Many-Sources Large Deviations for Max-Weight Scheduling

Vijay G. Subramanian, Member, IEEE, Tara Javidi, Member, IEEE, and Somsak Kittipiyakul, Member, IEEE

Abstract—In this paper, a many-sources large deviations principle (LDP) for the transient workload of a multi-queue single-server system is established where the service rates are chosen from a compact, convex and coordinate-convex rate region and where the service discipline is the max-weight policy. Under the assumption that the arrival processes satisfy certain mixing conditions, we derive the probability of buffer overflow, when the server scheduling follows a max-weight policy. More specifically, for a given finite value \( B > 0 \), we consider two buffer overflow quantities. First, we consider \( P(\mathbf{W}_{0,T} \geq B\mathbf{1}_K) \) where \( \mathbf{W}_{0,T} \in \mathbb{R}_+^K \) is the transient workload (to be formally defined later) at time \( 0 \) with “zero” initial workload at time \( -T \) and \( \mathbf{1}_K \in \mathbb{R}_+^K \) is the vector of all 1s. The second quantity we study is the overflow probability for the limiting workload vector as \( T \to \infty \), i.e., \( P(\mathbf{W} \geq B\mathbf{1}_K) \). Since these probabilities are, in general, very hard to compute exactly, we consider logarithmic asymptotics to the probabilities of interest using the theory of large deviations. In particular, this paper, under a “many-sources” scaling regime, establishes logarithmic asymptotics for 1) the transient workload for a compact, convex, and coordinate-convex rate region; and 2) the stationary workload for a simplex rate region using a work conserving scheduler.

In the classical many-sources asymptotic, one considers a sequence of queueing systems indexed by the number of independent sources multiplexed (or averaged) over a particular queue, i.e., the arrival process to each queue is the average of independent) sources multiplexed (or averaged) over a particular rate region, \( L \) processes. The analysis focuses on the asymptotic behavior of the systems when \( L \to \infty \). In our work, we consider a generalization of the many sources asymptotic in which the input to the queueing system \( L \) exhibits a sample path large deviations property (LDP) similar to that of the average of

I. INTRODUCTION

The drive to achieve maximum efficiency in wireless data networks and high-speed switches has lead to many advances in the design of good scheduling policies. One such family of good scheduling policies is an online policy called the maximum weight (max-weight) scheduling policy. For the traditional multi-class queue setting: each service option is a point within a compact, convex and coordinate-convex rate region. In this setting, the max-weight policy naturally generalizes to finding a richer choice of service options when compared to a traditional multi-class queue setting: each service option is a point within a compact, convex and coordinate-convex rate region. Thus, the server has access to a richer choice of service options when compared to a traditional multi-class queue setting: each service option is a point within a compact, convex and coordinate-convex rate region. In this setting, the max-weight policy naturally generalizes to finding an operating point within the rate region that has the maximum projection along the workload vector. Note that in this setting the traditional single-server multi-class queue has a rate-region given by a simplex; this is the convex hull of the discrete operating points corresponding to serving one user at a time.

In the present article, with \( K \) independent queues we seek to derive the probability of buffer overflow, when the server scheduling follows a max-weight policy. More specifically, for a given finite value \( B > 0 \), we consider two buffer overflow quantities. First, we consider \( P(\mathbf{W}_{0,T} \geq B\mathbf{1}_K) \) where \( \mathbf{W}_{0,T} \in \mathbb{R}_+^K \) is the transient workload (to be formally defined later) at time \( 0 \) with “zero” initial workload at time \( -T \) and \( \mathbf{1}_K \in \mathbb{R}_+^K \) is the vector of all 1s. The second quantity we study is the overflow probability for the limiting workload vector as \( T \to \infty \), i.e., \( P(\mathbf{W} \geq B\mathbf{1}_K) \). Since these probabilities are, in general, very hard to compute exactly, we consider logarithmic asymptotics to the probabilities of interest using the theory of large deviations. In particular, this paper, under a “many-sources” scaling regime, establishes logarithmic asymptotics for 1) the transient workload for a compact, convex, and coordinate-convex rate region; and 2) the stationary workload for a simplex rate region using a work conserving scheduler.

In the classical many-sources asymptotic, one considers a sequence of queueing systems indexed by the number of independent sources multiplexed (or averaged) over a particular queue, i.e., the arrival process to each queue is the average of \( L \) processes. The analysis focuses on the asymptotic behavior of the systems when \( L \to \infty \). In our work, we consider a generalization of the many sources asymptotic in which the input to the queueing system \( L \) exhibits a sample path large deviations property (LDP) similar to that of the average of

\( a \) online policies are those than can only use the past history of the arrivals, workloads and service decisions to decide on the scheduling choice; for example, these policies are not even aware of average arrival rates.

\( 1 \) Online policies are those than can only use the past history of the arrivals, workloads and service decisions to decide on the scheduling choice; for example, these policies are not even aware of average arrival rates.

\( 2 \) In many ways cross-layer optimization has resulted in some firm strides towards a union of information theory and communication networks of the sort that was sought in [11].

\( 3 \) The appellation “classical” is taken from [53] where a general scaling framework is presented that encompasses in a single setting all the different scalings used in the “many-sources” scaling regime.
l independent arrival streams (See Assumption 1). Given a sample path large deviations principle (see Definition 3) for the arrival processes, we derive a large deviations principle for the workload under the max-weight scheduling policy. In particular, we first show that the finite-horizon workload is a quasi-continuous map of the arrival process, for both the regular version of the max-weight policy and for a work-conserving/rate-maximizing version of it. Then the first contribution of the paper is that the finite-horizon workloads satisfy an LDP. This is obtained using a recent extension of the contraction principle by Garcia [41]. Restricting our attention to the simplex rate region (corresponding to the traditional multi-class single server queue), the second contribution of the paper is to use Garcia’s extended contraction principle along with a mixing condition assumption on the arrival process to establish an LDP for the stationary workload. Note that in contrast to related “many-sources’’ LDP results on FCFS and Priority policies that can be shown to be continuous, the challenge here is to establish LDP for the inherently discontinuous map that results from the max-weight scheduling policy over finite and infinite horizons.

The LDP results (Theorems 1, 2 and 3) directly imply that the probability of buffer overflow has an exponential tail whose decay rate is dictated by a good rate function determined by the statistics of the arrival process. This rate function can be expressed as a solution to a finite-dimensional optimization problem which has the same flavor of a deterministic optimal control problem. The final contribution of our work is to provide a simplified form for the corresponding rate functions when the arrival process has i.i.d. increments.

The outline of the paper is as follows. In Section II we briefly motivate and contextualize our work in the larger body of literature on LDP analysis of queues as well as cross layer scheduling. The problem formulation is given in Section III. Section IV provides background and preliminary results. The main results of the paper, which are the LDPs of the workloads, are given in Section V and proved in Section VI. Section VII gives simplified expressions of the rate functions. We conclude in Section VIII with a discussion of future work.

We close this section with a summary of various notation used in this work. We use bold letters to discriminate vectors from scalar quantities as well as their components. We denote the set of natural and non-negative real numbers by $\mathbb{N}$ and $\mathbb{R}_+$, respectively. We take $\otimes$ to represent the Kronecker product. For $0 \leq m_1 \leq m_2$ integers and a vector sequence $\{A_t, t \in \mathbb{N}\}$ where $A_t \in \mathbb{R}_{+}^K$ for $K \in \mathbb{N}$, we define $A(m_1,m_2) := \sum_{t=m_1+1}^{m_2} A_t$ as the cumulative arrivals from $m_1+1$ until timeslot $m_2$ where addition applies coordinate-wise. For vector-valued sequence $A$ we write $A_{\{m_1,m_2\}}$ to denote the finite subsequence $\{A_{m_1+1}, \ldots, A_{m_2}\}$. For a vector $x \in \mathbb{R}_{+}^K$ and set $B \subset \mathbb{R}_{+}^K$, $\text{Proj}_B(x)$ denotes the projection of vector $x$ on the set $B$, $\text{int}(B) := \{ \lambda \in B : \exists \lambda' \in B \text{ s.t. } \lambda < \lambda' \}$ denotes the set of points inside $B$ (i.e., not on the outer north-west boundary), and $|x|_+ := \max\{0, x\}$ where the function applies coordinate-wise. Lastly, for any given function $F : \mathcal{X} \mapsto \mathcal{Y}$ on metric spaces $\mathcal{X}$ and $\mathcal{Y}$, and $x \in \mathcal{X}$, we use the notation

$$xF := \{ y \in \mathcal{Y} : (\exists x_n \rightarrow x) \text{ such that } F(x_n) \rightarrow y \},$$

to denote the set of all cluster points (in $\mathcal{Y}$) of the images of sequences in $\mathcal{X}$ converging to a point $x \in \mathcal{X}$. Note that $xF$, in general, is a correspondence (also called a set-valued function) from $\mathcal{X}$ to $\mathcal{P}(\mathcal{Y})$ (the power set of $\mathcal{Y}$). However, $xF$ is single-valued at $x$, i.e., $xF = \{ F(x) \}$, if and only if $F$ is continuous at $x$.

II. RELATED WORK

In recent years, cross-layer scheduling has become a major focus of research in queuing and information theory due to its potential applications in communication networks. For brevity we do not list many important works, and instead mention only those closely related to this work.

In our LDP analysis, we follow the lead of many recent papers on the analysis of scheduling algorithms [9], [10], [12]–[16], [18] by considering logarithmic asymptotics to the probabilities of certain rare events. Maximum weight scheduling policy falls under class of the generalized $\text{cp}$-rule policies and is known to be stabilizing under very mild conditions [1]–[5]. A refined analysis of this policy shows that it minimizes the workload in the heavy traffic regime [6]–[8] over a large class of stationary online policies. This optimality of the max-weight policies also carries over to Large Deviations based tail asymptotes: the work-conserving/rate-maximizing version of these policies is known to minimize the exponent of the tail asymptote of the stationary workload over a large class of stationary, online and work-conserving policies [10].

