On the Queue-Overflow Probability of Wireless Systems: A New Approach Combining Large Deviations with Lyapunov Functions

V. J. Venkataramanan and Xiaojun Lin, *Senior Member, IEEE*

***Abstract*—In this paper we study the queue-overflow probability of wireless scheduling algorithms. In wireless networks operated under queue-length-based scheduling algorithms, there often exists a tight coupling between the service-rate process, the system backlog process, the arrival process, and the stochastic process governing channel variations. Although one can use sample-path large-deviations techniques to form an estimate of the queue-overflow probability, the formulation leads to a difficult multi-dimensional calculus-of-variations problem. In this paper, we present a new technique to address this complexity issue. Using ideas from the Lyapunov function approach in control theory, this technique maps the complex multi-dimensional calculus-of-variations problem to a one-dimensional calculus-of-variations problem, and the latter is often much easier to solve. Further, under appropriate conditions, we show that when a scheduling algorithm minimizes the drift of a Lyapunov function at each point of every fluid sample path, the algorithm will be optimal in the sense that it maximizes the asymptotic decay-rate of the probability that the Lyapunov function value exceeds a given threshold. We believe that these results can potentially be used to study the queue-overflow probability of a large class of wireless scheduling algorithms and to design new scheduling algorithms with optimal overflow probabilities.**

***Index Terms*—Drift-minimizing algorithms. Lyapunov functions, multi-dimensional calculus-of-variations, queue-overflow probabilities, sample-path large deviations, wireless scheduling algorithms.**

I. Introduction

A wireless network may be modeled as a system of queues with time-varying service rates. The variability in service rates is due to a number of factors. First, channel fading and mobility can lead to variations in the link capacity even if the transmission power is fixed. Second, the transmission power can vary over time according to the power control policy. Third, due to radio interference, it is usually preferable to schedule only a subset of links to be active at each time, and to alternate the subset of activated links over time.

When we study the performance of any system that involves queues, the first question that we can ask is whether the system is *stable* or not. Here, *stability* means that all queue backlog (or equivalently, the delay experienced by the packets) remains finite. Conversely, we can ask the question that, in order to maintain stability, what is the largest offered load that the system can carry. For wireless networks, these questions have led to results on *throughput-optimal* scheduling and routing algorithms for managing wireless network resources. Here, we use the term *scheduling* in the broader sense, i.e., it can include various control mechanisms at the MAC/PHY layer, such as link scheduling, power control, and adaptive coding/modulation. In addition, for multi-hop wireless networks, the routing functionality determines the path that each packet traverses, which also plays a key role in determining the capacity of the network. A scheduling and routing algorithm is *throughput-optimal* if, for any offered load under which any other scheduling and routing algorithm can stabilize the system, this algorithm can stabilize the system as well. One example of a throughput-optimal algorithm is the so-called "maximum-weight" and "back-pressure" algorithm proposed in the seminal work by Tassiulas and Ephremides in [1]. This algorithm chooses at each time, among all possible scheduling and routing decisions, the one that maximizes the sum of the link rates weighted by the differential backlog. This algorithm has been shown to be throughput-optimal, and it has been the basis for many other throughput-optimal algorithms for both cellular and multihop wireless networks.

Once we know about stability, we are then tempted to ask further questions regarding the distribution of queue length (or delay). For example, at a given offered load, what is the probability that the queue length at any link exceeds a given threshold (or, that the delay experienced by a packet is greater than a given threshold)? Conversely, what is the largest offered load that the system can support subject to a given constraint on the queue-overflow probability or delay-violation probability? (In other words, what is the *effective capacity region* of the system under such constraints?) Clearly, these question are important for applications that require more stringent delay guarantees than just stability.

Such problems for characterizing the queue-overflow probability or delay-violation probability of wireless networks can be difficult to solve. Here we draw a comparison to similar queueing problems in wireline networks. In wireline networks, even though the exact queue distribution can be difficult to obtain, there has been a large body of work, especially those using large-deviations techniques, to obtain sharp estimates of the queue-overflow probability [2]–[7]. Essentially, we can compute the asymptotic decay-rate with which the queue-

V. J. Venkataramanan is with Qualcomm Inc., 5775 Morehouse Dr., San Diego, CA 92121 USA (email: venkatar@qti.qualcomm.com). This work was carried out when he was a Ph.D. student at the School of Electrical and Computer Engineering, Purdue University.

X. Lin is with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA (email: linx@purdue.edu).

This work has been partially supported by the National Science Foundation through grants CNS-0643145, CNS-0721484, CNS-0813000, and CCF-0635202, and by Purdue Research Foundation.

overflow probability approaches zero as the overflow threshold approaches infinity. We can then compare the queueing performance of different systems by their corresponding asymptotic decay-rates, and we can ask questions regarding the largest offered load subject to the constraint that the decay-rate must be no smaller than a given threshold value. Most results along this line assume that the service rate of the queue is fixed (i.e., time-invariant), and the packet arrival process is known. These results have enabled us to define the notion of *effective bandwidth* of the arrival process based on its (known) statistics [2]–[7], which can then be used to determine the traffic carrying capability of the system at a given queue-overflow constraint. In contrast, in wireless networks, the service rate is time-varying. If the service rate process is again known *a priori*, large-deviations techniques can still be used to compute the *effective capacity* of the service rate process [8], [9], which is a notion similar to the *effective bandwidth* of the arrival process. This effective capacity can again be used to determine the traffic carrying capability of the system subject to a given constraint on the decay-rate of the queue-overflow probability. Unfortunately, under many wireless resource-allocation algorithms of interest, even the service rate process is unknown *a priori*. For example, for any queue-length based algorithms such as the throughput-optimal back-pressure algorithm of [1], the service rates depend on the queue lengths, which in turn depend on the past history of the arrival process and the channel state. Hence, the statistics of the service rate process is unknown *a priori*. In this case, even the computation of the asymptotic decay-rate of the queue-overflow probability becomes a very difficult problem. For these systems, although it is still possible to use sample-path large-deviations techniques to form an estimate of the decay-rate of the queue-overflow probability [10]–[12], such a formulation leads to a multi-dimensional calculus-of-variations problem. Due to the complex coupling between the service rate, the queue length, the arrival process, and the channel state, this multi-dimensional calculus-of-variations problem is known to be very difficult [10]–[13].

Motivated by the Lyapunov function approach for proving stability of complex systems, in this paper we provide a technique that addresses the complexity of the multi-dimensional calculus-of-variations problem that arises in sample-path large-deviations studies. In essence, through the use of a Lyapunov function, we map the multi-dimensional calculus-of-variations problem into a one-dimensional calculus-of-variations problem, and the latter is often much easier to solve. The solution to the one-dimensional calculus-of-variations problem will then provide us with a lower-bound estimate of the decay-rate of the queue-overflow probability, and consequently, a lower-bound estimate of the effective capacity region of the system. For many practical applications, the resulting effective capacity region is useful because the queue-overflow constraint is known to be satisfied (in the large-deviations sense). We emphasize that, unlike most prior large-deviations work on wireless systems that has focused on single-hop systems (e.g. the cellular downlink), the results described in this paper apply to both single-hop and multi-hop settings.

In addition to the above lower-bound on the asymptotic decay-rate under a given wireless control algorithm, we also provide a useful condition under which a scheduling algorithm is optimal in terms of the asymptotic decay-rate of the overflow probability. Specifically, we show that under suitable conditions, if a scheduling algorithm minimizes the drift of the Lyapunov function at each point in every fluid sample path, then the algorithm is in fact optimal in maximizing the asymptotic decay-rate of the probability that the Lyapunov function value overflows. This is a powerful result that can be used for both analysis and design of scheduling policies. For example, this result can be immediately used to draw the conclusion that the back-pressure algorithm [1] maximizes the asymptotic decay-rate of the probability that the sum of the squares of the per-flow queues exceeds a threshold. In a recent work [14], this result has been used to design a class of asymptotically decay-rate optimal algorithms for multi-hop networks. In another recent work [15], we have used this result to design a simple priority-based algorithm that is sum-queue decay-rate optimal for a tree network, which generalized the algorithm of [16] for a tandem network. For a more detailed discussion of the applications of our main results, please refer to Section IV-B.

The rest of the paper is organized as follows. We present the network model in Section II and introduce the large-deviations preliminaries in Section III. For convenience to the readers, in Section IV we list all main results and provide examples for their applications. The detailed technical proofs are provided in Sections V-VII. Specifically, in Section V we provide a general lower bound for the asymptotic decay-rate of the queue-overflow probability using sample-path large-deviations theory. However, this bound involves a difficult multi-dimensional calculus-of-variations problem. Then, in Section VI, we provide a Lyapunov function based approach to address the complexity issue, which provides a much simpler lower bound on the asymptotic decay-rate of the queue-overflow probability. In Section VII, we provide a drift-minimizing condition under which this lower bound is tight and the corresponding drift-minimizing algorithm is decay-rate optimal. In Section VIII, we provide a detailed example to show how these results can be applied to a specific network setting. Then we conclude.

Throughout the paper, we use x to denote a real number and $\vec{x}$ to denote a vector. For convenience, when we refer to a vector-valued stochastic process $\vec{x}(t)$ over a certain time-interval, we often drop the index t and denote it by a boldface symbol $\mathbf{x}$. In other words, $\vec{x}(t)$ denotes the value of the stochastic process $\mathbf{x}$ at time t. Unless stated otherwise, we use right derivatives throughout this paper, i.e., $\frac{d}{dt}x(t) = \lim_{\delta \downarrow 0^+} \frac{x(t+\delta)-x(t)}{\delta}$.

II. The System Model

We assume the following model for a wireless system with N nodes and L links. A diagram that illustrates some of the key variables is provided in Fig. 1.

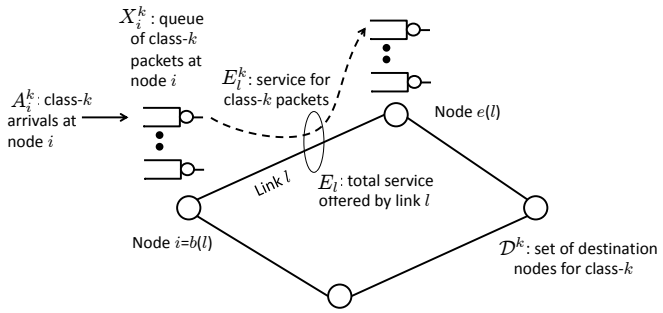

Fig. 1. The system model.

A. The Channel

We assume that time is divided into time-slots of unit length. In order to model channel fading, we assume that at any time-slot τ the state of the wireless channel, denoted by $C(\tau)$, can be in one of S (channel) states $j = 1, 2, ..., S$. We assume that the channel states $C(\tau), \tau = 1, 2, ...$ are *i.i.d.* across time. Let $p_j = \mathbf{P}[C(\tau) = j], j = 1, 2, ..., S.$ and $\vec{p} = [p_1, ..., p_S]$.

B. The Arrival

The system serves packets from multiple classes $k = 1, ..., K$. Each class k corresponds to a set $\mathcal{D}^k$ of destination nodes. In other words, once the class-k packets arrive at any node in $\mathcal{D}^k$, they will leave the system. Let $A_i^k(\tau)$ denote the number of class-k packets arriving at node i at time-slot τ. We assume that $[A_i^k(\tau)]$ is *i.i.d.* across time and independent across arriving nodes and across classes. In addition, we assume that $A_i^k(\tau)$ is bounded by M for all i, k and τ. Let $\lambda_i^k = \mathbf{E}[A_i^k(\tau)]$.

C. The Queue

Let $X_i^k(\tau)$ denote the backlog of class-k packets at node i at time τ, and let $\vec{X}(\tau) = [X_i^k(\tau), i = 1, ..., N, k = 1, ..., K]$. Let $b(l)$ and $e(l)$ denote the source-node and end-node, respectively, of link l. Each link l then corresponds to a server with time-varying service rate, which serves packets at node $b(l)$ and transfers them to node $e(l)$. The service offered by link l is determined by the scheduling and routing algorithm. In general, this service rate may depend on the global system backlog and the global channel state, and hence may correlate with the service at other links/nodes. Let $E_l^k(j, \vec{X})$ denote the service offered by link l to class-k packets at node $b(l)$, when the state of the system is j and the global backlog is $\vec{X}$. We impose the additional constraint that $E_l^k(j, \vec{X}) \leq X_{b(l)}^k$. (In other words, the service offered by link l to class-k packets is no greater than the backlog of class-k packets at the source node $b(l)$.) By definition of $\mathcal{D}^k$ (the set of destination nodes of class-k), $X_i^k(\tau) = 0$ for all nodes $i \in \mathcal{D}^k$. For a node i not in $\mathcal{D}^k$, the evolution of the class-k backlog is given by

$$X_i^k(\tau+1) = X_i^k(\tau) + A_i^k(\tau) - \sum_{j=1}^{S} \mathbf{1}_{\{C(\tau)=j\}} \sum_{l=1}^{L} R_{il} E_l^k(j, \vec{X}(\tau)), \quad (1)$$

for all nodes $i \notin \mathcal{D}^k$, where R_{il} denotes the connectivity matrix, i.e.,

$$R_{il} = \begin{cases} 1, & \text{if } i = b(l) \\ -1, & \text{if } i = e(l) \\ 0, & \text{otherwise.} \end{cases}$$

Let $E_l(j, \vec{X}) = \sum_{k=1}^{K} E_l^k(j, \vec{X})$, which denotes the aggregate service offered by link l. We assume that, for each state j, the service-rate vector $[E_l(j, \vec{X}), l = 1, ..., L]$ must belong to a set $\mathcal{E}_j$ of feasible service-rate vectors. We assume that for all j, the convex hull of $\mathcal{E}_j$, denoted Conv($\mathcal{E}_j$), is closed and bounded, and contains the intersection of a neighborhood of the origin and the positive quadrant.

D. The Performance Measure

Assume that the offered load $\vec{\lambda} = [\lambda_i^k, i = 1, ..., N, k = 1, .., K]$ is such that the system is stationary and ergodic. In this paper, we will focus on the stationary probability that some chosen norm of the system backlog exceeds a certain threshold B. In particular, let ϵ denote our target value of the overflow probability, we would like to ensure that

$$\mathbf{P}[||\vec{X}(\infty)|| \geq B] \leq \epsilon, \quad (2)$$

where $||\cdot||$ is an appropriately chosen norm, and B is the overflow threshold. Note that we have used $\vec{X}(\infty)$ to represent the stationary distribution of the stochastic process $\vec{X}(\tau)$. Unfortunately, the problem of calculating the exact probability $\mathbf{P}[||\vec{X}(\infty)|| \geq B]$ is often mathematically intractable. Instead, we will be interested in the decay rate of the queue-overflow probability, which is defined by

$$I_0(\vec{\lambda}) \triangleq -\lim_{B\to\infty} \frac{1}{B} \log \mathbf{P}[||\vec{X}(\infty)|| \geq B], \quad (3)$$

whenever the limit on the right-hand-side exists. Note that if (3) holds, then when B is large, the overflow probability can be approximated by

$$\mathbf{P}[||\vec{X}(\infty)|| \geq B] \approx g(B) \exp(-BI_0(\vec{\lambda})).$$

Here, $g(B)$ captures the terms that may grow sub-exponentially as $B \to \infty$. We refer to $I_0(\vec{\lambda})$ as the asymptotic decay-rate of the queue-overflow probability. Using the above approximation, in order to approximately satisfy the constraint (2), we only need to ensure that

$$I_0(\vec{\lambda}) - \frac{\log g(B)}{B} \geq \theta \triangleq -\frac{\log \epsilon}{B}.$$

For a fixed θ, assuming $\log g(B)/B \approx 0$, we approximate the above expression by:

$$I_0(\vec{\lambda}) \geq \theta \triangleq -\frac{\log \epsilon}{B}. \quad (4)$$

We can then define the *effective capacity region* as the set of arrival rates $\vec{\lambda}$ such that the above inequality holds. In more general settings, the limit in (3) may not exist or may not be easily computed. Still, we may be able to find a lower bound $I_0'(\vec{\lambda})$ on the asymptotic decay-rate of the overflow probability such that

$$\limsup_{B\to\infty} \frac{1}{B} \log \mathbf{P}[||\vec{X}(\infty)|| \geq B] \leq -I_0'(\vec{\lambda}). \quad (5)$$

Then $\exp(-BI_0'(\vec{\lambda}))$ provides an (approximate) upper bound on the overflow probability $\mathbf{P}[||\vec{X}(\infty)|| \geq B]$, and we can find a lower bound on the effective capacity region as the set of $\vec{\lambda}$ such that $I_0'(\vec{\lambda}) \geq \theta \triangleq -\frac{\log \epsilon}{B}$.

E. Generalizing the Channel and Arrival Processes

For ease of exposition, throughout this paper we have assumed that the channels and the arrival processes are *i.i.d.* across time. We now briefly comment on how the results of this paper can also be generalized to the case with time-correlated channel fluctuations and time-correlated arrivals. As readers will see, the key requirement for these results to hold is that the channel and arrival processes satisfy a sample-path large-deviations principle as will be described in Section III-A (although with different expressions for the integrands of the rate functions $I_s^T(\cdot)$ and $I_a^T(\cdot)$ there). Note that finite-state irreducible Markov chains also satisfy a sample-path large-deviations principle [17, Exercise 5.1.27]. Hence, our results are valid even if we assume that the channel state, $C(\tau)$, does not behave in an *i.i.d.* fashion from time-slot to time-slot but behaves according to a Markov chain. Similarly, we can also assume that each arrival process $A_i^k(\tau), i = 1, ..., N, k = 1, ..., K$, evolves over time-slots according to a Markov chain. The proof of Theorem 4 in Section V will have to be modified since now the evolution of the queue depends not only on the current state of the queue but also on the underlying state of the channel and arrival processes. Please see the beginning of the Appendix for details.

III. Preliminaries

In this section we will introduce some large-deviations preliminaries that are used in the rest of the paper, which include sample-path large deviations of the channel and arrival processes, fluid sample paths, and fluid limits.

A. Large Deviations of Channel and Arrival Processes

For a fixed B, define the scaled channel state process and the scaled arrival process on the time interval $t \in [0, T]$ as

$$s_j^B(t) = \frac{1}{B}\sum_{\tau=0}^{Bt} \mathbf{1}_{\{C(\tau)=j\}}, \quad (6)$$

$$a_i^{k,B}(t) = \frac{1}{B}\sum_{\tau=0}^{Bt} A_i^k(\tau) \quad (7)$$

for $t = \frac{m}{B}$, $m = 0, ..., \lceil BT \rceil$, and by linear interpolation otherwise. The parameter B is a scaling factor: in (6) and (7) we have compressed both time and magnitude by B. To see this, note that (6) and (7) are of the form $\frac{1}{B}f(Bt)$.

The quantity $s_j^B(t)$ can be interpreted as the sum of the (scaled) time in $[0, t]$ that the system is in channel state j. It is easy to check that $\sum_{j=1}^S s_j^B(t) = t$ for all $t \in [0, T]$. Let $\vec{s}^B(t) = [s_1^B(t), ...s_S^B(t)]$. Further, let $\phi_j^B(t) = \frac{d}{dt}s_j^B(t)$. (Note again that we use right-derivatives, and that the right-derivative of $s_j^B(t)$ is well defined almost everywhere on $[0, T]$ except when $t = m/B$ for some integer m.) Let $\vec{\phi}^B(t) = [\phi_1^B(t), ..., \phi_S^B(t)]$. Note that $\sum_{j=1}^S \phi_j^B(t) = 1$ for almost all $t \in [0, T]$. Similarly, let $\vec{a}^B(t) = [a_i^{k,B}(t), i = 1, ..., N, k = 1, ..., K]$ and $\vec{f}^B(t) = \frac{d}{dt}\vec{a}^B(t)$.

Remark: We make a few remarks before we proceed. First, for any fixed T, as $B \to \infty$, we have $BT \to \infty$. Hence, when B is large (as in the large-deviations study below), we aim to capture the infinite-time behavior of the original stochastic processes. (The same intuition also applies to the scaled backlog process $\vec{x}^B(t)$ defined shortly in Section III-B.) Second, for any fixed t, when $B \to \infty$, by a Law of Large Numbers argument, we would expect that *with probability* 1 the value of $\phi_j^B(t)$ will converge to p_j (i.e., the probability that the channel is in state j). However, since we are interested in the large-deviations decay-rate of the overflow probability, we are in fact interested in events with probability approaching 0 as $B \to \infty$. Hence, we have to also study sample paths such that $\phi_j^B(t)$ does not converge to p_j. This is precisely the difference between a *fluid sample path (FSP)* and a *fluid limit*, which will be defined shortly. Third, note that in both (6) and (7) we scale time linearly in B. This scaling is important to obtain the large-deviations decay rate. For instance, if we were to scale time differently than linearly in B, e.g., if we were to replace the upper limits of the summations by B^2T or $\sqrt{B}T$, as $B \to \infty$ the corresponding expressions will approach either ∞ or 0 with very high probability. Thus, they will not correspond to the "most likely path to overflow" from a decay-rate point of view (see also the discussions on the "most likely path to overflow" in Section IV-A2).

It is well-known that the scaled channel state process and the scaled arrival process satisfy large-deviations principles (LDPs). First, we describe the LDP for the scaled channel state process. For any $\vec{\phi} = [\phi_j, j = 1, ..., S] \geq 0$ and $\sum_{j=1}^S \phi_j = 1$, define $H(\vec{\phi}||\vec{p}) = \sum_{j=1}^S \phi_j \log \frac{\phi_j}{p_j}$. (Here we use the convention that $0 \log 0 = 0$.) Let $\Phi_s[0, T]$ denote the space of functions $\vec{s}(t)$ on $[0, T]$ that satisfy the following conditions: $\vec{s}(t)$ is component-wise non-decreasing on $[0, T]$, $\vec{s}(0) = 0$, and $\sum_{j=1}^S s_j(t) = t$ for all t. Let this space be equipped with the essential supremum norm [17, p176, p352], denoted here by $||\cdot||_\infty^T$. The sequence of scaled channel-state processes $\mathbf{s}^B = (\vec{s}^B(t), t \in [0, T])$ is known to satisfy a sample-path large deviations principle [17, p176] with large-deviations rate-function $I_s^T(\mathbf{s})$ given by:

$$I_s^T(\mathbf{s}) = \int_0^T H\left(\frac{d}{dt}\vec{s}(t)||\vec{p}\right) dt,$$

if $\mathbf{s} \in \Phi_s[0, T]$ is absolutely continuous, and $I_s^T(\mathbf{s}) = +\infty$ otherwise. (Note that $\frac{d}{dt}\vec{s}(t)$ is well defined almost everywhere on $[0, T]$ when $\mathbf{s} = (\vec{s}(t), t \in [0, T])$ is absolutely continuous.) Such a large-deviations principle means that, for any set Γ of trajectories in $\Phi_s[0, T]$, the probability that the sequence of scaled channel state processes $\mathbf{s}^B$ fall into Γ must satisfy

$$\begin{aligned} -\inf_{\mathbf{s}\in\Gamma^o} I_s^T(\mathbf{s}) &\leq \liminf_{B\to\infty} \frac{1}{B}\log \mathbf{P}[\mathbf{s}^B \in \Gamma] \quad (8)\\ &\leq \limsup_{B\to\infty} \frac{1}{B}\log \mathbf{P}[\mathbf{s}^B \in \Gamma] \leq -\inf_{\mathbf{s}\in\overline{\Gamma}} I_s^T(\mathbf{s}), \end{aligned}$$

where Γ^o and $\overline{\Gamma}$ denote the interior and closure, respectively, of the set Γ. In addition, if Γ is a *continuity* set [17, p5],

equality is achieved and we then have,

$$\lim_{B\to\infty}\frac{1}{B}\log\mathbf{P}[\mathbf{s}^B\in\Gamma]=-\inf_{\mathbf{s}\in\Gamma}I_s^T(\mathbf{s}). \tag{9}$$

Next, we describe the LDP for the scaled arrival process. We need to first define a few terms. Define

$$L_i^k(f)=\sup_\theta\left\{\theta f-\log\mathbf{E}[\exp(\theta A_i^k(0))]\right\}.$$

Note that this form of rate-function, i.e., as the Legendre transform of the cumulant generating function $\log\mathbf{E}[\exp(\theta A_i^k(0)]$, is standard in large-deviations theory [17, p176]. In particular, we have $L_i^k(f)\geq 0$ for all f, and $L_i^k(f)=0$ when $f=\mathbf{E}[A_i^k(0)]$. For any $\vec{f}=[f_i^k,i=1,...,N,k=1,...,K]$, let $L(\vec{f})=\sum_{i=1}^N\sum_{k=1}^K L_i^k(f_i^k)$. Further, let $\Phi_a[0,T]$ be the space of component-wise non-decreasing functions $\vec{a}(t)$ on $[0,T]$ with $\vec{a}(0)=0$. Let this space also be equipped with the essential supremum norm [17, p176, p352], denoted here by $||\cdot||_\infty^T$. Since the arrivals $[A_i^k(t)]$ are *i.i.d.* in time, the sequence of scaled arrival-processes $\mathbf{a}^B=(\vec{a}^B(t),t\in[0,T])$ also satisfies a sample-path large-deviations principle [17, p176] with large-deviations rate-function $I_a^T(\mathbf{a})$ given as follows:

$$I_a^T(\mathbf{a})=\int_0^T L\left(\frac{d}{dt}\vec{a}(t)\right)dt,$$

if $\mathbf{a}\in\Phi_a[0,T]$ is absolutely continuous, and $I_a^T(\mathbf{a})=+\infty$ otherwise. (Note that $\frac{d}{dt}\vec{a}(t)$ is well-defined almost everywhere on $[0,T]$ when $\mathbf{a}=(\vec{a}(t),t\in[0,T])$ is absolutely continuous.) A similar interpretation as in Equations (8) and (9) holds for $I_a^T(\cdot)$ as well.

