A Large Deviations Analysis of Scheduling in Wireless Networks

Lei Ying, R. Srikant, A. Eryilmaz, and G. E. Dullerud
Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
{lying, rsrikant, eryilmaz, dullerud}@uiuc.edu

***Abstract*— We consider a cellular network consisting of a base station and N receivers. The channel states of the receivers are assumed to be identical and independent of each other. The goal is to compare the throughput of two different scheduling policies (a queue-length-based policy and a greedy policy) given an upper bound on the queue overflow probability or the delay violation probability. We consider a multi-state channel model, where each channel is assumed to be in one of L states. Given an upper bound on the queue overflow probability or an upper bound on the delay violation probability, we show that the total network throughput of the queue-length-based policy is no less than the throughput of the greedy policy for all N. We also obtain a lower bound on the throughput of the queue-length-based policy. For sufficiently large N, the lower bound is shown to be tight, strictly increasing with N, and strictly larger than the throughput of the greedy policy. Further, for a simple multi-state channel model — ON-OFF channel, we prove that the lower bound is tight for all N.**

***Index Terms*— Multiuser wireless scheduling, Large deviations, Wireless networks, Queue-length-based policy, Greedy policy.**

I. INTRODUCTION

Multiuser wireless scheduling has received much attention in recent years. Consider a cellular network consisting of a base station and N users (receivers), where the base station maintains N separate queues, one corresponding to each user. Assume time is slotted and the channel states of the receivers at each time slot are known at the base station. Then, the base station can decide which queues to serve according to their channel states. In this paper, we assume that the base station operates in a TDMA fashion, i.e., the base station can serve only one queue in each time slot. Two scheduling policies have been widely studied in the literature: (i) the base station serves the user with the best (weighted) channel state (opportunistic scheduling) [16], [8]; or (ii) serve the one with the best queue-length-weighted channel state (queue-length based (QLB) scheduling) [15], [6], [10], [11], [4], [1], [9]. While the QLB scheduling is throughput optimal (i.e., can stabilize any set of user throughputs that can be stabilized by any other algorithm), opportunistic scheduling maximizes the total network throughput if all queues are continuously backlogged. If the arrival rates to the users are identical and the channel state distributions to the receivers are identical, then these two scheduling policies have the same stability region.

While stability is the first concern of scheduling policies, quality-of-service (QoS) is equally important in applications. For example, we may require the queue overflow probability to be small or require small delays. The performance of different scheduling policies under QoS constraints has received much attention recently. For reasons of analytical tractability, much of the prior work assumes that the channels to all the receivers are independent and statistically identical. Under this assumption, and assuming identical user utilities, opportunistic scheduling policies become greedy policies in which the base station transmits to the receiver with the best channel state. In [12], the author studies a simple network consisting of two users where the channels are assumed to be independent, identically distributed ON-OFF channels. Using large-deviations techniques, it is shown that the total network throughput of the QLB policy is larger than the throughput of the greedy policy under the queue overflow constraint. In [6], a wireless network with N users and ON-OFF channels is considered. It is assumed that the arrivals are identical and Poisson, and the capacity when the channel is ON is one packet per time slot. It is then shown that, when the number of users increases from N to $2N$, the expected sum of queue lengths is non-increasing under the QLB policy, while it increases linearly under the greedy policy. Further, in [5], the behavior of the greedy policy for Rayleigh fading channels is studied and it is shown that under a delay constraint, the total network throughput of the greedy policy increases initially with the number of users, but eventually decreases and goes to zero when the number of users is sufficiently large.

Motivated by these prior results, in this paper, we study the performance of the two scheduling policies (greedy and QLB) for a wireless network with multi-state channels and constant arrivals. Using sample-path large-deviations techniques that have been used in [2], [12] and [13], we obtain the following results:

1) Assuming a multi-state channel model and a constant arrival rate in each time slot, under the QLB policy, we compute a lower bound on the large-deviations exponent of the probability that at least one queue in the network exceeds a large threshold. We obtain lower bounds on the maximum network throughput under the QoS constraints, and for large N, the lower bounds are tight, strictly increasing, and strictly greater than the throughput of the greedy policy. For the ON-OFF channel model, we prove that the lower bounds are tight for all N. It was conjectured that in [12] that, for the ON-

OFF channel model, the complexity of the calculation of the large-deviations exponent increases exponentially with increasing N, but we show here that a simple closed-form expression can be obtained.

2) Consider ON-OFF channels and the QLB policy. In [6], under the assumption that the channel capacity is one packet per time slot, for a different model, it is shown the expected sum of the queue lengths is nondecreasing when the number of users increases from N to $2N$. In this paper, for the ON-OFF channel model, we show that the maximum network throughput is strictly increasing in N under the delay-violation constraint or queue overflow constraint. Our result does not only compare performance with N users and $2N$ users, but at all intermediate values as well. Our result also holds even when the capacity of the network is greater than one packet-per-slot. Further, for the general multi-state channel model, the maximum throughput is shown to be strictly increasing with N for large N.
3) For the greedy policy, we analytically show that the throughput goes to a constant under the queue overflow constraint, and decreases to zero under the delay violation constraint. This result holds for the general multi-state channel model, and is consistent with the numerical results for Rayleigh fading channels in [5].
4) Under the QoS constraints, we show that the throughput of the QLB scheduling policy is no less than the throughput of the greedy policy. This conclusion was also obtained in [12] for a two-user system and under the queue overflow constraint. Here, we prove that it is true for networks with N users ($N \geq 2$) and multi-state channels.

The rest of the paper is organized as follows: In Section II, we describe our system model in detail. In Section III, we study the QLB policy. Then, in Section IV, the greedy policy is investigated. In Section V, we compare the performance of the QLB policy and the greedy policy. In Section VI, we study the performance of the QLB policy and the greedy policy for Rayleigh fading channels using simulations. Finally, concluding remarks are provided in Section VII.

II. Basic Model

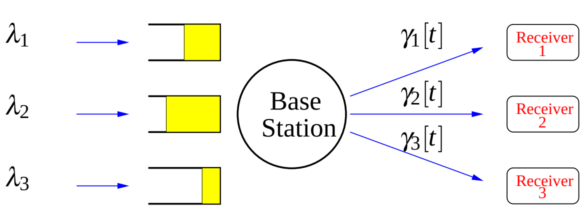


Fig. 1. Single-hop Network

Consider the downlink of a time-slotted cellular network shared by N users, where only one user is allowed to transmit in each time slot. An example with $N = 3$ is shown in Figure 1. Each user is associated with a channel and all channel-state processes $\gamma_i[t]$ are independent and statistically identical. We consider multi-state channels — where each channel has L states $\{0, \ldots, L-1\}$, the probability that the channel is in state l is p_l^c (the superscript "c" indicates it is the probability distribution of the channel states), we can transmit at most F_l bits/slot when a channel is in state l, and $F_k > F_l$ if $k > l$. Also assume that the arrival rate is equal to λ/N bits/slot for each user and the bits are deposited in the queue at the beginning of each time slot. In this paper, we will also analyze a special case of the multi-state channel model — the ON-OFF channel model – in detail, where $L = 2$ and $F_0 = 0$.

We will study the total network throughput of two different scheduling policies under two different quality-of-service (QoS) constraints. The two scheduling policies that will be investigated are:

1) Queue-length based (QLB) policy: Choose user i^* to transmit if

$$i^* \in \arg\max_i \gamma_i[t] Q_i[t],$$

where $Q_i[t]$ is the queue length of user i at the beginning of time slot t. In the ON-OFF channel model, this policy chooses the user with the largest queue length from ON channels.

2) Greedy policy: Choose user i^* if

$$i^* \in \arg\max_i \gamma_i[t].$$

If more than one user has the best channel state, we assume that the base station is equally likely to choose any one of those users.

The two QoS constraints that will be considered are:

1) Queue overflow constraint:

$$\Pr\left(\max_i Q_i(0) > B\right) \leq \varepsilon,$$

where $Q_i(0)$ is the stationary queue length. So this QoS constraint requires the steady-state probability that the queue length is larger than B to be small. Instead of studying this constraint as above, we study the approximation to the constraint given by

$$\theta_B(N, \lambda) := \lim_{B\to\infty} -\frac{1}{B} \log \Pr\left(\max_i Q_i(0) > B\right) \geq \delta, \quad (1)$$

where the large-deviations exponent θ_B is a function of the number of users and the total arrival rate. The exponent δ can be related to ε for large B using the approximation $e^{-\delta B} = \varepsilon$.

2) Delay violation constraint: Define $D(t)$ to be the maximum delay experienced so far by any bit in any of the queues in slot t, and $D(0)$ to be the stationary maximum delay. Assuming that the system starts at time $-\infty$, the steady-state delay violation constraint that we consider can be expressed as follows:

$$\Pr(D(0) > D) \leq \varepsilon.$$

Since the arrival rate is constant, it is easily seen that

$$\Pr(D(0) > D) = \Pr\left(\max_i Q_i(0) > \frac{\lambda}{N} D\right).$$

Thus, the delay violation constraint can be expressed as:

$$\Pr\left(\max_i Q_i(0) > \frac{\lambda}{N} D\right) \le \varepsilon.$$

As before, we study the following approximation to the constraint:

$$\theta_D(N,\lambda) := \lim_{D\to\infty} -\frac{1}{D}\log\Pr\left(\max_i Q_i(0) > \frac{\lambda D}{N}\right) \ge \delta. \quad (2)$$

Note that the delay as defined above is the "virtual delay" and not the actual delay of a packet. From the above description of the two quantities, it is clear that the two large-deviations exponents are related as follows:

$$\theta_D(N,\lambda) = \frac{\lambda\,\theta_B(N,\lambda)}{N}. \quad (3)$$

Thus, we will primarily consider the queue overflow problem when we analyze the wireless system using large deviations.

In this paper, we use θ to denote the large-deviations exponents, where the subscript indicates the QoS constraints ("B" indicates the buffer overflow constraint and "D" indicates the delay-violation constraint), and the superscript indicates the scheduling policy used ("Greedy" is for the greedy policy and "QLB" is for the queue-length based policy).