The present paper is closely related to [14], [15], where the buffer overflow probability for the workload processes of a single-server multi-queue queuing system under max-weight policies and general compact, convex and coordinate-convex capacity regions was established. While [14], [15] addresses the “large-buffer” scaling regime, this paper establishes similar logarithmic asymptotics results under the “many-sources” scaling regime (see [16]–[26], [30]–[39]). As the body of work on the “many-sources” scaling regime has grown, results have been established for many different scheduling policies for a single-server queue and also for networks of queues, namely, FCFS [20]–[22], [24], [26]–[28], [30]–[33], [38], [39], priority queuing [18], [24], [26], [36], [38], GPS [37], and SRPT and similar policies [16], [17], [34]. Our LDP analysis of max-weight scheduling is strongly motivated and complemented by these papers.

Finally, we close this section with a discussion of our motivation to consider the “many-sources” scaling studied in this paper. The interest in the “many-sources” asymptotic is known to be best motivated by a recent and practical interest in applications when there are large number of flows to each user or node. This asymptotic approach usually gives a more refined approximation to the probabilistic quantities of interest by incorporating the impact of the multiplexing gain [16]–[39] obtained by averaging many traffic sources together. However, our interest in the “many-sources” asymptotic has also been fueled by our earlier work [40] on a cross-layer optimization
of the PHY layer parameters, e.g., duration of the finite code blocks or cooperative cluster size when the fading channel is operated at high signal-to-noise-ratio (SNR). The high SNR regime is a very natural setting for the many-sources scaling since the capacity of the channel typically scales to infinity as \( \log(\text{SNR}) \), and therefore it is natural to scale the arrival rate of the flows with the same parameter, which is best accomplished by multiplexing more sources; in other words by setting \( L \propto \log(\text{SNR}) \). The present work on the many sources large deviation analysis of max-weight provides a first step in extending the above cross-layer optimization to a multi-user setting, an important topic for future research. We would like to end this discussion by noting that the time-invariance of rate/capacity region, which is assumed in this study, does not necessary exclude fading or time variations in the channel. Indeed, it is common for various notions of time-invariant capacity, such as ergodic and outage capacity regions, to characterize the information rate of channels with fading in the absence of channel state information at the transmitter (CSI-Tx). In the absence of CSI-Tx, the channel variations are considered and dealt with in a statistical sense in that various channel coding (FEC) solutions are carefully designed to combat the statistical variations in the channel quality. In such a setting, the maximum rate of reliable information delivery is a function of the (long term) statistical description of the channel (CDI) and not the instantaneous channel conditions (CSI), hence, a time-invariant channel capacity/rate region.

III. Problem Formulation

We consider a discrete-time queuing system with \( K \in \mathbb{N} \) independent queues and one server. We are interested in the statistical properties of the unfinished workload in queue \( k \) at time \( -t \) (for \( t \in \mathbb{N} \)) under a max-weight server allocation policy; note that we follow the Lindley-Loynes convention \([26]\) of working in negative time. Let \( W^k_T \in \mathbb{R}_+ \) be the unfinished workload (queue length) of queue \( k \) at the beginning of time \( -t \) and \( R^k_t \) be the amount of service allocated to queue \( k \) during time \((-t,-t+1)\). Let \( W^k_t := (W^k_T, k \in K) \) be the corresponding workload vector and \( R^k_t := (R^k_t, k \in K) \) be the rate vector. For every queue \( k \in K := \{1, \ldots, K\} \) we assume that work (in bits) arrives into the queue given by a sequence \((A^k_t, t \in \mathbb{N})\) where \( A^k_t \in \mathbb{R}_+ \) is the work brought in at time \(-t\). For \( t \in \mathbb{N} \), the dynamics of the workloads of queue \( k \in K \) is

\[
W^k_{t-1} = [W^k_t - R^k_t]^+ + A^k_t. \tag{1}
\]

Note that we assume that the arrivals \( A^k_t \) happen any time in \((-t,-t+1)\) but cannot be served in timeslot \(-t\).

The set of server’s operating points is restricted to a compact and convex set \( R \subset \mathbb{R}^K_+ \) known as the capacity or rate region of the server, i.e., \( R_t \subset R_t \), for all time \( t \). We make the simplifying assumption that bits are infinitely divisible so that the rate allocations can be assumed to be real numbers. Furthermore, we also assume that \( R \) is coordinate-convex, i.e., if \( \mu_1 \in R_t \), then every \( \mu_2 \in R_t \) such that \( \mu_2 \leq \mu_1 \) is also in \( R \) where the inequalities apply along each coordinate. We are interested in the max-weight scheduler, and its closely related work-conserving and rate-maximizing version. At the beginning of timeslot \(-t\), the rate vector \( R^*_t \in R \) is selected by a max-weight scheduler in response to the current workload \( W_t \). Specifically, under max-weight scheduler and in response to the current workload \( W_t \), the rate vector \( R^*_t \) is chosen such that

\[
R^*_t \in \text{arg max}_{R \in R} < R, W_t >. \tag{2}
\]

As later established by Lemma \( \boxed{\text{1}} \) it is possible to construct a quasi-continuous (see Defn \( \boxed{\text{1}} \) in Section \( \text{IV} \)) function \( H \) such that \( R^*_t = H(W_t) \). We call this construction the (max-weight) scheduling function. We also define a non-idling/rate-maximizing modification of max-weight scheduler for which the rate of service \( R^*_t \) such that it splits the service when the unfinished workload in each queue \( k \) is less than \( C^k = \max \{ R^k : R \in R \} \). Lemma \( \boxed{\text{1}} \) also shows that it is possible to construct a quasi-continuous function \( H^{wc} \) such that \( R^*_t = H^{wc}(W_t) \) and

\[
H^{wc}(W_t) = \begin{cases} \text{Proj}_R(W_t) & \text{if } W_t \in \Pi^K_{k=1} [0, C^k]; \\ H(W_t) & \text{otherwise.} \end{cases} \tag{3}
\]

When necessary we will distinguish the workload vectors that result from the work-conserving max-weight by labeling them as \( W^{wc} \).

As mentioned earlier, in this paper we are interested in the probability distributions for the finite-horizon and infinite-horizon workloads. The finite-horizon workload, denoted by \( W_{0,T} \), is the workload at time \( 0 \), assuming the initial condition at time \(-T\) is \( W_T = 0 \). The index \( T \) in \( W_{0,T} \) reminds us of this initial condition.\( \boxed{\text{2}} \) The infinite-horizon workload, \( W \), is defined as \( W := \lim_{T \to \infty} W_{0,T} \). We assume that the limit exists but may be infinite. Note that our results for the infinite-horizon workload are obtained in the restricted setting of a work conserving max-weight scheduler operating on a simplex rate region. For this work-conserving max-weight scheduler, it is known that \( W^{wc} \) is the stationary workload when the system is stable.

We will use functions \( G_T \) and \( G_T^{wc} \) to relate the arrival process \( A_{(0,T)} \) with the unfinished work under max-weight scheduling, \( W_{0,T} \) and under work conserving max-weight \( W_{0,T}^{wc} \), i.e. \( W_{0,T} = G_T(A_{(0,T)}) \) and \( W_{0,T}^{wc} = G_T^{wc}(A_{(0,T)}) \).

Similarly, we also define function \( G^{wc} \) to describe the stationary workload under work-conserving max-weight when the arrival sequence is given, i.e. \( W^{wc} = G^{wc}(A) \); we do not indicate the rate-region here as it is implicitly understood to be the simplex rate-region.

For each user \( k \in K \) and system indexed by \( L \in \mathbb{N} \), we will assume a stationary arrival process of work brought into the system given by a sequence \( A^{k,L}_{L} := (A^{k,L}_{L}, t \in \mathbb{N}) \) where \( A^{k,L}_{L} \in \mathbb{R}_+ \) is the work (in bits) brought in at time \(-t\) into the queue of user \( k \). The arrivals to different queues/users are assumed to be mutually independent. Also let \( L^*: = (A^{k,L}, k \in K) \) be the sequence of arrival vectors. In our large deviation analysis, we characterize the

\footnote{Note that the result remains valid even when the initial condition is within \( R \) with the work-conserving scheduler. With \( W_T \in R \), we always have the\( \boxed{\text{3}} \) at time \(-T+1\) be \( W_{T+1} = \left[ W_T - H^{wc}(W_T) \right]^+ + A_T = A_T \) from the non-idling condition that we imposed on the server allocation mechanism as \( \text{Proj}_R(W_T) = W_T \).}
asymptotic probability distributions for the finite-horizon and infinite-horizon workloads, $G^T(\mathbf{A}^L(t, T))$ and $G^\infty(\mathbf{A}^L(t, T))$, as $L \to \infty$.

We close this section by noting that in its typical avatar a max-weight scheduler is also accompanied by a set of non-zero weights $\beta \in \mathbb{R}^K_+$ to weight the workload vectors while determining the max-weight service vector. Using the observations in [14, 15] it can be seen that there is no loss of generality in assuming that the weight vector is $1_K$.

IV. BACKGROUND AND PRELIMINARIES

In this section, we provide a brief review of fundamental definitions, concepts, and relevant results in large deviations theory that are essential for understanding our paper. Except for Lemma 1, which establishes the quasi-continuity of the scheduling functions $H$ and $H^{\text{wc}}$, the material in this section can be found in [24, 26, 41, 51].

A. Quasi-Continuity and Almost Compactness

In this section, we recall two important analytic properties for functions on metric spaces: quasi-continuity and almost compactness. These properties allow for an extension of the contraction principle to which we later appeal. First, let us provide the definition of quasi-continuity on metric spaces:

**Definition 1**: [41, Theorem 3.2] Let $\mathcal{X}, \mathcal{Y}$ be complete metric spaces. Function $F : \mathcal{X} \mapsto \mathcal{Y}$ is quasi-continuous at $x \in \mathcal{X}$ if and only if there is a sequence $\{x_n\}$ such that $x_n \to x, F(x_n) \to F(x)$, and such that for all $n$, $F$ is continuous at $x_n$. If quasi-continuity holds for each $x \in \mathcal{X}$, then $F$ is deemed quasi-continuous.

**Remark 1**: Obviously, every continuous function is quasi-continuous. A step function $F : \mathbb{R} \mapsto \mathbb{R}$, where $F(x) = 0$ for $x < 0, F(x) = 1$ for $x \geq 0$, is quasi-continuous. However, if $F(0) = 1/2$, then $F$ is not quasi-continuous at $x = 0$.