Remark: The large-deviations rate-functions $I_s^T(\cdot)$ and $I_a^T(\cdot)$ characterize how rare the occurrence of trajectories $\mathbf{s}^B$ and $\mathbf{a}^B$ are, respectively. Note that $I_s^T(\mathbf{s})\geq 0$ for all trajectories $\mathbf{s}$ and $I_a^T(\mathbf{a})\geq 0$ for all trajectories $\mathbf{a}$. The larger the value of $I_s^T(\mathbf{s})$ is, the further the "empirical probability distribution" $\frac{d}{dt}\vec{s}(t)$ deviates from the underlying probability distribution $\vec{p}$. Hence, it is less likely that trajectory $\mathbf{s}$ will occur. Likewise, the larger the value of $I_a^T(\mathbf{a})$ is, the further the "empirical arrival rate" f_i^k deviates from the mean arrival rate λ_i^k. Equation (9) reflects the well-known large-deviations philosophy that "rare events occur in the most-likely way." Precisely, when B is large, the probability that the scaled channel-state process $\mathbf{s}^B$ falls into a set Γ is determined by the trajectory in Γ that is most likely to occur, i.e., with the smallest $I_s^T(\mathbf{s})$.

B. Fluid Sample Path

Similar to the scaling used in (6) and (7), define the scaled backlog process as,

$$x_i^{k,B}(t)=\frac{1}{B}X_i^k(Bt), \tag{10}$$

for $t=\frac{m}{B}$, $m=0,...,\lceil BT\rceil$, and by linear interpolation otherwise. Let $\vec{x}^B(t)=[x_i^{k,B}(t),i=1,...,N,k=1,...,K]$. Again, note that for any fixed T, by taking $B\to\infty$ (see below) we will in fact look at the infinite-time behavior of the original backlog process $\vec{X}(\tau)$ as $\tau\to\infty$.

We define the notion of a *Fluid Sample Path* (FSP). Note that according to (1), given an initial condition on $\vec{x}^B(0)$, the scaled backlog process $\mathbf{x}^B=(\vec{x}^B(t),t\in[0,T])$ is related to the scaled channel-state process $\mathbf{s}^B=(\vec{s}^B(t),t\in[0,T])$ and the scaled arrival process $\mathbf{a}^B=(\vec{a}^B(t),t\in[0,T])$ by

$$\begin{aligned}&\frac{x_i^{k,B}(t+1/B)-x_i^{k,B}(t)}{1/B}\\&=\left[\frac{a_i^{k,B}(t)-a_i^{k,B}(t-\frac{1}{B})}{1/B}\right]\\&\quad-\sum_{j=1}^S\sum_{l=1}^L R_{il}E_l^k(j,B\vec{x}^B(t))\left[\frac{s_j^B(t)-s_j^B(t-\frac{1}{B})}{1/B}\right],\end{aligned} \tag{11}$$

for $i\notin\mathcal{D}^k$, $t=\frac{m}{B}$, $m=1,...,\lceil BT\rceil$ and by linear interpolation otherwise. Thus, given any T and any initial condition $\vec{x}^B(0)$, Equation (11) defines a mapping from the scaled channel-state process $\mathbf{s}^B$ and the scaled arrival process $\mathbf{a}^B$ to the scaled backlog process $\mathbf{x}^B$ over the time-interval $[0,T]$.

As $B\to\infty$, take any sequence of $\mathbf{s}^B$ and $\mathbf{a}^B$ (which may come from *different* sample paths). They map to a sequence of scaled backlog processes $\mathbf{x}^B$. Note that $\mathbf{s}^B$, $\mathbf{a}^B$ and $\mathbf{x}^B$ are all Lipschitz-continuous. Hence, for any such sequence $(\mathbf{s}^B,\mathbf{a}^B,\mathbf{x}^B)$, there must exist a subsequence that converges to a limiting process $(\mathbf{s},\mathbf{a},\mathbf{x})$ uniformly over compact intervals. We define any such limiting process as a *Fluid Sample Path*, or an FSP, which we denote as $(\mathbf{s},\mathbf{a},\mathbf{x})_T$, where the subscript denotes the ending time T. Such an FSP often satisfies the following differential equation obtained by letting $B\to\infty$ on Equation (11):

$$\frac{d}{dt}x_i^k(t)=\frac{d}{dt}a_i^k(t)-\sum_{j=1}^S\frac{d}{dt}s_j(t)\sum_{l=1}^L R_{il}e_l^k(j,\vec{x}(t)). \tag{12}$$

where $e_l^k(j,\vec{x}(t))$ is some appropriately-chosen limiting value of $E_l^k(j,B\vec{x}^B(t))$ (an example is given later in Equation (58)). In the rest of the paper, we sometimes denote FSPs as $(\mathbf{s},\mathbf{a},\mathbf{x})$, i.e., without the time subscript T. In doing so, we mean that there is some finite time T such that $(\mathbf{s},\mathbf{a},\mathbf{x})_T$ is an FSP.

C. Fluid Limits

Related to the notion of fluid-sample-paths is the notion of *fluid limits* [18], [19]. Take the scaled queue-evolution equation (11), and take a sequence of $(\mathbf{s}^B,\mathbf{a}^B,\mathbf{x}^B)$ with $B\to+\infty$. For a large class of dynamic systems, one can show that, *with probability* 1, there must exist a subsequence $(\mathbf{s}^{B_n},\mathbf{a}^{B_n},\mathbf{x}^{B_n})$ such that $\vec{s}^{B_n}(t)\to\vec{p}t,\vec{a}^{B_n}(t)\to\vec{\lambda}t$ and $\vec{x}^{B_n}(t)\to\vec{x}(t)$ uniformly over compact intervals. Any such limit $\vec{x}(t)$ is called the *fluid limit* of the system [18]. This fluid limit can often be written in the form

$$\frac{d}{dt}x_i^k(t)=\lambda_i^k-\sum_{j=1}^S p_j\sum_{l=1}^L R_{il}e_l^k(j,\vec{x}(t)). \tag{13}$$

where $e_l^k(j,\vec{x}(t))$ is some appropriate limit of $E_l^k(j,\vec{X}(t))$.

Remark: Readers may note the similarity between Equation (13) and Equation (12). We now briefly comment on the difference between a fluid limit in (13) and a fluid sample

path (FSP) in (12). In essence, a fluid limit is a fluid sample path with $\frac{d}{dt}\vec{s}(t) = \vec{p}$ and $\frac{d}{dt}\vec{a}(t) = \vec{\lambda}$ for all t, which is why in Equation (13) λ_i^k and p_j replace $\frac{d}{dt}a_i^k(t)$ and $\frac{d}{dt}s_j(t)$, respectively, from (12). This is the case because the fluid limit is the *almost sure* limit of the scaled process, and hence a "Law of Large Numbers" type of argument can be invoked. In other words, the fluid limit dynamics can be viewed as the *mean* behavior of the system. In contrast, for an FSP we are interested in rare events, and hence the values of $\frac{d}{dt}\vec{a}(t)$ and $\frac{d}{dt}\vec{s}(t)$ can deviate from their mean values.

Readers may also note that the definitions of the scaled processes and the FSP are both over a finite time horizon $[0, T]$, and we allow them to start from some given initial condition $\vec{x}^B(0)$. This formulation may seem contradictory to our initial goal of studying the *stationary* overflow probability of $\vec{X}(\infty)$. It turns out that such constructions are common in sample-path large-deviations theory for studying stationary overflow probabilities (see [23, Chapter 6] and [13]). Such a sample-path LDP analysis typically takes the following two steps. In the first step, for a finite T, we start the system empty at time 0, and study the probability that $\vec{X}(BT)$ exceeds the threshold B. We will find the large-deviations decay rate of this overflow probability when $B \to \infty$ (see Propositions 1 and 2 in Section V-A). Note that the corresponding scaled version of the stochastic processes $\mathbf{a}^B, \mathbf{s}^B$ and $\mathbf{x}^B$ will be over a finite interval $[0, T]$ as we defined earlier. Then, in the second step, we increase T and take the infimum of the decay-rate over all T (see Theorems 3 and 4 in Secton V-B). The intuition that this two-step process may work is as follows. For any fixed B, the larger T is, the less $\vec{X}(BT)$ will depend on the initial condition at time 0 (provided that the system is stable). Hence, by taking $T \to \infty$, we hope to capture the stationary overflow probability of $\vec{X}(\infty)$. However, we caution that it requires additional technical conditions and steps in order to establish this intuition rigorously. For example, if the system was not stable, such passing of the limit $T \to \infty$ would not produce the right results (see the example at the beginning of Section V-B). Technically speaking, a switch of limit is involved here: To study the stationary overflow probability, we need to take $T \to \infty$ first, and then study the asymptotic decay rate when $B \to \infty$; In contrast, the above two-step process takes $B \to \infty$ first for a finite T, and then take $T \to \infty$. In order to prove that the second step is indeed correct, we need to use the Freidlin-Wentzell theory (see [23, Chapter 6] and [13]), which is accomplished by the (rather technical) proof of Theorem 4 in the Appendix. Roughly speaking, the key reasons that the argument holds are because (i) the arrivals, the channel, and the service rates are all assumed to be bounded, and (ii) the system is assumed to be stable. We refer interested readers to more detailed discussions at the beginning of Section V-B and the Appendix.

IV. Main Results and Implications

For the convenience to the readers, in this section we list our main results along with the assumptions that are necessary for the results to hold. We then discuss the applications of these main results. The detailed proofs will be presented in Sections V-VII.

A. A List of Main Assumptions and Results

1) Main Assumptions: The first set of assumptions basically assume that the system has a Lyapunov function with negative-drift, and thus must be stable. Note that these assumptions are mild because most wireless control algorithms that can achieve provable capacity regions have well-known Lyapunov functions.

Assumption 1: For the system being studied, there exists a Lyapunov function $V(\vec{x})$, defined for $\vec{x} \geq 0$, that satisfies the following:

(a) $V(\vec{x})$ is a continuous function of $\vec{x}$.
(b) $V(\vec{x}) \geq 0$ for all $\vec{x}$ and $V(\vec{x}) = 0$ if and only if $\vec{x} = 0$.
(c) $V(\vec{x}) \to \infty$ if $||\vec{x}|| \to \infty$.
(d) $\min_{||\vec{x}||\geq 1} V(\vec{x}) \geq 1$. Further, there exists a number $\tilde{C}$ such that $\max_{||\vec{x}||\leq 1} V(\vec{x}) \leq \tilde{C}$.
(e) For any $\mathcal{B} > 0$, there exists a constant $\mathcal{L}$ that may depend on $\mathcal{B}$, such that for any $||\vec{x}_1|| \leq \mathcal{B}$ and $||\vec{x}_2|| \leq \mathcal{B}$,
$$|V(\vec{x}_1) - V(\vec{x}_2)| \leq \mathcal{L}||\vec{x}_1 - \vec{x}_2||.$$
(f) Either of the following holds (for a fixed arrival rate $\vec{\lambda}$ and a fixed channel state distribution $\vec{p}$ assumed in the system model). For all fluid limits $\mathbf{x}$,
$$\frac{d}{dt}V(\vec{x}(t)) \triangleq \left(\frac{\partial V}{\partial \vec{x}}\right)^T \frac{d\vec{x}}{dt} \leq -\eta V^\alpha(\vec{x}(t)) \qquad (14)$$
for almost all t such that $V(\vec{x}(t)) > 0$, where $0 < \alpha < 1$ and η is a positive constant. Or, for all fluid limits $\mathbf{x}$,
$$\frac{d}{dt}V(\vec{x}(t)) \triangleq \left(\frac{\partial V}{\partial \vec{x}}\right)^T \frac{d\vec{x}}{dt} \leq -\eta \qquad (15)$$
for almost all t such that $V(\vec{x}(t)) > 0$, where η is a positive constant.

Remark: Parts (a)-(c) and (f) of the assumption are typical when one uses Lyapunov functions to establish stability. Although Parts (d) and (e) are not standard, with a proper scaling of the Lyapunov function they will hold for many Lyapunov functions that have been used for wireless systems. Specifically, Part (d) holds (after proper scaling) when $\max_{||\vec{x}||\leq C} V(\vec{x})$ is upper bounded for some constant $C > 0$, and Part (e) holds when the Lyapunov function has bounded gradients in any finite set. To see how these conditions imply the stability of the fluid limit, note that starting from any initial $\vec{x}(0)$ with $||\vec{x}(0)|| = 1$, we must have $||V(\vec{x}(0))|| \leq \tilde{C}$ from Part (d) of the assumption. Then, using the drift condition (f), we can find a value of T such that for all fluid limits with $||V(\vec{x}(0))|| \leq \tilde{C}$, we must have $V(\vec{x}(T)) = 0$. Using Part (b), this then implies that $\vec{x}(T) = 0$. Hence, the fluid limit model of the system is stable. By [18], [19], it implies that the original system is positive Harris recurrent. Note that the two drift conditions in (14) and (15) are in fact equivalent. If $V(\cdot)$ satisfies (14), then $U(\vec{x}) = \frac{V^{1-\alpha}(\vec{x})}{1-\alpha}$ satisfies (15).

Remark: For part (f) we would like to point out a small but important difference between Lyapunov functions for the original discrete-time systems and Lyapunov functions for fluid-limit models. In part (f), we state the negative-drift assumption using Lyapunov functions for the fluid-limit model. In the literature, e.g. [1], [20], the negative-drift property is

often established through a Lyapunov function for the original discrete-time system. Extra care may be required to prove the negative-drift property for the fluid limit as needed by part (f). In the example in Section VIII, we will illustrate how to establish part (f).

The second set of assumptions, which slightly strengthens Assumption 1, is needed to establish the large deviations of the stationary overflow probability (in particular, the Freidlin-Wentzell construction in the Appendix.)

Assumption 2: If the Lyapunov function $V(\cdot)$ satisfies (14), then we further assume that

(a) There exists $\epsilon > 0$ such that for all FSP$(\mathbf{s}, \mathbf{a}, \mathbf{x})_T$ and for all time t such that $V(\vec{x}(t)) > 0$, if $||\frac{d}{dt}\vec{s}(t) - \vec{p}|| \leq \epsilon$ and $||\frac{d}{dt}\vec{a}(t) - \vec{\lambda}|| \leq \epsilon$, the following holds:

$$\frac{d}{dt}V(\vec{x}(t)) \leq -\frac{\eta}{2}V^{\alpha}(\vec{x}(t)), \tag{16}$$

where $0 < \alpha < 1$ and $\eta > 0$ are the same constants as in (14).

(b) For any $\delta > 0$, there exists $M_1 \geq 0$ such that for all FSP$(\mathbf{s}, \mathbf{a}, \mathbf{x})_T$ and for all time t such that $V(\vec{x}(t)) > 0$, if $||\frac{d}{dt}\vec{s}(t) - \vec{p}|| \geq \delta$ or $||\frac{d}{dt}\vec{a}(t) - \vec{\lambda}|| \geq \delta$, the following holds,

$$\frac{d}{dt}V(\vec{x}(t)) \leq M_1 V^{\alpha}(\vec{x}(t)). \tag{17}$$

On the other hand, if the Lyapunov function $V(\cdot)$ satisfies (15), then we further assume that

(a) There exists $\epsilon > 0$ such that for all FSP$(\mathbf{s}, \mathbf{a}, \mathbf{x})_T$ and for all time t such that $V(\vec{x}(t)) > 0$, if $||\frac{d}{dt}\vec{s}(t) - \vec{p}|| \leq \epsilon$ and $||\frac{d}{dt}\vec{a}(t) - \vec{\lambda}|| \leq \epsilon$, the following holds:

$$\frac{d}{dt}V(\vec{x}(t)) \leq -\frac{\eta}{2}, \tag{18}$$

where $\eta > 0$ is the same constant as in (15).

(b) For any $\delta > 0$, there exists $M_1 \geq 0$ such that for all FSP$(\mathbf{s}, \mathbf{a}, \mathbf{x})_T$ and for all time t such that $V(\vec{x}(t)) > 0$, if $||\frac{d}{dt}\vec{s}(t) - \vec{p}|| \geq \delta$ or $||\frac{d}{dt}\vec{a}(t) - \vec{\lambda}|| \geq \delta$, the following holds,

$$\frac{d}{dt}V(\vec{x}(t)) \leq M_1. \tag{19}$$

Remark: We will need these assumptions when we study the large-deviations properties of the stationary queue-overflow probability. Essentially, they imply that the drift behaves nicely not only for the fluid limits but also for FSP. Specifically, (a) even if we perturbed the channel distribution and the distribution of the arrival process slightly from $(\vec{p}, \vec{\lambda})$, the drift of the Lyapunov function still remains negative; (b) if the perturbation is large, although the drift could become positive, it is upper-bounded by a constant M_1 (or this constant multiplied by $V^{\alpha}(\vec{x}(t))$). Again, note that the two parts of Assumption 2 are also equivalent. If $V(\cdot)$ satisfies the first part of the assumption, then $U(\vec{x}) = \frac{V^{1-\alpha}(\vec{x})}{1-\alpha}$ satisfies the latter part. We will provide an example in Section VIII how these conditions can be easily verified.

2) A General Lower Bound on the Decay-Rate of the Queue-Overflow Probability: In Theorems 3 and 4 (which will be shown in Section V), under the assumption that there exists a Lyapunov function $V(\cdot)$ that satisfies Assumptions 1 and 2, we establish an upper bound on the overflow probabilities and hence a lower bound on their large deviations decay rate. The bound is expressed in terms of a multi-dimensional calculus-of-variations problem of finding the minimum-cost-to-overflow. Recall that $\vec{X}(\infty)$ denotes the stationary distribution of the stochastic process $\vec{X}(\cdot)$.

Theorem 3 Assume that there exists a Lyapunov function $V(\cdot)$ that satisfies Assumptions 1 and 2. Then the following holds,

$$\begin{aligned} &\limsup_{B\to\infty} \frac{1}{B}\log \mathbf{P}[||\vec{X}(\infty)/B|| \geq 1] \\ \leq -\inf_{T\geq 0,\mathbf{s},\mathbf{a},\mathbf{x}} \quad &\int_0^T \left[H\left(\frac{d}{dt}\vec{s}(t)||\vec{p}\right) + L\left(\frac{d}{dt}\vec{a}(t)\right)\right] dt \\ \text{subject to} \quad &(\mathbf{s},\mathbf{a},\mathbf{x})_T \text{ is an FSP} \\ &\vec{x}(0) = 0, ||\vec{x}(T)|| \geq 1. \end{aligned} \tag{20}$$

Note that the right-hand-side of (20) takes the infimum not only over all FSP's $(\mathbf{s}, \mathbf{a}, \mathbf{x})_T$, but also over all $T > 0$.

The result again reflects the large deviations philosophy that "rare events occur in the most likely way". The FSP that attains the infimum on the right-hand-side of (20) (if such an FSP exists) is usually called the "most-likely path to overflow," and the corresponding infimum is called the "minimum cost to overflow." Theorem 3 states that the decay-rate of the queue-overflow probability is lower bounded by the cost of the most-likely path to overflow. In the standard large-deviations literature, a technique called the "contraction principle" is often used to establish a result like this. However, to apply the contraction principle one must first establish that the mapping from $\vec{s}(t)$ and $\vec{a}(t)$ to $\vec{x}(t)$ is continuous with respect to properly-chosen topologies for the corresponding functional spaces. Unfortunately, for the general models of wireless networks and control algorithms that we are interested in, it seems difficult to establish the required continuity. The significance of Theorem 3 is that, as far as a lower bound on the decay-rate is concerned, one does not even need continuity of the above mapping. We note that Theorem 3 is comparable to Theorem 7.1 of [13] for a refined LDP.

The next theorem is similar to Theorem 3 except that the overflow metric is changed from $||\cdot||$ to $V(\cdot)$.

Theorem 4 Assume that there exists a Lyapunov function $V(\cdot)$ that satisfies both Assumption 1 and Assumption 2. Then the following holds,

$$\begin{aligned} &\limsup_{B\to\infty} \frac{1}{B}\log \mathbf{P}[V(\vec{X}(\infty)/B) \geq 1] \\ \leq -\inf_{T\geq 0,\mathbf{s},\mathbf{a},\mathbf{x}} \quad &\int_0^T \left[H\left(\frac{d}{dt}\vec{s}(t)||\vec{p}\right) + L\left(\frac{d}{dt}\vec{a}(t)\right)\right] dt \\ \text{subject to} \quad &(\mathbf{s},\mathbf{a},\mathbf{x})_T \text{ is an FSP} \\ &\vec{x}(0) = 0, V(\vec{x}(T)) \geq 1. \end{aligned} \tag{21}$$

3) A Much Simpler Lower Bound on the Decay Rate of the Queue Overflow Probability: Unfortunately, solving the minimum-cost-to-overflow in (20) and (21) is a difficult multi-dimensional calculus-of-variations problem. The following

Theorem 5, which is the first main result of the paper and will be shown in Section VI, provides a much simpler lower bound on the large-deviations decay-rate. The key idea is to use the Lyapunov function $V(\vec{x})$ to map the multi-dimensional calculus-of-variations problem with respect to $\vec{x}(t)$ to a one-dimensional calculus-of-variations problem with respect to $V(t) = V(\vec{x}(t))$. Specifically, let

$$\begin{aligned} l_V(v,w) \triangleq \inf_{\mathbf{s},\mathbf{a},\mathbf{x}} \quad & H(\vec{\phi}||\vec{p}) + L(\vec{f}) \\ \text{subject to} \quad & (\mathbf{s},\mathbf{a},\mathbf{x}) \text{ is an FSP} \\ & \text{such that for some } t \\ & \frac{d}{dt}\vec{s}(t) = \vec{\phi} \\ & \frac{d}{dt}\vec{a}(t) = \vec{f} \\ & V(\vec{x}(t)) = v \\ & \frac{d}{dt}V(\vec{x}(t)) = w. \end{aligned}$$

Note that the quantity $l_V(v,w)$ can be viewed as the smallest *local* cost (among all FSPs) for the Lyapunov function $V(t) = V(\vec{x}(t))$ to pass the value $V(t) = v$ with the slope w at a given time t. Readers can refer to Section VI for details.