III. Queue-Length Based Policy

In this section, we will investigate the performance of the wireless system under the QLB policy. Consider a multi-state-channel system and define the state of the system using the composite state of all the channels, so there are L^N system states. Each system state can be represented as an N-tuple in $\{0,\dots,L-1\}^N$. To simplify the notation, we use integer $j\in\{0,\dots,L^N-1\}$ to represent the system state, and define the system state variable $S(t)$ as follows:

$$S(t) := \sum_{i=0}^{N-1}\gamma_i(t)L^i. \quad (4)$$

Sometimes, we will also use the N-tuple representation of the system state, and define an N-tuple S^j in $\{0,\dots,L-1\}^N$ to denote the state j. Let S_i^j be the i^{th} entry of S^j, it is also the state of channel i when the system is in system state j. Further, define a probability vector $\mathbf{p}$ where p_j is the probability the system is in state j, and

$$\bar{\mu} = \sum_{j=0}^{L^N-1}\left(\max_i F_{S_i^j}\right)p_j, \quad (5)$$

which is the maximum throughput the network can support without the QoS constraints. Note that the system is unstable if $\lambda \ge \bar{\mu}$; and large buffer overflow and large delay violation will not happen if $\lambda \le F_0$. Thus, we assume $F_0 < \lambda < \bar{\mu}$.

For sufficiently large T_s, we define $\mathbf{s}^{(B)}(t)$ on $[-T_s,0]$ using $\mathbf{S}(t)$ on $[0,BT_s]$ as follows:

$$s_j^{(B)}(t) := \frac{1}{B}\sum_{k=0}^{B(T_s+t)} 1_{S(k)=j}, \text{ for } t=\frac{k}{B}-T_s, \text{ and } k=\{0,\dots,BT_s\}$$

where for values of t which are not of the form k/B, define $s_j^B(t)$ by linear interpolation. Notice that we have scaled and shifted time so that the discrete time units $0,1,\dots,BT_s$ have now become the continuous time interval $[-T_s,0]$. For each t, the variable $s_j^{(B)}(t)$ is the amount of (scaled) time in the interval $[-T_s,t]$ that the system is in state j. Next, define the system channel rate processes using a L^N-tuple — $\mathbf{u}(t)$, where $\mathbf{u}(t)$ is nonnegative, integrable, and $\sum_{j=0}^{L^N-1}u_j(t)=1$. Further, for B large enough, we have for any $t_1<t_2$

$$\max_j\left|s_j^{(B)}(t_2)-s_j^{(B)}(t_1)-\int_{t_1}^{t_2}u_j(s)\,ds\right| \le \frac{1}{B}.$$

Next, define the Kullback-Leibler distance [3]

$$D(\mathbf{u}(t)\|\mathbf{p}) = \sum_{j=0}^{L^N-1}u_j(t)\log\frac{u_j(t)}{p_j}.$$

Refer to $\int_{-T}^0 D(\mathbf{u}(s)||\mathbf{p})\,ds$ as the cost function, and we define following minimum cost problem:

$$\theta_B^{\text{QLB}}(N,\lambda) = \inf_{T,\mathbf{u}}\int_{-T}^0 D(\mathbf{u}(s)||\mathbf{p})\,ds, \quad (6)$$

where $T\ge 0$, $q_i(-T)=0$ for all i, $\max_i q_i(0)=1$, and the QLB policy is used. We use T^* and $\mathbf{u}^*(t)$ to denote the optimal solution, and use $\mathbf{q}^*(t)$ to denote the corresponding trajectory. It is obvious that the scaled time T_s should be chosen such that $T^*\le T_s$. In Lemma 3, it will be shown that T^* is finite for $F_0<\lambda<\bar{\mu}$, so we assume $T_s\ge T^*$ in this paper.

Theorem 1:

$$\theta_B^{\text{QLB}}(N,\lambda) = \lim_{B\to\infty}\frac{-1}{B}\log\Pr\left(\max_i q_i(0)\ge 1\right),$$

where $\theta_B^{\text{QLB}}(N,\lambda)$ is defined as (6), and queues are scheduled according to the QLB policy.

Proof: The proof is a straightforward extension of Theorem 6.1 in [12]. ■

Note that the minimum cost problem is intuitively obvious: among all possible channel state trajectories that could lead to overflow, we pick the one that is "closest" to the mean value $\mathbf{p}$. Given $\mathbf{u}(t)$, we call $\int_{-T}^0 D(\mathbf{u}(s)||\mathbf{p})\,ds$ the cost of the trajectory generated by $\mathbf{u}(t)$. Thus, Theorem 1 tells us that the probability of the QoS violation is related to the minimum cost problem (6).

In general, the minimum cost problem (6) can be hard to solve. In this paper, since the channels are symmetric, a lower bound can be obtained for the general multi-state channel model, and further for the ON-OFF channel model, the lower bound turns out to be tight. To obtain the lower bound, we first show the following important property of the optimal trajectory.

Lemma 2 (Order Property): Under the QLB policy, given any trajectory, we can find another trajectory that has the same cost and the property such that if $i\ge k$, then

$$q_i(t)\ge q_k(t).$$

Proof: This lemma holds because the channels are symmetric, and the QLB policy is based on the queue lengths and the channel states. So the indices of two queues can be swapped after the two queue lengths become equal, without affecting the cost of the trajectory. The proof is given in Appendix A. ■

In the next lemma, we prove that T^* is finite when $F_0 < \lambda < \bar{\mu}$.

Lemma 3: The optimal solution T^* of the minimum cost problem (6) is finite when $F_0 < \lambda < \bar{\mu}$.

Proof: The proof is available in Appendix B. ■

We have shown T^* is finite and derived a useful property of the optimal trajectory — the order property in Lemma 2. Next, define $\mathscr{A}_{M,l}$ to be the set of system states j such that

$$\max_{i \geq N-M} S_i^j = l = \max_i S_i^j,$$

and let $\mathscr{P}_{M,l}$ denote the probability that the system state is in $\mathscr{A}_{M,l}$,

$$\mathscr{P}_{M,l} = \left(\left(\sum_{k=0}^{l} p_k^c \right)^M - \left(\sum_{k=0}^{l-1} p_k^c \right)^M \right) \left(\sum_{k=0}^{l} p_k^c \right)^{N-M}. \quad (7)$$

Further, define $\mathscr{A}_M = \left(\bigcup_{l=0}^{L-1} \mathscr{A}_{M,l} \right)^c$, and

$$\mathscr{P}_M = 1 - \sum_{l=0}^{L-1} \mathscr{P}_{M,l}. \quad (8)$$

Now, we define optimization problem $\mathrm{OP}(M,N,h)$ and show that the solution of this optimization problem provides us a lower bound to the objective in (6).

$$\mathrm{OP}(M,N,h): \quad C_M^N(h) = \inf_{\mathbf{u},T} TD(\mathbf{u}||\mathbf{p}) \quad (9)$$

$$\text{Subject to}: \quad T \left(\frac{M}{N}\lambda - \sum_{l=0}^{L-1} \left(F_l \sum_{j \in \mathscr{A}_{M,l}} u_j \right) \right) = Mh \quad (10)$$

$$\sum_{j=0}^{L^N-1} u_j = 1 \quad (11)$$

$$u_j \geq 0 \quad \forall j, \quad (12)$$

Notice from (10) that

$$T = \frac{Mh}{\frac{M}{N}\lambda - \sum_{l=0}^{L-1} \left(F_l \sum_{j \in \mathscr{A}_{M,l}} u_j \right)},$$

so

$$C_M^N(h) = \inf_{\mathbf{u}} \frac{MhD(\mathbf{u}||\mathbf{p})}{\frac{M}{N}\lambda - \sum_{l=0}^{L-1} \left(F_l \sum_{j \in \mathscr{A}_{M,l}} u_j \right)} = hC_M^N(1). \quad (13)$$

Further, $D(\mathbf{u}\|\mathbf{p})$ is a strictly convex function of $\mathbf{u}$, so for a fixed T, the solution of optimization problem (9) is unique. We use $\mathbf{u}_M^*$ and T_M^* to denote the optimal solution of problem (9).

In the optimization problem (9)-(12), the schedule feasibility constraint (i.e., schedule the queue with the largest value of the product of queue length and available service rate) has been removed, which gives a lower bound. Further, in this optimization problem, we only serve the M queues when they have the best channel state (recall the definition of $\mathscr{A}_{M,l}$), thus reducing the service rate available to them. This reduces the overflow time which gives a further lower bound.

In the following theorem, we show that $\min_M C_M^N(1)$ is a lower bound to the cost in (6).