**Remark 2**: An important property that we will later user is that if $F$ is a continuous function and $G$ is a quasi-continuous function, then $F \circ G$ is quasi-continuous. However, $G \circ F$ is not necessarily quasi-continuous [41].

From Lemma 1 below max-weight and its work conserving/rate-maximizing variation allow for quasi-continuous function selections.

**Lemma 1**: There exist quasi-continuous functions $H$ and $H^{\text{wc}}$ such that

$$H(\mathbf{W}_t) \in \arg\max_{\mathbf{R}^n \in \mathbb{R}} < \mathbf{R}, \mathbf{W}_t >,$$

and

$$H^{\text{wc}}(\mathbf{W}_t) = \begin{cases} \text{Proj}_{\mathbf{R}}(\mathbf{W}_t), & \mathbf{W}_t \in \Pi^K = [0, C^K] \\ H(\mathbf{W}_t), & \text{otherwise} \end{cases}.$$

**Proof**: See Appendix A. The proof relies on structural properties of the scheduling maps.

Having provided the definition of quasi-continuity, we are ready to define almost compactness for a function:

**Definition 2**: [41, Lemma 6.1] If $\mathcal{X}, \mathcal{Y}$ are complete metric spaces, a function $F : \mathcal{X} \mapsto \mathcal{Y}$ is almost compact at $x \in \mathcal{X}$ if for every sequence $x_n$ converging to $x$, there is a subsequence along which $F$ converges to a point $y \in \mathcal{Y}$. We say that $F$ is almost compact if it is almost compact at every $x \in \mathcal{X}$.

B. Topology for Sample Paths

Since a large deviations principle is defined using topological entities and since we will deal with continuity and convergence of the workload mappings, we need to precisely specify the topology for the space of the arrival sample paths. We use the scaled-uniform norm/weighted supremum norm topology \[^\text{used in [24] for our analysis.}\]

Let $D \subset \{x : \mathbb{N} \mapsto \mathbb{R}_+\}$ denote the space of non-negative sequences such that $\sup_{t \in \mathbb{N}} \frac{\left| x(t) \right|}{t} < +\infty$, and let $D^K$ be the $K$ Cartesian product of $D$. Let $|| \cdot ||_u$ be the scaled uniform norm on $D$, i.e., $||x||_u := \sup_{t \in \mathbb{N}} \frac{\left| x(t) \right|}{t}$ for all $x \in D$, while for all $a = ( \{a^k\}^K_{k=1} ) \in D^K$, the scaled uniform norm of $a$ is $||a||_u := \max_{k \in K} ||a^k||_u$. The space $D$ is metrizable via the scaled uniform norm $|| \cdot ||_u$, i.e., for all $x, y \in D$, the distance between them is $||x - y||_u = \sup_{t \in \mathbb{N}} \frac{\left| x(t) - y(t) \right|}{t}$.

Define a subspace $D_\mu$ of $D$ which contains all the arrival paths whose average rate is equal to the expected rate $\mu$, i.e., $D_\mu := \{ x \in D : \lim_{t \to \infty} \frac{x(t)}{t} = \mu \}$. Also for a vector $\mu = (\mu_1, \ldots, \mu^K)$ define $D^K_\mu$ to be the product space of $D^K_{\mu^k}$ for all $k \in K$. We equip $D_\mu$ and $D^K_\mu$ with the appropriate subspace and product topologies [42]. For finite dimensional metric spaces like $\mathbb{R}^n_+$, $\mathbb{N} \in \mathbb{N}$, we use the square uniform topology (same as the product topology) with the square metric $\rho[42]$, where $\rho(x, y) := \max_{i \in \{1, \ldots, n\}} |x^i - y^i|$. From [24], [26] it is also clear that the scaled uniform norm topology is stronger than the point-wise convergence topology, hence the projection and shift operators are continuous under the scaled uniform norm topology.

C. Large Deviations Principle

The following definition of a large deviations principle is taken from [24]. For a complete reference to the theory, definitions, and tools, see [51].

**Definition 3 (Large deviations principle)**: A sequence of random variables $X^L$ in a Hausdorff space $\mathcal{X}$ with $\sigma$-algebra $\mathcal{B}$ is said to satisfy a large deviations principle (LDP) \[^\text{with good rate function} I \text{ if, for any} B \in \mathcal{B}, \]

$$- \inf_{x \in B^o} I(x) \leq \liminf_{L \to \infty} \frac{1}{L} \log P(X^L \in B) \leq \limsup_{L \to \infty} \frac{1}{L} \log P(X^L \in B) \leq - \inf_{x \in B} I(x),$$

where $B^o$ and $\bar{B}$ are the interior and the closure of $B$, respectively, and if the rate function $I : \mathcal{X} \mapsto \mathbb{R}_+ \cup \{\infty\}$

\[^{1}\text{In the theory of weak convergence of probability measures this topology was first proposed in} \[43, 45\] \text{for continuous functions vanishing at infinity with a continuous index set. In the same context it was then generalized to cadlag functions with a continuous index set in} \[44\]. \text{The most general setting of this topology for discrete-time processes can be found in} \[46, 47\], \text{and the corresponding setting for continuous-time processes can be found in} \[48\]. \text{The usage of this topology in the context of Large Deviations can be found in} \[49\] \text{and} \[50\]. \text{Finally, the central theme in} \[24, 26\], \[50\] \text{is to demonstrate how this is a natural topology to use in the queuing context.} \]

\[^{2}\text{Often} \mathcal{B} \text{is taken to be the Borel} \sigma\text{-algebra, and a rate function is, by definition, non-negative and lower semicontinuous.} \]
has compact level sets, where the level sets are defined as \( \{ x : I(x) \leq \alpha \} \), for \( \alpha \in \mathbb{R} \).

If \( X^L \) is a mapping from \( \mathbb{N} \) to \( \mathbb{R} \) describing the sample path of a random sequence, the LDP is referred to as a sample path LDP.

D. Garcia’s Extended Contraction Principle

The contraction principle (see [51, p. 126]) says that if we have an LDP for a sequence of random variables, we can effortlessly obtain LDP’s for a whole other class of random sequences that are obtained via continuous transformations. However, due to the inherent discontinuity in the max-weight scheduling function, this (regular) contraction principle fails to provide sufficient structure in the setting of our interest. Instead, we will utilize the following powerful extension of the contraction principle for quasi-continuous transformations.

**Fact 1:** Assume that \( \Omega \overset{X^L}{\rightarrow} \mathcal{X} \overset{F}{\rightarrow} \mathcal{Y}, \mathcal{X}, \mathcal{Y} \) are metric spaces, and \( \{ X^L \} \) satisfies an LDP in \( \mathcal{X} \) with rate function \( I^L \). If at every \( x \) with \( I^L(x) < \infty \), the following hold:

1) \( F \) is almost compact; and
2) for all \( y \in xF \), there exists a sequence \( \{ x_n \} \) converging to \( x \) such that \( F(x_n) \to y \), \( F \) is continuous at \( x_n \), and \( I^L(x_n) \to I^L(x) \),

then \( \{ F(X^L) \} \) satisfies an LDP with rate function given by

\[
I(y) = \inf \{ I^L(x) : y \in xF \}.
\]

**Remark 3:** Whenever \( I^L(\cdot) \) is continuous, then the second condition is reduced to confirming that \( F \) is a quasi-continuous function.

V. Assumptions and Overview of the Results

Garcia’s extended contraction principle together with Lemma \( 1 \) suggests the following road map to obtaining an LDP for the finite and infinite horizon workload processes under max-weight scheduling. The large deviation property for sequences of finite- and infinite-horizon workloads would follow as a direct consequence of the sample-path LDP of the arrival process, as soon as one establishes the quasi-continuity and almost compactness of the mappings \( G_1, G^w_1, \) and \( G^w \) along with some continuity properties of the rate function obtained from the sample-path LDP assumption on the arrival processes. As a result, our first task is to restrict our attention to arrival streams that satisfy the sample-path LDP as stated by Assumptions \( 1 \) in Section V-A. There we also discuss a family of arrival processes which satisfies these assumptions.

A. Sample Path LDP of Arrival Processes

The following sample path LDP for the sequence of arrival processes \( A^L \) is the starting point of our analysis.

**Assumption 1 (Many-sources sample path LDP):** The sequence \( \{ A^L \} \) satisfies a sample path LDP in \( \mathcal{D}^K_\mu \) equipped with the scaled uniform topology with rate function \( I^L \), where the rate function \( I^L \) is given as

\[
I^L(a) := \sup_{t \in \mathbb{N}} I^L_t(a_{[0,t]}) = \lim_{t \to \infty} I^L_t(a_{[0,t]})
\]

for \( a \in \mathcal{D}^K_\mu \), where for every \( t \in \mathbb{N} \) we also assume that \( \{ A^L_{[0,t]} \} \) satisfies an LDP with rate function \( I^L_t(\cdot) \) in the product topology.

**Remark 4:** The most general conditions for Assumption \( 1 \) to be satisfied are given in [24, Thm. 3] (also stated in [26, Thm. 7.1, pg. 156]). There it is also shown how several standard stationary processes used for traffic modeling, such as i.i.d. increment processes, Markov-modulated processes, a general class of Gaussian and fractional Brownian processes (for long-range dependent or heavy-tailed traffic), satisfy Assumption \( 1 \). The conditions of [24, Thm. 3] also imply that the sequence \( \{ A^L \} \) also satisfies an LDP on \( \mathcal{D}^K_\mu \) equipped with the scaled uniform topology, with rate function \( I^L \), where \( I^L(a) = \infty \) for \( a \in \mathcal{D}^K_\mu \) [24, 26]. Finally, it is shown in [26, Lemma 7.8] that under the conditions of [24, Thm. 3], for all \( t \in \mathbb{N} \) we have \( I^L_t(\mu \otimes \Pi_t) = 0 \) and also that \( I^L_t(\cdot) \) is convex.

In this paper, we further assume the following continuity conditions on the rate functions:

**Assumption 2:** We assume that \( I^L_t(\cdot) \) is continuous in the product topology on \( \mathbb{R}^{K \times t}_+ \).