The following one-dimensional calculus-of-variations problem then finds a lower bound on the minimum cost for $V(t)$ to overflow. Define

$$\begin{aligned} \theta_0 = \inf_{T>0} \quad & \int_0^T l_V(V(t), \frac{d}{dt}V(t))dt \\ \text{subject to} \quad & V(t) \text{ is continuous and} \\ & \text{almost-everywhere differentiable,} \\ & V(0) = 0 \text{ and } V(T) \geq 1. \end{aligned} \tag{22}$$

Theorem 5 Assume that there exists a Lyapunov function $V(\cdot)$ that satisfies Assumption 1 and Assumption 2. Then θ_0 in (22) is a lower bound on the decay-rate of the queue-overflow probability. In other words,

$$\begin{aligned} &\limsup_{B\to\infty} \frac{1}{B}\log \mathbf{P}[||\vec{X}(\infty)/B|| \geq 1] \\ &\quad \leq \limsup_{B\to\infty} \frac{1}{B}\log \mathbf{P}[V(\vec{X}(\infty)/B) \geq 1] \leq -\theta_0. \end{aligned}$$

The lower bound θ_0 in (22) is simpler than (21) because we now only need to solve a *one-dimensional* calculus-of-variations problem. This lower bound can be further simplified under the following scale-linearity assumption on the Lyapunov function.

Assumption 3: The Lyapunov function $V(\cdot)$ is linear in scale, i.e., $V(c\vec{x}) = cV(\vec{x})$ for all $c \geq 0$.

Under Assumption 3, we will show that

$$\begin{aligned} \theta_0 \quad = \quad & \inf_{w>0,\mathbf{s},\mathbf{a},\mathbf{x}} \frac{1}{w}\left[H(\vec{\phi}||\vec{p}) + L(\vec{f})\right] \\ \text{subject to} \quad & (\mathbf{s},\mathbf{a},\mathbf{x}) \text{ is an FSP} \\ & \text{such that for some } t \\ & \frac{d}{dt}\vec{s}(t) = \vec{\phi} \\ & \frac{d}{dt}\vec{a}(t) = \vec{f} \\ & V(\vec{x}(t)) = 1 \\ & \frac{dV(\vec{x}(t))}{dt} = w. \end{aligned} \tag{23}$$

The details are provided in Section VI-A. The lower bound θ_0 is even simpler because, unlike the calculus-of-variations problems in (20) and (21) where one must search for the path to overflow with the *global* minimum cost, in θ_0 we only focus on the *local* cost and the dynamics of the path at some arbitrary time t.

4) A Condition for the Minimum-Cost-to-Overflow to be Exact: Our second set of main results concern with the conditions under which the above lower bounds become tight. Note that many scheduling and routing policies are designed to minimize the drift of the respective Lyapunov function, as stated in the following assumption. For any $\vec{f}, \vec{\phi}$ and $\vec{e}$, let $\delta_i^k = f_i^k - \sum_{j=1}^S \phi_j \sum_{l=1}^L R_{il} e_{lj}^k$. Define

$$\tilde{V}(\tau, \vec{e}|\vec{x}, \vec{\phi}, \vec{f}) \triangleq V([\vec{x} + \vec{\delta}\tau]^+). \tag{24}$$

Then $\frac{\partial}{\partial\tau}\tilde{V}(\tau, \vec{e}|\vec{x}, \vec{\phi}, \vec{f})\Big|_{\tau=0}$ can be viewed as the drift of the Lyapunov function from $\vec{x}(t) = \vec{x}$ if the service-rate vector is chosen as $\vec{e}$, conditioned on that the channel state process satisfies $\frac{d}{dt}\vec{s}(t) = \vec{\phi}$ and the arrival process satisfies $\frac{d}{dt}\vec{a}(t) = \vec{f}$. Recall that throughout this paper, we use right-derivatives unless otherwise stated.

Assumption 4: For any FSP$(\mathbf{s},\mathbf{a},\mathbf{x})$, the following holds for all t:

$$\frac{d}{dt}V(\vec{x}(t)) = \min_{e_{lj}^k \in \text{Conv}(\mathcal{E}_j)} \frac{\partial}{\partial\tau}\tilde{V}(\tau, \vec{e}|\vec{x}(t), \vec{\phi}(t), \vec{f}(t))\Big|_{\tau=0}. \tag{25}$$

where $\vec{\phi}(t) = \frac{d}{dt}\vec{s}(t)$, $\vec{f}(t) = \frac{d}{dt}\vec{a}(t)$.

This assumption states that at any point t of an FSP$(\mathbf{s},\mathbf{a},\mathbf{x})$, the scheduling and routing algorithm minimizes the drift of the Lyapunov function over all possible decisions.

In addition, we assume the following for the Lyapunov function.

Assumption 5: $V(\vec{x})$ is increasing in each component x_i.

Assumption 6: $V(\vec{x}_1 + \vec{x}_2) \leq V(\vec{x}_1) + V(\vec{x}_2)$ for any two vectors $\vec{x}_1 \geq 0$ and $\vec{x}_2 \geq 0$,

Note that Assumptions 3 and 6 combined imply that the Lyapunov function $V(\vec{x})$ almost behaves as a norm except that it may not be defined for negative values of the variable $\vec{x}$.

In the next main result Theorem 8 (which will be shown in Section VII), we prove that if the Lyapunov function $V(\cdot)$ satisfies Assumptions 1, 2, 3, 4, 5 and 6, then the scheduling and routing algorithm achieves the best possible large-deviations

decay rate of the overflow probability $\mathbf{P}(V(\vec{X}(\infty)) \geq B)$ as $B \to \infty$, and the lower bound θ_0 is tight.

Theorem 8 Suppose that a scheduling and routing policy satisfies Assumptions 1, 2, 3, 4, 5 and 6. Then under this policy the value of θ_0 (given in (23)) is the exact decay-rate of the overflow probability according to the Lyapunov function metric, i.e.,

$$\lim_{B\to\infty} \frac{1}{B} \log \mathbf{P}[V(\vec{X}(\infty)/B) \geq 1] = -\theta_0. \qquad (26)$$

Further, this drift-minimizing policy (according to Assumption 4) is optimal in maximizing this decay-rate. In other words, for any policy π we must have

$$\liminf_{B\to\infty} \frac{1}{B} \log \mathbf{P}^{\pi}[V(\vec{X}(\infty)/B) \geq 1] \geq -\theta_0, \qquad (27)$$

where $\mathbf{P}^{\pi}$ denote the stationary distribution under the policy π.

Remark: Although the conclusion of Theorem 8 may seem very intuitive, we emphasize that it is not an obvious result. Minimizing the drift of the Lyapunov function at a given time instant t is a *local* and *myopic* property. Minimizing the probability that the Lyapunov function overflows is a *global* property. It is not uncommon for a myopic policy to only attain suboptimal global performance. Hence, the fact that Theorem 8 holds is in fact quite remarkable. A main contribution of this paper is to quantify the precise conditions for Theorem 8 to hold. As readers can see from Assumptions 3 and 6, the shape of the Lyapunov function is very important.

B. Applications of the Main Results

Before we present the proofs of these results, we would like to use some examples to illustrate their significance in applications. We will focus on Theorem 8 since it is the most convenient result to use.

1) Analysis: Firstly, Theorem 8 can be very useful for analyzing the overflow probabilities of known scheduling and routing algorithms because many known scheduling and routing algorithms are designed by minimizing the drift of a Lyapunov function (in the fluid limit). For example, consider a single cell in a cellular network or an access-point based network. Each user communicates directly with the base-station. Further, only one user can be selected for service at a time. Let us focus on the downlink from the base-station to the users (the uplink can be treated analogously). Since this is a single-hop model, to map to the system model in Section II we can identify each link l with a particular user/class, and hence we can drop the index k for traffic class. In other words, we use $A_l(\tau)$ to denote the packets generated for the user associated with link l at time slot τ, and use $X_l(\tau)$ to denote the backlog of link l at time τ. Let $E_l(j, \vec{X})$ be the service offered to link l when the channel state is j and the global backlog is $\vec{X} = [X_l, l = 1, ..., L]$. Imposing the constraint that $E_l(j, \vec{X}) \leq X_l$, the evolution of the queue-length is then given by

$$X_l(\tau+1) = X_l(\tau)+A_l(\tau)-\sum_{j=1}^{S} \mathbf{1}_{\{C(\tau)=j\}} E_l(j, \vec{X}(\tau)). \quad (28)$$

Note that this equation is a simplified version of (1).

When the channel state is j, let r_{lj} denote the capacity of link l if it is selected for transmission. Then for each state j the service-rate vector $[E_l(j, \vec{X}), l = 1, ..., L]$ must belong to the set $\mathcal{E}_j$ given by

$$\mathcal{E}_j = \{[t_l r_{lj}, l = 1, ..., L] : t_l \in [0, 1] \text{ for all } l, \text{ and only one element of } [t_l] \text{ is non-zero}\}.$$

Max-Weight Scheduling Policy: One important policy for choosing $E_l(j, \vec{X})$ (often referred to as the *max-weight* policy in the literature [1], [21]), is to serve the link l such that the weighted queue-length $r_{lj}X_l$ is the largest among all users (ties can be broken arbitrarily). Let $l^* = \text{argmax}_l\, r_{lj}X_l$. The policy can then be written as:

$$E_l(j, \vec{X}) = \begin{cases} \min\{r_{lj}, X_l\}, & \text{if } l = l^* \\ 0, & \text{otherwise,} \end{cases}$$

It is well-known that this class of policies are throughput-optimal, i.e., they can stabilize the system under the largest set of offered loads [21]. Further, if we instead choose $l^* = \text{argmax}_l\, r_{lj}X_l^{\alpha}$, $\alpha > 0$ (note that the queue length is raised to the power α), the corresponding policy (referred to as the α-algorithms in [22]) is also throughput-optimal.

It can be shown that any α-algorithm ($\alpha > 0$) minimizes the drift of the Lyapunov function $V_\alpha(\vec{X}) = ||\vec{X}||_{\alpha+1}$ at every time in an FSP, where $||\cdot||_{\alpha+1}$ denotes the $(\alpha+1)$-norm [22]. Using Theorem 8, we can immediately draw the conclusion that, for any $\alpha > 0$, the α-algorithm is large-deviations decay-rate optimal for minimizing the overflow probability $\mathbf{P}(||\vec{X}(\infty)||_{\alpha+1} \geq B)$ (As a special case, the standard max-weight algorithm, i.e., $\alpha = 1$, is large-deviations decay-rate optimal for the overflow probability $\mathbf{P}(||\vec{X}(\infty)||_2 \geq B)$). We can see how easily Theorem 8 can be used to analyze large-deviations optimality of known algorithms. In Section VIII, we will provide another detailed example where we show that the max-weight algorithm is large deviations decay-rate optimal in terms of the overflow probability of the max-queue ($\mathbf{P}(||\vec{X}(\infty)||_\infty \geq B)$) when we consider a special case with on-off channels. We will further show how to use this result to characterize the effective capacity of the system.

Theorem 8 is not only useful for single-hop networks, it also applies to multihop networks. Note that unlike the previous cellular example where only one user/link can be selected at each time, in multi-hop wireless networks it is often possible to activate multiple links simultaneously. Depending on the interference and transmission model, the activated links must satisfy certain interference constraints, and their rates depend on the power and interference levels. We give two special cases of the model in Section II.

Case 1: Assume that the scheduling policy can decide whether to activate or inactivate a link, but cannot change the transmission power of a link. Let $\pi_l = 1$ if link l is activated, and $\pi_l = 0$, otherwise. Let $\vec{\pi} = [\pi_l, l = 1, ..., L]$, and let Π denote the set of feasible activation vectors $\vec{\pi}$. Let $r_{lj}(\vec{\pi})$ denote the rate of link l if the activation vector $\vec{\pi}$ is applied at state j. Then at each channel state j, the set $\mathcal{E}_j$ of feasible

service-rate vectors can be written as:

$$\mathcal{E}_j = \{[r_{lj}, l = 1, ..., L] : \text{ there exists } \vec{\pi} \in \Pi \text{ such that } r_{lj} \leq r_{lj}(\vec{\pi}) \text{ for all links } l\}. \quad (29)$$

Case 2: Assume that the scheduling policy can decide both the activation pattern and the power assignments. Let π_l denote the power assignment of link l. $\pi_l = 0$ if the link is not activated. Let $\vec{\pi} = [\pi_l, l = 1, ..., L]$, and let Π denote the set of feasible power-assignment vectors $\vec{\pi}$. Then at each channel state j, each vector $\vec{\pi}$ can again be mapped to a rate-vector $[r_{lj}(\vec{\pi}), l = 1, ..., L]$. The set $\mathcal{E}_j$ of feasible service-rate vectors can also be written as in (29).

The Back-Pressure Algorithm: For both cases, the scheduling and routing algorithm proposed in [1], [20], which is often referred to as the "back-pressure" algorithm, is known to be throughput-optimal. At each time-slot τ, for each link l, first find the class $k_l^*(\tau)$ with the maximum differential backlog, i.e.,

$$k_l^*(\tau) = \operatorname*{argmax}_k (X^k_{b(l)}(\tau) - X^k_{e(l)}(\tau)).$$

Let the corresponding differential backlog be

$$w_l(\tau) = \max\{0, X^{k_l^*(\tau)}_{b(l)}(\tau) - X^{k_l^*(\tau)}_{e(l)}(\tau)\}.$$

Then, when the channel state at time τ is j, compute the schedule $\vec{\pi}_j^*(\tau)$ that maximizes the sum of the rates weighted by $w_l(\tau)$, i.e.,

$$\vec{\pi}_j^*(\tau) = \operatorname*{argmax}_{\vec{\pi} \in \Pi} w_l(\tau) r_{lj}(\vec{\pi}).$$

The scheduling and routing decision is then given by the following: if the channel state at time-slot τ is j,

- *Scheduling:* use the activation vector $\vec{\pi}_j^*(\tau)$.
- *Routing:* on each link l, only serve the packets belonging to class $k_l^*(\tau)$.

In other words, the service rate vector is given by

$$E_l^k(j, \vec{X}) = \begin{cases} \min\{X^k_{b(l)}(\tau), r_{lj}(\vec{\pi}_j^*(\tau))\}, & \text{if } k = k_l^*(\tau) \\ 0 & \text{otherwise.} \end{cases}$$

It can be shown that the Back-Pressure algorithm minimizes the drift of the Lyapunov function $V(\vec{x}) = ||\vec{x}||_2$ at every time in an FSP [14]. Hence, using Theorem 8, we can immediately draw the conclusion that the Back-Pressure algorithm is large-deviations decay-rate optimal for the overflow probability $\mathbf{P}(||\vec{X}(\infty)||_2 \geq B)$. Similarly, if we replace X_i^k by $(X_i^k)^\alpha$, $\alpha > 0$, in the definition of the differential backlog, the modified Back-Pressure algorithm with parameter α is large-deviations decay-rate optimal for the overflow probability $\mathbf{P}(||\vec{X}(\infty)||_{\alpha+1} \geq B)$.

2) Design: Theorem 8 is not only useful for analysis, it is also useful for designing new large-deviations optimal scheduling algorithms. Suppose that we are interested in minimizing the probability that $\mathbf{P}(f(\vec{X}) \geq B)$. First, if the function $f(\cdot)$ can be shown to be a Lyapunov function of the system, and we can find an algorithm that minimizes its drift at every time in fluid-sample-paths, then this algorithm is exactly the large-deviations decay-rate optimal algorithm that we need. The detailed example in Section VIII falls into this category. Secondly, even if $f(\cdot)$ is not a Lyapunov function of the system, we can use other Lyapunov functions to approximate it. Then the corresponding drift-minimizing algorithm is approximately the algorithm that we need. For example, suppose that we are interested in the overflow probability $\mathbf{P}(\max_l X_l \geq B)$ in the cellular downlink example in Section IV-B1. For non-ON-OFF models, $\max_l X_l$ is usually not a Lyapunov function of the system. We can instead use $V_\alpha(\vec{X}) = ||\vec{X}||_{\alpha+1}$ with a large α to approximate it. Since each of the α-algorithms is large-deviations decay-rate optimal for the overflow probability $\mathbf{P}(||\vec{X}||_{\alpha+1} \geq B)$, and $||\vec{X}||_{\alpha+1} \to \max_l X_l$ as $\alpha \to \infty$, we can draw the conclusion that as $\alpha \to \infty$, the α-algorithm asymptotically achieves the optimal large-deviations decay-rate of the overflow probability $\mathbf{P}(\max_l X_l \geq B)$. This conclusion recovers the result in [22], where we also demonstrate how to use the insight to develop algorithms with low overflow-probabilities in practice. Finally, the above methodology also applies to multi-hop wireless networks. We refer the readers to our more recent work [14], [15] for further details.

V. A General Lower Bound on the Decay-Rate of the Queue-Overflow Probability

In this section, we first prove Theorems 3 and 4, which provide a lower bound $I_0'(\vec{\lambda})$ on the decay-rate of the queue-overflow probability as defined in (5). Prior work [10], [11] has derived the decay-rate of the queue-overflow probability as the cost of the most-likely-path to overflow. As we discussed in Section IV-A, the approach there requires that the limiting mapping from the scaled channel-state process to the scaled queue-backlog process be unique and continuous with respect to a suitably chosen topological space, so that the contraction principle [17, p131] can be invoked to establish a sample-path LDP for the queue-backlog process. However, for general wireless systems, the mapping from the channel-state process to the queue-backlog process may not be continuous, and hence the approach in [10], [11] can not be applied. Nonetheless, in this section we show that, for a large class of wireless systems, the cost of the most-likely-path to overflow turns out to be a lower bound on the decay-rate of the queue-overflow probability. Specifically, let

$$I_0'(\vec{\lambda}) = \inf_{T \geq 0, \mathbf{s}, \mathbf{a}, \mathbf{x}} \int_0^T \left[H\left(\frac{d}{dt}\vec{s}(t) || \vec{p}\right) + L\left(\frac{d}{dt}\vec{a}(t)\right)\right] dt$$
$$\text{subject to } (\mathbf{s}, \mathbf{a}, \mathbf{x})_T \text{ is an FSP}$$
$$\vec{x}(0) = 0, ||\vec{x}(T)|| \geq 1. \quad (30)$$

The FSP that attains the infimum on the right-hand-side of (30), if such an FSP exists, is usually called the "most-likely path to overflow." Our goal in this section to establish that $I_0'(\vec{\lambda})$ satisfies (5), i.e., it is a lower-bound on the decay rate of the stationary queue-overflow probability. Recall that such a lower bound implies that we can then use $\exp(-BI_0'(\vec{\lambda}))$ as an (approximate) upper bound on the overflow probability $\mathbf{P}[||\vec{X}(\infty)|| \geq B]$.

We will derive the result in two steps. First, in Section V-A we consider a system that starts at time 0, and derive a lower bound on the decay rate of the overflow probability at time BT. Then, in Section V-B we let $T \to \infty$ and derive a lower bound for the stationary distribution. In the literature, such a limiting argument is usually carried out using the so-called Freidlin-Wentzell theory (see [23, Chapter 6] and [13]). Often, to apply the Freidlin-Wentzell theory, one will need to impose additional restrictions on the system model [23, p133]. One of our contributions in this section is to provide a fairly general condition for this result to hold. As readers will see soon, our condition essentially requires that there exists a Lyapunov function that satisfies Assumptions 1 and 2, which are introduced in Section IV-A.

A. Bounds for a Finite Time System

As a step towards proving the result on the stationary overflow probability, we first consider the probability of overflow at time BT, $\mathbf{P}(||\vec{X}(BT)|| \geq B)$, for a system that starts from $\vec{X}(0) = 0$. Note that according to the transformation $\vec{x}^B(t) = \frac{1}{B}\vec{X}(Bt)$, we have $\vec{x}^B(0) = 0$ and the above overflow probability can be rewritten as $\mathbf{P}[||\vec{x}^B(T)|| \geq 1]$. Let $\mathbf{P}_0^{B,T}$ denote the probability distribution conditioned on $\vec{x}^B(0) = 0$. We have the following bound on the probability of overflow.

Proposition 1: Fix $T > 0$. The following holds,

$$\begin{aligned} &\limsup_{B\to\infty} \frac{1}{B}\log \mathbf{P}_0^{B,T}[||\vec{x}^B(T)|| \geq 1] \\ \leq -\inf_{\mathbf{s},\mathbf{a},\mathbf{x}} \quad &\int_0^T \left[H\left(\frac{d}{dt}\vec{s}(t)||\vec{p}\right) + L\left(\frac{d}{dt}\vec{a}(t)\right)\right] dt \\ \text{subject to} \quad &(\mathbf{s},\mathbf{a},\mathbf{x})_T \text{ is an FSP} \\ &\vec{x}(0) = 0 \text{ and } ||\vec{x}(T)|| \geq 1. \end{aligned} \tag{31}$$

Instead of proving Proposition 1, we will prove a generalized version in Proposition 2. The extra effort will serve useful in proving the stationary overflow probability.

Fix $\vec{x}_0$. For the more general version, consider a system that starts with $\vec{X}(0) = B\vec{x}_0$ at time 0 (i.e., $\vec{x}^B(0) = \vec{x}_0$). Let $\mathbf{P}_{\vec{x}_0}^{B,T}$ denote the probability distribution conditioned on $\vec{x}^B(0) = \vec{x}_0$. Let $\Phi_x[0,T]$ denote the space of non-negative Lipschitz-continuous functions on the interval $[0,T]$, equipped with the essential supremum norm. We can then show the following result, which is comparable to Theorem 7.1 of [13] for a refined LDP.