Theorem 4: For an N-user system with multi-state channels, the objective of the minimum cost problem (6) is lower bounded by $\min_M C_M^N(1)$, thus

$$\theta_B^{\text{QLB}}(N,\lambda) \geq \min_M C_M^N(1) \quad \text{and} \quad \theta_D^{\text{QLB}}(N,\lambda) \geq \frac{\lambda}{N} \min_M C_M^N(1).$$

Proof: From Lemma 2, we only need to consider ordered trajectories. Given any ordered trajectory, we segment $[-T,0]$ into small intervals $\{[t_m,t_{m+1})\}$. For each interval $[t_m,t_{m+1})$, there exists an M_m such that for any $t \in (t_m,t_{m+1})$, we have $q_{N-1}(t) = \ldots = q_{N-M_m}(t) > q_{N-M_m-1}(t)$. Now define an L^N-tuple $\mathbf{K}^m$ such that

$$K_j^m = \frac{1}{t_{m+1}-t_m} \int_{t_m}^{t_{m+1}} u_j(s)\,ds.$$

Since $D(\mathbf{u}(t)||\mathbf{p})$ is convex in $\mathbf{u}(t)$, from Jensen's inequality [14], we have

$$\int_{t_m}^{t_{m+1}} D(\mathbf{u}(s)||\mathbf{p})\,ds \geq (t_{m+1}-t_m) D(\mathbf{K}^m||\mathbf{p}).$$

Define $\mathscr{B}_{M_m}(\mathbf{q}(t))$ to be the set of the system states such that one of users $\{N-M_m,\ldots,N-1\}$ will be scheduled under the QLB policy if the queue lengths are $\mathbf{q}(t)$. Thus, if $q_{N-1}(t_{m+1}) - q_{N-1}(t_m) = h_m$, we have

$$M_m h_m = (t_{m+1}-t_m)\frac{M_m\lambda}{N} - \int_{t_m}^{t_{m+1}} \sum_{j \in \mathscr{B}_{M_m}(\mathbf{q}(s))} \left(F_{l_{j,M_m}} \right) u_j(s)\,ds,$$

where $l_{j,M_m} = \max_{i \geq N-M_m} S_i^j$. Now, consider system state j such that $j \in \bigcup_{l=0}^{L-1} \mathscr{A}_{M_m,l}$. From the definition of $\mathscr{A}_{M_m,l}$, one of users $\{N-M_m,\ldots,N-1\}$ will be scheduled, so we have $\bigcup_{l=0}^{L-1} \mathscr{A}_{M_m,l} \subseteq \mathscr{B}_{M_m}(\mathbf{q}(t))$, $l_{j,M_m} = l$ for $j \in \mathscr{A}_{M_m,l}$, and

$$\begin{aligned} M_m h_m &\leq (t_{m+1}-t_m)\frac{M_m\lambda}{N} - \int_{t_m}^{t_{m+1}} \sum_{l=0}^{L-1} \left(F_l \sum_{j \in \mathscr{A}_{M_m,l}} u_j(s) \right) ds \\ &= (t_{m+1}-t_m) \left(\frac{M_m\lambda}{N} - \sum_{l=0}^{L-1} \left(F_l \sum_{j \in \mathscr{A}_{M_m,l}} K_j^m \right) \right). \end{aligned}$$

Now choose $\hat{T}_m$ such that

$$M_m h_m = \hat{T}_m \left(\frac{M_m\lambda}{N} - \sum_{l=0}^{L-1} \left(F_l \sum_{j \in \mathscr{A}_{M_m,l}} K_j^m \right) \right),$$

then $\hat{T}_m \leq (t_{m+1}-t_m)$. Since $\{\hat{T}_m, \mathbf{K}^m\}$ is a feasible solution of $\mathrm{OP}(M_m,N,h_m)$, we have

$$\begin{aligned} \int_{t_m}^{t_{m+1}} D(\mathbf{u}(s)||\mathbf{p})\,ds &\geq (t_{m+1}-t_m)D(\mathbf{K}^m\|\mathbf{p}) \geq \hat{T}_m D(\mathbf{K}^m||\mathbf{p}) \\ &\geq C_{M_m}^N(h_m) \geq \min_M C_M^N(h_m), \end{aligned}$$

and

$$\int_{-T}^{0} D(\mathbf{u}(s)||\mathbf{p})\,ds \geq \sum_m \min_M C_M^N(h_m) = \min_M C_M^N(1), \quad (14)$$

where the last equality follows from (13). Inequality (14) holds for any ordered trajectory, so $\theta_B^{\text{QLB}}(N,\lambda) \geq \min_M C_M^N(1)$.

From relation (3), we also have

$$\theta_D^{\text{QLB}}(N,\lambda) \geq \frac{\lambda}{N} \min_M C_M^N(1).$$

■

Suppose $K \in \arg\min_M C_M^N(1)$, the lower bounds are tight if the overflow time of the trajectory generated by channel processes $\mathbf{u}_K^*$ is T_K^*. This is not true for the multi-state channel system in general. However, we will show later that these lower bounds are tight for large N. But first, we prove that the lower bounds are tight for all N for the ON-OFF channel model.

Theorem 5: For an N-user network with ON-OFF channels, the optimal channel rate processes that solve (6) are constant, and thus

$$\theta_B^{\text{QLB}}(N,\lambda) = \min_M C_M^N(1) \quad \text{and} \quad \theta_D^{\text{QLB}}(N,\lambda) = \frac{1}{N}\min_M C_M^N(1).$$

Proof: Define $U = \{M : C_M^N(1) = \min_M C_M^N(1)\}$. We will show that there exists $K \in U$ such that under channel rate processes $\mathbf{u}_K^*$ and the QLB policy, the overflow time is T_K^*.

First, it is obvious that if $N \in U$, $\mathbf{u}_N^*$ generates a feasible trajectory under the QLB policy because all queues are identical under $\mathbf{u}_N^*$. So we only need to consider the case where $N \notin U$. Suppose $K \in U$ and

$$\frac{1}{K}\sum_{j=2^{N-K}}^{2^N-1} u_{K,j}^* < \frac{1}{N-K}\sum_{j=1}^{2^{N-K}-1} u_{K,j}^*, \tag{15}$$

where $u_{K,j}^*$ is the j^{th} entry of $\mathbf{u}_K^*$. Then, we argue that the queue length dynamics will satisfy

$$\sum_{i=N-K}^{N-1} \dot{q}_i = \frac{K}{N}\lambda - F_1 \sum_{j=2^{N-K}}^{2^N-1} u_{K,j}^*; \tag{16}$$

$$\sum_{i=0}^{N-K-1} \dot{q}_i = \frac{N-K}{N}\lambda - F_1 \sum_{j=1}^{2^{N-K}-1} u_{K,j}^*. \tag{17}$$

To see this, note that from the symmetric structure of the optimal problem (9), queue i and queue k have the same channel state processes if $i,k < N-K$ or $i,k \geq N-K$. Thus, under dynamics (16) and (17), we will have $q_i(t) = q_k(t)$ if $i,k < N-K$ or $i,k \geq N-K$. If condition (15) holds, then

$$q_{N-1}(t) = \ldots = q_{N-K}(t) > q_{N-K-1}(t) = \ldots = q_0(t).$$

Then, this trajectory generated by (16) and (17) is a feasible trajectory under the QLB policy (i.e., when a user in $\{N-K,\ldots,N-1\}$ is ON and a user in $\{0,\ldots,N-K-1\}$ is also ON, then the user in $\{N-K,\ldots,N-1\}$ is scheduled as the QLB policy dictates), and the cost of the trajectory is $C_K^N(1)$.

Otherwise, if $K \in U$ and condition (15) doesn't hold, then we have

$$\frac{1}{K}\sum_{j=2^{N-K}}^{2^N-1} u_{K,j}^* \geq \frac{1}{N}\sum_{j=1}^{2^N-1} u_{K,j}^*. \tag{18}$$

Choose $\hat{T}$ such that

$$\hat{T}\left(\lambda - F_1 \sum_{j=1}^{2^N-1} u_{K,j}^*\right) = N. \tag{19}$$

Since

$$T_K^*\left(\frac{K}{N}\lambda - \sum_{j=2^{N-K}}^{2^N-1} u_{K,j}^*\right) = K,$$

from inequality (18), we can conclude that $\hat{T} \leq T_K^*$, which implies

$$\hat{T}D(\mathbf{u}_K^*\|\mathbf{p}) \leq T_K^* D(\mathbf{u}_K^*\|\mathbf{p}) = C_K^N(1).$$

Furthermore, from (19) and (10), we know that $(\hat{T},\mathbf{u}_K^*)$ is a feasible solution of OP$(N,N,1)$. Thus,

$$C_N^N(1) \leq \hat{T}D(\mathbf{u}_K^*\|\mathbf{p}) \leq C_K^N(1),$$

which contradicts $N \notin U$.

Recall that $K \in U$, so from all of the above, we can conclude that there exists a trajectory under the QLB policy which has the minimum cost $C_K^N(1)$. Thus, $\theta_B^{\text{QLB}}(N,\lambda) = C_K^N(1) = \min_M C_M^N(1)$. ■

Now, we focus on the optimization problem (9). It is hard to obtain a closed-form expression for $C_M^N(1)$. Instead, we solve the following problem:

$$\text{Rate }(N,M) \quad R_M^N(1) = \inf_{\mathbf{u},T}\left(\frac{N}{T} + \frac{N}{M}\sum_{l=0}^{L-1}\left(F_l \sum_{j\in\mathscr{A}_{M,l}} u_j\right)\right) \tag{20}$$

$$\text{Subject to}: \quad TD(\mathbf{u}\|\mathbf{p}) = \theta \tag{21}$$

$$\textstyle\sum_{j=0}^{L^N-1} u_j = 1 \tag{22}$$

$$u_j \geq 0 \quad \forall j, \tag{23}$$

Note that the objective in (20) is the expression for λ obtained from the constraint (10) in problem OP(M,N,h) by letting $h=1$. Further, (21) in Rate(N,M) is a constraint on the objective in OP(M,N,h). Thus, if OP$(M,N,1)$ is viewed as an optimization problem for computing a lower bound on the large-deviations exponent given an arrival rate, then Rate(N,M) can be viewed as an optimization problem for computing a lower bound on the maximum arrival rate that the network can support given a QoS constraint expressed in terms of the large-deviations exponent. It can be verified that the optimal solution $(\mathbf{u}_M^*, T_M^*)$ of (9) is also the optimal solution of (20). Let $\lambda_B^{\text{QLB}}(N,\theta)$ denote the maximum throughput the system can support given the constraint $\Pr(\max_i Q_i(0) > B) \leq e^{-\theta B}$, and define

$$\underline{\lambda}_B^{\text{QLB}}(N,\theta) = \min_M R_M^N(1).$$

From Theorem 4, we have

$$\lambda_B^{\text{QLB}}(N,\theta) \geq \underline{\lambda}_B^{\text{QLB}}(N,\theta).$$

The lower bound is tight if $N \in \arg\min_M R_M^N(1)$, but not so in general. In the next corollary, we obtain the closed-from expression of $R_M^N(1)$.

