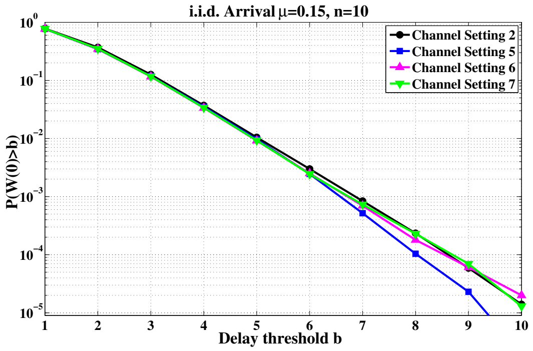

Fig. 5. Performance comparison under different channel settings, with $\mu = 0.15$, $n = 10$. Channels are *i.i.d.* in channel setting 2 and from channel setting 7 down to 5, channels become more *negatively correlated*

Next, we would like to explore the story under negatively correlated channels. As before, we plot the delay violation probability against different b under channel settings 2, 5, 6, and 7 for $\mu = 0.15$. As we can see from Fig. 5, when channel becomes more *negatively correlated*, the system has a smaller delay. An extreme example is channel setting 5, where alternating ON-OFF-ON... will be observed with high probability. Once the initial state is determined, the service rate of the system is almost deterministic. According to basic queueing theory, smaller service variation should give us a smaller delay. However, when we look at channel setting 6 and 7, there is no big difference between itself and the *i.i.d.* channel setting. Therefore, there is still some space for us to find a better scheduling policy under the *negatively correlated* channel model.

VII. Conclusion

In this paper, we considered the scheduling problem of an OFDM downlink system with multiple users and multiple subcarriers with time-correlated channels. Our theoretical result shows that the class of *oldest packets first* (OPF) policies, which give a higher priority to large delay packets, is delay rate-function optimal when two conditions are both satisfied: 1) The channel is *non-negatively correlated*, and 2) The distribution of "OFF" period has a memoryless property. An open problem for future work is to consider multi-rate channels rather than ON/OFF channels with a unit capacity. In this multi-rate channel model, a lexicographically-optimal algorithm that makes the HOL delays most balanced over all the queues is expected to achieve good delay performance. However, the channel-rate heterogeneity introduces a new trade-off between maximizing instantaneous throughput and balancing delays. Nonetheless, we believe that the results in this paper will provide useful insights for designing high-performance scheduling policies for more general scenarios.

Appendix A
Proof of Lemma 1

Applying the law of total probability to different values of $C_{i,j}(t-1)$, we have:

$$\begin{aligned}&\mathbb{P}(C_{i,j}(t)=1|\mathcal{S}(t_1),\mathcal{S}(t_2),\cdots,\mathcal{S}(t_d),\mathcal{S}(t-1))\\&=\mathbb{P}(C_{i,j}(t)=1|\mathcal{S}(t_1),\cdots,\mathcal{S}(t_d),C_{i,j}(t-1)=1)\\&\quad\cdot\mathbb{P}(C_{i,j}(t-1)=1|\mathcal{S}(t_1),\cdots,\mathcal{S}(t_d),\mathcal{S}(t-1))\\&+\mathbb{P}(C_{i,j}(t)=1|\mathcal{S}(t_1),\cdots,\mathcal{S}(t_d),C_{i,j}(t-1)=0)\\&\quad\cdot\mathbb{P}(C_{i,j}(t-1)=0|\mathcal{S}(t_1),\cdots,\mathcal{S}(t_d),\mathcal{S}(t-1)).\end{aligned}\tag{37}$$

Recall that $\hat{T}(t,C_{i,j})$ is the length of the time period from the beginning of its last state-change ("ON" to "OFF" or "OFF" to "ON") time-slot before time-slot t to the beginning of time-slot $t-1$. Summing up all possible values for $\hat{T}(t,C_{i,j})$, we have:

$$\begin{aligned}&\mathbb{P}(C_{i,j}(t)=1|\mathcal{S}(t_1),\mathcal{S}(t_2),\cdots,\mathcal{S}(t_d),C_{i,j}(t-1)=1)\\&=\sum_{k=0}^{\infty}\Big(\mathbb{P}(C_{i,j}(t)=1|\mathcal{S}(t_1),\cdots,\mathcal{S}(t_d),C_{i,j}(t-1)=1,\\&\hat{T}(t,C_{i,j})=k)\\&\times\mathbb{P}(\hat{T}(t,C_{i,j})=k|\mathcal{S}(t_1),\cdots,\mathcal{S}(t_d),C_{i,j}(t-1)=1)\Big).\end{aligned}\tag{38}$$