Proposition 2: Fix $T > 0$. Let $\mathcal{X}$ denote a closed set in the NK-dimensional real space $\Re^{NK}$. Let Γ denote a closed set of trajectories $\mathbf{x} = (\vec{x}(t), t \in [0,T])$ from the topological space $\Phi_x[0,T]$ that satisfies $\vec{x}(0) \in \mathcal{X}$. The following holds,

$$\begin{aligned} &\limsup_{B\to\infty} \frac{1}{B}\log \sup_{\vec{x}_0\in\mathcal{X}} \mathbf{P}_{\vec{x}_0}^{B,T}[\mathbf{x}^B \in \Gamma] \\ \leq -\inf_{\mathbf{s},\mathbf{a},\mathbf{x}} \quad &\int_0^T \left[H\left(\frac{d}{dt}\vec{s}(t)||\vec{p}\right) + L\left(\frac{d}{dt}\vec{a}(t)\right)\right] dt \\ \text{subject to} \quad &(\mathbf{s},\mathbf{a},\mathbf{x})_T \text{ is an FSP} \\ &\mathbf{x} \in \Gamma. \end{aligned} \tag{32}$$

Proof: Note that $\mathbf{x}^B$ is related to $\mathbf{s}^B$ and $\mathbf{a}^B$ according to Equation (11). Let $\tilde{\Gamma}^B$ denote the set of all $(\mathbf{s}^B, \mathbf{a}^B)$ on the interval $[0,T]$ such that there exists $\vec{x}_0 \in \mathcal{X}$ with which $(\mathbf{s}^B, \mathbf{a}^B)$ maps to a backlog process $\mathbf{x}^B$ that starts from $\vec{x}^B(0) = \vec{x}_0$ and satisfies $\mathbf{x}^B \in \Gamma$. Then for every $\vec{x}_0 \in \mathcal{X}$,

$$\mathbf{P}_{\vec{x}_0}^{B,T}[\mathbf{x}^B \in \Gamma] \leq \mathbf{P}[(\mathbf{s}^B, \mathbf{a}^B) \in \tilde{\Gamma}^B].$$

Note that the right-hand-side does not depend on $\vec{x}_0$. Further, for any fixed n, $\tilde{\Gamma}^B \subset \cup_{B'=n}^{\infty}\tilde{\Gamma}^{B'}$ when $B \geq n$. Hence, we have, for any fixed n,

$$\begin{aligned} &\limsup_{B\to\infty} \frac{1}{B}\log \sup_{\vec{x}_0\in\mathcal{X}} \mathbf{P}_{\vec{x}_0}^{B,T}[\mathbf{x}^B \in \Gamma] \\ &\qquad \leq \limsup_{B\to\infty} \frac{1}{B}\log \mathbf{P}[(\mathbf{s}^B, \mathbf{a}^B) \in \cup_{B'=n}^{\infty}\tilde{\Gamma}^{B'}]. \end{aligned}$$

Using the sample-path LDP of $\mathbf{s}^B$ and $\mathbf{a}^B$ (see (8)) and the fact that they are independent, we have

$$\begin{aligned} &\limsup_{B\to\infty} \frac{1}{B}\log \mathbf{P}[(\mathbf{s}^B, \mathbf{a}^B) \in \cup_{B'=n}^{\infty}\tilde{\Gamma}^{B'}] \\ \leq -\inf_{(\mathbf{s},\mathbf{a})} \quad &\int_0^T \left[H\left(\frac{d}{dt}\vec{s}(t)||\vec{p}\right) + L\left(\frac{d}{dt}\vec{a}(t)\right)\right] dt, \\ \text{subject to} \quad &(\mathbf{s},\mathbf{a}) \in \overline{\cup_{B=n}^{\infty}\tilde{\Gamma}^B} \end{aligned}$$

where $\overline{\cup_{B=n}^{\infty}\tilde{\Gamma}^B}$ denotes the closure of the set $\cup_{B=n}^{\infty}\tilde{\Gamma}^B$. Note that this inequality holds for all n. Further, since the set $\overline{\cup_{B=n}^{\infty}\tilde{\Gamma}^B}$ is decreasing in n, the right-hand-side is decreasing in n as well. Therefore, we can tighten the bound by letting $n \to \infty$ as follows. To simplify notation, for any $(\mathbf{s},\mathbf{a})$, define its cost by

$$J_T(\mathbf{s},\mathbf{a}) \triangleq \int_0^T \left[H\left(\frac{d}{dt}\vec{s}(t)||\vec{p}\right) + L\left(\frac{d}{dt}\vec{a}(t)\right)\right] dt.$$

We then have,

$$\begin{aligned} &\limsup_{B\to\infty} \frac{1}{B}\log \sup_{\vec{x}_0\in\mathcal{X}} \mathbf{P}_{\vec{x}_0}^{B,T}[\mathbf{x}^B \in \Gamma] \\ &\qquad \leq -\lim_{n\to\infty} \inf_{(\mathbf{s},\mathbf{a})\in\overline{\cup_{B=n}^{\infty}\tilde{\Gamma}^B}} J_T(\mathbf{s},\mathbf{a}). \end{aligned}$$

Let

$$Y = \lim_{n\to\infty} \inf_{(\mathbf{s},\mathbf{a})\in\overline{\cup_{B=n}^{\infty}\tilde{\Gamma}^B}} J_T(\mathbf{s},\mathbf{a}). \tag{33}$$

It remains to show that

$$\begin{aligned} Y \geq \inf_{\mathbf{s},\mathbf{a},\mathbf{x}} \quad &J_T(\mathbf{s},\mathbf{a}) \\ \text{subject to} \quad &(\mathbf{s},\mathbf{a},\mathbf{x})_T \text{ is an FSP} \\ &\mathbf{x} \in \Gamma. \end{aligned}$$

To see this, note that by (33), there must exist a sequence $(\mathbf{s}_n, \mathbf{a}_n), n = 1, 2, ...$ such that

$$(\mathbf{s}_n, \mathbf{a}_n) \in \overline{\cup_{B=n}^{\infty}\tilde{\Gamma}^B} \text{ for all } n,$$

and

$$\lim_{n\to\infty} J_T(\mathbf{s}_n, \mathbf{a}_n) = Y.$$

Since both $\mathbf{s}_n$ and $\mathbf{a}_n$ are non-decreasing and Lipschitz-continuous, there must exist a further subsequence that converges uniformly over compact intervals. Without loss of

generality, we can abuse notation and denote this subsequence also as $(\mathbf{s}_n, \mathbf{a}_n)$, and let $(\underline{\mathbf{s}}, \underline{\mathbf{a}})$ be the corresponding limit. Then, due to the lower-semicontinuity of the large-deviation rate function $J_T(\cdot,\cdot)$ [17, p4], we must have

$$J_T(\underline{\mathbf{s}}, \underline{\mathbf{a}}) \leq Y.$$

We now show that we must then be able to find an FSP$(\underline{\mathbf{s}}, \underline{\mathbf{a}}, \underline{\mathbf{x}})_T$ with $\underline{\mathbf{x}} \in \Gamma$. To see this, note that by definition each $(\mathbf{s}_n, \mathbf{a}_n)$ also corresponds to a sequence $(\mathbf{s}_{n,m}, \mathbf{a}_{n,m})$ such that $(\mathbf{s}_{n,m}, \mathbf{a}_{n,m}) \in \cup_{B=n}^{\infty} \tilde{\Gamma}^B$ for all $m = 1, 2, ...$, and $(\mathbf{s}_{n,m}, \mathbf{a}_{n,m})$ converges to $(\mathbf{s}_n, \mathbf{a}_n)$ uniformly over compact intervals. Assign any sequence $\epsilon_n > 0, n = 1, 2, ...$, such that $\lim_{n\to\infty} \epsilon_n = 0$. For each n, we can then find an element $(\tilde{\mathbf{s}}_n^{B_n}, \tilde{\mathbf{a}}_n^{B_n})$ from the sequence $(\mathbf{s}_{n,m}, \mathbf{a}_{n,m})$ such that

$$||\tilde{\vec{s}}_n^{B_n}(t) - \vec{s}_n(t)||_\infty + ||\tilde{\vec{a}}_n^{B_n}(t) - \vec{a}_n(t)||_\infty \leq \epsilon_n,$$

and $(\tilde{\mathbf{s}}_n^{B_n}, \tilde{\mathbf{a}}_n^{B_n}) \in \tilde{\Gamma}^{B_n}$ for some $B_n \geq n$. Since the sequence $(\mathbf{s}_n, \mathbf{a}_n)$ converges to $(\underline{\mathbf{s}}, \underline{\mathbf{a}})$ uniformly over compact intervals, we must have that $(\tilde{\mathbf{s}}_n^{B_n}, \tilde{\mathbf{a}}_n^{B_n})$ also converges to $(\underline{\mathbf{s}}, \underline{\mathbf{a}})$ uniformly over compact intervals. Further, since each $(\tilde{\mathbf{s}}_n^{B_n}, \tilde{\mathbf{a}}_n^{B_n}) \in \tilde{\Gamma}^{B_n}$, there must exists a corresponding backlog process $\tilde{\mathbf{x}}_n^{B_n}$ such that $\tilde{\mathbf{x}}_n^{B_n} \in \Gamma$. Take a further subsequence of $(\tilde{\mathbf{s}}_n^{B_n}, \tilde{\mathbf{a}}_n^{B_n})$ such that the corresponding subsequence of $\tilde{\mathbf{x}}_n^{B_n}$ converges uniformly over compact intervals to a limiting backlog process $\underline{\mathbf{x}}$. Then, since the set Γ is closed, this limiting process $\underline{\mathbf{x}}$ must also satisfy $\underline{\mathbf{x}} \in \Gamma$. Hence, $(\underline{\mathbf{s}}, \underline{\mathbf{a}}, \underline{\mathbf{x}})_T$ is an FSP, and it satisfies the constraints used to define the right hand side of (32). We then have

$$\begin{aligned} Y \geq J_T(\underline{\mathbf{s}}, \underline{\mathbf{a}}) \geq \inf_{\mathbf{s},\mathbf{a},\mathbf{x}} \quad & J_T(\mathbf{s}, \mathbf{a}) \\ \text{subject to} \quad & (\mathbf{s}, \mathbf{a}, \mathbf{x})_T \text{ is an FSP} \\ & \mathbf{x} \in \Gamma. \end{aligned}$$

The result then follows. *Q.E.D.*

By setting $\mathcal{X} = \{0\}$, and

$$\Gamma = \{\mathbf{x} : \vec{x}(0) = 0 \text{ and } ||\vec{x}(T)|| \geq 1\},$$

we then recover the result of Proposition 1.

B. The Stationary Overflow Probability

Since Propositions 1 and 2 hold for any $T > 0$, one may then be tempted to let $T \to \infty$, and claim a lower bound on the large-deviations decay-rate of the stationary overflow probability $\mathbf{P}(||\vec{X}(\infty)|| \geq B)$. This argument, however, does not always hold. For example, consider a queueing system in which $||\vec{X}(t)||$ grows sub-linearly (e.g., $||\vec{X}(t)|| = \sqrt{t}$). For any finite $T > 0$, the probability of overflow $\mathbf{P}(||\vec{X}(BT)|| \geq B)$ will be 0 for $B > T$. Hence, the large-deviations decay-rate (as $B \to \infty$) for any finite $T > 0$ is ∞, i.e., $\lim_{B\to\infty} \frac{1}{B} \log \mathbf{P}_0^{B,T}[||\vec{x}^B(T)|| \geq 1] = -\infty$. However, since the system is clearly unstable, the "stationary" overflow probability is 1, and hence its decay-rate (as $B \to \infty$) is 0. Clearly, passing limit as $T \to \infty$ would not produce the correct large-deviations decay-rate of the stationary overflow probability. Fortunately, for our system model, due to Assumptions 1 and 2, the system must be stable and hence the above scenario cannot occur.

Specifically, we now use the Freidlin-Wentzell theory (see [23, Chapter 6] and [13]) to derive a lower bound on the decay-rate of the stationary queue-overflow probability. The following are the main results of the section.

Theorem 3: Assume that there exists a Lyapunov function $V(\cdot)$ that satisfies Assumptions 1 and 2. Then the following holds,

$$\begin{aligned} & \limsup_{B\to\infty} \frac{1}{B} \log \mathbf{P}[||\vec{X}(\infty)/B|| \geq 1] \\ \leq - \inf_{T\geq 0,\mathbf{s},\mathbf{a},\mathbf{x}} \quad & \int_0^T \left[H\left(\frac{d}{dt}\vec{s}(t)||\vec{p}\right) + L\left(\frac{d}{dt}\vec{a}(t)\right)\right] dt \\ \text{subject to} \quad & (\mathbf{s}, \mathbf{a}, \mathbf{x})_T \text{ is an FSP} \\ & \vec{x}(0) = 0, ||\vec{x}(T)|| \geq 1. \end{aligned} \tag{34}$$

This theorem provides a result similar to Proposition 1 but now is for the stationary overflow probability. Note that since Proposition 1 provides a lower bound for a finite time interval, the infimum in (31) is for a fixed T while the infimum in (34) is taken over all $T > 0$.

The proof of Theorem 3 is very similar to the proof of Theorem 4 that follows. Hence, in order to avoid repetition we omit its proof.

Theorem 4: Assume that there exists a Lyapunov function $V(\cdot)$ that satisfies both Assumption 1 and Assumption 2. Then the following holds,

$$\begin{aligned} & \limsup_{B\to\infty} \frac{1}{B} \log \mathbf{P}[V(\vec{X}(\infty)/B) \geq 1] \\ \leq - \inf_{T\geq 0,\mathbf{s},\mathbf{a},\mathbf{x}} \quad & \int_0^T \left[H\left(\frac{d}{dt}\vec{s}(t)||\vec{p}\right) + L\left(\frac{d}{dt}\vec{a}(t)\right)\right] dt \\ \text{subject to} \quad & (\mathbf{s}, \mathbf{a}, \mathbf{x})_T \text{ is an FSP} \\ & \vec{x}(0) = 0, V(\vec{x}(T)) \geq 1. \end{aligned} \tag{35}$$

Note that the statements of the two theorems are very similar. The difference is that Theorem 3 considers the overflow event $||\vec{X}(\infty)/B|| \geq 1$, whereas Theorem 4 considers the overflow event $V(\vec{X}(\infty)/B) \geq 1$. The importance of Theorem 4 will become clear in the later sections. Specifically, it is needed in the proof of Theorem 8.

The proof of Theorem 4 uses the Freidlin-Wentzell Theory. It is fairly technical and is provided in the Appendix. We emphasize that Theorems 3 and 4 provide a lower bound on the decay-rate of the stationary queue-overflow probability under very general assumptions.

VI. Combining Large Deviations with Lyapunov Functions: A Much-Simpler Lower Bound on the Decay-Rate

In the previous section, we have derived a lower bound (Theorems 3 and 4) on the decay-rate of the stationary queue-overflow probability for a wireless system under fairly general assumptions. The infimum on the right-hand-side of (34) and (35) is often referred to as the "minimum-cost-to-overflow," and the fluid sample path (FSP) that attains the infimum (if such an FSP exists) is referred to as the "most-likely path to

overflow." As we discussed in Section IV-A, searching for the most-likely path to overflow is a multi-dimensional calculus-of-variations problem, which is unfortunately very difficult to solve. To view this difficulty in another way, suppose now we want to verify that θ lower-bounds the minimum-cost-to-overflow (or, equivalently, the probability of overflow is approximately upper bounded by $\exp(-B\theta)$ when B is large). We then need to ensure that

$$\int_0^T \left[H\left(\frac{d}{dt}\vec{s}(t)||\vec{p}\right) + L\left(\frac{d}{dt}\vec{a}(t)\right)\right] dt \geq \theta \qquad (36)$$

for *all* FSPs $(\mathbf{s},\mathbf{a},\mathbf{x})_T$ that go from $\vec{x}(0)=0$ to $||\vec{x}(T)|| \geq 1$. For advanced wireless resource-allocation algorithms like the max-weight algorithm [1], the complexity of enumerating all such paths soon becomes prohibitive except for some restrictive cases [10]–[12].

In this section, we develop a new technique to address this difficulty. Our new technique combines the large-deviation lower bound in Theorem 4 with Lyapunov functions to derive another even-simpler lower-bound on the decay-rate of the queue-overflow probability. The reason that we seek help from a Lyapunov function approach is actually very simple and intuitive. Note that the above-mentioned difficulty of evaluating all FSPs is in fact not unique. A similar scenario also arises when we want to prove stability of a dynamic system. For example, recall that in the fluid limit approach [18], [19], in order to show that the fluid limit model of a system is stable, we need to show that there exists a $T>0$, such that for *all* fluid limits with $||\vec{x}(0)|| = 1$, we must have $\vec{x}(T)=0$. Again, it would have been very difficult if one attempts to evaluate all possible multi-dimensional fluid limits. The Lyapunov function approach is indeed developed to address this complexity issue. The basic idea of a Lyapunov function approach is to map each multi-dimensional path $\vec{x}(t)$ to a one-dimensional path $V(\vec{x}(t))$. Recall from part (d) of Assumption 1 that such a function maps $||\vec{x}(0)|| = 1$ to $V(\vec{x}(0)) \leq \tilde{C}$. By establishing that $V(\vec{x}(t))$ has a negative drift, we can show that $V(\vec{x}(t))$ must go from $V(\vec{x}(0)) \leq \tilde{C}$ to $V(\vec{x}(T)) = 0$, which then implies that $\vec{x}(T)=0$. In other words, the key idea of the Lyapunov function approach is this mapping from a *multi-dimensional* space to a *one-dimensional* space, which greatly reduces the complexity for proving stability.

Can we use a similar Lyapunov function approach to characterize the decay-rate of the queue-overflow probability under wireless resource-allocation algorithms? Indeed, Lyapunov functions have been used to solve other calculus-of-variations problems in the control literature. We next demonstrate how such an approach can be used to derive an even simpler lower bound on the minimum-cost-to-overflow (i.e., the infimum given in Theorem 4). Recall that in Assumption 1, part (d), the Lyapunov function $V(\cdot)$ is chosen such that $||\vec{x}|| \geq 1$ implies $V(\vec{x}) \geq 1$. For any $v \geq 0$ and w, define

$$\begin{aligned} l_V(v,w) \triangleq \inf_{\mathbf{s},\mathbf{a},\mathbf{x}} \quad & H(\vec{\phi}||\vec{p}) + L(\vec{f}) \\ \text{subject to} \quad & (\mathbf{s},\mathbf{a},\mathbf{x}) \text{ is an FSP} \\ & \text{such that for some } t \\ & \frac{d}{dt}\vec{s}(t) = \vec{\phi} \\ & \frac{d}{dt}\vec{a}(t) = \vec{f} \\ & V(\vec{x}(t)) = v \\ & \frac{d}{dt}V(\vec{x}(t)) = w. \end{aligned} \qquad (37)$$

According to (14) (or (15), correspondingly), for any $w > -\eta V(\vec{x}(t))$ (or $w > -\eta$, correspondingly), the trajectory with $\frac{d}{dt}V(\vec{x}(t)) = w$ becomes a "rare" event. The function $l_V(v,w)$ provides a lower bound on the *local rate-function* [23, p71] for $V(\vec{x}(t))$, i.e., it bounds how *rarely* that, given $V(\vec{x}(t)) = v$ at some time t, the trajectory $V(\vec{x}(t))$ will move in the direction $\frac{d}{dt}V(\vec{x}(t)) = w$ immediately after t. Note that the infimum in (37) is taken over all possible FSPs such that the corresponding trajectory $V(\vec{x}(t))$ passes through v with slope w.

For any FSP$(\mathbf{s},\mathbf{a},\mathbf{x})_T$, since both the arrival rate and the service rate are bounded, the function $\vec{x}(t)$ must be Lipschitz-continuous. Further, $\vec{x}(t)$ must be bounded over any finite interval. Hence, due to Part (e) of Assumption 1, the function $V(\vec{x}(t))$ must also be Lipschitz-continuous over any finite interval, and thus it must be differentiable almost everywhere. Using the definition of $l_V(\cdot,\cdot)$, we then have the following inequality for any FSP$(\mathbf{s},\mathbf{a},\mathbf{x})_T$:

$$\begin{aligned} \int_0^T H\left(\frac{d}{dt}\vec{s}(t)||\vec{p}\right) + L\left(\frac{d}{dt}\vec{a}(t)\right) dt \\ \geq \int_0^T l_V(V(\vec{x}(t)), \frac{d}{dt}V(\vec{x}(t)))dt. \end{aligned} \qquad (38)$$

Let

$$\begin{aligned} \theta_0 = \inf_{T>0} \quad & \int_0^T l_V(V(t), \frac{d}{dt}V(t))dt \\ \text{subject to} \quad & V(t) \text{ is continuous and} \\ & \text{almost-everywhere differentiable,} \\ & V(0)=0 \text{ and } V(T) \geq 1. \end{aligned} \qquad (39)$$

Note that in (39) we are optimizing over all possible trajectories that the Lyapunov function can take. Hence, we abuse notation and use $V(t)$ to represent any continuous and almost-everywhere differentiable function that satisfies the constraints as mentioned above.

We then obtain a lower bound for the calculus-of-variations problem in (35), as stated in the following Theorem.

Theorem 5: Assume that there exists a Lyapunov function $V(\cdot)$ that satisfies Assumption 1 and Assumption 2. Then θ_0 in (39) is a lower bound on the decay-rate of the queue-overflow probability. In other words,

$$\begin{aligned} \limsup_{B\to\infty} \frac{1}{B}\log \mathbf{P}[||\vec{X}(\infty)/B|| \geq 1] \\ \leq \limsup_{B\to\infty} \frac{1}{B}\log \mathbf{P}[V(\vec{X}(\infty)/B) \geq 1] \leq -\theta_0. \end{aligned} \qquad (40)$$

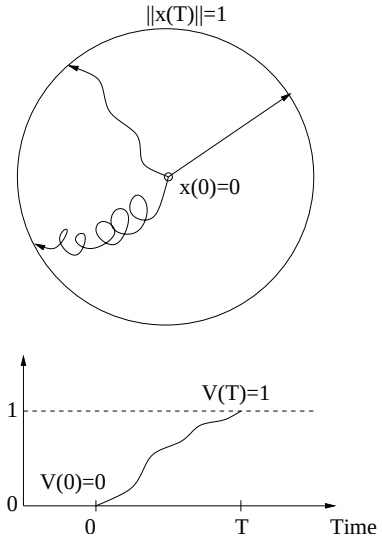

Fig. 2. Top: The overflow probability $\mathbf{P}[||\vec{X}(\infty)|| \geq B]$ is related to the most likely path to overflow. Bottom: The technique that we presented maps any multi-dimensional path $\vec{x}(t)$ to a one-dimensional path $V(t)$.

Proof:

First, by Assumption 1, the following is true

$$\mathbf{P}[||\vec{X}(\infty)/B|| \geq 1] \leq \mathbf{P}[V(\vec{X}(\infty)/B) \geq 1].$$

This proves the first inequality. To show the second inequality, by Theorem 4, we only need to show that, for all FSPs $(\mathbf{s}, \mathbf{a}, \mathbf{x})_T$ that go from $\vec{x}(0) = 0$ to $V(\vec{x}(T)) \geq 1$, the following must hold,

$$\int_0^T H\left(\frac{d}{dt}\vec{s}(t)||\vec{p}\right) + L\left(\frac{d}{dt}\vec{a}(t)\right) dt \geq \theta_0.$$

Using (38), it suffices to show that, for all such FSPs ,

$$\int_0^T l_V(V(\vec{x}(t)), \frac{d}{dt}V(\vec{x}(t))) \geq \theta_0. \tag{41}$$

Note that, for all FSP$(\mathbf{s}, \mathbf{a}, \mathbf{x})_T$ that goes from $\vec{x}(0) = 0$ to $V(\vec{x}(T)) \geq 1$, inequality (41) must hold due to the definition of θ_0 in (39). The result of the Theorem then follows. *Q.E.D.*

It is also easy to see that a sufficient condition for all fluid sample paths to satisfy the constraint (36) is to ensure that

$$\int_0^T l_V(V(t), \frac{d}{dt}V(t))dt \geq \theta \tag{42}$$

holds for all *one-dimensional* paths $V(t)$ that go from $V(0) = 0$ to $V(T) = 1$. Again, we have successfully reduced the original multi-dimensional calculus-of-variations problem to a one-dimensional calculus-of-variations problem. The one-dimensional calculus-of-variations problem in (39) and (42) is usually much easier to solve (Fig. 2).

Remark: Lyapunov functions have been used in the control literature to solve other calculus-of-variations problems. Often, the key to success of such an approach is to find the right Lyapunov function. The unique feature of the scheduling and routing problem studied in this paper is that the Lyapunov function for stability automatically becomes the suitable Lyapunov function for the calculus-of-variations problem. Hence, for any scheduling and routing algorithm that is provably stable, which usually means that there exists a Lyapunov function for stability, we may then apply the above techniques to characterize the queue-overflow probability. In other words, the difficulty level of characterizing the queue-overflow probability is reduced to that of a stability problem. Since (42) is a sufficient condition to (36), we can obtain an upper bound on the overflow probability, and correspondingly, if a constraint on the overflow probability is imposed, we obtain a lower bound on the effective capacity region. The hope of this approach is that, if the function $V(\cdot)$ is appropriately chosen, we may recover a large fraction of, or even the entire effective capacity region.

A. Scale-Linear Lyapunov Functions

In this section, we consider the special case when the Lyapunov function is linear in scale as defined in Assumption 3 in Section IV-A. In this case, we can show that the solution to θ_0 in (39) can be further simplified. This is possible because the function $l_V(v, w)$ turns out to be independent of v. We need the following simple lemma.