Corollary 6: Suppose the QLB policy is used. For the multi-state channel model, given the queue-overflow constraint $\theta_B^{\text{QLB}}(N,\lambda) = \theta$, the maximum total throughput of the network satisfies

$$\underline{\lambda}_B^{\text{QLB}}(N,\theta) = \min_{1\leq M\leq N}\left(-\frac{N}{\theta}\log\left(\sum_{l=0}^{L-1} e^{-\frac{F_l\theta}{M}}\mathscr{P}_{M,l} + \mathscr{P}_M\right)\right); \tag{24}$$

given the delay-violation constraint $\theta_D^{\text{QLB}}(N,\lambda) = \theta$, the maximum total throughput of the network satisfies

$$\lambda_D^{\text{QLB}}(N,\theta) \geq \underline{\lambda}_D^{\text{QLB}}(N,\theta) = \min_M \lambda_M, \tag{25}$$

where λ_M is the positive solution of

$$1 = -\frac{\log\left(\sum_{l=0}^{L-1} e^{-\frac{NF_l\theta}{M\lambda_M}}\mathscr{P}_{M,l} + \mathscr{P}_M\right)}{\theta}. \tag{26}$$

Further, the lower bounds are tight for the ON-OFF channel model.

Proof: The proof is available in Appendix C. ■

We have obtained the lower bounds on the maximum throughput under the QoS constraints. In the next theorem, by analyzing the closed-form expressions, we are able to prove that the lower bounds are tight for large N. The main idea is to show that $N \in \arg\min_M R_M^N(1)$ when N is large. Since from (36), all users are symmetric under $(\mathbf{u}_N^*, T_N^*)$, $R_N^N(1)$ is achievable and

$$\lambda_B^{\text{QLB}}(N,\theta) = \underline{\lambda}_B^{\text{QLB}}(N,\theta) = R_N^N(1)$$

for large N.

Before we present the theorem, we note from (26) that, it requires $\min_M \log \frac{1}{\mathscr{P}_M} > \theta$ to guarantee $\underline{\lambda}_D^{\text{QLB}}(N,\theta) > 0$. From the definition of $\mathscr{P}_M$, we have $\mathscr{P}_M \le 1 - p_{L-1}^c$ and $\lim_{N\to\infty} \mathscr{P}_1 = 1 - p_{L-1}^c$, which implies that $\theta < \log \frac{1}{1-p_{L-1}^c}$ is a necessary and sufficient condition to guarantee $\underline{\lambda}_D^{\text{QLB}}(N,\theta) > 0$ for all N. So in the following theorem, we investigate the delay violation constraint θ such that $\theta < \log \frac{1}{1-p_{L-1}^c}$.

Theorem 7: Suppose the QLB policy is used. For the multi-state channel model, given the buffer overflow constraint $\theta > 0$, there exists positive N_B^*, such that for any $N \ge N_B^*$,

$$\lambda_B^{\text{QLB}}(N,\theta) = \underline{\lambda}_B^{\text{QLB}}(N,\theta) \quad \text{and} \quad \lambda_B^{\text{QLB}}(N,\theta) < \lambda_B^{\text{QLB}}(N+1,\theta);$$

given the delay constraint $\log \frac{1}{1-p_{L-1}^c} > \theta > 0$, there exists positive N_D^*, such that for any $N \ge N_D^*$,

$$\lambda_D^{\text{QLB}}(N,\theta) = \underline{\lambda}_D^{\text{QLB}}(N,\theta) \quad \text{and} \quad \lambda_D^{\text{QLB}}(N,\theta) < \lambda_D^{\text{QLB}}(N+1,\theta).$$

For the ON-OFF channel model, $N_B^* = N_D^* = 1$. Another consequence is that, for $N \ge N_B^*$ ($N \ge N_D^*$), the lower bound on the large-deviations exponent $\underline{\theta}_B^{\text{QLB}}(N,\lambda)$ ($\underline{\theta}_D^{\text{QLB}}(N,\lambda)$) given in Theorem 4 is the large-deviations exponent itself.

Proof: We first study the network throughput under the buffer overflow constraint. Define $\tilde{p}_{-1} = 0$ and $\tilde{p}_l = \sum_{k=0}^{l} p_k^c$ for $l \ge 0$, note that $\tilde{p}_l$ is the probability that a channel is in one of states $\{0,\dots,l\}$. Then, from (24),

$$\begin{aligned}
\underline{\lambda}_B^{\text{QLB}}(N,\theta) &= \min_{1\le M\le N} \left(-\frac{N}{\theta}\log\left(\sum_{l=0}^{L-1} e^{-\frac{F_l\theta}{M}}\rho_l + 1 - \sum_{l=0}^{L-1}\rho_l\right)\right) \\
&= \min_{1\le M\le N} \left(-\frac{N}{\theta}\log\left(1 + \sum_{l=0}^{L-1}\left(e^{-\frac{F_l\theta}{M}} - 1\right)\rho_l\right)\right),
\end{aligned}$$

where $\rho_l = \left(\tilde{p}_l^M - \tilde{p}_{l-1}^M\right)\tilde{p}_l^{N-M}$. Define $f_B(x)$ for $x > 0$ as follows:

$$f_B(x) = 1 + \sum_{l=0}^{L-1}\left(e^{-\frac{F_l\theta}{x}} - 1\right)\tilde{p}_l^N\left(1 - \left(\frac{\tilde{p}_{l-1}}{\tilde{p}_l}\right)^x\right),$$

so the derivative of $f_B(x)$ is

$$\begin{aligned}
f_B'(x) &= \frac{F_0\theta}{x^2}e^{-\frac{F_0\theta}{x}}\tilde{p}_0^N + \sum_{l=1}^{L-1}\frac{e^{-\frac{F_l\theta}{x}}}{x^2}\tilde{p}_l^N\left(F_l\theta - F_l\theta\left(\frac{\tilde{p}_{l-1}}{\tilde{p}_l}\right)^x\right. \\
&\quad \left. - x^2\left(\frac{\tilde{p}_{l-1}}{\tilde{p}_l}\right)^x\left(1 - e^{\frac{F_l\theta}{x}}\right)\log\frac{\tilde{p}_{l-1}}{\tilde{p}_l}\right).
\end{aligned}$$

Since $\left(\frac{\tilde{p}_{l-1}}{\tilde{p}_l}\right)^x$ and $x^2\left(\frac{\tilde{p}_{l-1}}{\tilde{p}_l}\right)^x$ converge to zero when x goes to infinity, there exists $x_B^* > 0$, which is independent on N, such that $f_B'(x) > 0$ for any $x > x_B^*$. So the lower bound can be rewritten as

$$\begin{aligned}
\underline{\lambda}_B^{\text{QLB}}(N,\theta) = \quad \min\Big\{&\min_{1\le M\le\min\{x_B^*, N-1\}}\left(-\tfrac{N}{\theta}\left(\log f_B(M)\right)\right), \\
&-\tfrac{N}{\theta}\left(\log f_B(N)\right)\Big\}.
\end{aligned}$$

Since $1 - e^{-x} \ge 0$ and $\tilde{p}_{L-1} = 1$, we have

$$-\frac{N}{\theta}\left(\log f_B(M)\right) \ge -\frac{N}{\theta}\log\left(1 + \left(e^{-\frac{F_{L-1}\theta}{M}} - 1\right)\left(1 - \tilde{p}_{L-2}^M\right)\right).$$

Note that x_B^* is independent on N, so for any $M \in [1, x_B^*]$,

$$\lim_{N\to\infty} -\frac{N}{\theta}\left(\log f_B(M)\right) = \infty.$$

Further, $\lambda_B^{\text{QLB}}(N,\theta) \le F_{L-1}$ since F_{L-1} is the maximum service rate one user can receive at each time slot. So there exists N_B^* such that for any $N \ge N_B^*$,

$$\underline{\lambda}_B^{\text{QLB}}(N,\theta) = -\frac{N}{\theta}\left(\log f_B(N)\right) = R_N^N(1).$$

It is obvious that $R_N^N(1)$ is achievable since all users are symmetric under $(\mathbf{u}_N^*, T_N^*)$, so the lower bound is tight for $N \ge N_B^*$, and

$$\lambda_B^{\text{QLB}}(N,\theta) = \underline{\lambda}_B^{\text{QLB}}(N,\theta) = -\frac{N}{\theta}\log\left(\sum_{l=0}^{L-1} e^{-\frac{F_l\theta}{N}}\mathscr{P}_{N,l}\right),$$

where $\mathscr{P}_N = 0$. Further, we have

$$\begin{aligned}
&\lambda_B^{\text{QLB}}(N+1,\theta) \\
>\ & \lambda_B^{\text{QLB}}\left(N+1, \frac{(N+1)\theta}{N}\right) \\
=\ & -\frac{N}{\theta}\log\left(e^{-\frac{F_{L-1}\theta}{N}}\tilde{p}_{L-1}^{N+1} + \sum_{l=1}^{L-1}\left(e^{-\frac{F_{l-1}\theta}{N}} - e^{-\frac{F_l\theta}{N}}\right)\tilde{p}_{l-1}^{N+1}\right) \\
>\ & -\frac{N}{\theta}\log\left(e^{-\frac{F_{L-1}\theta}{N}}\tilde{p}_{L-1}^{N} + \sum_{l=1}^{L-1}\left(e^{-\frac{F_{l-1}\theta}{N}} - e^{-\frac{F_l\theta}{N}}\right)\tilde{p}_{l-1}^{N}\right) \\
=\ & \lambda_B^{\text{QLB}}(N,\theta).
\end{aligned}$$

Next, we investigate the network throughput under the delay constraint. First from (26), we have

$$\begin{aligned}
e^{-\theta} &= \sum_{l=0}^{L-1} e^{-\frac{NF_l\theta}{M\lambda_M}}\mathscr{P}_{M,l} + \mathscr{P}_M \le 1 - \left(1 - e^{-\frac{NF_{L-1}\theta}{M\lambda_M}}\right)\mathscr{P}_{M,L-1} \\
&\le 1 - \left(1 - e^{-\frac{NF_{L-1}\theta}{M\lambda_M}}\right)p_{L-1}^c,
\end{aligned}$$

where the last inequality holds since $\mathscr{P}_{M,L-1} \ge p_{L-1}^c$. Since $e^{-\theta} > 1 - p_{L-1}^c$, we can conclude that

$$\lambda_M \ge \frac{NF_{L-1}\theta}{M\left(\log p_{L-1}^c - \log\left(e^{-\theta} - (1 - p_{L-1}^c)\right)\right)}. \tag{27}$$