Note that $C_{i,j}(t)$ only depends on the last known state (here is $C_{i,j}(t-1)$) and the last state-change time-slot before time-slot t, thus, we can simplify the above equation as:

$$\begin{aligned}&\mathbb{P}(C_{i,j}(t)=1|\mathcal{S}(t_1),\mathcal{S}(t_2),\cdots,\mathcal{S}(t_d),C_{i,j}(t-1)=1)\\&=\sum_{k=0}^{\infty}\Big(\mathbb{P}(C_{i,j}(t)=1|C_{i,j}(t-1)=1,\hat{T}(t,C_{i,j})=k)\\&\times\mathbb{P}(\hat{T}(t,C_{i,j})=k|\mathcal{S}(t_1),\cdots,\mathcal{S}(t_d),C_{i,j}(t-1)=1)\Big)\\&=\sum_{k=0}^{\infty}\frac{\mathbb{P}(U\geq k+2)}{\mathbb{P}(U\geq k+1)}\\&\cdot\mathbb{P}(\hat{T}(t,C_{i,j})=k|\mathcal{S}(t_1),\cdots,\mathcal{S}(t_d),C_{i,j}(t-1)=1)\\&\geq\min_{k\in\{0,1,\cdots\}}\frac{1-F_U(k+1)}{1-F_U(k)}.\end{aligned}\tag{39}$$

Applying the same method, we have:

$$\begin{aligned}&\mathbb{P}(C_{i,j}(t)=1|\mathcal{S}(t_1),\mathcal{S}(t_2),\cdots,\mathcal{S}(t_d),C_{i,j}(t-1)=0)\\&\geq\min_{k\in\{0,1,\cdots\}}\frac{\mathbb{P}(D=k+1)}{1-F_D(k)}.\end{aligned}\tag{40}$$

Substitute (39) and (40) into (37),

$$\begin{aligned}&\mathbb{P}(C_{i,j}(t)=1|\mathcal{S}(t_1),\mathcal{S}(t_2),\cdots,\mathcal{S}(t_d),\mathcal{S}(t-1))\\&\geq\min_{k\in\{0,1,\cdots\}}\frac{1-F_U(k+1)}{1-F_U(k)}\\&\cdot\mathbb{P}(C_{i,j}(t-1)=1|\mathcal{S}(t_1),\cdots,\mathcal{S}(t_d),\mathcal{S}(t-1))\\&+\min_{k\in\{0,1,\cdots\}}\frac{\mathbb{P}(D=k+1)}{1-F_D(k)}\\&\cdot\mathbb{P}(C_{i,j}(t-1)=0|\mathcal{S}(t_1),\cdots,\mathcal{S}(t_d),\mathcal{S}(t-1))\\&=\min\{\min_{k\in\{0,1,\cdots\}}\frac{1-F_U(k+1)}{1-F_U(k)},\min_{k\in\{0,1,\cdots\}}\frac{\mathbb{P}(D=k+1)}{1-F_D(k)}\}\\&=\hat{q}.\end{aligned}\tag{41}$$

The above result gives us the lower bound on the conditional probability that $C_{i,j}(t)=1$ given $\mathcal{S}(t_1),\cdots,\mathcal{S}(t_d),\mathcal{S}(t-1)$, thus, by simply replacing q with $\hat{q}$ in the proof of Lemma 6 in [7], the result stated in the lemma follows.

Appendix B
Proof of Lemma 2

From Lemma 1, there exists an $N_F>0$, for all $n>N_F$, the probability that $X_F(t)=0$ occurs given the connectivity at time-slots $t_1,\cdots,t_d,t-1$ can be bounded as,

$$\begin{aligned}&\mathbb{P}(X_F(t)=0|\mathcal{S}(t_1),\cdots,\mathcal{S}(t_d),\mathcal{S}(t-1))\\&\leq(\frac{n}{1-\hat{q}})^{7H}e^{-n\log\frac{1}{1-\hat{q}}}.\end{aligned}\tag{42}$$

Now, we are seeking an upper bound on the probability that there are exactly $t+a$ time-slots that satisfy $X_F(t)=1$ among all $t+b$ time-slots during the time interval $[-t-b,-1]$.

$$\begin{aligned}&\mathbb{P}(\sum_{\tau=-t-b}^{-1}X_F(\tau)=t+a)\\&\leq\mathbb{P}\Big(\bigcup_{t_1<t_2<\cdots<t_{b-a}}X_F(t_1)=0,\cdots,X_F(t_{b-a})=0\Big)\\&\leq\binom{t+b}{t+a}\max_{t_1<t_2<\cdots<t_{b-a}}\mathbb{P}\Big(X_F(t_1)=0,\cdots,X_F(t_{b-a})=0\Big).\end{aligned}\tag{43}$$