Lemma 6: If $(\mathbf{s}, \mathbf{a}, \mathbf{x})_T$ is an FSP, then for any given $\hat{t} \in [0, T]$ and for any $c > 0$, there exists another FSP $(\underline{\mathbf{s}}, \underline{\mathbf{a}}, \underline{\mathbf{x}})_{cT}$ such that

$$\frac{d}{dt}\underline{\vec{s}}(t)\Big|_{c\hat{t}} = \frac{d}{dt}\vec{s}(t)\Big|_{\hat{t}} \tag{43}$$

$$\frac{d}{dt}\underline{\vec{a}}(t)\Big|_{c\hat{t}} = \frac{d}{dt}\vec{a}(t)\Big|_{\hat{t}} \tag{44}$$

$$\underline{\vec{x}}(c\hat{t}) = c\vec{x}(\hat{t}),\ \frac{d}{dt}\underline{\vec{x}}(t)\Big|_{c\hat{t}} = \frac{d}{dt}\vec{x}(t)\Big|_{\hat{t}}. \tag{45}$$

Proof: Let $\vec{s}^B(t), \vec{a}^B(t), \vec{x}^B(t)$ be the sequence of scaled processes that converge to the FSP $(\mathbf{s}, \mathbf{a}, \mathbf{x})_T$. Then for each B, the "unscaled" processes [i.e., before taking the scaling in (6), (7), and (10)] are correspondingly $B\vec{s}^B(\frac{t}{B})$, $B\vec{a}^B(\frac{t}{B})$ and $B\vec{x}^B(\frac{t}{B})$. Consider the new sequence $Bc\vec{s}^{Bc}(\frac{t}{Bc})$, $Bc\vec{a}^{Bc}(\frac{t}{Bc})$ and $Bc\vec{x}^{Bc}(\frac{t}{Bc})$. In other words, we are choosing a sub-sequence from the original "unscaled" sequence. Then, perform the scaling in (6), (7) and (10) again on this subsequence. The corresponding scaled processes are $c\vec{s}^{Bc}(\frac{t}{c})$, $c\vec{a}^{Bc}(\frac{t}{c})$ and $c\vec{x}^{Bc}(\frac{t}{c})$. Taking the limit as $B \to \infty$, we get the FSP $(\underline{\mathbf{s}}, \underline{\mathbf{a}}, \underline{\mathbf{x}})_{cT} = (c\vec{s}(\frac{t}{c}), c\vec{a}(\frac{t}{c}), c\vec{x}(\frac{t}{c}))$. It is easy to verify that this FSP satisfies the conditions in (43), (44) and (45). *Q.E.D.*

Next we prove that under Assumption 3, $l_V(v, w)$ is independent of v.

Proposition 7: When Assumption 3 holds, the function $l_V(v, w)$ is independent of v, i.e.,

$$l_V(v, w) = l_V(cv, w)$$

for all $c > 0$.

Proof: Consider a fixed $c > 0$. According to definition (37),

$$\begin{aligned} l_V(cv, w) = \inf_{\mathbf{s},\mathbf{a},\mathbf{x}} \quad & H(\vec{\phi}||\vec{p}) + L(\vec{f}) \\ \text{subject to} \quad & (\mathbf{s}, \mathbf{a}, \mathbf{x}) \text{ is an FSP} \\ & \text{such that for some } t \\ & \frac{d}{dt}\vec{s}(t) = \vec{\phi} \\ & \frac{d}{dt}\vec{a}(t) = \vec{f} \\ & V(\vec{x}(t)) = cv \\ & \frac{d}{dt}V(\vec{x}(t)) = w. \end{aligned}$$

For any FSP$(\mathbf{s}, \mathbf{a}, \mathbf{x})_T$ and $\hat{t} \in [0, T]$ such that $\frac{d}{dt}\vec{s}(t)\big|_{\hat{t}} = \vec{\phi}$, $\frac{d}{dt}\vec{a}(t)\big|_{\hat{t}} = \vec{f}$, $V(\vec{x}(\hat{t})) = v$, $\frac{d}{dt}V(\vec{x}(t))\big|_{\hat{t}} = w$, according to Lemma 6, there must exist another FSP$(\underline{\mathbf{s}}, \underline{\mathbf{a}}, \underline{\mathbf{x}})_{cT}$ such that $\frac{d}{dt}\underline{\vec{s}}(t)\big|_{c\hat{t}} = \vec{\phi}$, $\frac{d}{dt}\underline{\vec{a}}(t)\big|_{c\hat{t}} = \vec{f}$, $\underline{\vec{x}}(c\hat{t}) = c\vec{x}(\hat{t})$, and $\frac{d}{dt}\underline{\vec{x}}(t)\big|_{c\hat{t}} = \frac{d}{dt}\vec{x}(t)\big|_{\hat{t}}$. Using Assumption 3, we then have

$$V(\underline{\vec{x}}(c\hat{t})) = cv.$$

Further,

$$\begin{aligned} & \frac{d}{dt}V(\underline{\vec{x}}(t))\Big|_{c\hat{t}} \\ = & \lim_{\tau\to 0} \frac{V(\underline{\vec{x}}(c\hat{t}+\tau)) - V(\underline{\vec{x}}(c\hat{t}))}{\tau} \\ = & \lim_{\tau\to 0} \left[\frac{V\left(\underline{\vec{x}}(c\hat{t}) + \frac{d\underline{\vec{x}}(t)}{dt}\Big|_{c\hat{t}}\tau\right) - V(\underline{\vec{x}}(c\hat{t}))}{\tau} \right. \\ & \left. + \frac{V(\underline{\vec{x}}(c\hat{t}+\tau)) - V\left(\underline{\vec{x}}(c\hat{t}) + \frac{d\underline{\vec{x}}(t)}{dt}\Big|_{c\hat{t}}\tau\right)}{\tau} \right]. \end{aligned}$$

According to Assumption 3, the first term is equal to

$$\begin{aligned} & \lim_{\tau\to 0} \frac{V\left(\underline{\vec{x}}(c\hat{t}) + \frac{d\underline{\vec{x}}(t)}{dt}\Big|_{c\hat{t}}\tau\right) - V(\underline{\vec{x}}(c\hat{t}))}{\tau} \\ = & \lim_{\tau\to 0} \frac{V\left(\vec{x}(\hat{t}) + \frac{d\vec{x}(t)}{dt}\Big|_{\hat{t}}\frac{\tau}{c}\right) - V(\vec{x}(\hat{t}))}{\tau/c} \\ = & \frac{d}{dt}V(\vec{x}(t))\Big|_{\hat{t}}. \end{aligned}$$

According to Assumption 1, the second term satisfies,

$$\begin{aligned} & \lim_{\tau\to 0} \frac{\left|V\left(\underline{\vec{x}}(c\hat{t})+\tau\right) - V\left(\underline{\vec{x}}(c\hat{t}) + \frac{d\underline{\vec{x}}(t)}{dt}\Big|_{c\hat{t}}\tau\right)\right|}{\tau} \\ \le & \lim_{\tau\to 0} \frac{\mathcal{L}\left\|\underline{\vec{x}}(c\hat{t}+\tau) - \underline{\vec{x}}(c\hat{t}) - \frac{d\underline{\vec{x}}(t)}{dt}\Big|_{c\hat{t}}\tau\right\|}{\tau} \\ = & 0. \end{aligned}$$

Hence, we have,

$$\frac{d}{dt}V(\underline{\vec{x}}(t))\Big|_{c\hat{t}} = \frac{d}{dt}V(\vec{x}(t))\Big|_{\hat{t}} = w.$$

This implies that the FSP$(\underline{\mathbf{s}}, \underline{\mathbf{a}}, \underline{\mathbf{x}})_{cT}$ satisfies the constraint in the definition of $l_V(cv, w)$. Hence,

$$l_V(cv, w) \le l_V(v, w)$$

A similar argument proves the opposite direction that $l_V(cv, w) \ge l_V(v, w)$. Since $c > 0$ is arbitrary, the result then follows. *Q.E.D.*

When the function $l_V(v, w)$ is independent of v, the trajectory $V(\cdot)$ that attains the infimum in (39) is in fact very easy to solve [23, p520], and the infimum is equal to $\inf_{w>0} \frac{l_V(1,w)}{w}$, i.e.,

$$\begin{aligned} \theta_0 \quad = \quad & \inf_{w>0,\mathbf{s},\mathbf{a},\mathbf{x}} \frac{1}{w}\left[H(\vec{\phi}||\vec{p}) + L(\vec{f})\right] \qquad (46) \\ \text{subject to} \quad & (\mathbf{s}, \mathbf{a}, \mathbf{x}) \text{ is an FSP} \\ & \text{such that for some } t \\ & \frac{d}{dt}\vec{s}(t) = \vec{\phi} \\ & \frac{d}{dt}\vec{a}(t) = \vec{f} \\ & V(\vec{x}(t)) = 1 \\ & \frac{dV(\vec{x}(t))}{dt} = w. \end{aligned}$$

The value of θ_0 has an intuitive interpretation. If we interpret w as the rate of increase of the value of the Lyapunov function, then the objective function in (46) can be viewed as the minimum *per-unit* cost to increase the Lyapunov function, where the minimization is taken over all backlog levels $\vec{x}(t)$, channel states $\vec{s}(t)$, and arrivals $\vec{a}(t)$. In order to overflow, we must lift the value of $V(\vec{x}(t))$ from zero to one. Hence, θ_0 becomes a lower bound on the minimum cost to overflow. According to Theorem 5, θ_0 then corresponds to a lower bound on the decay rate of the overflow probability.

VII. A Condition For The Minimum-Cost-To-Overflow To Be Exact

In the previous section, we have shown that θ_0 is a lower bound on the decay rate of the queue overflow probability (see Theorem 5). In this section, we provide a condition under which a drift-minimizing algorithm is large-deviations decay-rate optimal and the value of θ_0 becomes the *exact* decay-rate of the overflow probability.

We are ready for the following Theorem.

Theorem 8: Suppose that a scheduling and routing policy satisfies Assumptions 1, 2, 3, 4, 5 and 6. Then under this policy the value of θ_0 is the exact decay-rate of the overflow probability according to the Lyapunov function metric, i.e.,

$$\lim_{B\to\infty} \frac{1}{B} \log \mathbf{P}[V(\vec{X}(\infty)/B) \ge 1] = -\theta_0. \qquad (47)$$

Further, this drift-minimizing policy (according to Assumption 4) is optimal in maximizing this decay-rate. In other words, for any policy π we must have

$$\liminf_{B\to\infty} \frac{1}{B} \log \mathbf{P}^{\pi}[V(\vec{X}(\infty)/B) \ge 1] \ge -\theta_0, \qquad (48)$$

where $\mathbf{P}^{\pi}$ denote the stationary distribution under the policy π.

The proof of Theorem 8 contains two parts. First, we show that the decay rate of the probability of overflow, in terms of the Lyapunov metric, is bounded from above for all scheduling

policies. Then we show that under the assumptions on the Lyapunov function, this bound matches with the lower bound θ_0.

Consider the following optimization problem:

$$\begin{aligned}\tilde{w}(\vec{\phi},\vec{f}) = \min_{\vec{x}} \quad & V(\vec{x}) \\ \text{subject to} \quad & x_i^k = [f_i^k - \sum_{j=1}^{S}\phi_j \sum_{l=1}^{L} R_{il}e_{lj}^k]^+ \\ & [e_{lj}^k] \in \text{Conv}(\mathcal{E}_j) \text{ for all } j.\end{aligned}$$

The function $\tilde{w}(\vec{\phi},\vec{f})$ can be viewed as the minimum rate of increase of the Lyapunov function if the channel state process satisfies $\vec{s}(t) = t\vec{\phi}$ and the arrival process satisfies $\vec{a}(t) = t\vec{f}$. Let

$$\tilde{\theta}_0 = \inf_{\{\vec{\phi},\vec{f}:\tilde{w}(\vec{\phi},\vec{f})>0\}} \frac{1}{\tilde{w}(\vec{\phi},\vec{f})}\left[H(\vec{\phi}||\vec{p}) + L(\vec{f})\right]. \quad (49)$$

We first show the following.

Proposition 9: For any policy π we must have

$$\liminf_{B\to\infty} \frac{1}{B}\log \mathbf{P}^{\pi}[V(\vec{X}(\infty)/B) \geq 1] \geq -\tilde{\theta}_0, \quad (50)$$

where $\mathbf{P}^{\pi}$ denotes the stationary distribution under the policy π.

Proof: By the definition of $\tilde{\theta}_0$, for any $\delta \in (0,1)$ there exists $\vec{\phi}_0$ and $\vec{f}_0$ such that

$$\frac{1}{\tilde{w}(\vec{\phi}_0,\vec{f}_0)}\left[H(\vec{\phi}_0||\vec{p}) + L(\vec{f}_0)\right] \leq \tilde{\theta}_0 + \delta.$$

Further, it is easy to show that the function $\tilde{w}(\cdot,\cdot)$ is continuous with respect to $\vec{\phi}$ and $\vec{f}$. Hence, there exists ϵ such that for any $|\vec{\phi}-\vec{\phi}_0| \leq \epsilon$ and $|\vec{f}-\vec{f}_0| \leq \epsilon$, the following holds

$$\tilde{w}(\vec{\phi},\vec{f}) \geq \tilde{w}(\vec{\phi}_0,\vec{f}_0)(1-\delta). \quad (51)$$

Let $\gamma > 0$ be a small number and let $T = \frac{1+\gamma}{\tilde{w}(\vec{\phi}_0,\vec{f}_0)(1-\delta)}$. Define a channel-state process $\vec{s}_0(\cdot)$ and an arrival process $\vec{a}_0(\cdot)$ on the interval $[0,T]$ as follows:

$$\vec{s}_0(t) = t\vec{\phi}_0 \text{ and } \vec{a}_0(t) = t\vec{f}_0.$$

Let $B_T(\vec{s}_0(\cdot))$ denote an ϵ-ball around $\vec{s}_0(\cdot)$, i.e., it contains all $\vec{s}(\cdot)$ such that $\vec{s}(0) = 0$ and

$$||\vec{s}(t) - \vec{s}_0(t)||_{\infty}^{T} < \epsilon.$$

Similarly, define an ϵ-ball $B_T(\vec{a}_0(\cdot))$ around $\vec{a}_0(\cdot)$. Consider that the queue process has reached stationarity at time 0. Then the queue is evolving according to its stationary distribution at any time $T > 0$ and we have $\mathbf{P}^{\pi}[V(\vec{X}(\infty)/B) \geq 1] = \mathbf{P}^{\pi}[V(\vec{x}^B(T)) \geq 1]$. We will now show that, as long as ϵ is sufficiently small, any $\vec{s}^B(\cdot) \in B_T(\vec{s}_0(\cdot))$ and $\vec{a}^B(\cdot) \in B_T(\vec{a}_0(\cdot))$ imply that $V(\vec{x}^B(T)) \geq 1$, regardless of the scheduling and routing policy π used.

Towards this end, consider any $\vec{s}^B(\cdot)$ and $\vec{a}^B(\cdot)$ in these ϵ-balls. From the mapping in (11), we have,

$$\begin{aligned}& x_i^{k,B}(\frac{\lfloor BT\rfloor}{B}) - x_i^{k,B}(\frac{1}{B}) \\ & \quad = a_i^{k,B}(\frac{\lfloor BT\rfloor - 1}{B}) - \sum_{j=1}^{S}\sum_{l=1}^{L} R_{il}\mathcal{A}_{lj}^k, \quad (52)\end{aligned}$$

where we have used $\mathcal{A}_{lj}^k$ to denote

$$\begin{aligned}& E_l^k(j, B\vec{x}^B(\frac{1}{B}))\left[s_j^B(\frac{1}{B}) - s_j^B(0)\right] + \ldots \\ & + E_l^k(j, B\vec{x}^B(\frac{\lfloor BT\rfloor - 1}{B})) \\ & \qquad \times \left[s_j^B(\frac{\lfloor BT\rfloor - 1}{B}) - s_j^B(\frac{\lfloor BT\rfloor - 2}{B})\right].\end{aligned}$$

We make the following observations to simplify (52).

$$\begin{aligned} x_i^{k,B}(T) - x_i^{k,B}(\frac{\lfloor BT\rfloor}{B}) &= O(\frac{1}{B}) \\ x_i^{k,B}(0) - x_i^{k,B}(\frac{1}{B}) &= O(\frac{1}{B}) \\ a_i^{k,B}(T) - a_i^{k,B}(\frac{\lfloor BT\rfloor - 1}{B}) &= O(\frac{1}{B}) \\ s_j^B(T) - s_j^B(\frac{\lfloor BT\rfloor - 1}{B}) &= O(\frac{1}{B}).\end{aligned}$$

Further, we can write $\mathcal{A}_{lj}^k = \left[s_j^B(\frac{\lfloor BT\rfloor - 1}{B}) - s_j^B(0)\right] e_{lj}^k$ where $[e_{lj}^k] \in \text{Conv}(\mathcal{E}_j)$. This follows because $[E_l^k(j, B\vec{x}^B(\cdot))]$ is in the set $\mathcal{E}_j$ and the terms $\frac{s_j^B(\frac{1}{B}) - s_j^B(0)}{s_j^B(\frac{\lfloor BT\rfloor - 1}{B}) - s_j^B(0)}, \ldots, \frac{s_j^B(\frac{\lfloor BT\rfloor - 1}{B}) - s_j^B(\frac{\lfloor BT\rfloor - 2}{B})}{s_j^B(\frac{\lfloor BT\rfloor - 1}{B}) - s_j^B(0)}$ can be thought of as weights that sum to 1. Since $\vec{s}^B(0) = 0$, we have $\mathcal{A}_{lj}^k = s_j^B(\frac{\lfloor BT\rfloor - 1}{B})e_{lj}^k$. Equation (52) then simplifies to

$$\begin{aligned} x_i^{k,B}(T) - x_i^{k,B}(0) &= a_i^{k,B}(T) - \sum_{j=1}^{S}\sum_{l=1}^{L} R_{il}s_j^B(T)e_{lj}^k \\ &\quad + O(\frac{1}{B}).\end{aligned}$$

Since $x_i^{k,B}(0) \geq 0$ and $x_i^{k,B}(T) \geq 0$, we can show that

$$x_i^{k,B}(T) \geq \left[a_i^{k,B}(T) - \sum_{j=1}^{S}\sum_{l=1}^{L} R_{il}s_j^B(T)e_{lj}^k\right]^+ + O(\frac{1}{B}).$$

Rewriting the equation using $\vec{\phi} \triangleq \frac{\vec{s}^B(T)}{T}$ and $\vec{f} \triangleq \frac{\vec{a}^B(T)}{T}$, we have

$$x_i^{k,B}(T) + O(\frac{1}{B}) \geq T[f_i^k - \sum_{j=1}^{S}\phi_j \sum_{l=1}^{L} R_{il}e_{lj}^k]^+,$$

where $[e_{lj}^k] \in \text{Conv}(\mathcal{E}_j)$. Using Assumption 5 and Assumption 6 on this inequality, we obtain

$$\begin{aligned} V(\vec{x}^B(T)) + V(O(\frac{1}{B})) &\geq V(T\vec{x}) \\ \text{where} \quad & x_i^k = [f_i^k - \sum_{j=1}^{S}\phi_j \sum_{l=1}^{L} R_{il}e_{lj}^k]^+.\end{aligned}$$

This provides a bound on $V(\vec{x}^B(T))$. However, this bound depends on the particular value of e_{lj}^k, which in turn depends on the scheduling and routing policy π. To obtain a bound that is independent of the policy used, we take the minimum

on the right-hand-side over all $[e_{lj}^k] \in \text{Conv}(\mathcal{E}_j)$. Therefore,

$$\begin{aligned} V(\vec{x}^B(T)) + V(O(\frac{1}{B})) & \\ \geq T \min \quad & V(\vec{x}) \\ \text{subject to} \quad & x_i^k = [f_i^k - \sum_{j=1}^{S} \phi_j \sum_{l=1}^{L} R_{il} e_{lj}^k]^+ \\ & [e_{lj}^k] \in \text{Conv}(\mathcal{E}_j) \text{ for all } j. \end{aligned}$$

This implies that

$$V(\vec{x}^B(T)) + V(O(\frac{1}{B})) \geq T\tilde{w}(\vec{\phi}, \vec{f}) \geq 1 + \gamma,$$

where in the last step we have used the definition of T and the fact that (51) holds by choosing sufficiently small ϵ.

Therefore, under any policy π, there exists B_γ such that for all $B > B_\gamma$, if $\vec{s}^B(\cdot) \in B_T(\vec{s}_0(\cdot))$ and $\vec{a}^B(\cdot) \in B_T(\vec{a}_0(\cdot))$ then

$$V(\vec{x}^B(T)) \geq 1.$$

Now, using the LDP for $\vec{s}^B(\cdot)$ and $\vec{a}^B(\cdot)$, we complete the proof as follows,

$$\begin{aligned} & \liminf_{B\to\infty} \frac{1}{B} \log \mathbf{P}^\pi[V(\vec{X}(\infty)/B) \geq 1] \\ = \; & \liminf_{B\to\infty} \frac{1}{B} \log \mathbf{P}^\pi[V(\vec{x}^B(T)) \geq 1] \\ \geq \; & \liminf_{B\to\infty} \frac{1}{B} \big\{\log \mathbf{P}[\vec{s}^B(\cdot) \in B_T(\vec{s}_0(\cdot))] \\ & \qquad + \log \mathbf{P}[\vec{a}^B(\cdot) \in B_T(\vec{a}_0(\cdot))]\big\} \\ = \; & -\inf_{\vec{s}(\cdot)\in B_T(\vec{s}_0(\cdot))} \int_0^T H(\frac{d}{dt}\vec{s}(t)||\vec{p})dt \\ & \qquad - \inf_{\vec{a}(\cdot)\in B_T(\vec{a}_0(\cdot))} \int_0^T L(\frac{d}{dt}\vec{a}(t))dt \\ \geq \; & -\int_0^T H(\frac{d}{dt}\vec{s}_0(t)||\vec{p})dt - \int_0^T L(\frac{d}{dt}\vec{a}_0(t))dt \\ = \; & -T[H(\vec{\phi}_0||\vec{p}) + L(\vec{f}_0)] \\ = \; & -\frac{1+\gamma}{\tilde{w}(\vec{\phi}_0, \vec{f}_0)(1-\delta)} \left[H(\vec{\phi}_0||\vec{p}) + L(\vec{f}_0)\right] \\ \geq \; & -\frac{1+\gamma}{1-\delta}(\tilde{\theta}_0 + \delta). \end{aligned}$$

Since δ and γ can be arbitrarily small, we conclude that

$$\liminf_{B\to\infty} \frac{1}{B} \log \mathbf{P}^\pi[V(\vec{X}(\infty)/B) \geq 1] \geq -\tilde{\theta}_0. \qquad (53)$$

Q.E.D.

We are now ready to prove Theorem 8.

Proof of Theorem 8 : By Theorem 5,

$$\limsup_{B\to\infty} \frac{1}{B} \log \mathbf{P}[V(\vec{X}(\infty)/B) \geq 1] \leq -\theta_0,$$

where θ_0 is given by (46). By Proposition 9, we have

$$\liminf_{B\to\infty} \frac{1}{B} \log \mathbf{P}^\pi[V(\vec{X}(\infty)/B) \geq 1] \geq -\tilde{\theta}_0, \qquad (54)$$

for any policy π. The two inequalities combined imply that $\theta_0 \leq \tilde{\theta}_0$. Hence, to show Theorem 8, it only remains to show that $\theta_0 \geq \tilde{\theta}_0$. Consider any FSP$(\mathbf{s}, \mathbf{a}, \mathbf{x})$ that satisfies the constraint in the definition of θ_0 (see Equation (46)). Define $\vec{\phi} \triangleq \frac{d}{dt}\vec{s}(t)$, $\vec{f} \triangleq \frac{d}{dt}\vec{a}(t)$ and $w \triangleq \frac{dV(\vec{x}(t))}{dt}$. By Assumption 4 in Section IV-A, we must have

$$w \;\leq\; \left.\frac{\partial \tilde{V}(\tau, \vec{e}|\vec{x}(t), \vec{\phi}, \vec{f})}{\partial \tau}\right|_{\tau=0} \qquad (55)$$

for any feasible $\vec{e}$.