**Assumption 3:** For every point \( a \) in the effective domain of \( I^L \), i.e., \( \{ a \in \mathcal{D}^K_\mu : I^L(a) < +\infty \} \), we assume that for every sequence \( \{ a^n \} \) converging to \( a \) in \( \mathcal{D}^K_\mu \) such that there exists \( t \in \mathbb{N} \) so that \( a^n_s = a_s \) for all \( s > t \) (for all \( n \)), we have \( I^L(a^n) \to I^L(a) \).

**Remark 5:** Assumption \( 3 \) is a manifestation of a mixing condition that provides a certain independence of the long-term behavior of the process with respect to any finite initial block/window.

In Proposition \( 4 \) below we demonstrate that processes with i.i.d. increments naturally satisfy Assumptions \( 1 \) and \( 2 \) before proving this we define a coercive function as follows.

**Definition 4:** A function \( f : \mathbb{R}_+ \rightarrow \mathbb{R}_+ \cup \{ +\infty \} \) with domain \( \text{Dom}(f) := \{ x : f(x) < +\infty \} \) is defined to be **coercive** if for every \( y \in \mathbb{R}_+ \), there is compact set \( C \subseteq \text{Dom}(f) \) such that \( f(x) > y \) for all \( x \in \text{Dom}(f) \setminus C \).

**Proposition 1:** For scale parameter \( L \) assume that the arrival process has i.i.d. increments while the arrival processes for different users are independent. Furthermore, assume that for every \( k \), \( \{ A^L_k \} \) satisfies the conditions of the Gärtner-Ellis Theorem [51, Thm. 2.3.6] and that the limiting rate function is either coercive or has domain \( \mathbb{R}_+ \). Then Assumptions \( 1 \) and \( 2 \) hold.

**Proof:** The fact the above class of processes satisfy Assumption \( 1 \) is immediate from [24, Thm. 3]. Similarly the Gärtner-Ellis Theorem [51, Thm. 2.3.6] yields convexity and from the coercivity of the rate function or with its domain being \( \mathbb{R}_+ \), Assumption \( 2 \) follows. When the arrival process has i.i.d. increments, then it also follows that for \( y \in \mathbb{R}^{k \times t}_+ \)
and \( x \in D^K \),

\[
I_t^s(y) = \sum_{k=1}^{K} \sum_{s=1}^{t} \Lambda_{s,k}^k(y_k^s),
\]

and

\[
I_t(x) = \sum_{k=1}^{K} \sum_{s=1}^{+\infty} \Lambda_{s,k}^k(x_k^s),
\]

where \( \Lambda_{s,k}^k \) is the Fenchel-Legendre transform of \( \Lambda_{s,k}^k(\theta) := \lim_{L \to \infty} \frac{1}{L} \log E e^{\theta L A_{s,k}^k} \). Note that that \( I_t(x) = +\infty \) if \( x \in D^K \). Now for every sequence \( \{x^n\} \) converging to \( x \) in \( D^K_{\mu} \) with \( I_t(x) < +\infty \) such that there exists \( t \in \mathbb{N} \) so that \( x_n^s = x_s \) for all \( s > t \) (for all \( n \)), we have \( I_t(x^n) < +\infty \) for all \( n \) large enough and

\[
I_t^s(x^n) - I_t^s(x) = I_t^s(x^n|_{0,t}) - I_t^s(x|_{0,t}) \xrightarrow{n \to +\infty} 0.
\]

**Remark 6:** A more general characterization, especially of Assumption 3 remains an important area of future work. However, to provide insight about Assumptions 2 and 3 and their relationship to Assumption 1 we provide the following simple example. Our example will be constructed as averages of \( L \) stationary i.i.d. random processes so it is sufficient to describe the underlying stochastic process: We let the random process be \( X_\kappa = \gamma(1 + (-1)^{k+1}V) \), where \( V \) is a Uniform \([-1, 1] \) random variable and \( \gamma > 0 \) is the long-term mean for all our sequences. Here it can be verified that \( I(x) < +\infty \) if and only if \( x_k = \gamma (1 + (-1)^{k+1} \alpha) \), where \( \alpha \in [-1, 1] \). It can also be verified that \( I(x) = 0 \) for the all-\( \gamma \) sequence, i.e., \( x_k \equiv \gamma \). However, it is clear that one can easily construct sequences as required in Assumption 3 with the rate function being infinite which, nevertheless, converge to the all-\( \gamma \) sequence. In other words, Assumptions 2 and 3 require more than the regularity conditions from [24], [26] that guarantee validity of Assumption 1.

Given the above assumptions on the arrival processes, we are now ready to provide an overview of the main results of the paper.

**B. Main Results: An Overview**

Assuming that the sequence of the arrival processes \( \{A^L\} \) satisfies a many-sources sample-path LDP with a “continuous” rate function (Assumptions 11, 13), LDPs for the finite and infinite-horizon workloads will be a direct consequence of Garcia’s extended contraction principle once the required quasi-continuity and almost compactness properties are shown. We will demonstrate that quasi-continuity and almost compactness of the finite horizon workload mappings are inherited from the quasi-continuity of the schedulers \( H \) and \( H^{wc} \).

**Theorem 1:** Under Assumptions 12 and for all \( t \in \mathbb{N} \), the sequence of the finite-horizon workloads under a max-weight scheduler \( \{W_t^{L}: \mu \} := G_t(A^L_{\mu}|_{0,t}) \) satisfies an LDP on \( R^K_+ \) with the rate function \( I_t \), where for \( b \in R^K_+ \)

\[
I_t(b) = \inf_{x \in R^{\times N_+:} \gamma G \ni b} I_t^s(x).
\]

**Theorem 2:** Under Assumptions 12 and for all \( t \in \mathbb{N} \), the sequence of the finite-horizon workloads under the work-conserving max-weight scheduler, \( \{W_t^{L}: \mu \} := G_t^{wc}(A^L_{\mu}|_{0,t}) \), satisfies an LDP on \( R^K_+ \) with the rate function \( I_t \), where for \( b \in R^K_+ \)

\[
I_t(b) = \inf_{x \in R^{\times N_+:} \gamma G \ni b} I_t^s(x).
\]

The quasi-continuity and almost compactness of the stationary workload mapping, however, requires slightly more work as shown in Section VI. In fact, unlike the finite-horizon LDP results of Theorems 1 and 2, which hold for general rate regions and for both max-weight and its work conserving version, the infinite horizon LDP result of Theorem 3 is established only under the work conserving max-weight policy and only with a simplex rate region. For a vector \( x \in R^K_+ \) define \( \hat{x} := \sum_{k=1}^{K} x_k/C^k \), then the simplex rate region is given by

\[
R_a := \{ r \in R^K_+ : r \leq \hat{x} \}.
\]

**Theorem 3:** Consider a work conserving max-weight scheduler with a simplex rate region \( R_a \). Let \( \mu \in R^K_+ \) be an admissible arrival rate vector inside this simplex rate region, i.e., \( \mu \in \text{int}(R_a) \) and \( \{A^L\} \) be a sequence of arrival processes that satisfies Assumptions 12 and 3 in \( D^K_{\mu} \). Then the sequence of infinite-horizon workloads \( \{W_t^{L} := G^{wc}(A^L_{\mu}|_{0,t}) \} \) satisfies an LDP on \( R^K_+ \) with good rate function \( J \), where for \( b \in R^K_+ \)

\[
J(b) = \inf_{a \in R^{\times N_+:} \gamma G^{wc} \ni b} I_t^s(a).
\]

Note that all the rate functions have the appearance of being what one would naturally expect, i.e., among all arrival sequences that result in the required workload at the required epoch, find the one with the least cost to deduce the rate function. However, the discontinuity of the queueing map makes this a non-trivial assertion and also enlarges the set of allowed arrival sequences.

**VI. ANALYSIS: LARGE DEVIATIONS FOR WORKLOADS**

In this section, we prove the main results of the paper: establishing LDP for the sequences of the finite-horizon and infinite-horizon workloads. We first delineate the proof for the LDPs for the sequence of the finite-horizon workloads.

**A. LDP for Finite-Horizon Workloads**

In this section, for \( t \in \mathbb{N} \), we establish an LDP for finite-horizon workloads \( \{W_t^{L} := G^{wc}(A^L_{\mu}|_{0,t}) \} \) and \( \{W_t^{L} := G_t(A^L_{\mu}|_{0,t}) \} \). The approach is to first show that the mappings \( G^{wc} : R^K_+ \to R^K_+ \) and \( G_t : R^K_+ \to R^K_+ \) are quasi-continuous and almost compact and use Garcia’s extended contraction principle to obtain an LDP for the finite-horizon workloads from the LDP assumption for \( A^L_{\mu}|_{0,t} \). From Fact 1, the almost compactness and quasi-continuity of workload mappings are sufficient to establish an LDP for finite-horizon workload, since according to Assumption 2, \( I_t^s(.) \) is continuous (in the product topology).

Using the quasi-continuity of the scheduling function we now prove the required quasi-continuity of the workload maps.
First we consider the work-conserving max-weight scheduler.

**Lemma 2:** For \( t \in \mathbb{N} \), \( G_t^{wc} \) is quasi-continuous with respect to the scaled uniform topology.

**Proof:** See Appendix [3]. The induction-based proof uses the quasi-continuity of the scheduler \( H \) and the linear dependence of the workload \( W_s \) at time \(-s \) on \( A_{s+1} \) for all \( s \in [0, t - 1] \).

Similarly, in order to establish an LDP for the workload process under the maximum weight scheduler, we show the quasi-continuity of function \( G_t \).

**Lemma 3:** For \( t \in \mathbb{N} \), \( G_t \) is quasi-continuous with respect to the scaled uniform topology.

**Proof:** See Appendix [3]. In this induction-based proof we establish certain analytical properties of the workload map for all possible rate-regions that then allows us to obtain a quasi-continuous selection.

Having established quasi-continuity, the next result demonstrates the almost compactness of the workload maps.

**Lemma 4:** For \( t \in \mathbb{N} \), both \( G_t \) and \( G_t^{wc} \) are almost compact on \( \mathbb{R}_+^K \) with respect to the scaled uniform topology.