Applying the chain rule of conditional probability, we have:

$$\begin{aligned}&\mathbb{P}\Big(X_F(t_1)=0,\cdots,X_F(t_{b-a})=0\Big)\\&=\mathbb{P}(X_F(t_1)=0)\mathbb{P}(X_F(t_2)=0|X_F(t_1)=0)\times\cdots\\&\times\mathbb{P}(X_F(t_{b-a})=0|X_F(t_1)=0,\cdots,X_F(t_{b-a-1})=0).\end{aligned}\tag{44}$$

Next, we consider the R. H. S. of (44). The upper bound on the first term is quite obvious: substituting q by the stationary probability $1-\pi_0$ in lemma 6 in [7], we have:

$$\mathbb{P}(X_F(t_1)=0)\leq(\frac{n}{\pi_0})^{7H}e^{n\log\pi_0}.\tag{45}$$

For the $d^{th} (d > 1)$ term, it is the probability of $\{X_F(t_d) = 0\}$ happens given $\{X_F(t_1) = 0, \cdots, X_F(t_{d-1}) = 0\}$ occurs. Now, we want to obtain a bound for $\mathbb{P}(X_F(t_d) = 0|X_F(t_1) = 0, \cdots, X_F(t_{d-1}) = 0)$:

$$\begin{aligned}
&\mathbb{P}(X_F(t_d) = 0|X_F(t_1) = 0, \cdots, X_F(t_{d-1}) = 0) \\
&= \mathbb{P}(X_F(t_d) = 0|X_F(t_1) = 0, \cdots, \\
&\qquad\qquad X_F(t_{d-1}) = 0, X_F(t_d - 1) = 0) \\
&\times \mathbb{P}(X_F(t_d - 1) = 0|X_F(t_1) = 0, \cdots, X_F(t_{d-1}) = 0) \\
&+ \mathbb{P}(X_F(t_d) = 0|X_F(t_1) = 0, \cdots, \\
&\qquad\qquad X_F(t_{d-1}) = 0, X_F(t_d - 1) = 1) \\
&\times \mathbb{P}(X_F(t_d - 1) = 1|X_F(t_1) = 0, \cdots, X_F(t_{d-1}) = 0).
\end{aligned} \tag{46}$$

We use $\mathbf{S}$ to represent the connectivity $\mathcal{S}(\cdot)$ at each time-slot, and define $\mathcal{F}$ to be a collection of all possible vectors $\mathbf{S}$ such that $X_F(t_1) = 0, \cdots, X_F(t_{d-1}) = 0, X_F(t_d - 1) = 0$. Then we evaluate the following term:

$$\begin{aligned}
&\mathbb{P}(X_F(t_d) = 0|X_F(t_1) = 0, \cdots, \\
&\qquad\qquad X_F(t_{d-1}) = 0, X_F(t_d - 1) = 0) \\
&= \sum_{\mathbf{S}\in\mathcal{F}} \mathbb{P}(X_F(t_d) = 0|\mathbf{S}) \cdot \mathbb{P}(\mathbf{S}|\mathcal{F}) \\
&= \sum_{\mathbf{S}\in\mathcal{F}} \mathbb{P}(X_F(t_d) = 0|\mathcal{S}(t_1), \cdots, \mathcal{S}(t_{d-1}), \mathcal{S}(t_d - 1)) \cdot \mathbb{P}(\mathbf{S}|\mathcal{F}) \\
&\le \big(\frac{n}{1-\hat{q}}\big)^{7H} e^{-n\log\frac{1}{1-\hat{q}}} \sum_{\mathbf{S}\in\mathcal{F}} \mathbb{P}(\mathbf{S}|\mathcal{F}) \\
&= \big(\frac{n}{1-\hat{q}}\big)^{7H} e^{-n\log\frac{1}{1-\hat{q}}}.
\end{aligned} \tag{47}$$

where the inequality comes from Lemma 1. Similarly, we have the same result for $\mathbb{P}(X_F(t_d) = 0|X_F(t_1) = 0, \cdots, X_F(t_{d-1}) = 0, X_F(t_d - 1) = 1)$, thus we have

$$\begin{aligned}
&\mathbb{P}(X_F(t_d) = 0|X_F(t_1) = 0, \cdots, X_F(t_{d-1}) = 0) \\
&\le \big(\frac{n}{1-\hat{q}}\big)^{7H} e^{-n\log\frac{1}{1-\hat{q}}}.
\end{aligned} \tag{48}$$

Combining what we have already derived in (45) and (48), we have:

$$\begin{aligned}
&\mathbb{P}\Big(X_F(t_1) = 0, \cdots, X_F(t_{b-a}) = 0\Big) \\
&\le \big(\frac{n}{\pi_0}\big)^{7H} \big(\frac{n}{1-\hat{q}}\big)^{7bH} e^{-n\{-\log\pi_0 + (b-a-1)\log\frac{1}{1-\hat{q}}\}}.
\end{aligned} \tag{49}$$