Define

$$\lambda_{\min} = \frac{F_{L-1}\theta}{\left(\log p_{L-1}^c - \log\left(e^{-\theta} - (1 - p_{L-1}^c)\right)\right)},$$

we have $\lambda_M \ge \lambda_{\min}$, which implies

$$\lambda_{\min} \le \lambda_D^{\text{QLB}}(N,\theta) \le F_{L-1}.$$

Now define

$$f_D(\lambda,x) = 1+\sum_{l=0}^{L-1}\left(e^{-\frac{NF_l\theta}{\lambda x}}-1\right)\tilde{p}_l^N\left(1-\left(\frac{\tilde{p}_{l-1}}{\tilde{p}_l}\right)^x\right),$$

then

$$\frac{\partial f_D}{\partial x} = \sum_{l=1}^{L-1}\frac{e^{-\frac{NF_l\theta}{\lambda x}}}{x^2}\tilde{p}_l^N\left(\frac{NF_l\theta}{\lambda}\left(1-\left(\frac{\tilde{p}_{l-1}}{\tilde{p}_l}\right)^x\right)\right.$$
$$\left.-x^2\left(\frac{\tilde{p}_{l-1}}{\tilde{p}_l}\right)^x\left(1-e^{\frac{NF_l\theta}{\lambda x}}\right)\log\frac{\tilde{p}_{l-1}}{\tilde{p}_l}\right)+\frac{NF_0\theta}{\lambda x^2}e^{-\frac{NF_0\theta}{\lambda x}}\tilde{p}_0^N.$$

Consider $\lambda_{\min}\le\lambda\le F_{L-1}$ and $x\ge\frac{\alpha N}{\log\log N}$, where $\alpha=\frac{F_{L-1}\theta}{\lambda_{\min}}$, we have

$$\begin{aligned}\frac{\partial f_D}{\partial x} \ge\ & \frac{NF_0\theta}{\lambda x^2}e^{-\frac{NF_0\theta}{\lambda x}}\tilde{p}_0^N+\sum_{l=1}^{L-1}\frac{Ne^{-\frac{NF_l\theta}{\lambda x}}}{x^2}\tilde{p}_l^N\left(\frac{F_l\theta}{F_{L-1}}\right.\\ &\left.-\frac{F_l\theta}{F_{L-1}}\left(\frac{\tilde{p}_{l-1}}{\tilde{p}_l}\right)^x-x^2\left(\frac{\tilde{p}_{l-1}}{\tilde{p}_l}\right)^x\left(\frac{\log N}{N}-\frac{1}{N}\right)\log\frac{\tilde{p}_l}{\tilde{p}_{l-1}}\right)\\ \ge\ & \frac{NF_0\theta}{\lambda x^2}e^{-\frac{NF_0\theta}{\lambda x}}\tilde{p}_0^N+\sum_{l=1}^{L-1}\frac{Ne^{-\frac{NF_l\theta}{\lambda x}}}{x^2}\tilde{p}_l^N\left(\frac{F_l\theta}{F_{L-1}}\right.\\ &\left.-\frac{F_l\theta}{F_{L-1}}\left(\frac{\tilde{p}_{l-1}}{\tilde{p}_l}\right)^x-x^2\left(\frac{\tilde{p}_{l-1}}{\tilde{p}_l}\right)^x\log\frac{\tilde{p}_l}{\tilde{p}_{l-1}}\right).\end{aligned}$$

Since $\left(\frac{\tilde{p}_{l-1}}{\tilde{p}_l}\right)^x$ and $x^2\left(\frac{\tilde{p}_{l-1}}{\tilde{p}_l}\right)^x$ converge to zero when x goes to infinity, there exists $x_D^*>0$, which is independent on N, such that $\frac{\partial f_D}{\partial x}>0$ for any $x>\max\{x_D^*,\frac{\alpha N}{\log\log N}\}$. Choose N_1 such that $\frac{\alpha N_1}{\log\log N_1}\ge x_D^*$, we have $\frac{\partial f_D}{\partial x}>0$ for $N\ge N_1$, $x\ge\frac{\alpha N}{\log\log N}$ and $\lambda_{\min}\le\lambda\le F_{L-1}$. Further, $f_D(\lambda,x)$ is a strictly increasing function of λ, so for any $N\ge N_1$,

$$\underline{\lambda}_D^{\text{QLB}}(N,\theta)=\min\left\{\lambda_N,\min_{1\le M\le\frac{\alpha N}{\log\log N}}\lambda_M\right\},$$

where λ_M and λ_N satisfy (26). From (27), it is easy to see that $\lim_{N\to\infty}\lambda_M=\infty$ for $M\le\alpha N/\log\log N$. So we can conclude that there exists a positive N_D^* such that $\underline{\lambda}_D^{\text{QLB}}(N,\theta)=\lambda_N$ for any $N>N_D^*$. Since λ_N is achievable, we further have for any $N\ge N_D^*$,

$$\lambda_D^{\text{QLB}}(N,\theta)=\lambda_N.$$

Next define

$$\begin{aligned}f(\lambda,N) &= \sum_{l=0}^{L-1}e^{-\frac{F_l\theta}{\lambda}}\mathscr{P}_{N,l}\\ &= e^{-\frac{F_{L-1}\theta}{\lambda}}\tilde{p}_{L-1}^N+\sum_{l=1}^{L-1}\left(e^{-\frac{F_{l-1}\theta}{\lambda}}-e^{-\frac{F_l\theta}{\lambda}}\right)\tilde{p}_{l-1}^N,\end{aligned}$$

so $f(\lambda,N)$ is strictly increasing with λ, and strictly decreasing with N. Consider λ_N and λ_{N+1} such that

$$\frac{-\log\left(f(\lambda_{N+1},N+1)\right)}{\theta}=\frac{-\log\left(f(\lambda_N,N)\right)}{\theta}=1,$$

it is easy to see that $\lambda_N<\lambda_{N+1}$, so

$$\lambda_D^{\text{QLB}}(N,\theta)<\lambda_D^{\text{QLB}}(N+1,\theta)$$

for $N\ge N_D^*$.

Finally, consider the ON-OFF channel model. We can solve λ from (25), and obtain

$$\lambda_D^{\text{QLB}}(N,\theta)=\min_{1\le M\le N}\left(-\frac{\theta NF_1}{M\left(\log\left(\frac{e^{-\theta}-(1-p_1^c)^M}{1-(1-p_1^c)^M}\right)\right)}\right).$$

Define

$$\mathscr{R}_M^N=\frac{\theta NF_1}{M\left(\log\left(e^{-\theta}-(1-p_1^c)^M\right)-\log\left(1-(1-p_1^c)^M\right)\right)},$$

we will show $\mathscr{R}_M^N<\mathscr{R}_M^{N+1}$ and $\mathscr{R}_N^N<\mathscr{R}_{N+1}^{N+1}$. It is trivial to show that $\mathscr{R}_M^N<\mathscr{R}_M^{N+1}$. Now consider $\mathscr{R}_{N+1}^{N+1}$, we have

$$\begin{aligned}\mathscr{R}_{N+1}^{N+1} &= \frac{\theta F_1}{\left(\log\left(e^{-\theta}-(1-p_1^c)^{N+1}\right)-\log\left(1-(1-p_1^c)^{N+1}\right)\right)}\\ &= \theta F_1\frac{1}{\log\left(1+\frac{e^{-\theta}-1}{1-(1-p_1^c)^{N+1}}\right)}\\ &> \theta F_1\frac{1}{\log\left(1+\frac{e^{-\theta}-1}{1-(1-p_1^c)^{N}}\right)}\\ &= \mathscr{R}_N^N.\end{aligned}$$

Thus, we can conclude that

$$\lambda_D^{\text{QLB}}(N,\theta)<\lambda_D^{\text{QLB}}(N+1,\theta),$$

which also implies $\theta_D^{\text{QLB}}(N,\lambda)<\theta_D^{\text{QLB}}(N+1,\lambda)$. Then, from relation (3), we have

$$\begin{aligned}\theta_B^{\text{QLB}}(N+1,\lambda) &= \frac{N+1}{\lambda}\theta_D^{\text{QLB}}(N+1,\lambda)>\frac{N}{\lambda}\theta_D^{\text{QLB}}(N,\lambda)\\ &= \theta_B^{\text{QLB}}(N,\lambda),\end{aligned}$$

and $\lambda_B^{\text{QLB}}(N,\theta)<\lambda_B^{\text{QLB}}(N+1,\theta)$ holds. Thus $N_D^*=N_B^*=1$ for the ON-OFF channel model. ■

Remark:

(1) The lower bounds are shown to be tight for sufficiently large N. From numerical examples, the lower bounds appear to be tight for small or moderate values of N. For example, consider a three-state channel system with $p_0^c=0.5$, $p_1^c=0.4$, $p_2^c=0.1$, $F_0=0$, $F_1=0.5$, and $F_2=1$. Fixing the buffer overflow constraint to be $\theta=2$, it can be shown that the lower bound is tight for $N\ge9$. Fixing the delay-violation constraint to be $\theta=0.15$, the lower bound is tight for $N\ge3$.