Define $\vec{\delta} = [\delta_i^k, i = 1, ..., N, k = 1, ..., K]$, where $\delta_i^k = f_i^k - \sum_{j=1}^{S} \phi_j \sum_{l=1}^{L} R_{il} e_{lj}^k$. We have

$$\begin{aligned} & \tilde{V}(\tau, \vec{e}|\vec{x}(t), \vec{\phi}, \vec{f}) - \tilde{V}(0, \vec{e}|\vec{x}(t), \vec{\phi}, \vec{f}) \\ = \; & V([\vec{x}(t) + \vec{\delta}\tau]^+) - V(\vec{x}(t)). \end{aligned}$$

Further, using Assumptions 3, 5 and 6, we have

$$\begin{aligned} V([\vec{x}(t) + \vec{\delta}\tau]^+) - V(\vec{x}(t)) \;&\leq\; V(\vec{x}(t) + [\vec{\delta}]^+\tau) - V(\vec{x}(t)) \\ &\leq\; V([\vec{\delta}]^+\tau) = \tau V([\vec{\delta}]^+). \end{aligned}$$

Hence, for any feasible $\vec{e}$, by (55), we must have

$$w \;\leq\; V([\vec{\delta}]^+).$$

Minimizing the right-hand-side as in the definition of $\tilde{w}(\vec{\phi}, \vec{f})$, we have

$$w \leq \tilde{w}(\vec{\phi}, \vec{f}),$$

from which we can conclude that $\{w > 0\} \subseteq \{\tilde{w}(\vec{\phi}, \vec{f}) > 0\}$. This leads to the following inequality

$$\begin{aligned} & \inf_{\{w>0, \mathbf{s},\mathbf{a},\mathbf{x}\}} \frac{1}{w} \left[H(\vec{\phi}||\vec{p}) + L(\vec{f})\right] \\ \geq \; & \inf_{\{\vec{\phi}, \vec{f}: \tilde{w}(\vec{\phi}, \vec{f}) > 0\}} \frac{1}{\tilde{w}(\vec{\phi}, \vec{f})} \left[H(\vec{\phi}||\vec{p}) + L(\vec{f})\right]. \end{aligned}$$

Hence, by the definitions of θ_0 in (46) and $\tilde{\theta}_0$ in (49), we have

$$\theta_0 \;\geq\; \tilde{\theta}_0.$$

Thus, $\theta_0 = \tilde{\theta}_0$ and the result of the Theorem then follows.

Q.E.D.

VIII. An Example

We have presented a set of powerful results (in particular Theorem 8) that can be applied to very general wireless systems. As we discussed in Section IV-B, these results can be very useful for both analyzing and designing wireless control algorithms with low overflow-probability. A delicate detail to use these results is to establish that an algorithm minimizes the drift of a Lyapunov function in every fluid sample path (i.e., it satisfies Assumption 4). This is related but different from showing that an algorithm minimizes the drift of a Lyapunov function *in every step in the original discrete-time system*. Specifically, an infinitesimal interval of length δ in the fluid sample path will correspond to an interval of $B\delta$ in the original system. Hence, one needs to be careful in analyzing the drift of the Lyapunov function in fluid sample paths. In this section, we present a detailed example to illustrate this point (see Proposition 10).

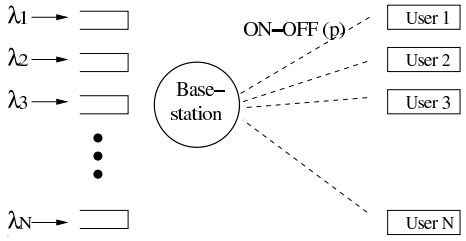

Fig. 3. The scheduling problem in cellular networks under fading channel.

Consider the following model for a base-station serving N users (See Fig. 3). We recall here the assumptions made on the arrival process and the channel model from Section II, and list some additional simplifying assumptions needed for our current purpose. $A_l(t)$ denotes the number of packets generated by user l. We assume that $A_l(t)$ is *i.i.d.* across time and independent across users. Let $\lambda_l = \mathbf{E}[A_l(t)]$. Only one user can be scheduled for transmission at any time. We assume an ON-OFF channel fading model between the base-station and the users. $C(t)$ denotes the channel state at time slot t and $\mathcal{S}$ denotes the set of all possible channel states. Each possible value $j \in \mathcal{S}$ of the channel state $C(t)$ can be thought of as mapping to a vector, each component of which represents whether the channel of a particular user is ON or OFF. We assume that the channel state $C(t)$ is *i.i.d.* across time. However, the channel may be correlated across users. The probability that the channel state $C(t)$ is equal to j at time t is p_j. For any subset $\mathcal{A} \subset \{1, 2, ...N\}$, $\mathcal{S}(\mathcal{A})$ denotes the set of states j such that the channel of some user $l \in \mathcal{A}$ is ON. We also use $\mathcal{S}(l)$ as the short-hand notation for $\mathcal{S}(\{l\})$. Let F denote the bandwidth of the system. Hence, if a user's channel is ON and it is scheduled for transmission, its service rate is F. (*Remark:* The above model is similar to the one in [11] although we do not assume identical arrival rates and *i.i.d.* channel state distribution across the users.) The throughput-optimal Tassiulas-Ephremides algorithm [1] in this case reduces to the QLB (Queue-Length Based) algorithm as follows.

QLB-scheduling policy: At each time-slot t, the base-station schedules the ON user with the largest backlog. If there are multiple ON-users that all have the largest backlog, the base-station can schedule any one of these users.

We assume that the system is stable, which requires that there exists $e_{lj} \geq 0$, $l = 1, \ldots, N, j = 1, \ldots, S$, such that $\sum_{l=1}^{N} e_{lj} = F$ for all j, and that the following holds for some $\hat{\epsilon} > 0$,

$$\lambda_l(1+\hat{\epsilon}) < \sum_{j \in \mathcal{S}(l)} p_j e_{lj} \text{ for all users } l = 1, \ldots, N. \quad (56)$$

The interpretation of e_{lj} is the long-term fraction of system bandwidth given to user l in channel state j. Note that the summation on the right-hand-side is only over those states j such that the channel is ON for user l.

In this section, we would like to characterize the decay-rate of the tail probability of any user's backlog exceeding a given threshold B, i.e., the decay-rate of the probability

$$\mathbf{P}[\max_{l=1,\ldots,N} X_l(\infty) \geq B] \quad (57)$$

when $B \to \infty$.

Define $\vec{s}^B(t), \vec{a}^B(t), \vec{x}^B(t)$ as in (6), (7) and (10), and define the fluid sample path (FSP) accordingly. For any FSP$(\mathbf{s}, \mathbf{a}, \mathbf{x})$, let $\mathcal{I}_1(\vec{x}(t)) = \{i | x_i(t) = \max_k x_k(t)\}$ be the set of users with the (identically) largest queue at time t. Further, let $\mathcal{I}_2(\vec{x}(t), \vec{x}'(t)) = \{i \in \mathcal{I}_1(\vec{x}(t)) | \frac{d}{dt}x_i(t) = \max_{k \in \mathcal{I}_1(\vec{x}(t))} \frac{d}{dt}x_k(t)\}$. That is, $\mathcal{I}_2(\vec{x}(t), \vec{x}'(t))$ is the set of users that, among those users with the largest queue at time t, also have the largest queue growth rate. In other words, these set of users will have the largest queue *immediately after time* t. Then, immediately after time t, as long as one user in $\mathcal{I}_2(\vec{x}(t), \vec{x}'(t))$ is ON, according to the QLB-policy this group of users collectively must receive the full service rate F. Therefore, we must have

$$\sum_{i \in \mathcal{I}_2(\vec{x}(t), \vec{x}'(t))} \frac{d}{dt}x_i(t) = \sum_{i \in \mathcal{I}_2(\vec{x}(t), \vec{x}'(t))} \frac{d}{dt}a_i(t) - F \sum_{j \in \mathcal{S}(\mathcal{I}_2(\vec{x}(t), \vec{x}'(t)))} \frac{d}{dt}s_j(t). \quad (58)$$

(*Remark:* Note that this is an example of Equation (12).)

Let $V(\vec{x}) = \max_{l=1,\ldots,N} x_l$. Note that we have chosen the Lyapunov function to be the same as the norm for the overflow metric (57). We now show the following properties of $V(\vec{x})$.

Proposition 10: The function $V(\vec{x})$ satisfies Assumptions 1, 2, 3, 4, 5 and 6.

Proof: Most of the conditions in the assumptions are easy to verify. Hence, we only provide proofs of Assumption 1 part (f), Assumption 2 and Assumption 4.

We first show Assumption 2 part (a). Consider an FSP$(\mathbf{s}, \mathbf{a}, \mathbf{x})$. Let $\vec{\phi}(t) = \frac{d}{dt}\vec{s}(t)$ and let $\vec{f}(t) = \frac{d}{dt}\vec{a}(t)$. According to the definition of $\mathcal{I}_2(\vec{x}(t), \vec{x}'(t))$, the Lyapunov drift for the QLB policy is given by

$$\begin{aligned} & \frac{d}{dt}V(\vec{x}(t)) \\ = & \frac{1}{|\mathcal{I}_2(\vec{x}(t), \vec{x}'(t))|} \sum_{l \in \mathcal{I}_2(\vec{x}(t), \vec{x}'(t))} \frac{d}{dt}x_l(t) \\ = & \frac{1}{|\mathcal{I}_2(\vec{x}(t), \vec{x}'(t))|} \left[\sum_{l \in \mathcal{I}_2(\vec{x}(t), \vec{x}'(t))} f_l(t) - F \sum_{j \in \mathcal{S}(\mathcal{I}_2(\vec{x}(t), \vec{x}'(t)))} \phi_j(t) \right]. \end{aligned} \quad (59)$$

Let $\epsilon \leq \hat{\epsilon}\frac{\min_{l=1,\ldots,N} \lambda_l}{2(N+FS)}$. Assume that $||\frac{d}{dt}\vec{s}(t) - \vec{p}|| < \epsilon$ and $||\frac{d}{dt}\vec{a}(t) - \vec{\lambda}|| < \epsilon$. Then

$$\frac{d}{dt}V(\vec{x}(t)) \leq \frac{1}{|\mathcal{I}_2(\vec{x}(t), \vec{x}'(t))|} \left[\sum_{l \in \mathcal{I}_2(\vec{x}(t), \vec{x}'(t))} \lambda_l - F \sum_{j \in \mathcal{S}(\mathcal{I}_2(\vec{x}(t), \vec{x}'(t)))} p_j \right] + \epsilon(N + FS). \quad (60)$$

By the stability condition (56), we have

$$\begin{aligned}
&\sum_{l\in\mathcal{I}_2(\vec{x}(t),\vec{x}'(t))}\lambda_l(1+\hat{\epsilon})\\
\leq\ &\sum_{l\in\mathcal{I}_2(\vec{x}(t),\vec{x}'(t))}\sum_{j\in\mathcal{S}(l)}p_je_{lj}\\
\leq\ &\sum_{j\in\mathcal{S}(\mathcal{I}_2(\vec{x}(t),\vec{x}'(t)))}p_jF.
\end{aligned}$$

Therefore, Equation (60) becomes

$$\begin{aligned}
\frac{d}{dt}V(\vec{x}(t)) &\leq \frac{-\hat{\epsilon}}{|\mathcal{I}_2(\vec{x}(t),\vec{x}'(t))|}\sum_{l\in\mathcal{I}_2(\vec{x}(t),\vec{x}'(t))}\lambda_l\\
&\quad +\epsilon(N+FS)\\
&\leq \frac{-\hat{\epsilon}\min_{l=1,\ldots,N}\lambda_l}{2}.
\end{aligned}$$

This shows Assumption 2 part (a). Note that Assumption 1 part (f) can be shown with a similar proof.

Now we show Assumption 2 part (b): By (59), and using the fact that the arrival process is bounded by M, we have

$$\frac{d}{dt}V(\vec{x}(t))\leq\frac{1}{|\mathcal{I}_2(\vec{x}(t),\vec{x}'(t))|}\left[\sum_{l\in\mathcal{I}_2(\vec{x}(t),\vec{x}'(t))}M\right]=M.$$

To show Assumption 4, we will first bound the Lyapunov drift for any scheduling policy and then show that the bound is in fact the drift for the QLB policy (59).

Fix any feasible value of $\vec{e}$, i.e., $e_{lj}\geq 0$, $l=1,\ldots,N, j=1,\ldots,S$; $e_{lj}>0$ only if the channel for user l is ON at channel state j; and $\sum_{l=1}^{N}e_{lj}\leq F$ for all j. Define $\delta_l=f_l-\sum_{j\in\mathcal{S}(l)}\phi_je_{lj}$. By definition in Equation (24),

$$\begin{aligned}
\tilde{V}(\tau,\vec{e}|\vec{x},\vec{\phi},\vec{f}) &= V([\vec{x}+\vec{\delta}\tau]^+)\\
&= \max_{l=1,\ldots,N}[x_l+\delta_l\tau]^+.
\end{aligned}$$

We must then have,

$$\begin{aligned}
&\frac{\partial}{\partial\tau}\tilde{V}(\tau,\vec{e}|\vec{x}(t),\vec{\phi}(t),\vec{f}(t))\Big|_{\tau=0}\\
\geq\ &\max_{l\in\mathcal{I}_1(\vec{x}(t))}(f_l(t)-\sum_{j\in\mathcal{S}(l)}\phi_j(t)e_{lj}).
\end{aligned}$$

Further, since $\mathcal{I}_2(\vec{x}(t),\vec{x}'(t))\subset\mathcal{I}_1(\vec{x}(t))$, we have

$$\begin{aligned}
&\frac{\partial}{\partial\tau}\tilde{V}(\tau,\vec{e}|\vec{x}(t),\vec{\phi}(t),\vec{f}(t))\Big|_{\tau=0}\\
\geq\ &\max_{l\in\mathcal{I}_2(\vec{x}(t),\vec{x}'(t))}(f_l(t)-\sum_{j\in\mathcal{S}(l)}\phi_j(t)e_{lj})\\
\geq\ &\frac{1}{|\mathcal{I}_2(\vec{x}(t),\vec{x}'(t))|}\left[\sum_{l\in\mathcal{I}_2(\vec{x}(t),\vec{x}'(t))}(f_l(t)\right.\\
&\left.-\sum_{j\in\mathcal{S}(l)}\phi_j(t)e_{lj})\right]. \qquad (61)
\end{aligned}$$

Now, note that

$$\begin{aligned}
&\sum_{l\in\mathcal{I}_2(\vec{x}(t),\vec{x}'(t))}\sum_{j\in\mathcal{S}(l)}\phi_j(t)e_{lj}\\
=\ &\sum_{j\in\mathcal{S}(\mathcal{I}_2(\vec{x}(t),\vec{x}'(t)))}\phi_j(t)\sum_{l\in\mathcal{I}_2(\vec{x}(t),\vec{x}'(t))}e_{lj}\\
\leq\ &F\sum_{j\in\mathcal{S}(\mathcal{I}_2(\vec{x}(t),\vec{x}'(t)))}\phi_j(t).
\end{aligned}$$

Therefore, inequality (61) reduces to

$$\begin{aligned}
&\frac{\partial}{\partial\tau}\tilde{V}(\tau,\vec{e}|\vec{x}(t),\vec{\phi}(t),\vec{f}(t))\Big|_{\tau=0}\\
\geq\ &\frac{1}{|\mathcal{I}_2(\vec{x}(t),\vec{x}'(t))|}\left[\sum_{l\in\mathcal{I}_2(\vec{x}(t),\vec{x}'(t))}f_l(t)\right.\\
&\left.-F\sum_{j\in\mathcal{S}(\mathcal{I}_2(\vec{x}(t),\vec{x}'(t)))}\phi_j(t)\right],
\end{aligned}$$

where the right-hand-side is exactly equal to the drift of the QLB scheduler (59). The inequality (25) then follows.

The other assumptions are easily verified. *Q.E.D.*

By Theorem 8 and Proposition 10, we then have

$$\lim_{B\to\infty}\frac{1}{B}\log\mathbf{P}[V(\vec{X}(\infty)/B)\geq 1]=-\theta_0.$$

where θ_0 is given by

$$\theta_0=\tilde{\theta}_0=\inf_{\{\vec{\phi},\vec{f}:\tilde{w}(\vec{\phi},\vec{f})>0\}}\frac{1}{\tilde{w}(\vec{\phi},\vec{f})}\left[H(\vec{\phi}||\vec{p})+L(\vec{f})\right], \quad (62)$$

where

$$\begin{aligned}
\tilde{w}(\vec{\phi},\vec{f})=\min_{[e_{lj}]}\quad &\max_{l=1,\ldots,N}x_l\\
\text{subject to}\quad &x_l=[f_l-\sum_{j=1}^{S}\phi_je_{lj}]^+\\
&[e_{lj}]\in\text{Conv}(\mathcal{E}_j)\text{ for all }j.
\end{aligned}$$

Unfortunately, (62) is not a convex program and thus is not easy to solve. To derive a simpler characterization of θ_0, we first introduce a decomposition. For any subset $\mathcal{M}\subset\{1,2,...,N\}$, define

$$\begin{aligned}
\theta_0(\mathcal{M}) &= \inf_{w>0,\vec{\phi},\vec{f}}\frac{1}{w}\left[H(\vec{\phi}||\vec{p})+L(\vec{f})\right] \qquad (63)\\
\text{subject to}\quad &w=\frac{1}{|\mathcal{M}|}\left[\sum_{l\in\mathcal{M}}f_l-F\sum_{j\in\mathcal{S}(\mathcal{M})}\phi_j\right].
\end{aligned}$$

We then have the following results.

Lemma 11:

$$\theta_0=\min_{\mathcal{M}\subset\{1,2,\ldots,N\}}\theta_0(\mathcal{M}).$$

Proof: Fix a set $\mathcal{M}$. For any $\vec{\phi}$, $\vec{f}$, and $w>0$ that satisfy the constraints in (63), we automatically have

$$\frac{1}{|\mathcal{M}|}\left[\sum_{l\in\mathcal{M}}f_l-F\sum_{j\in\mathcal{S}(\mathcal{M})}\phi_j\right]=w.$$

Now, for any feasible $\vec{e} \in \text{Conv}(\mathcal{E}_j)$, we must have $\sum_{j=1}^{S} e_{lj} \leq F$ for all j and $e_{lj} > 0$ only if the channel for user l is ON at channel state j. Thus, we have,

$$\begin{aligned}
&\max_{l=1,\ldots,N}[f_l - \sum_{j=1}^{S}\phi_j e_{lj}]^+ \\
\geq & \frac{1}{|\mathcal{M}|}\left[\sum_{l\in\mathcal{M}}[f_l - \sum_{j=1}^{S}\phi_j e_{lj}]^+\right] \\
\geq & \frac{1}{|\mathcal{M}|}\left[\sum_{l\in\mathcal{M}} f_l - F\sum_{j\in\mathcal{S}(\mathcal{M})}\phi_j\right] \\
= & w.
\end{aligned}$$

Hence, $\tilde{w}(\vec{\phi},\vec{f}) \geq w$ and

$$\theta_0 \leq \frac{1}{w}\left[H(\vec{\phi}||\vec{p}) + L(\vec{f})\right].$$

Since this is true for all $\vec{\phi},\vec{f}$ and $w > 0$ that satisfy the constraints in (63), we then have $\theta_0 \leq \theta_0(\mathcal{M})$.

To show the other direction, it is sufficient to show that for every $\vec{\phi}$ and $\vec{f}$, there is some $\mathcal{M}$ for which $w = \tilde{w}(\vec{\phi},\vec{f})$ satisfies the constraint (63). Let $w = \tilde{w}(\vec{\phi},\vec{f})$. By the definition of $\tilde{w}(\cdot,\cdot)$, there must exist $\vec{e} \in \text{Conv}(\mathcal{E}_j)$ such that

$$\max_{l=1,\ldots,N}[f_l - \sum_{j=1}^{S}\phi_j e_{lj}]^+ = w.$$

Let $\mathcal{M}$ be the set of l such that $f_l - \sum_{j=1}^{S}\phi_j e_{lj} = w$. Note that we must have $\sum_{l\in\mathcal{M}} e_{lj} = F$ for any state $j \in \mathcal{S}(\mathcal{M})$ because otherwise we should be able to further reduce $\max_{l=1,\ldots,N}[f_l - \sum_{j=1}^{S}\phi_j e_{lj}]^+$. Hence, we have

$$\sum_{l\in\mathcal{M}} f_l - F\sum_{j\in\mathcal{S}(\mathcal{M})}\phi_j = w|\mathcal{M}|.$$

This equation implies that $\vec{\phi},\vec{f}$ and w satisfies the constraint in (63). Hence, we must have $\theta_0 \geq \theta_0(\mathcal{M})$. The result of the lemma then follows. *Q.E.D.*

Next, consider the case when the arrivals are at a constant rate λ_l. In other words, $L_l(f_l) = +\infty$ except when $f_l = \lambda_l$. In this case, we can use Lemma 11 to obtain the following characterization of θ_0. For any subset $\mathcal{C}$ of the possible channel states, let $p(\mathcal{C}) = \sum_{j\in\mathcal{C}} p_j$.

Proposition 12: When the arrivals are at a constant rate λ_l,

$$\theta_0 = \min_{\mathcal{M}\subset\{1,\ldots,N\}} \inf_{0\leq u\leq\sum_{i\in\mathcal{M}}\frac{\lambda_i}{F}} \frac{|\mathcal{M}|D_{\mathcal{M}}(u||p)}{(\sum_{i\in\mathcal{M}}\lambda_i) - uF}$$

where

$$\begin{aligned}
D_{\mathcal{M}}(u||p) = & u\log\frac{u}{p(\mathcal{S}(\mathcal{M}))} \\
& +(1-u)\log\frac{(1-u)}{(1-p(\mathcal{S}(\mathcal{M})))}.
\end{aligned}$$

Proof: According to Lemma 11, it suffices to show that

$$\theta_0(\mathcal{M}) = \inf_{0\leq u\leq\sum_{i\in\mathcal{M}}\frac{\lambda_i}{F}} \frac{|\mathcal{M}|}{(\sum_{i\in\mathcal{M}}\lambda_i) - uF} D_{\mathcal{M}}(u||p).$$

Towards this end, note that for any fixed w, the sub-optimization problem of the one defined in (63) corresponds to a convex program. (The value of $\theta_0(\mathcal{M})$ will then correspond to the minimum over w among all of these sub-optimization problems.) Associate a Lagrange multiplier η for the constraint of (63), and a Lagrange multiplier γ for the constraint $\sum_{j=1}^{S}\phi_j = 1$. Ignoring the term $L(\vec{f})$ and letting $f_l = \lambda_l$, we can then construct the Lagrangian of the sub-optimization problem for a fixed w as

$$\begin{aligned}
& \mathcal{L}(\vec{\phi},\vec{f},w,\eta,\gamma) \\
= & \sum_{j=1}^{S}\phi_j\log\frac{\phi_j}{p_j} - \eta\left[\sum_{l\in\mathcal{M}}\lambda_l\right. \\
& \left. -F\sum_{j\in\mathcal{S}(\mathcal{M})}\phi_j - w|\mathcal{M}|\right] + \gamma\left[\sum_{j=1}^{S}\phi_j - 1\right] \\
= & \left[\sum_{j=1}^{S}\phi_j\log\frac{\phi_j}{p_j} + \eta F\sum_{j\in\mathcal{S}(\mathcal{M})}\phi_j + \gamma\phi_j\right] \\
& -\eta\left[\sum_{l\in\mathcal{M}}\lambda_l - w|\mathcal{M}|\right].
\end{aligned}$$

It is easy to verify that, in order to minimize the Lagrangian over all $\vec{\phi}$, we must have

$$\begin{aligned}
\phi_j &= p_j\exp[-(1+\eta F+\gamma)] \text{ if } j\in\mathcal{S}(\mathcal{M}), \\
\phi_j &= p_j\exp[-(1+\gamma)] \text{ if } j\notin\mathcal{S}(\mathcal{M}).
\end{aligned}$$

The optimal η and γ are such that the constraint of (63) and $\sum_{j=1}^{S}\phi_j = 1$ are both satisfied. Hence, we must have

$$\begin{aligned}
\sum_{j\in\mathcal{S}(\mathcal{M})}\phi_j &= \exp[-(1+\eta F+\gamma)]\sum_{j\in\mathcal{S}(\mathcal{M})}p_j \\
&= \frac{\sum_{l\in\mathcal{M}}\lambda_l - w|\mathcal{M}|}{F}.
\end{aligned}$$

Let $u = \frac{\sum_{l\in\mathcal{M}}\lambda_l - w|\mathcal{M}|}{F}$. We then have,

$$\exp[-(1+\eta F+\gamma)] = \frac{u}{p(\mathcal{S}(\mathcal{M}))},$$

and

$$\phi_j = p_j\frac{u}{p(\mathcal{S}(\mathcal{M}))} \text{ if } j\in\mathcal{S}(\mathcal{M}).$$

Similarly, we must have

$$\begin{aligned}
\sum_{j\notin\mathcal{S}(\mathcal{M})}\phi_j &= \exp[-(1+\gamma)]\sum_{j\notin\mathcal{S}(\mathcal{M})}p_j \\
&= 1 - \frac{\sum_{l\in\mathcal{M}}\lambda_l - w|\mathcal{M}|}{F} = 1-u.
\end{aligned}$$

Hence,

$$\phi_j = p_j\frac{(1-u)}{(1-p(\mathcal{S}(\mathcal{M})))} \text{ if } j\notin\mathcal{S}(\mathcal{M}).$$

We thus have, for any fixed w, the minimum value of the sub-optimization problem of (63) is equal to

$$\begin{aligned}
&\frac{1}{w}\sum_{j=1}^{S}\phi_j\log\frac{\phi_j}{p_j}\\
=\;&\frac{1}{w}\left[u\log\frac{u}{p(\mathcal{S}(\mathcal{M}))}+(1-u)\log\frac{(1-u)}{(1-p(\mathcal{S}(\mathcal{M})))}\right]\\
=\;&\frac{\sum_{l\in\mathcal{M}}\lambda_l-uF}{|\mathcal{M}|}D_{\mathcal{M}}(u||p),
\end{aligned}$$

where in the last step we have used the definition of $D_{\mathcal{M}}(u||p)$ and the assignment that $u=\frac{\sum_{l\in\mathcal{M}}\lambda_l-wM}{F}$. Taking a minimum over w, or equivalently over u, the result follows. *Q.E.D.*

Remark: Readers can verify that when the channel states are *i.i.d.* across users and when the arrivals from all users are at a constant rate λ, Proposition 12 reduces to Theorem 5 in [11].