**Proof:** Since the workload at any time cannot exceed the amount of work brought in from the last time the system was empty irrespective of the scheduler used, we automatically get the following bounds:

\[
G_t(a|0,t) \leq a(0,t) \quad \text{and} \quad G_t^{wc}(a|0,t) \leq a(0,t).
\]

This implies the almost compactness of \( G_t \) and \( G_t^{wc} \). Since every sequence \( \{a^n\} \) converging to \( a \) (in \( D^K_\mu \) in the scaled uniform norm topology) converges point-wise too, it follows that \( G_t(a^n|0,t) \) and \( G_t^{wc}(a^n|0,t) \) will lie in a bounded subset of \( \mathbb{R}_+^K \) so we can use the Bolzano-Weierstrass theorem.

Now, as already discussed, the proof of Theorems [1] and [2] is complete. Next, we discuss the LDP for the infinite-horizon workloads for the work-conserving max-weight scheduler with a simplex rate-region.

**B. LDP for Infinite-Horizon Workloads**

In this section, we establish an LDP of the sequence of the infinite-horizon workloads \( \{W^L = G^{wc}(A^L)\} \) where \( A^L \in D^K_\mu, \mu \in \text{int}(R_s), \) and \( \{A^L\} \) satisfies Assumptions [13]. Similar to the last section, we first show that the mapping \( G^{wc} \) is quasi-continuous and almost compact on \( D^K_\mu \) and then use Garcia’s extended contraction principle to establish the desired LDP.

The following lemmas establish the necessary steps to apply Garcia’s contraction principle to the stationary workload map. The first of these lemmas relates the infinite horizon workload mapping to that of a finite horizon.

**Lemma 5:** Consider an arrival process \( a \in D^K_\mu \). There exists a \( s^* = s^*(a) < \infty \) such that the workloads at time \(-s^* \) under \( a \) lies within the rate region \( R_s \), i.e., \( G^{wc}(a|s^*,\infty) \in R_s \). Furthermore, for any sequence of arrival processes \( \{a^n \in D^K_\mu\} \) converging to \( a \) (in scaled uniform topology), the workloads at time \(-s^* \) under \( a^n \), when \( n \) is large enough, also lie within the rate region \( R_s \), i.e., \( \exists n_0 \) such that \( G^{wc}(a^n|s^*,\infty) \in R_s \) for \( n > n_0 \).

**Proof:** See Appendix [3]. The main idea is to use the fact that a suitably normalized sum (over all queues) workload process behaves like the workload of a single server queue with only one flow and a work-conserving service discipline. This allows us to use the continuity of the workload mapping of a single server queue and the stability of the queue to arrive at the assertion of the lemma.

Now we are ready to verify the requirements of Garcia’s extended contraction principle.

**Lemma 6:** Let \( a \in D^K_\mu \) be an arrival process with rate \( \mu \in \text{int}(R_s) \), and \( G^{wc}(a) \) be the corresponding infinite-horizon workload. For any \( W \in aG^{wc} \) there exists a sequence of arrivals \( \{a^n \in D^K_\mu\} \) such that \( a^n \) converges to \( a \) in scaled uniform topology, \( G^{wc}(a^n) \to W \), \( G^{wc} \) is continuous at \( a^n \), and \( I^n(a^n) \to I(a) \).

**Proof:** See Appendix [3]. The idea is to relate the infinite horizon workload, using Lemma [5] to a finite-horizon workload mapping whose quasi-continuity was established earlier.

**Lemma 7:** The mapping \( G^{wc,\Lambda_s} \) is almost compact on \( D^K_\mu \) with respect to the scaled uniform topology.

**Proof:** See Appendix. The proof is along the same lines as that of Lemma [6].

Again, the above lemmas and Garcia’s extended contraction principle immediately yield the LDP for the sequence of the infinite-horizon workload in Theorem [3].

Let us now consider the problem of calculating the rate function. Equation (10) suggests that the rate function, \( J(b) = \inf_{a \in D^K_\mu, aG^{wc} \ni b} I^J(a) \), could be interpreted as the minimum-cost solution among all paths \( a \in D^K_\mu \) such that \( b \in aG^{wc} \), where the cost of the path \( a \) is \( I^J(a) \) and convex. Hence, the problem of finding the rate functions is a deterministic optimal control problem like those in [12], [14]. However, the expressions for the rate functions \( I_t \) and \( J \) in [7] and [10] are of little use in their current forms, as their computation is far from straightforward. In the next section, we simplify the rate functions when the arrival processes are limited to having i.i.d. increments.

**VIII. I.I.D. INCREMENTS: SIMPLIFIED RATE FUNCTIONS**

In this section, we give a calculation of the finite-horizon and infinite-horizon rate functions in the case when the arrivals have i.i.d. increments and are independent across queues. In this case, the cost of a sample path \( a \in D^K_\mu \), which is \( I^J(a) \), is additive and the total cost of any arrival sample path is the sum of the cost over all timeslots and queues. This property helps us to simplify the calculation of the rate functions.

Consider the underlying arrival process \( A \) to be a process with i.i.d. increments, e.g., a compound Poisson arrival process with exponential packet length (see [40]). For these i.i.d. increment arrival processes, as mentioned earlier, we have

\[
I^J(a|0,t) = \sum_{k=1}^{K} \sum_{i=1}^{t} \Lambda^s_{1,k}(x_t^i), \tag{11}
\]

and

\[
I^J(a) = \sum_{k=1}^{K} \sum_{i=1}^{\infty} \Lambda^s_{1,k}(x_t^i). \tag{12}
\]
Next, we simplify the rate functions that we have derived.

A. Infinite-Horizon Rate Function

The following lemma expresses the infinite-horizon rate function $J$ as the infimum of the finite-horizon rate functions $I_t$ over all time $t$.

**Lemma 8:** For i.i.d. increment arrival processes with $\mu \in \text{int}(\mathcal{R}_s)$, the infinite-horizon rate function $J$ is simplified as

$$J(b) = \inf_{t \geq 1} I_t(b). \quad (13)$$

**Proof:** The cost of a sample path over time is the sum of the cost of arrivals in all timeslots. As in the proof of Lemma 6 for $a \in D^K_{\mu}$ where $\mu \in \text{int}(\mathcal{R}_s)$, we can find $t := s^*(a)$ such that $W_t(a) \in \mathcal{R}_s$. Hence, for $a$ such that $b \in aG_{\mu}^{\infty}$, one can reduce the cost of the path by constructing a new sample-path $\tilde{a}$ by setting $\tilde{a}_v = \mu$ for all $v > t$ and $\tilde{a}_v = a_v$ for all $v \leq t$ while still satisfying $b \in \tilde{a}G_{\mu}^{\infty}$. This is because $I_t^2(\tilde{a}) = I_t^2(a|_{[0,t]}) \leq I_t^2(a)$. On the other hand, since $W_t(a) \in \mathcal{R}_s$, we can write $b \in aG_{\mu}^{\infty}$. All of these imply that

$$J(b) = \inf_{a \in D^K_{\mu}: aG_{\mu}^{\infty} \ni b} I_t^2(a) = \inf_{t \geq 1} \inf_{x \in D^K_{\mu}: xG_{\mu}^{\infty} \ni b} I_t^2(x) = \inf_{t \geq 1} I_t(b),$$

by the definition of $I_t(b)$ in (8).

With this simplification available, we now look at the finite-horizon rate function $I_t$ in more detail.

B. Finite-Horizon Rate Function

In this subsection, we provide a further simplified expression of the finite-horizon rate function $I_t$.

**Lemma 9:** For $t \in \mathbb{N}$, the finite-horizon rate function $I_t$ is simplified as

$$I_t(b) = \min \left( I^1_t(b), \inf_{u \in [1,t]} \inf_{x \in \mathcal{A}(u,b)} I^u_t(x) \right) \quad (14)$$

for $b \in \mathbb{R}^K_{+}$, where for $u > 1$

$$\mathcal{A}(u,b) := \{x \in \mathbb{R}^K_{+} \times \mathcal{L} : b \in xG_{\mu}^{\infty}, G_{\mu}^{\infty}(x|_{[t,u]}) \not\in \mathcal{R}_s, \forall v \in [1,u-1]\}. \quad (15)$$

**Proof:** This follows the idea from the proof of Lemma 6. Let $t \in \mathbb{N}$. For $a \in D^K_{\mu}$ such that $b \in aG_{\mu}^{\infty}$, we let

$$u = \min \{t, \min \{s \in [1,t-1] : W_s = G_{t-s}(a|_{[s,t]}) \in \mathcal{R}_s\}\}.$$

In other words, $-u$ is the last time the workload vector is inside the capacity region $\mathcal{R}_s$ before time 0. With this definition of $u$, for all $u > 1$ we have $W_u \not\in \mathcal{R}_s$ for all $v \in [1,u-1]$. By definition of $I_t$, we already know that the workload vector starts initially inside $\mathcal{R}_s$ at time $-t$. Therefore, we can find another path $\bar{a} \in \mathbb{R}^K_{+}$ with a reduced cost while keeping the workloads at time $-u+1$ to 0 (i.e., $W_{u-1} \rightarrow W_0$) intact by setting $\bar{a}_v = \mu, \forall v \in (u,t]$ and $\bar{a}_v = a_v$ otherwise. It is easy to see that we have $I^2_t(a|_{[0,t]}) \geq I^u_t(a|_{[0,u]}) = I^1_t(\bar{a}|_{[0,t]})$

and yet $b \in \hat{a}(0,u)G_{\mu}^{\infty} = \hat{\bar{a}}G_{\mu}^{\infty}$. The same logic also applies to the case when $u = 1$. However, in this case the only way that we can achieve a workload vector $b$ at 0 is if exactly that amount of work arrives, i.e., if $a_1 = b$.

Since by definition $W_v = G_{u-v}(a|_{[v,u]})$ for $v \in [1,u-1]$ (when $u > 1$), we have

$$I_t(b) = \inf_{u \in [1,t]} \inf_{x \in \mathcal{A}(u,b)} I^u_t(x) = \inf_{u \in [1,t]} \inf_{x \in \mathcal{A}(u,b)} I^u_t(x),$$

where $\mathcal{A}(u,b)$ is defined as in (15).