(2) When N goes to infinity, we have

$$\begin{aligned}\lim_{N\to\infty}\lambda_B^{\text{QLB}}(N,\theta) &= \lim_{N\to\infty}-\frac{N}{\theta}\log\left(\sum_{l=0}^{L-1}e^{-\frac{F_l\theta}{N}}\mathscr{P}_{N,l}\right)\\ &= \lim_{N\to\infty}-\frac{N}{\theta}\log\left(e^{-\frac{F_{L-1}\theta}{N}}\tilde{p}_{L-1}^N\right.\\ &\quad\left.+\sum_{l=1}^{L-1}\left(e^{-\frac{F_{l-1}\theta}{N}}-e^{-\frac{F_l\theta}{N}}\right)\tilde{p}_{l-1}^N\right)\\ &= F_{L-1},\end{aligned}$$

where the last equality holds since $\tilde{p}_{L-1}=1$ and $\lim_{N\to\infty}\tilde{p}_l^N=0$ for $l<L-1$. Similarly, we can obtain that

$$\lim_{N\to\infty}\lambda_D^{\text{QLB}}(N,\theta)=F_{L-1}.$$

Note that F_{L-1} is the maximum throughput the network can support because it is the maximum service rate the user can receive at each time slot. In the next section, we will investigate the greedy policy and show that $\lambda_B^{\text{Greedy}}(N,\theta)\le\frac{1-e^{-\theta F_{L-1}}}{\theta}$ and $\lambda_D^{\text{Greedy}}(N,\theta)=0$ for large N. So we can conclude that the QLB policy is better than the greedy policy for large N.

IV. Greedy Policy

In this section, the greedy policy is considered. Under the greedy policy, the probability $\hat{p}_l$ that a user is in state l and is picked to transmit is given by

$$\hat{p}_l = \frac{1}{N}\left(\left(1-\sum_{j=l+1}^{L-1} p_j^c\right)^N - \left(1-\sum_{j=l}^{L-1} p_j^c\right)^N\right), \qquad (28)$$

and the probability that a user is not selected is

$$\hat{p}_{\text{null}} = 1 - \sum_{l=0}^{L-1} \hat{p}_l. \qquad (29)$$

From the symmetry of the channel states and the greedy scheduling scheme, we know that

$$\Pr(q_i(0)\geq 1) \leq \Pr(\max_i q_i(0) > 1) \leq N\Pr(q_i(0)\geq 1),$$

which implies

$$-\lim_{B\to\infty}\frac{1}{B}\log\Pr(q_i(0)\geq 1) = -\lim_{B\to\infty}\frac{1}{B}\Pr(\max_i q_i(0) > 1).$$

Thus, we can obtain $\theta_B^{\text{Greedy}}(N,\lambda)$ by calculating $\lim_{B\to\infty}\frac{-1}{B}\Pr(q_i(0)\geq 1)$, which means that under the greedy policy, the large-deviations exponents of the system are same as the large-deviations exponents of a single user. So we can compute the large-deviations exponents under the greedy policy by considering a single user system with arrival rate λ/N and probability vector $\hat{\mathbf{p}}$, where $\hat{p}_l$ is defined by (28) and (29).

Theorem 8: Suppose the greedy policy is used. Given the buffer-overflow constraint $\theta_B^{\text{Greedy}}(N,\lambda) = \theta$, the maximum throughput the network can support is

$$\lambda_B^{\text{Greedy}}(N,\theta) = -\frac{N\log\left(\sum_{l=0}^{L-1} e^{-F_l\theta}\hat{p}_l + \hat{p}_{\text{null}}\right)}{\theta},$$

and

$$\lim_{N\to\infty}\lambda_B^{\text{Greedy}}(N) \leq \frac{1-e^{-\theta F_{L-1}}}{\theta}.$$

Given the delay-violation constraint $\theta_D^{\text{Greedy}}(N,\lambda)=\theta$, the maximum throughput $\lambda_D^{\text{Greedy}}(N,\theta)$ the network can support satisfies

$$1 = -\frac{\log\left(\sum_{l=0}^{L-1}\hat{p}_l e^{-\frac{NF_l\theta}{\lambda_D^{\text{Greedy}}(N,\theta)}} + \hat{p}_{\text{null}}\right)}{\theta},$$

and there exists $a>0$ such that if $N\geq a$,

$$\lambda_D^{\text{Greedy}}(N) = 0.$$

Proof: The expressions of the throughput of the greedy policy can be obtain by applying Corollary 6 to a single-user network with arrival rate λ/N and probability vector $\hat{\mathbf{p}}$, and it is obvious that the lower bounds are tight for a single-user system.

Next we consider the behavior of the throughput for large N. First we have

$$\begin{aligned}\lambda_B^{\text{Greedy}}(N,\theta) &= -\frac{N\log\left(\sum_{l=0}^{L-1} e^{-F_l\theta}\hat{p}_l + \hat{p}_{\text{null}}\right)}{\theta}\\ &\leq -\frac{N\log\left(\hat{p}_{\text{null}} + (1-\hat{p}_{\text{null}})e^{-F_{L-1}\theta}\right)}{\theta}\\ &= -\frac{N\log\left((1-\hat{p}_{\text{null}})(e^{-F_{L-1}\theta}-1)+1\right)}{\theta}.\end{aligned}$$

Since $1-\hat{p}_{\text{null}} = \frac{1}{N}$, we can conclude that

$$\begin{aligned}\lim_{N\to\infty}\lambda_B^{\text{Greedy}}(N,\theta) &\leq \lim_{N\to\infty} -\frac{N\log\left((1-\hat{p}_{\text{null}})(e^{-F_{L-1}\theta}-1)+1\right)}{\theta}\\ &= \frac{1-e^{-\theta F_{L-1}}}{\theta}.\end{aligned}$$

Now, suppose the delay-violation constraint $\theta_D^{\text{Greedy}}(N,\lambda)=\theta$, we have λ is supportable if

$$1 \leq -\frac{\log\left(\sum_{l=0}^{L-1}\hat{p}_l e^{-\frac{NF_l\theta}{\lambda}} + \hat{p}_{\text{null}}\right)}{\theta},$$

which requires

$$e^{-\theta} - \hat{p}_{\text{null}} \geq \sum_{l=0}^{L-1}\hat{p}_l e^{-\frac{NF_l\theta}{\lambda}} \geq 0.$$

Since $\lim_{N\to\infty}\hat{p}_{\text{null}} = 1$, there exists $a>0$ such that $e^{-\theta} < \hat{p}_{\text{null}}$ for any $N\geq a$, which implies $\lambda_D^{\text{Greedy}}(N,\theta)=0$ for $N\geq a$. ■

Note that $\frac{1-e^{-\theta F_{L-1}}}{\theta} < F_{L-1}$ for $\theta>0$. So from Theorem 7 and Theorem 8, we can conclude that the throughput of the QLB policy is strictly larger than the throughput of the greed policy for large N. Further, under the delay-violation constraint, since the maximum throughput of the greedy policy decreases to zero, but the throughput of the QLB converges to F_{L-1}, the performance of these two scheduling schemes are dramatically different when N is large.

V. QLB Policy vs Greedy Policy

We have shown that the QLB policy performs better than the greedy policy for large N. In this section, we show it is actually true for all N. Since the ON-OFF channel model is a simple multi-state channel model, we only prove the fact for the general multi-state channel model.

We first define the system states and the channel rate processes for an N-user system under the greedy policy. We use $\{(j,i)\}$ to indicate the state of the system, where $j = 0,\ldots,L^N-1$ represents the composite state of all the channels, and i indicates the channel which is chosen to transmit. Let l_j be the best channel state when the system is in state j, and m_j be the number of channels in state l_j when the system in state j. Then, we have

$$\tilde{p}_{j,i} = \begin{cases}\frac{1}{m_j}p_j, & \text{if } S_i^j = l_j;\\ 0, & \text{otherwise}.\end{cases}$$

Use $\tilde{\mathbf{p}}$ to denote the probability vector. Further, define $\{\tilde{u}_{j,i}(t)\}$ to be the channel rate processes under the greedy policy.

Theorem 9: For an N-user system, the maximum throughput under the QLB policy is no less than the throughput under the greedy policy:

$$\lambda_B^{\text{QLB}}(N,\theta) \geq \lambda_B^{\text{Greedy}}(N,\theta) \quad\text{and}\quad \lambda_D^{\text{QLB}}(N,\theta) \geq \lambda_D^{\text{Greedy}}(N,\theta).$$

Proof: Theorem holds if

$$\theta_B^{\text{QLB}}(N,\lambda) \geq \theta_B^{\text{Greedy}}(N,\lambda) \quad\text{and}\quad \theta_D^{\text{QLB}}(N,\lambda) \geq \theta_D^{\text{Greedy}}(N,\lambda).$$

Since $\theta_B(N,\lambda) = \frac{N}{\lambda}\theta_D(N,\lambda)$, we only need to show $\theta_B^{\text{QLB}}(N,\lambda) \geq \theta_B^{\text{Greedy}}(N,\lambda)$. The idea is that for any $\mathbf{u}(t)$ under the QLB policy, there exits a $\tilde{\mathbf{u}}(t)$ under the greedy policy such that

$$\int_{-T}^{0} D(\mathbf{u}(s)\|\mathbf{p})\, ds \geq \int_{-\tilde{T}}^{0} D(\tilde{\mathbf{u}}(s)\|\tilde{\mathbf{p}})\, ds, \tag{30}$$

where T and $\tilde{T}$ are the overflow times under $(\mathbf{u}(t), \text{QLB policy})$ and $(\tilde{\mathbf{u}}(t), \text{Greedy policy})$, respectively. Then, we have

$$\theta_B^{\text{QLB}}(N,\lambda) \geq \theta_B^{\text{Greedy}}(N,\lambda),$$

and from which the result follows.

For any channel rate processes $\mathbf{u}(t)$ under the QLB policy, we define a new channel rate processes $\tilde{\mathbf{u}}(t)$ for the system under the greedy policy such that

$$\tilde{u}_{j,i}(t) = \begin{cases} \frac{1}{m_j} u_j(t), & \text{if } S_i^j = l_j; \\ 0, & \text{otherwise.} \end{cases} \tag{31}$$

It is easy to verify that

$$D(\tilde{\mathbf{u}}(t)||\tilde{\mathbf{p}}) = D(\mathbf{u}(t)||\mathbf{p}),$$

so inequality (30) holds if $\tilde{T} \leq T$.