A. Effective Capacity

When the arrivals are time-varying, it is no longer possible to derive a simpler characterization of θ_0 as in Proposition 12. Instead, we can concentrate on the effective capacity region. Note that for any $\theta>0$, if we would like to ensure that

$$\lim_{B\to\infty}\frac{1}{B}\log\mathbf{P}[\max_l X_l(\infty)\geq B]\leq-\theta,$$

it is equivalent to require that $\theta\leq\theta_0$. We then have the following result. Let $M_l(\theta)=\log\mathbf{E}[\exp(\theta A_l(0))]$, and let $c_l(\theta)=\frac{M_l(\theta)}{\theta}$. The quantity $c_l(\theta)$ is typically referred to as the effective bandwidth of the arrival process to user l.

Proposition 13: $\theta\leq\theta_0$ is equivalent to the following condition:

$$\begin{aligned}
\sum_{l\in\mathcal{M}}\frac{|\mathcal{M}|}{\theta}M_l(\frac{\theta}{|\mathcal{M}|}) \;\leq\; &-\frac{|\mathcal{M}|}{\theta}\log[p(\mathcal{S}(\mathcal{M}))\exp(-\frac{\theta F}{|\mathcal{M}|})\\
&+1-p(\mathcal{S}(\mathcal{M}))],
\end{aligned}\tag{64}$$

for all $\mathcal{M}\subset\{1,2,...,N\}$.

Remark: Given θ, Proposition 13 provides a necessary and sufficient condition on the arrival process $A_l(t)$ for it to belong to the effective capacity region.

Proof: According to Lemma 11, we only need to show that (64) is equivalent to

$$\theta\leq\inf_{w>0}\frac{1}{w}\left[H(\vec{\phi}||\vec{p})+L(\vec{f})\right]\tag{65}$$

where $w=\frac{\sum_{l\in\mathcal{M}}f_l-F\sum_{j\in\mathcal{S}(\mathcal{M})}\phi_j}{|\mathcal{M}|}$. To see this, note that

$$\begin{aligned}
&\text{Inequality (65)}\\
\Leftrightarrow\;&\theta w\leq H(\vec{\phi}||\vec{p})+L(\vec{f})\text{ for all }\vec{\phi},\vec{f}\text{ and}\\
&\qquad w=\frac{\sum_{l\in\mathcal{M}}f_l-F\sum_{j\in\mathcal{S}(\mathcal{M})}\phi_j}{|\mathcal{M}|}\\
\Leftrightarrow\;&H(\vec{\phi}||\vec{p})+L(\vec{f})\geq\frac{\theta}{|\mathcal{M}|}\left[\sum_{l\in\mathcal{M}}f_l-F\sum_{j\in\mathcal{S}(\mathcal{M})}\phi_j\right]\\
&\qquad\text{for all }\vec{\phi},\vec{f}\\
\Leftrightarrow\;&\frac{\theta}{|\mathcal{M}|}\sum_{l\in\mathcal{M}}f_l-L(\vec{f})\leq\frac{\theta F}{|\mathcal{M}|}\sum_{j\in\mathcal{S}(\mathcal{M})}\phi_j+H(\vec{\phi}||\vec{p})\\
&\qquad\text{for all }\vec{\phi},\vec{f}\\
\Leftrightarrow\;&\sup_{\vec{f}}\frac{\theta}{|\mathcal{M}|}\sum_{l\in\mathcal{M}}f_l-L(\vec{f})\\
&\qquad\leq-\sup_{\vec{\phi}}\left[-\frac{\theta F}{|\mathcal{M}|}\sum_{j\in\mathcal{S}(\mathcal{M})}\phi_j-H(\vec{\phi}||\vec{p})\right].
\end{aligned}$$

Now by definition, $L_l(f_l)=\sup_\theta[\theta f_l-M_l(\theta)]$. Using properties of Legendre transforms (Lemma 4.5.8 in [24, p152]), we have

$$\sup_{\vec{f}}\frac{\theta}{|\mathcal{M}|}\sum_{l\in\mathcal{M}}f_l-L(\vec{f})=\sum_{l\in\mathcal{M}}M_l(\frac{\theta}{|\mathcal{M}|}).$$

Similarly, by definition

$$H(\vec{\phi}||\vec{p})=\sup_{\vec{\theta}}\left\{\sum_{j=1}^{S}\theta_j\phi_j-\log\mathbf{E}[\exp(\sum_{j=1}^{S}\theta_j\Phi_j)]\right\},$$

where $[\Phi_j]$ is a random vector such that $\Phi_j=\mathbf{1}_{\{C(t)=j\}}$. Note that $[\Phi_j]$ must have exactly one non-zero component and the non-zero component must be equal to 1, and Φ_j takes the value 1 with probability p_j. Hence, using the properties of Legendre transforms again, we have

$$\begin{aligned}
&\sup_{\vec{\phi}}\left[-\frac{\theta F}{|\mathcal{M}|}\sum_{j\in\mathcal{S}(\mathcal{M})}\phi_j-H(\vec{\phi}||\vec{p})\right]\\
=\;&\log\mathbf{E}[\exp(-\sum_{j\in\mathcal{S}(\mathcal{M})}\frac{\theta F}{|\mathcal{M}|}\Phi_j)]\\
=\;&\log[p(\mathcal{S}(\mathcal{M}))\exp(-\frac{\theta F}{|\mathcal{M}|})+(1-p(\mathcal{S}(\mathcal{M})))].
\end{aligned}$$

Hence, (65) is equivalent to

$$\begin{aligned}
\sum_{l\in\mathcal{M}}\frac{|\mathcal{M}|}{\theta}M_l(\frac{\theta}{|\mathcal{M}|}) \;\leq\; &\frac{-|\mathcal{M}|}{\theta}\log[p(\mathcal{S}(\mathcal{M}))\exp(-\frac{\theta F}{|\mathcal{M}|})\\
&+(1-p(\mathcal{S}(\mathcal{M})))].
\end{aligned}$$

Combining these conditions over all $\mathcal{M}$, the result then follows. *Q.E.D.*

The quantity $\frac{M_l(\theta)}{\theta}$ is typically referred to as the effective bandwidth of the arrival process to user l. The quantity on the right-hand-side of (64) can be interpreted as the effective

capacity available to the users in $\mathcal{M}$. Proposition 13 then carries the intuitive explanation that the total effective bandwidth of users in $\mathcal{M}$ must be no greater than the effective capacity available to them. Note that both the effective bandwidth and effective capacity can be computed independently from each other for any given set $\mathcal{M}$.

Remark: Readers can verify that, when the channel states are *i.i.d.* across users and when arrivals from all users are at a constant rate λ, Proposition 13 reduces to Corollary 6 in [11].

IX. Conclusions

In this paper we study the problem of characterizing the queue-overflow probability of complex wireless scheduling and routing algorithms. We present a new technique to address the complexity issue of the multi-dimensional calculus-of-variations problem involved in sample-path large-deviations. Our new technique combines sample-path large-deviations with Lyapunov stability, which may develop into a powerful method to study a large class of scheduling and routing algorithms. We also show that when a scheduling and routing algorithm minimizes the drift of the Lyapunov function at every time in every fluid-sample-path, it is optimal in maximizing the asymptotic decay-rate of the probability that the Lyapunov function value exceeds a threshold. We illustrate the potential of this approach through examples.

Acknowledgment

The authors would like to thank the anonymous reviewers for their constructive comments and suggestions that greatly enhance the quality of the paper.

Appendix

Proof of Theorem 4:

Pick any $0 < \rho < 1$. Pick another two positive constants $0 < \delta < \epsilon < \rho$. Let $\mathbf{P}$ denote the stationary probability distribution of the system. We are interested in the following quantity

$$\limsup_{B\to\infty} \frac{1}{B} \log \mathbf{P}[V(\vec{X}(\infty)/B) \geq 1].$$

Throughput this appendix, we will focus on the scaled version of $\vec{X}(\tau)$ such that

$$x_i^{k,B}(t) = \frac{1}{B} X_i^k(Bt),$$

for $t = \frac{m}{B}$, $m = 0, 1, 2, ...,$, and by linear interpolation otherwise. Let $\vec{x}^B(t) = [x_i^{k,B}(t), i = 1, ..., N, k = 1, ..., K]$. Note that this definition is almost identical to (10) in Section III-B, except that now $\vec{x}^B(t)$ is defined for infinitely large t. Further, since we are interested in the stationary distribution $\vec{X}(\infty)$, we can assume that $\vec{X}(\tau)$ starts with its stationary distribution at $\tau = 0$. Hence, $\vec{X}(\tau)$ will admit its stationary distribution at every time instant τ. As a result, $\vec{x}^B(t)$ will also admit its stationary distribution at every $t = \frac{m}{B}$, $m = 0, 1, 2, ...,$, which is the same as the distribution of $\vec{X}(\infty)/B$.

Let $\lceil t \rceil$ denote the smallest integer that is greater than or equal to t. For each B, consider the following sequence of stopping times defined on a sample path $\mathbf{x}^B$. (Here we use the notation from [13].)

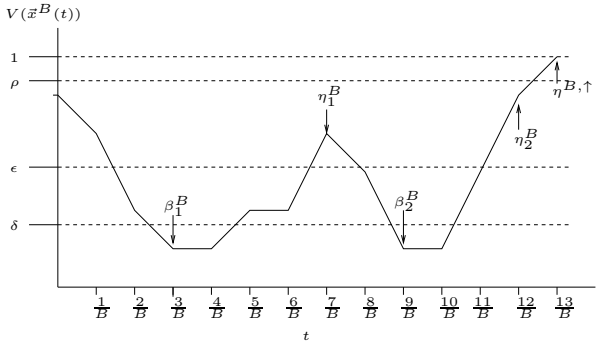

Fig. 4. An example of the trajectory of $V(\vec{x}^B(t))$ and the different stopping times.

$$\beta_1^B \triangleq \frac{\lceil \inf\{t \geq 0 : V(\vec{x}^B(t)) \leq \delta\} B \rceil}{B},$$

and

$$\eta_n^B \triangleq \frac{\lceil \inf\{t \geq \beta_n^B : V(\vec{x}^B(t)) \geq \epsilon\} B \rceil}{B}, \quad n = 1, 2, \ldots$$
$$\beta_n^B \triangleq \frac{\lceil \inf\{t \geq \eta_{n-1}^B : V(\vec{x}^B(t)) \leq \delta\} B \rceil}{B}, \quad n = 2, 3, \ldots$$

Fig. 4 provides an example of these stopping times.

Consider the Markov chain $\hat{\vec{x}}^B(n)$ obtained by sampling $\vec{x}^B(t)$ at the stopping times η_n^B, i.e,. $\hat{\vec{x}}^B(n) = \vec{x}^B(\eta_n^B)$, $n = 1, 2, ..., +\infty$. Since $\mathbf{x}^B$ is stationary, there must also exist a stationary distribution for the Markov chain $\hat{\vec{x}}^B(n)$. Denote this stationary distribution (of the Markov chain $\hat{\vec{x}}^B(n)$) by $\hat{\mathbf{P}}^B$. Further, let Θ^B denote the state space of the Markov chain $\hat{\vec{x}}^B(n)$. We can then express the stationary distribution of $\vec{X}(\infty)/B$ as (see [25, Lemma 10.1]):

$$\begin{aligned} &\mathbf{P}[V(\vec{X}(\infty)/B) \geq 1] \\ = \quad &\frac{\int_{\Theta^B} \hat{\mathbf{P}}^B(d\vec{x}) \mathbf{E}_{\vec{x}}(\int_0^{\eta_1^B} \mathbf{1}_{\{V(\vec{x}^B(t)) \geq 1\}} dt)}{\int_{\Theta^B} \hat{\mathbf{P}}^B(d\vec{x}) \mathbf{E}_{\vec{x}}(\eta_1^B)} \end{aligned} \tag{66}$$

where $\mathbf{E}_{\vec{x}}(\cdot)$ denotes the expectation conditioned on the event that $\vec{x}^B(0) = \vec{x}$.

Remark on generalizing the proof for Markovian channel and arrival processes: As discussed in Section II-E, the proof will have to be modified if one wishes to assume that the channel and arrival processes are Markov chains. In the proof, we make use of the fact that the queue process is Markovian. This enables us to claim that the sampled process, $\hat{\vec{x}}^B(n)$, is a Markov chain. This is no longer true if one uses channel and arrival processes that are Markov chains. However, it is relatively simple to fix the proof by considering the system state as the joint state of the channel, arrival and the queue backlog. For instance, if $\chi(t)$ represents the combined state of the channel, arrival and queue backlog processes, one can assume that $\chi(\eta_n^B)$ is a Markov chain. Then, our proof methodology can still apply by denoting $\hat{\mathbf{P}}^B$ as the stationary probability of $\chi(\eta_n^B)$ and denoting Θ^B as the state space of $\chi(\eta_n^B)$.

A. Bounding the Denominator of (66)

Consider first the denominator in (66). From (1) and the boundedness of both the arrival-rate $A_i^k(t)$ and the service-rate $E_l^k(j,\vec{X}(t))$, there exists an upper bound M_1 such that $||\vec{X}(t+1)-\vec{X}(t)|| \le M_1$ for all t. From Assumption 1, we have that $V(\vec{x}^B(\eta_1^B))-V(\vec{x}^B(\beta_1^B)) \le \mathcal{L}||\vec{x}^B(\eta_1^B)-\vec{x}^B(\beta_1^B)||$. Denote $M_0 \triangleq \mathcal{L}M_1$. Hence, we must have,

$$\begin{aligned} V(\vec{x}^B(\eta_1^B)) - V(\vec{x}^B(\beta_1^B)) &\le \mathcal{L}||\vec{x}^B(\eta_1^B) - \vec{x}^B(\beta_1^B)|| \\ &\le (\eta_1^B - \beta_1^B)M_0. \end{aligned}$$

Further, since η_1^B is the first time $V(\vec{x}^B(t))$ exceeds ϵ, we must have

$$V(\vec{x}^B(\eta_1^B)) \ge \epsilon - \frac{M_0}{B}.$$

Similarly, we have $V(\vec{x}^B(\beta_1^B)) \le \delta + \frac{M_0}{B}$. Therefore,

$$\begin{aligned} \mathbf{E}_{\vec{x}}(\eta_1^B) &\ge \mathbf{E}(\eta_1^B - \beta_1^B) \\ &\ge \frac{1}{M_0}\left(\epsilon - \delta - \frac{2M_0}{B}\right). \end{aligned}$$

Thus, there exists $B_1 > 0$ such that for all $B \ge B_1$, the denominator of (66) can be bounded from below by

$$\mathbf{E}_{\vec{x}}(\eta_1^B) \ge \frac{\epsilon-\delta}{2M_0}. \tag{67}$$

B. Bounding the Numerator of (66)

We next estimate the asymptotics of the numerator of (66). Recall that, by definition, each η_n^B is at most one $\frac{1}{B}$ time-unit after $V(\vec{x}^B(t))$ just exceeds ϵ. Since $\epsilon < \rho$, using the boundedness of the arrival rates and the service-rates, we can conclude that there exists $B_2 > 0$ (which depends on $\rho-\epsilon$), such that $V(\vec{x}^B(\eta_n^B)) \le \rho$ for all $B \ge B_2$.

We next define the following additional stopping time (see Fig. 4 for an example):

$$\eta^{B,\uparrow} \triangleq \frac{\lceil \inf\{t \ge 0 : V(\vec{x}^B(t)) \ge 1\}B \rceil}{B}.$$

Then, for any $\vec{x} \in \Theta^B$, we must have,

$$\begin{aligned} &\mathbf{E}_{\vec{x}}\left[\int_0^{\eta_1^B} \mathbf{1}_{\{V(\vec{x}^B(t))\ge 1\}} dt\right] \\ \le\ &\mathbf{E}_{\vec{x}}\left[\mathbf{1}_{\{\eta^{B,\uparrow} \le \beta_1^B\}}(\beta_1^B - \eta^{B,\uparrow})\right]. \end{aligned}$$

The above inequality holds because: (a) if β_1^B occurs before $\eta^{B,\uparrow}$, then both sides will be zero; and (b) if β_1^B occurs after $\eta^{B,\uparrow}$, then the amount of time $V(\vec{x}^B(t)) \ge 1$ must be no greater than $\beta_1^B - \eta^{B,\uparrow}$. Let $\mathbf{P}_{\vec{x}}$ denote the probability distribution conditioned on $\vec{x}^B(0) = \vec{x}$. We then have

$$\begin{aligned} &\mathbf{E}_{\vec{x}}\left[\int_0^{\eta_1^B} \mathbf{1}_{\{V(\vec{x}^B(t))\ge 1\}} dt\right] \\ \le\ &\mathbf{E}_{\vec{x}}\left[\beta_1^B - \eta^{B,\uparrow} | \eta^{B,\uparrow} \le \beta_1^B\right] \mathbf{P}_{\vec{x}}(\eta^{B,\uparrow} \le \beta_1^B) \end{aligned}$$

Now, $\mathbf{E}_{\vec{x}}\left[\beta_1^B - \eta^{B,\uparrow} | \eta^{B,\uparrow} \le \beta_1^B\right]$ is equal to $\mathbf{E}_{\vec{x}}\left\{\mathbf{E}_{\vec{x}}[\beta_1^B - \eta^{B,\uparrow} | \vec{x}^B(\eta^{B,\uparrow}), \eta^{B,\uparrow} \le \beta_1^B] | \eta^{B,\uparrow} \le \beta_1^B\right\}$ and, due to the Markovian property of $\mathbf{x}^B$, we must have $\mathbf{E}_{\vec{x}}[\beta_1^B - \eta^{B,\uparrow} | \vec{x}^B(\eta^{B,\uparrow}), \eta^{B,\uparrow} \le \beta_1^B] = \mathbf{E}_{\vec{x}^B(\eta^{B,\uparrow})}(\beta_1^B)$. Hence,

$$\begin{aligned} &\mathbf{E}_{\vec{x}}\left[\int_0^{\eta_1^B} \mathbf{1}_{\{V(\vec{x}^B(t))\ge 1\}} dt\right] \\ \le\ &\mathbf{E}_{\vec{x}}\left[\mathbf{E}_{\vec{x}^B(\eta^{B,\uparrow})}(\beta_1^B) | \eta^{B,\uparrow} \le \beta_1^B\right] \mathbf{P}_{\vec{x}}(\eta^{B,\uparrow} \le \beta_1^B). \end{aligned}$$

Let C be a number that is slightly larger than 1. Using the boundedness of the arrival rates and the service-rates again, we can find $B_3 > 0$ such that for all $B \ge B_3$, $V(\vec{x}^B(\eta^{B,\uparrow})) \le C$. Hence, we can bound the above quantity by

$$\begin{aligned} &\mathbf{E}_{\vec{x}}\left[\int_0^{\eta_1^B} \mathbf{1}_{\{V(\vec{x}^B(t))\ge 1\}} dt\right] \\ \le\ &\left[\sup_{\{\vec{y}:V(\vec{y})\le C\}} \mathbf{E}_{\vec{y}}(\beta_1^B)\right] \mathbf{P}_{\vec{x}}(\eta^{B,\uparrow} \le \beta_1^B). \end{aligned}$$

Let T be a positive number (which will be chosen later). Recall that $V(\vec{x}) \le \rho$ for all $\vec{x} \in \Theta^B$ when $B \ge B_2$. Hence, for any such $\vec{x} \in \Theta^B$, we have,

$$\begin{aligned} &\mathbf{E}_{\vec{x}}\left[\int_0^{\eta_1^B} \mathbf{1}_{\{V(\vec{x}^B(t))\ge 1\}} dt\right] \\ \le\ &\left[\sup_{\{\vec{y}:V(\vec{y})\le C\}} \mathbf{E}_{\vec{y}}(\beta_1^B)\right] \left[\mathbf{P}_{\vec{x}}(\eta^{B,\uparrow} \le T) \right. \\ &\qquad \left. + \mathbf{P}_{\vec{x}}(\beta_1^B \ge T)\right] \\ \le\ &\left[\sup_{\{\vec{y}:V(\vec{y})\le C\}} \mathbf{E}_{\vec{y}}(\beta_1^B)\right] \left[\sup_{\{\vec{x}:V(\vec{x})\le \rho\}} \mathbf{P}_{\vec{x}}(\eta^{B,\uparrow} \le T) \right. \\ &\qquad \left. + \sup_{\{\vec{x}:V(\vec{x})\le \rho\}} \mathbf{P}_{\vec{x}}(\beta_1^B \ge T))\right]. \end{aligned} \tag{68}$$

Substituting (67) and (68) into (66), we then have, for all $B \ge \max\{B_1, B_2, B_3\}$,

$$\begin{aligned} &\mathbf{P}[V(\vec{X}(\infty)/B) \ge 1] \\ \le\ &\frac{2M_0}{\epsilon-\delta}\left[\sup_{\{\vec{y}:V(\vec{y})\le C\}} \mathbf{E}_{\vec{y}}(\beta_1^B) \times \right. \\ &\qquad \left. \left(\sup_{\{\vec{x}:V(\vec{x})\le \rho\}} \mathbf{P}_{\vec{x}}(\eta^{B,\uparrow} \le T) \right.\right. \\ &\qquad \left.\left. + \sup_{\{\vec{x}:V(\vec{x})\le \rho\}} \mathbf{P}_{\vec{x}}(\beta_1^B \ge T)\right)\right]. \end{aligned} \tag{69}$$

We next study the asymptotics for each term in the above inequality.