**Remark 7:** The above lemma reduces the set of feasible sample paths to the set $\mathcal{A}(u,b)$ for $u \in (0,t]$. It is interesting to note the property of the sample-paths in this set. For any $x \in \mathcal{A}(u,b)$, we have $W_0(x) = x(0,u) - (u-1) = b$, recalling that the $\hat{\cdot}$ notation is the normalized sum of the elements of the vectors. There is no wastage of service capacity over the $u-1$ timeslots because $W_v = G_{u-v}(x|_{[v,u]}) \not\in \mathcal{R}_s$ and hence $W_v > 1$. That is, any sample path $x \in \mathcal{A}(u,b)$ has its normalized sum of the arrivals over time $(0,u]$ and queues equal to $x(0,u) = b + (u-1)$.

In addition, an immediate implication of Lemma 9 is that we can rewrite $J$ in (13) as

$$J(b) = \inf_{t \geq 1} I_t(b) = \inf_{t \geq 2} \left( I^1_t(b), \inf_{u \in [1,t]} I^1_t(b) \right) = \inf_{t \geq 2} \left( I^1_t(b), \inf_{u \in [1,t]} I^u_t(x|_{[0,u]}) \right), \quad (16)$$

where we also used the fact that $I_1(b) = I^1_t(b)$. If we denote $t^*$ as the optimizer of the last equation, then $t^*$ is called the critical timescale (see [24]). It can then be interpreted that $t^*$ is the most likely length of time it would take to “fill” the buffers to a given level $b$ from being “empty” (more precisely, anywhere within $\mathcal{R}_s$).

**Remark 8:** The induction-based proof of Lemma 2 reveals another important property of the sample-paths, namely, that at every stage it suffices to consider the quasi-continuous selection $H^{wc}(\cdot)$ of the scheduling function. This then implies that we need to consider all valid arrivals sequences that respect the constraints of the set $\mathcal{A}(\cdot, b)$ such that using any of the allowed (based on the workload vector at each time) scheduling actions given by $H^{wc}(\cdot)$ results in the workload vector $b$ at the required epoch.

Note that for fixed $u \in \mathbb{N}$, $\inf_{x \in \mathcal{A}(u,b)} I^u_t(x|_{[0,u]})$ is a optimization problem with a convex cost function $I^u_t(\cdot)$ and a set $\mathcal{A}(u,b)$ of allowable solutions, in general, a $(K(u-1)]$-dimensional set. This problem is difficult to solve analytically. Since the cost function $I^u_t(x|_{[0,u]}) = \sum_{i=1}^{u} \sum_{k=1}^{K} \Lambda^k(x_i)$ is additive, a possible numerical method is the numerical
C. Properties of the Minimum-Cost Sample Paths

Here, we see that the convexity of the cost function \( \Lambda_1^{*,k} \) for all \( k \in K \), induces two properties for the optimal paths.

Property 1: Constant-speed linear path is the cheapest. Among all arrival sample paths \( \mathbf{x} \in D^K_\mu \) with the only constraint that \( \mathbf{x}(0,\ell] = \mathbf{b} \), i.e., the total amount of arrivals at the end of time \( \ell \in \mathbb{N} \) equals \( \mathbf{b} \), the cheapest or minimum-cost path is the constant-speed linear path, where the arrival in each timeslot is equal to \( \mathbf{b}/\ell \).

Proof: This is because the path cost function is additive, i.e., \( \Lambda^*_i(\mathbf{x}) = \sum_{k=1}^K \sum_{t=1}^\ell \Lambda^{*,k}_1(x^k_t) \), and the per-timeslot cost function \( \Lambda^{*,k}_1 \) is convex. Applying Jensen’s inequality [54] gives

\[
\Lambda^*_i(\mathbf{x}) = \sum_{k=1}^K \sum_{t=1}^\ell \Lambda^{*,k}_1(x^k_t) \\
\geq \sum_{k=1}^K t \Lambda^{*,k}_1 \left( \frac{1}{\ell} \sum_{i=1}^\ell x^k_i \right) \\
= \sum_{k=1}^K t \Lambda^{*,k}_1 (b^k/\ell),
\]

with equality when \( x^k_i = b^k/\ell \) for all \( i \) and \( k \). See an illustration in Figure 1(a).

From now onwards, without loss of generality, we assume that the arrivals, workloads and service vectors are scaled by \( 1/C \) where \( C = (C^1, \ldots, C^K) \) is the vector of maximum service rates. In this normalized setting, the maximum service rate of all the users is 1. Assume also that \( \Lambda^{*,k}_1 \) for \( k \in K \) is suitably modified for this scaling, and, for ease of exposition, that with this scaling the processes are statistically identical, i.e., \( \Lambda^{*,k}_1 = \Lambda^*_1 \) for all \( k \in K \). We can then write down the following property.

Property 2: Constant-speed linear path closest to the equal line is the cheapest. For constant-speed linear paths \( \mathbf{a} \in \mathbb{R}^{Kt}_+ \) with the sum \( \mathbf{a}(0,\ell] = d \), the cost of the path is a Schur-convex function [55].

Proof: Since the arrival paths are constant-speed linear path, without loss of generality we can consider arrival paths in a single timeslot. Consider path \( \mathbf{x} \in \mathbb{R}^{Kt}_+ \), then the cost of the path is \( \sum_{k=1}^K \Lambda^*_i(x^k) \) where \( \Lambda^*_i(\cdot) \) is a convex function. Therefore from the results in [55] it follows that \( \sum_{k=1}^K \Lambda^*_i(x^k) \) is a Schur-convex function. In order words, if \( \mathbf{x} \) is majorized by \( \mathbf{y} \in \mathbb{R}^{Kt}_+ \) (denoted by \( \mathbf{x} \prec \mathbf{y} \)), i.e., \( \sum_{t=1}^\ell x^{[k]}_t \leq \sum_{t=1}^\ell y^{[k]}_t \) for \( j = 1, \ldots, K - 1 \) and \( \sum_{t=1}^\ell x^{[K]}_t = \sum_{t=1}^\ell y^{[K]}_t \) where \( x^{[k]}_t \) is the \( k \)-th largest component of \( \mathbf{x} \), then \( \sum_{k=1}^K \Lambda^*_i(x^k) \leq \sum_{k=1}^K \Lambda^*_i(y^k) \).}

This is easily appreciated when \( K = 2 \). Let \( \mathbf{x} = (x, d-x) \in \mathbb{R}_+^2 \) and \( \mathbf{y} = (y, d-y) \in \mathbb{R}_+^2 \), where \( y > x > d/2 \), then we have \( \Lambda^*_1(x) + \Lambda^*_1(d-x) \leq \Lambda^*_1(y) + \Lambda^*_1(d-y) \), and \( \mathbf{x} \) is cheaper than \( \mathbf{y} \). We illustrate this in Figure 1(b).

These properties are also used in [9], [12], [13] for large-deviations analysis of scheduling disciplines. Next, we use these properties to calculate \( I_2 \) and bounds on \( I_t \) for \( t \in \mathbb{N} \) for just the work-conserving scheduler operating on the simplex rate-region.

D. Example: Calculation of \( I_2 \)

Here we look at an example for calculation of the finite-horizon rate function \( I_t \) to illustrate the fact that the calculation continues to be rather involved. For simplicity, consider the case when \( t = 2, K = 2 \), and the setting of Property 2

![Diagram showing two properties of the minimum-cost sample paths: (a) Property 1: the minimum-cost path is the constant-speed linear path. (b) Property 2: Path 1 which is closer to the Equal Line has a lower cost than Path 2.](image-url)
From (14), $I_2(b)$ for $b \in \mathbb{R}^2_+$ can be written as

$$I_2(b) = \min \left\{ \sum_{i=1}^{2} \Lambda_i^\circ(b^i), \inf_{x \in [0,2]} \sum_{i=1}^{2} \Lambda_A^\circ(x^i) \right\},$$

(17)

where

$$A(2, b) = \left\{ a_{[0,2]} \in \mathbb{R}^4_+ : a_2 \notin R_s, b \in a_{[0,2]} G_{2^w} \right\}.$$

The workload at time zero is $W_0 = G_{2^w}(a_{[0,2]}) = a_1 + [a_2 - H_{2^w}(a_2)]^+$, which is equal to $a(0,2) - H_{2^w}(a_2)$ since $a_2 \notin R_s$. On the other hand, we require $W_0 = b$. Hence, using the scheduler $H_{2^w}$ given in (3), we can express $A(2, b)$ as $A(1) \cup A(2) \cup A(3)$, where $A(j) \subseteq \mathbb{R}^4_+$, $j = 1, 2, 3$, are defined as

$$A(1) := \{(a_1, a_2) \in \mathbb{R}^4_+ : a_1 \geq a_2, a_2 \geq 1, a_2 = b + (1,0)\},$$

$$A(2) := \{(a_1, a_2) \in \mathbb{R}^4_+ : a_1 \leq a_2, a_2 \geq 1, a_2 = b + (0,1)\},$$

$$A(3) := \{(a_1, a_2) \in \mathbb{R}^4_+ : a_1 \leq 1, a_2 \leq 1, a_2 + a_2 \geq 1, a_2 = b + \text{Proj}_{R_s}(a_2)\}.$$

Note that in the definition of $A(1)$ and $A(2)$ we have used the property highlighted in Remark 8. For the two user case the scheduling function $H_{2^w}(-)$ is only discontinuous at $x \in \mathbb{R}^4_+$ such that $x_1 = x_2 \geq 1$. Here one can either choose the service vector $(1,0)$ or $(0,1)$ to obtain a quasi-continuous selection.

Using these newly defined sets the second term in the RHS of (17) can be rewritten as

$$\inf_{x \in A(2, b)} \sum_{i=1}^{2} \sum_{k=1}^{2} \Lambda_A^\circ(x^i_k) = \min_{j \in [1,3]} \inf_{x \in A(j)} \sum_{i=1}^{2} \sum_{k=1}^{2} \Lambda_A^\circ(x^i_k).$$

(a) Trajectories of Arrival Paths

(b) Trajectories of Workload Paths

Fig. 2. Example of accumulated arrival and workload paths for calculation of $I_2(b)$.

give the bounds to $I_t$ and $J$. We focus on $K = 2$ but similar results can be obtained for general $K$. Again the setting is the same as for Property 2.