Define $\mathscr{A}$ to be the set of system states such that $S_{N-1}^j \geq S_i^j$ for all i. So user $N-1$ has the best channel state when the system is in state $j \in \mathscr{A}$, and under the greedy policy,

$$\dot{\tilde{q}}_{N-1}(t) = \frac{\lambda}{N} - \sum_{j \in \mathscr{A}} F_{l_j} \tilde{u}_{j,N-1}(t).$$

Further, the queue length of user $N-1$ is no less than other users, so under the QLB policy, user $N-1$ will be served at least $\frac{1}{m_j}$ of the time the system in state j, and

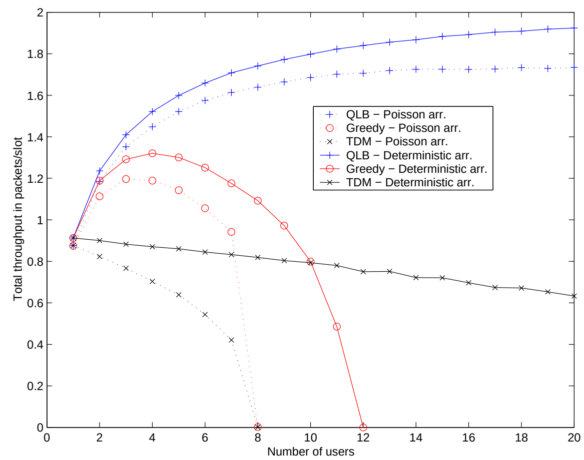
$$\dot{q}_{N-1}(t) \leq \frac{\lambda}{N} - \sum_{j \in \mathscr{A}} F_{l_j} \frac{u_j(t)}{m_j}.$$

From (31), we can conclude that $\dot{\tilde{q}}_{N-1}(t) \geq \dot{q}_{N-1}(t)$, so $\tilde{T} \leq T$, and the theorem holds. ■

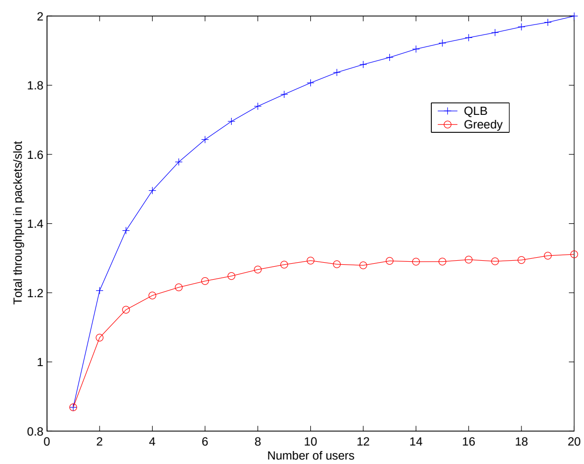
VI. SIMULATIONS

In this section we study the performance of the Greedy and the QLB policies for realistic channel models. We wish to compare the behavior of the schedulers with the analytical results, and investigate whether similar behaviors are observed under the realistic conditions. For this purpose, we consider a Rayleigh fading channel that is independently distributed across the N users and across the time slots. It is assumed that the time slots are of duration $T_c = 1$ msecs, and the available bandwidth for transmission is $W = 1.25$ MHz. The traffic entering the buffers is measured in *packets/slot* for convenience. Here, a *packet* is defined to be of size $W \times T_c = 1.25$ Knats.

EXPERIMENT 1: We assume symmetric channel conditions, where the Signal-to-Noise Ratio (*SNR*) between the base station and each of the users is equal to $3dB$. In addition to the QLB and Greedy Schedulers, we also include the performance of a Time-Division-Multiplexing (TDM) Scheduler, which provides service to the users in a periodic fashion regardless of the channel conditions of that user in the slot that it is served. It has been observed in [5] that the TDM scheduler can outperform the Greedy Scheduler if N is large enough. In Figure 2(a) we study the maximum throughput levels attainable by each of the schedulers when the delay constraints are $D = 100$ and $\varepsilon = 0.001$. The dotted lines correspond to the case when the arrivals to each queue is Poisson distributed with mean equal to λ packets/slot. The solid lines, on the other hand, corresponds to deterministic arrivals at rate λ packets/slot.



(a) The Maximum Throughput under the Delay Constraint



(b) The Maximum Throughput under the Queue Overflow Constraint

Fig. 2. The Maximum Throughput for Rayleigh Fading Channel

It is not surprising that the randomness in the arrivals cause a decrease in the total throughput achievable for each of the schedulers. However, each scheduler is affected to a different extent. For example, we observe that the TDM policy, contrary to the deterministic arrival case, no longer possesses an advantage over the Greedy policy for any N under the random arrival scenario. We also observe that for both the deterministic and Poisson distributed arrival processes, the behavior of the QLB and Greedy policies are very similar to those predicted in our analysis.

In the second part of the experiment, we study the maximum achievable total throughput levels of the QLB and Greedy policies under the buffer constraint. Figure 2(b) depicts the results when $B = 20$ and $\varepsilon = 0.001$.

Both of these simulations support the predicted behavior of the policies for channel processes other than the ON-OFF model, and even under the case of random arrivals. Next, we compare the QLB and Greedy policies under asymmetric channel conditions.

VII. Conclusions

In this paper, we used a large deviations analysis to investigate the performance of different scheduling policies for the downlink of a cellular network under QoS constraints. For a multi-state channel model, we proved that the throughput of QLB policy is no less than that of the greedy policy. Furthermore, we computed lower bounds on the network throughput of the QLB policy, and showed that for large N, the lower bounds are tight, strictly increasing with N, and strictly larger than the throughput of the greedy policy. Particularly under the delay violation constraint, with the number of users increases, the throughput of the QLB policy converges to F_{L-1}, but the throughput of the greedy policy decreases to zero.

Acknowledgment: The second author gratefully acknowledges a conversation with Prof. John Tsitsiklis, MIT, which stimulated this work.

Appendix A: Proof of Lemma 2

Proof: This lemma exploits the fact that the channels are symmetric. First we prove following statement: given the channel rate processes $\bar{\mathbf{u}}(t)$ and suppose $\bar{q}_i(\bar{t}) = \bar{q}_k(\bar{t})$ at time $\bar{t}$. Then, there exists another trajectory $\hat{\mathbf{q}}(t)$ such that $\hat{q}_i(t) = \bar{q}_k(t)$ and $\hat{q}_k(t) = \bar{q}_i(t)$ for $t \geq \bar{t}$, and two trajectories have the same cost. This claim essentially states that the indices of two queues can be swapped after the two queue lengths become equal, without affecting the cost of the trajectory.

To prove the above claim, define new channel rate processes $\hat{\mathbf{u}}(t)$ such that

$$\hat{u}_j(t) = \begin{cases} \bar{u}_j(t), & \text{if } t < \bar{t}; \\ \bar{u}_{d_j}(t), & \text{if } t \geq \bar{t}, \end{cases}$$

where d_j is obtained from j by exchanging the i^{th} and k^{th} elements of the N-tuple expression for j. For example, for a two-user system with ON-OFF channels, we will have $\bar{u}_{(0,1)}(t) = \hat{u}_{(1,0)}(t)$, $\bar{u}_{(1,0)}(t) = \hat{u}_{(0,1)}(t)$, $\bar{u}_{(0,0)}(t) = \hat{u}_{(0,0)}(t)$, and $\bar{u}_{(1,1)}(t) = \hat{u}_{(1,1)}(t)$. Then, since $\bar{q}_i(\bar{t}) = \bar{q}_k(\bar{t})$, the dynamics of queue i and queue j are swapped after $\bar{t}$ under the channel rate processes, and we will have $\hat{q}_i(t) = \bar{q}_k(t)$ and $\hat{q}_k(t) = \bar{q}_i(t)$ for $t \geq \bar{t}$ in the new trajectory.

Furthermore, since the channels are symmetric, it is easy to see $p_j = p_{d_j}$. So we can conclude that the new trajectory will have the same cost as the original one because

$$\int_{\bar{t}}^{0} D(\hat{\mathbf{u}}(s)\|\mathbf{p})\,ds = \int_{\bar{t}}^{0} \sum_j \hat{u}_j(s) \log \frac{\hat{u}_j(s)}{p_j}\,ds$$
$$= \int_{\bar{t}}^{0} \sum_j \bar{u}_{d_j}(s) \log \frac{\bar{u}_{d_j}(s)}{p_j}\,ds = \int_{\bar{t}}^{0} D(\bar{\mathbf{u}}(s)\|\mathbf{p})\,ds.$$

Now, we have proved that if two queues have the same queue length at time $\bar{t}$, there exists a new trajectory with the same cost such that the lengths of the two queues are swapped after $\bar{t}$. It is also easy to see that if we have more than two queues with the same length at time $\bar{t}$, then we can swap any two of them after $\bar{t}$ to get a new trajectory with the same cost. Thus, give any trajectory, we can get a new trajectory with the same cost such that $q_i(t) \geq q_k(t)$ if $i \geq k$. ■

Appendix B: Proof of Lemma 3

Proof: First consider channel rate processes $\bar{\mathbf{u}}(t)$ such that $\bar{u}_0 = 1$ and $\bar{u}_j = 0$ for $j \neq 0$. Under these channel rate processes, it is easy to see that

$$\bar{q}_0(t) = \ldots = \bar{q}_{N-1}(t),$$

and the overflow time $\bar{T} = \frac{N}{\lambda - F_0}$. Thus, we have

$$\theta_B^{\text{QLB}}(N,\lambda) \leq \bar{T} D(\bar{\mathbf{u}}\|\mathbf{p}) = \bar{\theta}, \tag{32}$$

where

$$\bar{\theta} = -\frac{N^2}{\lambda - F_0} \log p_0^c.$$

From [3, pp. 300-301],

$$D(\mathbf{u}\|\mathbf{p}) \geq \frac{1}{2}\left(\sum_{j=0}^{L^N-1} |u_j - p_j|\right)^2 \geq \frac{1}{2}\left(\max_j |u_j - p_j|\right)^2. \tag{33}$$