1) Bound for $\sup_{\{\vec{y}:V(\vec{y})\le C\}} \mathbf{E}_{\vec{y}}(\beta_1^B)$*:* We will show that $\sup_{\{\vec{y}:V(\vec{y})\le C\}} \mathbf{E}_{\vec{y}}(\beta_1^B)$ is bounded from above and hence does not affect the asymptotics of (69). Due to the continuity of the Lyapunov function and the assumption that $V(\vec{x}) = 0$ only if $||\vec{x}|| = 0$, there exists a $\gamma > 0$ such that $||\vec{x}|| < \gamma$ implies $V(\vec{x}) < \delta$. Further, we can find a $K > 0$ such that $V(\vec{x}) < C$ implies $||\vec{x}|| < K$.

Consider the following additional stopping time

$$\hat{\beta}_1^B \triangleq \frac{\lceil \inf\{t \ge 0 : ||\vec{x}^B(t)|| \le \gamma\}B \rceil}{B}.$$

It is easy to see that $\sup_{\{\vec{y}:||\vec{y}||\leq K\}} \mathbf{E}_{\vec{y}}(\hat{\beta}_1^B)$ is an upper bound on $\sup_{\{\vec{y}:V(\vec{y})\leq C\}} \mathbf{E}_{\vec{y}}(\beta_1^B)$.

We now proceed to show that $\sup_{\{\vec{y}:||\vec{y}||\leq K\}} \mathbf{E}_{\vec{y}}(\hat{\beta}_1^B)$ is bounded from above. From Assumption 1, the fluid limit of the system satisfies either (14) or (15). For both cases, it follows that there exists a constant t_0 such that for all fluid limits $\mathbf{x}$ with $||\vec{x}(0)|| \leq 1$, we must have $||\vec{x}(t_0)|| = 0$. This not only implies that the original system is stable (see [18, Theorem 4.2]), but also leads to the following limit:

$$\lim_{||\vec{x}||\to\infty} \frac{1}{||\vec{x}||}\mathbf{E}\left[\vec{X}(t_0||\vec{x}||)\Big|\vec{X}(0)=\vec{x}\right] = 0.$$

(See the proof of Theorem 4.2 in [18]). Then using the techniques in the proof of Theorem 3.1 in [18], there must exist numbers $\tilde{\epsilon} > 0$, $\kappa > 0$, $\tilde{b} \geq 0$, and a bounded set $\mathcal{B} \triangleq \{\vec{X} : ||\vec{X}|| \leq \kappa\}$ such that for all $\vec{x}$, conditioned on $\vec{X}(0) = \vec{x}$, the following holds,

$$\mathbf{E}[\tau_{\mathcal{B}}(t_0)|\vec{X}(0)=\vec{x}] \leq \frac{t_0}{\tilde{\epsilon}}(||\vec{x}|| + \tilde{b}),$$

where $\tau_{\mathcal{B}}(t_0) \triangleq \inf\{t \geq t_0 : \vec{X}(t) \in \mathcal{B}\}$ is the first time after t_0 when $\vec{X}(t)$ returns to the set $\mathcal{B}$.

Recall the transformation $\vec{x}^B(t) = \frac{1}{B}\vec{X}(Bt)$, $t = 0, \frac{1}{B}, \frac{2}{B}, ...$ For all $B \geq \frac{\kappa}{\gamma}$, as long as $||\vec{X}(Bt)|| \leq \kappa$, it implies that $||\vec{x}(t)|| \leq \gamma$ and thus $\hat{\beta}_1^B \leq t$. Hence, for any $\vec{y}$ such that $||\vec{y}|| \leq K$ and for any $B \geq \frac{\kappa}{\gamma}$, the following holds

$$\begin{aligned}\mathbf{E}_{\vec{y}}(\hat{\beta}_1^B) &\leq \frac{1}{B}\mathbf{E}[\tau_{\mathcal{B}}(t_0)|\vec{X}(0)=B\vec{y}] \\ &\leq \frac{t_0}{\tilde{\epsilon}B}(B||\vec{y}|| + \tilde{b}).\end{aligned}$$

Let $B_5 = \max\{\frac{\kappa}{\gamma}, \frac{\tilde{b}}{K}\}$. Then, for all $B \geq B_5$, we have,

$$\sup_{\{\vec{y}:V(\vec{y})\leq C\}} \mathbf{E}_{\vec{y}}(\beta_1^B) \leq \sup_{\{\vec{y}:||\vec{y}||\leq K\}} \mathbf{E}_{\vec{y}}(\hat{\beta}_1^B) \leq 2\frac{Kt_0}{\tilde{\epsilon}}. \quad (70)$$

2) Asymptotics for $\sup_{\{\vec{x}:V(\vec{x})\leq\rho\}} \mathbf{P}_{\vec{x}}(\eta^{B,\uparrow} \leq T)$*:* Let

$$\begin{aligned}\Gamma_{\leq\rho} \triangleq \{\mathbf{x} : &V(\vec{x}(0)) \leq \rho \text{ and } V(\vec{x}(t)) \geq 1 \\ &\text{for some } t \in (0,T]\}\end{aligned}$$

Then, by Proposition 2, we have

$$\begin{aligned}&\limsup_{B\to\infty} \frac{1}{B}\log \sup_{\{\vec{x}:V(\vec{x})\leq\rho\}} \mathbf{P}_{\vec{x}}(\eta^{B,\uparrow} \leq T) \\ =\ &\limsup_{B\to\infty} \frac{1}{B}\log \sup_{\{\vec{x}:V(\vec{x})\leq\rho\}} \mathbf{P}_{\vec{x}}(\mathbf{x}^B \in \Gamma_{\leq\rho}) \\ \leq\ &-\inf_{\text{all FSP}_{(\mathbf{s},\mathbf{a},\mathbf{x})_T:\mathbf{x}\in\Gamma_{\leq\rho}}} \int_0^T \left[H\left(\frac{d}{dt}\vec{s}(t)||\vec{p}\right)\right. \\ &\left. + L\left(\frac{d}{dt}\vec{a}(t)\right)\right] dt. \quad (71)\end{aligned}$$

3) Asymptotics for $\sup_{\{\vec{x}:V(\vec{x})\leq\rho\}} \mathbf{P}_{\vec{x}}[\beta_1^B \geq T]$*:* Let

$$\begin{aligned}\Upsilon_{\leq\rho} \triangleq \{\mathbf{x} : &V(\vec{x}(0)) \leq \rho \text{ and } V(\vec{x}(t)) > \delta \\ &\text{for all } t \in [0,T-1]\}.\end{aligned}$$

Then, by Proposition 2, we have

$$\begin{aligned}&\limsup_{B\to\infty} \frac{1}{B}\log \sup_{\{\vec{x}:V(\vec{x})\leq\rho\}} \mathbf{P}_{\vec{x}}[\beta_1^B \geq T] \\ \leq\ &\limsup_{B\to\infty} \frac{1}{B}\log \sup_{\{\vec{x}:V(\vec{x})\leq\rho\}} \mathbf{P}_{\vec{x}}[\mathbf{x}^B \in \Upsilon_{\leq\rho}] \\ \leq\ &-\inf_{\text{all FSP}_{(\mathbf{s},\mathbf{a},\mathbf{x})_T:\mathbf{x}\in\Upsilon_{\leq\rho}} \int_0^{T-1} \left[H\left(\frac{d}{dt}\vec{s}(t)||\vec{p}\right)\right. \\ &\left. + L\left(\frac{d}{dt}\vec{a}(t)\right)\right] dt. \quad (72)\end{aligned}$$

For any FSP $(\mathbf{s},\mathbf{a},\mathbf{x})_T$ such that $\mathbf{x} \in \Upsilon_{\leq\rho}$, we have

$$\begin{aligned}\delta &\leq V(\vec{x}(0)) + \int_0^{T-1} \frac{d}{dt}V(\vec{x}(t))dt \\ &\leq \rho + \int_0^{T-1} \frac{d}{dt}V(\vec{x}(t))dt.\end{aligned}$$

We now need to use Assumption 2. Since the two parts in Assumption 2 are equivalent, in the following we will assume the latter part holds*. Let η be defined as in Assumptions 1 and 2. Then, according to Assumption 2, there exists $\epsilon' > 0$ such that for all FSPs $(\mathbf{s},\mathbf{a},\mathbf{x})_T$, if at any time t we have $||\frac{d}{dt}\vec{s}(t) - \vec{p}|| \leq \epsilon'$ and $||\frac{d}{dt}\vec{a}(t) - \vec{\lambda}|| \leq \epsilon'$, then the following holds

$$\frac{d}{dt}V(\vec{x}(t)) \leq -\eta/2.$$

Further, there exists $M_1 \geq 0$ such that if at any time t we have $||\frac{d}{dt}\vec{s}(t) - \vec{p}|| \geq \epsilon'$ or $||\frac{d}{dt}\vec{a}(t) - \vec{\lambda}|| \geq \epsilon'$, then the following holds

$$\frac{d}{dt}V(\vec{x}(t)) \leq M_1.$$

Let $\mathcal{M}$ denote the set of $(\vec{\phi},\vec{f})$ such that $||\vec{\phi} - \vec{p}|| \leq \epsilon'$ and $||\vec{f} - \vec{\lambda}|| \leq \epsilon'$. We then have,

$$\begin{aligned}\delta \leq \rho + \int_0^{T-1} \Big[&-\frac{\eta}{2}\mathbf{1}_{\{(\frac{d}{dt}\vec{s}(t),\frac{d}{dt}\vec{a}(t))\in\mathcal{M}\}} \\ &+M_1\mathbf{1}_{\{(\frac{d}{dt}\vec{s}(t),\frac{d}{dt}\vec{a}(t))\notin\mathcal{M}\}}\Big] dt.\end{aligned}$$

Hence,

$$\begin{aligned}&\left(M_1 + \frac{\eta}{2}\right)\int_0^{T-1} \mathbf{1}_{\{(\frac{d}{dt}\vec{s}(t),\frac{d}{dt}\vec{a}(t))\notin\mathcal{M}\}} dt \\ \geq\ &(T-1)\frac{\eta}{2} + \delta - \rho. \quad (73)\end{aligned}$$

Let

$$\begin{aligned}J_{\min} = \min_{\vec{\phi},\vec{f}}\quad &J(\vec{\phi},\vec{f}) \\ \text{subject to}\quad &||\vec{\phi} - \vec{p}|| \geq \epsilon' \text{ or } ||\vec{f} - \vec{\lambda}|| \geq \epsilon'.\end{aligned}$$

*If the first part of Assumption 2 holds, the following proof can be easily modified by using the Lyapunov function $U(\vec{x}) = \frac{V^{1-\alpha}(\vec{x})}{1-\alpha}$.

It is easy to see that $J_{\min}$ is positive. Thus, for any FSP $(\mathbf{s},\mathbf{a},\mathbf{x})_T$ such that $\mathbf{x}\in\Upsilon_{\leq\rho}$, we have

$$\begin{aligned}&\int_0^T\left[H\left(\frac{d}{dt}\vec{s}(t)||\vec{p}\right)+L\left(\frac{d}{dt}\vec{a}(t)\right)\right]dt \quad (74)\\ \geq\ &\int_0^{T-1}J_{\min}\mathbf{1}_{\{(\frac{d}{dt}\vec{s}(t),\frac{d}{dt}\vec{a}(t))\notin\mathcal{M}\}}dt\\ \geq\ &\frac{(T-1)\frac{\eta}{2}+\delta-\rho}{M_1+\eta/2}J_{\min}.\end{aligned}$$

where the last inequality follows from (73). Substituting into (72), we then have

$$\begin{aligned}&\limsup_{B\to\infty}\frac{1}{B}\log\sup_{\{\vec{x}:V(\vec{x})\leq\rho\}}\mathbf{P}_{\vec{x}}[\beta_1^B\geq T] \quad (75)\\ &\qquad\leq-\frac{(T-1)\frac{\eta}{2}+\delta-\rho}{M_1+\eta/2}J_{\min}.\end{aligned}$$

Clearly, for fixed δ and ρ, by choosing large T, we can make the right-hand-side arbitrarily small.

C. Completing the Proof of Theorem 4

We are now ready to prove the statement of Theorem 4. Pick any FSP $(\mathbf{s},\mathbf{a},\mathbf{x})_{T_0}$ such that $\vec{x}(0)=0$, $V(\vec{x}(T_0))\geq 1$. Suppose the cost of this FSP is $\tilde{J}$, i.e.

$$\int_0^{T_0}\left[H\left(\frac{d}{dt}\vec{s}(t)||\vec{p}\right)+L\left(\frac{d}{dt}\vec{a}(t)\right)\right]dt=\tilde{J}.$$

Clearly, for any $T\geq T_0$, the right-hand-side of (71) must be no smaller than $-\tilde{J}$. According to (75), for fixed δ and ρ there must exist $T_1>T_0$ (which is independent of ρ), such that for all $T\geq T_1$, the right-hand-side of (75) is smaller than $-\tilde{J}$. Fix such a $T\geq T_1$. Substituting (70), (71) and (75) into (69), and taking the appropriate limits, we then have,

$$\begin{aligned}&\limsup_{B\to\infty}\frac{1}{B}\log\mathbf{P}[V(\vec{X}(\infty)/B)\geq 1]\\ \leq\ &-\inf_{\text{all FSP}_{(\mathbf{s},\mathbf{a},\mathbf{x})_T:\mathbf{x}\in\Gamma_{\leq\rho}}}\int_0^T H\left(\frac{d}{dt}\vec{s}(t)||\vec{p}\right)\\ &\qquad+L\left(\frac{d}{dt}\vec{a}(t)\right)dt.\end{aligned}$$

Note that the above inequality holds for all $\rho>0$. Let J_ρ denote the infimum on the right-hand-side. As $\rho\to 0$, let J^* denote the limit, i.e., $J^*=\lim_{\rho\to 0}J_\rho$. We then have

$$\limsup_{B\to\infty}\frac{1}{B}\log\mathbf{P}[V(\vec{X}(\infty)/B)\geq 1]\quad\leq\quad -J^*.$$

Let

$$\begin{aligned}J_0\quad=\quad&\inf_{\substack{\text{all FSP}_{(\mathbf{s},\mathbf{a},\mathbf{x})_T:\\ \vec{x}(0)=0,V(\vec{x}(T))\geq 1}}\int_0^T H\left(\frac{d}{dt}\vec{s}(t)||\vec{p}\right)\\ &\qquad+L\left(\frac{d}{dt}\vec{a}(t)\right)dt.\end{aligned}$$

It only remains to show that $J^*\geq J_0$. To see this, take a sequence $\rho_n\to 0$. There must exist a sequence of FSPs $(\mathbf{s}_n,\mathbf{a}_n,\mathbf{x}_n)_T$ such that $V(\vec{x}_n(0))\leq\rho_n$ and $V(\vec{x}_n(T))\geq 1$ for each n, and

$$\lim_{n\to\infty}\int_0^T\left[H\left(\frac{d}{dt}\vec{s}_n(t)||\vec{p}\right)+L\left(\frac{d}{dt}\vec{a}_n(t)\right)\right]dt=J^*.$$

Take a further subsequence that converges uniformly over compact intervals. Without loss of generality, we can denote this subsequence also by $(\mathbf{s}_n,\mathbf{a}_n,\mathbf{x}_n)_T$, and let $(\underline{\mathbf{s}},\underline{\mathbf{a}},\underline{\mathbf{x}})_T$ be the corresponding limit. Then, using the lower-semicontinuity of the cost function, we must have

$$\int_0^T H\left(\frac{d}{dt}\underline{\vec{s}}(t)||\vec{p}\right)+L\left(\frac{d}{dt}\underline{\vec{a}}(t)\right)dt\leq J^*.$$

Using a similar argument as in the proof of Proposition 2, we can show that $(\underline{\mathbf{s}},\underline{\mathbf{a}},\underline{\mathbf{x}})_T$ is also an FSP, and it satisfies the condition that $\underline{\mathbf{x}}(0)=0$ and $V(\underline{\mathbf{x}}(T))\geq 1$. Hence, it belongs to the set of FSP in the constraint set in (76). We thus have $J_0\leq J^*$. In other words, we have shown that for all $T\geq T_1$,

$$\begin{aligned}&\limsup_{B\to\infty}\frac{1}{B}\log\mathbf{P}[V(\vec{X}(\infty)/B)\geq 1]\\ \leq\ &-\inf_{\substack{\text{all FSP}_{(\mathbf{s},\mathbf{a},\mathbf{x})_T:\\ \vec{x}(0)=0\text{ and }V(\vec{x}(T))\geq 1}}\int_0^T H\left(\frac{d}{dt}\vec{s}(t)||\vec{p}\right)\\ &\qquad+L\left(\frac{d}{dt}\vec{a}(t)\right)dt.\end{aligned}$$

Note that the above inequality is for a fixed $T\geq T_1$. Note that the infimum on the right-hand-side decreases as T increases. Hence, taking another infimum over all $T>0$, we must then have

$$\begin{aligned}&\limsup_{B\to\infty}\frac{1}{B}\log\mathbf{P}[V(\vec{X}(\infty)/B)\geq 1]\\ \leq\ &-\inf_{\substack{\text{all FSP}_{(\mathbf{s},\mathbf{a},\mathbf{x})_T,T>0:\\ \vec{x}(0)=0\text{ and }V(\vec{x}(T))\geq 1}}\int_0^T H\left(\frac{d}{dt}\vec{s}(t)||\vec{p}\right)\\ &\qquad+L\left(\frac{d}{dt}\vec{a}(t)\right)dt.\end{aligned}$$

REFERENCES

[1] L. Tassiulas and A. Ephremides, "Stability Properties of Constrained Queueing Systems and Scheduling Policies for Maximum Throughput in Multihop Radio Networks," *IEEE Transactions on Automatic Control*, vol. 37, no. 12, pp. 1936–1948, December 1992.

[2] F. P. Kelly, "Effective Bandwidth in Multiclass Queues," *Queueing Systems*, vol. 9, pp. 5–16, 1991.

[3] A. I. Elwalid and D. Mitra, "Effective Bandwidth of General Markovian Traffic Sources and Admission Control of High Speed Networks," *IEEE/ACM Transactions on Networking*, vol. 1, no. 3, pp. 329–343, June 1993.

[4] G. Kesidis, J. Walrand, and C.-S. Chang, "Effective Bandwidth for Multiclass Markov Fluids and other ATM Sources," *IEEE/ACM Transactions on Networking*, vol. 1, no. 4, pp. 424–428, Aug. 1993.

[5] D. D. Botvich and N. G. Duffield, "Large Deviations, the Shape of the Loss Curve, and Economies of Scale in Large Multiplexers," *Queueing Systems*, vol. 20, pp. 293–320, 1995.

[6] N. G. Duffield and N. O'Connell, "Large deviations and overflow probabilities for the general single server queue, with application," *Math. Proc. Cambr. Phil. Soc.*, vol. 118, pp. 363–374, 1995.

[7] C. Courcoubetis and R. Weber, "Buffer Overflow Asymptotics for a Buffer Handling Many Traffic Sources," *Journal of Applied Probability*, vol. 33, pp. 886–903, 1996.

[8] D. Wu and R. Negi, "Effective Capacity: A Wireless Link Model for Support of Quality of Service," *IEEE Transactions on Wireless Communications*, vol. 2, no. 4, pp. 630–643, July 2003.
[9] A. Eryilmaz and R. Srikant, "Scheduling with Quality of Service Constraints over Rayleigh Fading Channels," in *Proceedings of the IEEE Conference on Decision and Control*, December 2004.
[10] S. Shakkottai, "Effective Capacity and QoS for Wireless Scheduling," *IEEE Transactions on Automatic Control*, vol. 53, no. 3, April 2008.
[11] L. Ying, R. Srikant, A. Eryilmaz, and G. E. Dullerud, "A Large Deviations Analysis of Scheduling in Wireless Networks," *IEEE Transactions on Information Theory*, vol. 52, no. 11, November 2006.
[12] D. Bertsimas, I. C. Paschalidis, and J. N. Tsitsiklis, "Asymptotic Buffer Overflow Probabilities in Multiclass Multiplexers: An Optimal Control Approach," *IEEE Transactions on Automatic Control*, vol. 43, no. 3, pp. 315–335, March 1998.
[13] A. L. Stolyar, "Large Deviations of Queues Sharing a Randomly Time-varying Server," *Queueing Systems*, vol. 59, 2008.
[14] V. J. Venkataramanan, X. Lin, L. Ying, and S. Shakkottai, "On Scheduling for Minimizing End-to-End Buffer Usage over Multihop Wireless Networks," in *Proceedings of IEEE INFOCOM*, March 2010.
[15] V. J. Venkataramanan and X. Lin, "Low-Complexity Scheduling Algorithm for Sum-Queue Minimization in Wireless Convergecast," in *Proceedings of IEEE INFOCOM*, April 2011.
[16] L. Tassiulas and A. Ephremides, "Dynamic Scheduling for Minimum Delay in Tandem and Parallel Constrained Queueing Models," *Annals of Operation Research*, vol. 48, pp. 333–355, 1994.
[17] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. New York: Springer-Verlag, 1998.
[18] J. G. Dai, "On Positive Harris Recurrence of Multiclass Queueing Networks: A Unified Approach via Fluid Limit Models," *Annals of Applied Probability*, vol. 5, no. 1, pp. 49–77, 1995.
[19] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, R. Vijayakumar, and P. Whiting, "Scheduling in a queuing system with asynchronously varying service rates," *Probability in the Engineering and Informational Sciences*, vol. 18, pp. 191–218, 2004.
[20] M. J. Neely, E. Modiano, and C. E. Rohrs, "Dynamic Power Allocation and Routing for Time Varying Wireless Networks," *IEEE Journal on Selected Areas in Communications, Special Issue on Wireless Ad-Hoc Networks*, vol. 23, no. 1, pp. 89–103, January 2005.
[21] A. Eryilmaz, R. Srikant, and J. Perkins, "Stable Scheduling Policies for Fading Wireless Channels," *IEEE/ACM Transactions on Networking*, vol. 13, no. 2, pp. 411–424, April 2005.
[22] V. J. Venkataramanan and X. Lin, "On Wireless Scheduling Algorithms for Minimizing the Queue-Overflow Probability," *IEEE/ACM Transactions on Networking*, vol. 18, no. 3, pp. 788–801, Jun. 2010.
[23] A. Shwartz and A. Weiss, *Large Deviations for Performance Analysis: Queues, Communications, and Computing*. London: Chapman & Hall, 1995.
[24] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Boston: Jones and Bartlett, 1993.
[25] A. L. Stolyar, "Control of End-to-End Delay Tails in a Multiclass Networks: LWDF Discipline Optimality," *Annals of Applied Probability*, vol. 13, no. 3, pp. 1151–1206, 2003.

V. J. Venkataramanan received his B.Tech degree in Electrical Engineering from the Indian Institute of Technology Madras, India in 2006 and his Ph.D degree in Electrical and Computer Engineering from Purdue University, West Lafayette, IN in 2010. He is currently an engineer at Qualcomm Research. His research interests are in modeling and evaluation of communication networks.

Xiaojun Lin (S'02 / M'05 /SM'12) received his B.S. from Zhongshan University, Guangzhou, China, in 1994, and his M.S. and Ph.D. degrees from Purdue University, West Lafayette, Indiana, in 2000 and 2005, respectively. He is currently an Associate Professor of Electrical and Computer Engineering at Purdue University.

Dr. Lin's research interests are in the analysis, control and optimizatoin of wireless and wireline networks. He received the IEEE INFOCOM 2008 best paper award and 2005 best paper of the year award from *Journal of Communications and Networks*. His paper was also one of two runner-up papers for the best-paper award at IEEE INFOCOM 2005. He received the NSF CAREER award in 2007. He was the Workshop co-chair for IEEE GLOBECOM 2007, the Panel co-chair for WICON 2008, the TPC co-chair for ACM MobiHoc 2009, and the Mini-Conference co-chair for IEEE INFOCOM 2012. He is currently serving as an Area Editor for (Elsevier) Computer Networks journal and an Associate Editor for IEEE/ACM Transactions on Networking, and has also served as a Guest Editor for (Elsevier) Ad Hoc Networks journal.