**Lemma 10:** For $K = 2$, $b \in \mathbb{R}^2_+$, $I_t(b)$ can be bounded as

$$I_t(b) \geq \min_{u \notin \{0, t\}} \sum_{k=1}^{K} \Lambda_A^\circ \left( \frac{1}{u} \left( \text{Proj}_{X(u,b)}(0) \right)^k \right),$$

(18)

and when $b \notin [0,1]^2$,$$
I_t(b) \leq \min_{u \notin \{0, t\}} \sum_{k=1}^{K} \Lambda_A^\circ \left( \frac{1}{u} \left( b^k + (u-1)H(b)^k \right) \right),$$

(19)

where the convex set $X(u,b) \subseteq \mathbb{R}^2_+$ is defined as

$$X(u,b) := \{b + v : v^1 + v^2 = (u-1), v^1, v^2 \geq 0\}.$$

**Proof:** See Appendix D.

Next we look at the tightness of the above bounds for an example of compound Poisson source process with exponential packet size. We expect the tightness to depend on the traffic load.

E. Bounds on $I_t$

In this subsection, we find simple expressions that give lower or upper bounds to $\inf_{x \in A(u,b)} I_t^\circ(x)$, which in turn...
F. Numerical Examples

Here we compare the finite-horizon rate functions $I_2$ for three schedulers, namely, max-weight, GPS with equal weights, and a priority scheduler that gives higher priority to queue 1. We also examine the tightness of the bounds given in Lemma 10. Both of these are for an average of i.i.d. compound Poisson source processes with exponential packet size where the packet arrivals follow Poisson distribution of rate $\lambda$ and the average packet size is $1/\mu$ (see [40]). The function $\Lambda^*_1$ for this process is given by

$$\Lambda^*_1(x) = \mu(\sqrt{x} - \sqrt{\lambda})^2,$$

for $x \in \mathbb{R}_+$. Note that this has domain $\mathbb{R}_+$ and therefore satisfies Proposition 1. We once again make the simplifying assumption that the processes for the different users are statistically identical and that the rate-region is the unit simplex.

First we present results for the comparison of the different scheduling policies. Here we set $\lambda = 0.1$ or 0.3 respectively. However, these calculations are best appreciated when we compare them to the rate functions of other well-known scheduling policies. Fig. 4(a) compares the max-weight scheduler with a GPS scheduler with equal weights both at $\lambda = 0.1$ and 0.3 respectively. However, these calculations are best appreciated when we compare them to the rate functions of other well-known scheduling policies. Fig. 4(a) compares the max-weight scheduler with a GPS scheduler with equal weights both at $\lambda = 0.1$. One can see that the rate function for the max-W scheduler is greater than the rate-function of the GPS scheduler for some range of $b$, i.e., where $b^1$ is much greater than $b^2$ and vice versa. This means that for this range of $b$ and in the many-sources asymptotic sense, the work-conserving MW scheduler performs better than the work-conserving GPS scheduler (only considering the two time-step workload). The reason is that with $b^1$ much greater than $b^2$, MW can serve queue 1 at full capacity (1), whereas GPS has to serve both queues equally at $1/2$. Hence, the workload average packet size of $1/\mu = 100$. Fig. 3(a) and 3(b) show the finite-horizon (two-timestep) rate function $I_2(b)$ for $\lambda = 0.1$ and 0.3 respectively. However, these calculations are best appreciated when we compare them to the rate functions of other well-known scheduling policies. Fig. 4(a) compares the max-weight scheduler with a GPS scheduler with equal weights both at $\lambda = 0.3$. One can see that the rate function for the max-W scheduler is greater than the rate-function of the GPS scheduler for some range of $b$, i.e., where $b^1$ is much greater than $b^2$ and vice versa. This means that for this range of $b$ and in the many-sources asymptotic sense, the work-conserving MW scheduler performs better than the work-conserving GPS scheduler (only considering the two time-step workload). The reason is that with $b^1$ much greater than $b^2$, MW can serve queue 1 at full capacity (1), whereas GPS has to serve both queues equally at $1/2$. Hence, the workload average packet size of $1/\mu = 100$. Fig. 3(a) and 3(b) show the finite-horizon (two-timestep) rate function $I_2(b)$ for $\lambda = 0.1$ and 0.3 respectively. However, these calculations are best appreciated when we compare them to the rate functions of other well-known scheduling policies. Fig. 4(a) compares the max-weight scheduler with a GPS scheduler with equal weights both at $\lambda = 0.3$. One can see that the rate function for the max-W scheduler is greater than the rate-function of the GPS scheduler for some range of $b$, i.e., where $b^1$ is much greater than $b^2$ and vice versa. This means that for this range of $b$ and in the many-sources asymptotic sense, the work-conserving MW scheduler performs better than the work-conserving GPS scheduler (only considering the two time-step workload). The reason is that with $b^1$ much greater than $b^2$, MW can serve queue 1 at full capacity (1), whereas GPS has to serve both queues equally at $1/2$. Hence, the workload average packet size of $1/\mu = 100$. Fig. 3(a) and 3(b) show the finite-horizon (two-timestep) rate function $I_2(b)$ for $\lambda = 0.1$ and 0.3 respectively. However, these calculations are best appreciated when we compare them to the rate functions of other well-known scheduling policies. Fig. 4(a) compares the max-weight scheduler with a GPS scheduler with equal weights both at $\lambda = 0.3$. One can see that the rate function for the max-W scheduler is greater than the rate-function of the GPS scheduler for some range of $b$, i.e., where $b^1$ is much greater than $b^2$ and vice versa. This means that for this range of $b$ and in the many-sources asymptotic sense, the work-conserving MW scheduler performs better than the work-conserving GPS scheduler (only considering the two time-step workload). The reason is that with $b^1$ much greater than $b^2$, MW can serve queue 1 at full capacity (1), whereas GPS has to serve both queues equally at $1/2$. Hence, the workload
under MW is less likely to reach \((b^1, b^2)\) at the end of the two time-slots when the system starts from being empty. Similarly \(J(b)\) compares the rate-functions for the max-weight scheduler and a priority scheduler where user 1 has higher priority. Since the max-weight policy does not discriminate between the two users, we find that it is less likely for \(b^2\) to be large for the max-weight policy in comparison to the priority policy with the reverse being true for \(b^1\).

Next we compare our bounds for the rate function of the max-weight policy with the exact rate function in the scenarios where we can calculate it by brute force. Figure 5(a) shows the upper and lower bounds and the actual values of \(I_t\), for \(t = 10\), at \(\mu = 0.01\), and various values of \(\lambda = 0.1, 0.2, 0.3\) and when \(b = (b^1, b^2 = 1)\) for various values of \(b^1\). Figure 5(b) shows the corresponding minimizing \(t^*\) for the bounds and the actual expression of \(I_t\). We note that for all \(b\) in this example, \(J(b)\) is actually equal to \(I_t(b)\) for \(t = 10\) since all optimizing \(t^*\) values are less than 10 (see 16). This example shows that, in the range of \(b\) in consideration, both bounds are tight and almost coincide when the traffic load is small, i.e., \(\lambda = 0.1\). However, when the traffic load is higher, the lowerbound becomes loose while the upperbound is still considerably tight.

It is interesting to note the optimal timescale \(t^*\) which the queues most likely to take to reach the level \(b\). Figure 5(b) shows that, for example, it is most likely to take only two timeslots for CPE process with \(\lambda = 0.2\) to reach the buffer level \(b = (3, 1)\), while the most likely timescale is four timeslots when the traffic load is higher (\(\lambda = 0.3\)). Figure 6(c) and Figure 6(d) show the optimal trajectories of the accumulated arrival process and the workload process for \(\lambda = 0.2\) and 0.3, respectively.

Note that despite a potential exponential growth in computation of \(I_t\), such computations commonly reduce to simpler cases. For instance, consider the calculation of \(I_{10}\) in Figs. 6(c) and 6(d), in which we sequentially calculate \(I_1, I_2, I_3\), etc. However, this sequential computation stops as soon as one reaches the optimal timescale \(t^* < t\). For the values of the vector \(b\) in Figs. 6(c) and 6(d) for instance, \(t^*\) was at most 4,
making the calculations of $I_6$-$J_{10}$ unnecessary.

VIII. CONCLUDING REMARKS

In this paper, we have established a many-sources LDP for the stationary (infinite-horizon) workload for multi-queue single-server system with simplex rate region and under maximum-weight scheduling, when the arrival processes are assumed to satisfy a certain many-sources sample path LDP. To extend the LDP of the arrival processes to the LDP of the workloads, we employed Garcia’s extended contraction principle, which applies to quasi-continuous mappings. Along the way, we also establish an LDP for the finite-horizon workload in a very general setting of arbitrary compact, convex, and coordinate convex rate region under max-weight scheduling.

We gave the associated rate functions and the expression of the infinite-horizon rate function in term of the finite-horizon ones, when the arrival processes have i.i.d. increments.

Next, we catalogue some interesting areas of future research. The extension of our LDP result for the infinite-horizon workload in the case of an arbitrary compact, convex, and coordinate convex rate region remains open by and large. The main difficulty in establishing an LDP for the infinite-horizon workload is in showing the quasi-continuity of the infinite-horizon workload mapping. In the case of simplex rate region, the infinite-horizon workload mapping was shown to be reducible to a finite-horizon mapping whose quasi-continuity was established via induction. Another question that has only been partially explored in the current paper is the nature of Assumptions 2-3. In particular, beyond Proposition 1, the relationship between stochastic mixing properties of the arrival process and the analytical properties of $I^2_t$ and $I^1_t$ remains an important area of future research.

APPENDIX A

PROOF OF LEMMA 1

Lemma 1 states that it is possible to construct quasi-continuous functions $H$ and $H^\text{wuc}$ such that

$$H(W_t) \in \arg \max_{R \in \mathcal{R}} <R, W_t>,$$

and

$$H^\text{wuc}(W_t) = \begin{cases} \text{Proj}_\mathcal{R}(W_t) & W_t \in \Pi^K_{k=1} [0, C^k] \\ H(W_t) & \text{otherwise} \end{cases}.$$

We do this using selection theorems (see [58] and [59]) for correspondences associated with max-weight scheduling:

$$\mathcal{H}(W_t) := \arg \max_{R \in \mathcal{R}} <R, W_t>.$$ (20)

First we define the following analytical properties of correspondences.