This inequality holds for any $t_1, t_2, \cdots, t_{b-a}$, hence,

$$\begin{aligned}
&\max_{t_1,\cdots,t_{b-a}} \mathbb{P}\Big(X_F(t_1) = 0, \cdots, X_F(t_{b-a}) = 0\Big) \\
&\le \big(\frac{n}{\pi_0}\big)^{7H} \big(\frac{n}{1-\hat{q}}\big)^{7bH} e^{-n\{-\log\pi_0 + (b-a-1)\log\frac{1}{1-\hat{q}}\}}.
\end{aligned} \tag{50}$$

Thus, we have for all $a \le b - 1$:

$$\begin{aligned}
&\mathbb{P}(\sum_{\tau=-t-b}^{-1} X_F(\tau) = t + a) \\
&\le \binom{t+b}{t+a} \big(n \cdot \frac{1}{\pi_0}\big)^{7H} \big(\frac{n}{1-\hat{q}}\big)^{7bH} e^{-n\{-\log\pi_0 + (b-a-1)\log\frac{1}{1-\hat{q}}\}} \\
&\le 2^{t+b} \big(\frac{n}{\pi_0}\big)^{7H} \big(\frac{n}{1-\hat{q}}\big)^{7bH} e^{-n\{-\log\pi_0 + (b-a-1)\log\frac{1}{1-\hat{q}}\}}.
\end{aligned} \tag{51}$$

Acknowledgment

This work is funded in part by NSF grants CNS-1446582, CNS-1421576, and CNS-1518829, ONR grant N00014-15-1-2166, and ARO grant W911NF-14-1-0368.

References

[1] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Trans. Automatic Control*, vol. 37, no. 12, pp. 1936-1948, 1992.

[2] S. Bodas, S. Shakkotai, L. Ying, and R. Srikant, "Scheduling in Multi-Channel Wireless Networks: rate-function Optimality in the Small Buffer Regime," in *Proceedings of ACM SIGMETRICS*, 2009.

[3] S. Kittipiyakul and T. Javidi, "Delay-optimal server allocation in multiqueue multiserver systems with time-varying connectivities," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2319-2333, 2009.

[4] S. Bodas, S. Shakkottai, L. Ying, and R. Srikant, "Scheduling in multi-channel wireless networks: rate-function optimality in the small-buffer regime," in *ACM Proceedings of the eleventh international joint conference on Measurement and modeling of computer systems (SIGMETRICS)*, 2009, pp. 121-132.

[5] S. Bodas and T. Javidi, "Scheduling for multi-channel wireless networks: Small delay with polynomial complexity," in *2011 International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*. IEEE, 2011, pp. 78-85.

[6] M. Sharma and X. Lin, "OFDM downlink scheduling for delay-optimality: Many-channel many-source asymptotics with general arrival processes," *IEEE Information Theory and Applications Workshop (ITA)*, 2011.

[7] M. Sharma and X. Lin, "OFDM downlink scheduling for delay-optimality: Many-channel many-source asymptotics with general arrival processes," Purdue University, Tech. Rep., 2011. [Online]. Available: https://engineering.purdue.edu/%7elinx/papers.html

[8] B. Ji, C. Joo, and N. B. Shroff, "Delay-Based Back-Pressure Scheduling in Multihop Wireless Networks," *IEEE/ACM Transactions on Networking*, vol. 21, no. 5, pp. 1539-1552, 2013.

[9] V. G. Kulkarni, *Modeling and Analysis of Stochastic Systems*, CRC Press, 1996.

[10] B. Ji, G. Gupta, X. Lin and N. B. Shroff, "Low-Complexity Scheduling Policies for Achieving Throughput and Asymptotic Delay Optimality in Multi-Channel Wireless Networks," *IEEE/ACM Transactions on Networking*, vol. 22, no. 6, pp. 1911-1924, 2014.

[11] B. Ji, G. Gupta, M. Sharma, X. Lin and N. B. Shroff, "Achieving Optimal Throughput and Near-Optimal Asymptotic Delay Performance in Multi-Channel Wireless Networks with Low Complexity: A Practical Greedy Scheduling Policy," *IEEE/ACM Transactions on Networking*, vol. 23, no. 3, pp. 880-893, 2015.

[12] D. P. Dubhashi and A. Panconesi, *Concentration of Measure for the Analysis of Randomized Algorithms*, Cambridge University Press, 2009.

[13] D. Dubhashi, D. Ranjan, "Balls and bins: A study in negative dependence," *BRICS Report Series 3*, no. 25, 1996.

[14] Z. Qian, B. Ji, K. Srinivasan and N. B. Shroff, "Achieving Delay Rate-function Optimality in OFDM Downlink with Time-correlated Channels," Arxiv Preprint arXiv:1601.06241, Jan. 2016 [Online]. Available: http://arxiv.org/abs/1601.06241