Recall that T^* and $\mathbf{u}^*(t)$ are the optimal solutions of problem (6), and $\mathbf{q}^*(t)$ is the corresponding trajectory. Let $I_\varepsilon \subset [-T^*, 0]$ be the set such that $\max_j |u_j^*(t) - p_j| \geq \varepsilon$ for all $t \in I_\varepsilon$. From (32) and (33), we have

$$\bar{\theta} \geq \int_{-T^*}^{0} D(\mathbf{u}^*(t)\|\mathbf{p})\,ds \geq \int_{I_\varepsilon} D(\mathbf{u}^*(t)\|\mathbf{p})\,ds \geq \frac{1}{2}\varepsilon^2 \int_{I_\varepsilon} ds. \tag{34}$$

On the other hand, since $q_{N-1}^*(t) \geq q_i^*(t)$ for $i < N-1$, it is easy to show that when $\max_j |u_j^*(t) - p_j| \leq \varepsilon$, the service rate received by user $N-1$ is no less than

$$\frac{1}{N}\left(\sum_{j=0}^{L^N-1}\left(\max_i F_{S_i^j}\right)(p_j - \varepsilon)\right) = \frac{\bar{\mu}}{N} - \frac{\varepsilon}{N}\sum_{j=0}^{L^N-1}\left(\max_i F_{S_i^j}\right).$$

Choose

$$\varepsilon = \frac{\bar{\mu} - \lambda}{2\sum_{j=0}^{L^N-1}\left(\max_i F_{S_i^j}\right)}$$

and consider the dynamics of queue $N-1$. Since $q_{N-1}^*(0) = 1$, we have

$$\begin{aligned} 1 \leq\ & \int_{I_\varepsilon} \tfrac{\lambda}{N}\,ds + \int_{[-T^*,0]\setminus I_\varepsilon} \left(\tfrac{\lambda}{N} - \left(\tfrac{\bar{\mu}}{N} - \tfrac{\varepsilon}{N}\textstyle\sum_{j=0}^{L^N-1}\left(\max_i F_{S_i^j}\right)\right)\right) ds \\ =\ & \int_{I_\varepsilon} \tfrac{\lambda}{N}\,ds - \int_{[-T^*,0]\setminus I_\varepsilon} \tfrac{1}{2}\left(\tfrac{\bar{\mu}-\lambda}{N}\right) ds, \end{aligned}$$

which implies

$$\int_{[-T^*,0]\setminus I_\varepsilon} ds \leq \frac{2N}{\bar{\mu} - \lambda}\left(\int_{I_\varepsilon} \frac{\lambda}{N}\,ds - 1\right). \tag{35}$$

Thus, from (34) and (35), we can conclude that

$$T^* \leq \int_{I_\varepsilon} ds + \frac{2\left(\lambda \int_{I_\varepsilon} ds - N\right)}{\bar{\mu} - \lambda} \leq \frac{2\bar{\theta}}{\varepsilon^2} + \frac{2\left(\lambda \frac{2\bar{\theta}}{\varepsilon^2} - N\right)}{\bar{\mu} - \lambda} < \infty.$$

■

APPENDIX C: PROOF OF COROLLARY 6

Proof: Consider optimization problem Rate(N,M) defined in (20). Since $T=\theta/D(\mathbf{u}\|p)$ from (21), the problem can be written as

$$\text{Rate}(N,M) \quad R_M^N(1)=\inf_{\mathbf{u}}\left(\frac{N}{\theta}D(\mathbf{u}||\mathbf{p})+\frac{N}{M}\sum_{l=0}^{L-1}\left(F_l\sum_{j\in\mathscr{A}_{M,l}}u_j\right)\right)$$

$$\text{Subject to}: \quad \sum_{j=0}^{L^N-1}u_j=1$$
$$u_j\geq 0 \quad \forall j,$$

Define the Lagrangian

$$L=\frac{N}{\theta}\sum_{j=0}^{L^N-1}u_j\log\frac{u_j}{p_j}+\frac{N}{M}\sum_{l=0}^{L-1}\left(F_l\sum_{j\in\mathscr{A}_{M,l}}u_j\right)+\chi\left(1-\sum_{j=0}^{L^N-1}u_j\right).$$

The first-order optimality conditions obtained by differentiating the Lagrangian and setting it equal to zero yield

$$\begin{aligned}\frac{N}{\theta}\log\frac{u_j}{p_j}+\frac{N}{\theta}+\frac{N}{M}F_l &= \chi \text{ for } j\in\mathscr{A}_{M,l};\\ \frac{N}{\theta}\log\frac{u_j}{p_j}+\frac{N}{\theta} &= \chi \text{ for } j\in\mathscr{A};\\ \sum_{j=0}^{L^N-1}u_j &= 1.\end{aligned}$$

Solve $\{u_j^*\}$ from above equations, we get

$$u_j^*=\begin{cases}\frac{p_j}{R}e^{-\frac{\theta F_l}{M}} & \text{for } j\in\mathscr{A}_{M,l};\\ \frac{p_j}{R}, & \text{for } j\in\mathscr{A}\end{cases}, \qquad (36)$$

where

$$R=\sum_{l=0}^{L-1}\mathscr{P}_{M,l}e^{-\frac{F_l\theta}{M}}+\mathscr{P}_M.$$

So

$$R_M^N(1)=-\frac{N}{\theta}\log\left(\sum_{l=0}^{L-1}e^{-\frac{F_l\theta}{M}}\mathscr{P}_{M,l}+\mathscr{P}_M\right),$$

and

$$\lambda_B^{\text{QLB}}(N,\theta)\geq\min_{1\leq M\leq N}\left(-\frac{N}{\theta}\log\left(\sum_{l=0}^{L-1}e^{-\frac{F_l\theta}{M}}\mathscr{P}_{M,l}+\mathscr{P}_M\right)\right). \quad (37)$$

From the relation (3) and inequality (37), we also have that, given $\theta_D^{\text{QLB}}(N,\lambda)=\theta$, λ is supportable under the QLB policy if

$$\lambda\geq-\max_{1\leq M\leq N}\frac{\lambda}{\theta}\log\left(\sum_{l=0}^{L-1}e^{-\frac{F_lN\theta}{\lambda M}}\mathscr{P}_{M,l}+\mathscr{P}_M\right).$$

Thus, λ is supportable if

$$1=-\max_{1\leq M\leq N}\frac{\lambda}{\theta}\log\left(\sum_{l=0}^{L-1}e^{-\frac{F_lN\theta}{\lambda M}}\mathscr{P}_{M,l}+\mathscr{P}_M\right),$$

which implies $\lambda_D^{\text{QLB}}(N,\theta)\geq\min_M\lambda_M$.

Further, from Theorem 5, these lower bounds are tight for the ON-OFF channel model. ■

REFERENCES

[1] M. Andrews, K. Kumaran, K. Ramanan, A.L. Stolyar, R. Vijayakumar, and P. Whiting. CDMA data QoS scheduling on the forward link with variable channel conditions. Bell Laboratories Tech. Rep., April 2000.

[2] D. Bertsimas, I. Ch. Paschalidis, and J. N. Tsitsiklis. Asymptotic buffer overflow probabilities in multiclass multiplexers: An optimal control approach. In *IEEE Transactions on Automatic Control,* 43:315-335, March 1998.

[3] T. M. Cover and J. A. Thomas. Elements of Information Theory. Wiley, 1991.

[4] A. Eryilmaz and R. Srikant, and J. Perkins. Stable scheduling policies for fading wireless channels. In *IEEE/ACM Transactions on Networking,* April 2005, pp. 411-424.

[5] A. Eryilmaz and R. Srikant. Scheduling with Quality of Service Constraints over Rayleigh Fading Channels In *Proceedings of IEEE Conference on Decision and Control,* Dec. 2004.

[6] A. Ganti, E. Modiano and J. Tsitsiklis. Optimal Transmission Scheduling in Symmetric Communication Models with Intermittent Connectivity, 2004 Preprint.

[7] P. W. Glynn and W. Whitt. Logarithmic asymptotics for steady-state tail probabilities in single-server queues. In *Journal of Applied Probability,* 31A:131C156, 1994.

[8] X. Liu, E. Chong, and N. Shroff. Opportunistic transmission scheduling with resource-sharing constraints in wireless networks. In *IEEE Journal on Selected Areas in Communications,* 19(10):2053 – 2064, October 2001.

[9] M. J. Neely, E. Modiano, and C. E. Rohrs. Power and server allocation in a multi-beam satellite with time varying channels. In *Proceedings of IEEE Infocom,* New York, NY. June 2002

[10] S. Shakkottai, R. Srikant, and A. Stolyar. Pathwise optimality of the exponential scheduling rule for wireless channels. In *Advances in Applied Probability,* 1021-1045, Dec. 2004.

[11] S. Shakkottai and A. Stolyar. Scheduling for multiple flows sharing a time-varying channel: The exponential rule. In *Translations of the American Mathematical Society,* Series 2,A volume in memory of F. Karpelevich, Yu. M. Suhov, Editor, Vol. 207, 2002.

[12] S. Shakkottai. Effective Capacity and QoS for Wireless Scheduling, 2004 Preprint.

[13] A. Stolyar and K. Ramanan. Largest weighted delay first scheduling: Large deviations and optimality. In *Ann. Appl. Probab.* 11:1-48, 2001.

[14] D. Stroock. Probability Theory: An Analytic View. Cambridge University Press, 2000.

[15] L. Tassiulas and A. Ephremides. Dynamic server allocation to parallel queues with randomly varying connectivity. In *IEEE Transactions on Information Theory,* 39:466-478, March 1993.

[16] P. Viswanath, D. Tse, and R. Laroia. Opportunistic beamforming using dumb antennas. In *IEEE Transactions on Information Theory,* 48(6):1277C1294, June 2002.