Definition 5 (Local-boundedness): A correspondence $F : \mathcal{X} \rightrightarrows \mathcal{Y}$ is locally-bounded (lb) [56, Defn. 5.15, pp. 157–158] at point $x \in \mathcal{X}$ if there exists a neighborhood $V$ of $x$ such that $F(V) := \cup_{a \in V} F(a)$ is bounded in $\mathcal{Y}$. Correspondence $F$ is deemed lb if it is lb for every point $x \in \mathcal{X}$.

Definition 6 (Outer semicontinuity): For a correspondence $F : \mathcal{X} \rightrightarrows \mathcal{Y}$ define the outer-limit at $x \in \mathcal{X}$ to be

$$\limsup_{a \rightarrow x} F(a) := \cup_{a \rightarrow x} \limsup_{n \rightarrow \infty} F(x^n)$$

$$= \{y | \exists x^n \rightarrow x, \exists y^n \rightarrow y \text{ with } y^n \in F(x^n) \}.$$  

Then $F$ is outer semicontinuous (osc) [56, Defn. 5.4, pg. 152] at $x \in \mathcal{X}$ if $\limsup_{a \rightarrow x} F(a) \subset F(x)$. Correspondence $F$ is deemed osc if it is osc for every point $x \in \mathcal{X}$.

Definition 7 (Upper semicontinuity): A correspondence $F : \mathcal{X} \rightrightarrows \mathcal{Y}$ is upper semicontinuous (usc) ( [56, Thm. 5.19, pp. 160] and [57]) at point $x \in \mathcal{X}$ if for any open neighborhood $V$ of $F(x)$, there exists a neighborhood $U$ of $x$ such that $F(U) \subseteq V$ for all $a \in U$. Alternatively, $F$ is said to be upper semicontinuous at $x \in \mathcal{X}$ if whenever we have sequences $\{x_m\} \in \mathcal{X}$ and $\{y_m\} \in \mathcal{Y}$ such that $y_m \in F(x_m)$ for all $m$ with $x_m \rightarrow x$ and $y_m \rightarrow y$, then $y \in F(x)$. Correspondence $F$ is deemed usc if it is usc for every point $x \in \mathcal{X}$.

Now we can prove Lemma 1.

Proof of Lemma 1: The scheduling function $H(\cdot) : \mathbb{R}^K_+ \mapsto \mathcal{P}(\mathcal{R})$ is a maximal monotone correspondence [56, Thm. 12.17, pp. 542–543] that picks closed and convex subsets of $\mathcal{R}$ for every $x \in \mathbb{R}^K_+$; thus, compact subsets of $\mathbb{R}^K_+$. It is, therefore, both lb and usc from [56, Ex. 12.8b, pg. 536]. Now it follows that we get a quasi-continuous selection by [58, Thm. 2] and [59, Thms. 2.4 & 2.5] since $H(W_t(\cdot))$ is usc and has compact values.

To prove that $H^\text{wuc}$ is a quasi-continuous selection we first use the same steps as above but with a restricted domain, i.e., for $\mathcal{H}(\cdot) : \mathbb{R}^K_+ \setminus \prod_{k \in \mathcal{K}} [0, C^k] \mapsto \mathcal{P}(\mathcal{R})$. Since $\mathbb{R}^K_+ \setminus \prod_{k \in \mathcal{K}} [0, C^k]$ is a Baire space [42] the results from [58], [59] still apply. From the definition of the work-conserving/rate-maximizing max-weight scheduler, if $x \in \prod_{k \in \mathcal{K}} [0, C^k]$, then we get continuity from within this set from the properties of $\text{Proj}_{\mathcal{R}}(\cdot)$. Thus, we can satisfy the definition of quasi-continuity from Defn. 1.

We refer the reader to [60, Thm. 2.2] and [61, Thm. 3.4] for an exposition and for generalizations of the result from [58], [59].

APPENDIX B

PROOF OF LEMMAS 2-3

Next we prove Lemma 2 by using the following fact:

Fact 2: Assume $\mathcal{X}, \mathcal{Y}$ are metric spaces and $F_1$ and $F_2$ are functions from $\mathcal{X}$ onto $\mathcal{Y}$. If $F_1$ is quasi-continuous at $x \in \mathcal{X}$ and $F_2$ is continuous at $x$, then $F_1 + F_2$ is quasi-continuous at $x$.

Below we state and prove Lemma 2 for a simplex rate region $R_s$ to avoid unnecessary notational complexities; the result readily extends to a general rate region as discussed in Remark 3.

Lemma 2: For $t \in \mathbb{N}$, $G^\text{wuc}_t$ for $R_s$ is quasi-continuous on $\mathbb{R}^{K_t}_+$ with respect to the scaled uniform norm topology.

Proof: Using our queuing equation we first observe the following recursive relation between $G^\text{wuc}_t$ and $G^\text{wuc}_{t-1}$ for any
where we used the fact that \( W^{wc}(x_{i0}) = G^{wc}(x_{i0}) \), and \( W^{wc}(x_{i1}) = G^{wc}(x_{i1}) \) when the initial backlog at time \(-t\) is 0.

Equation \((21)\) says that \( G^{wc}(x_{i0}) \) depends linearly on \( x_1 \). This implies the following simple but consequential observation:

**Observation 1:** If \( G^{wc} \) is quasi-continuous at \( x_{i0} \), then it is quasi-continuous at \( \hat{x}_{i0} := (x_1, x_2, \ldots, x_k) \) for any \( x_1 \in \mathbb{R}^K \), and if \( G^{wc} \) is continuous at \( x_{i0} \), then it is also continuous at \( \hat{x}_{i0} \).

**Observation 2:** If \( G^{wc}(x^n_{i0}) \rightarrow G^{wc}(x_{i0}) \) for a sequence \( \{x^n_{i0}\} \) such that \( x^n_{i0} \rightarrow x_{i0} \), then for any sequence \( \{x^n_{i1}\} = (x^n_{i1}, x^n_{i2}, \ldots, x^n_{ik}) \) where \( x^n_{i1} \rightarrow x_1 \), we also have \( G^{wc}(x^n_{i1}) \rightarrow G^{wc}(x_{i1}) \).

Using the recursive relation in \((21)\), we prove this lemma by induction on \( t \in \mathbb{N} \). For every \( t \in \mathbb{N} \) we assume that the system is empty at \(-t\). Therefore, the arrival sample-path prior to \(-t\) has no influence on \( W^{wc}_{0:t} \). Thus, we will only specify the values of the constructed sequences up until time \( t \); the extension to sample-paths in \( D^\mu \) while ensuring that the system is empty at \(-t\) is trivial. For \( t = 1 \), \( G^{wc}(a_1) = a_1 \), hence \( G^{wc} \) is continuous on \( \mathbb{R}^K \). Assuming that \( G^{wc} \) is quasi-continuous on \( \mathbb{R}^{K(t+1)} \) we want to show that \( G^{wc} \) is quasi-continuous on \( \mathbb{R}^{K(t+1)} \). Using the fact that the \(|\cdot|\) function is continuous, Remark \(2\) and Fact \(2\) it suffices to show that the function \( F_t := G^{wc} - H^{wc} \circ G^{wc} \) is quasi-continuous on \( \mathbb{R}^{K(t+1)} \) to show that \( G^{wc} \) is quasi-continuous. In particular, for any arrival sample path \( a \in D^\mu \), we need to show that \( F_t \) is quasi-continuous at \( a_{i0} \). It suffices to show that it is possible to select a sequence \( \hat{a}^n \rightarrow a \) (in \( D^\mu \)) for which

\[
G^{wc}(\hat{a}^n_{i0}) \rightarrow G^{wc}(a_{i0}), \quad H^{wc} \circ G^{wc}(\hat{a}^n_{i0}) \rightarrow H^{wc} \circ G^{wc}(a_{i0}),
\]

such that both \( G^{wc}(\cdot) \) and \( H^{wc} \circ G^{wc}(\cdot) \) are continuous at every \( \hat{a}^n_{i0} \). Note that in contrast to Fact \(2\) we are adding two quasi-continuous functions and showing that the sum is quasi-continuous; the key to our proof is to ensure that we use the same sequence for both functions.

We show this by first noting that the induction hypothesis, i.e., quasi-continuity of \( G^{wc} \), and the definition of quasi-continuity ensure that there exists a sequence \( \{a^n\} \) such that \( a^n \rightarrow a \) in the scaled uniform topology, such that \( G^{wc}(a^n_{i0}) \rightarrow G^{wc}(a_{i0}) \), and \( G^{wc}(\cdot) \) is continuous at \( a^n_{i0} \) for all \( n \). We will construct the desired sequence \( \{\hat{a}^n\} \) by modifying \( a^n \) appropriately. We proceed by considering the following two cases, depending on the value of \( a_1 \).

**Case 1:** \( a_1 > 0 \), i.e., every component of the \( a_1 \in \mathbb{R}^K \) is positive. Let \( \epsilon > 0 \) be the smallest component of \( a_1 \), i.e., \( \epsilon = \min_{k \in K} a_k^1 \). Since \( H(\cdot) \) is quasi-continuous, it is possible to choose a sequence of (workload) vectors \( \{w^n\} \) such that \( w^n \rightarrow G^{wc}(a_{i0}) \), and \( H \) is continuous at \( w^n \) for all \( n \).

Now, we define

\[
\hat{a}_1^n = w^n - [F_{t-1}(a^n_{1})]^+ = w^n - G^{wc}(a^n_{10}) + a_1^n = (w^n - G^{wc}(a^n_{10})) + (G^{wc}(a^n_{10}) - G^{wc}(a_{10})) + (a_1^n - a_1) + a_1.
\]

It is clear from the last relationship that \( \hat{a}_1^n \rightarrow a_1 \). We still need to ensure that \( \hat{a}_1^n \geq 0 \) since negative quantities are involved in the definition. We do this by using the facts that every component of \( a_1 \in \mathbb{R}^K \) is greater than or equal to \( \epsilon > 0 \), and that \( w^n \rightarrow G^{wc}(a_{10}) \), \( G^{wc}(a_{10}) \rightarrow G^{wc}(a_{10}) \), and \( a_1^n \rightarrow a_1 \). Thus, we will only have non-negativity of all components. Since we construct a new sequence \( \{\hat{a}^n_{i0}\} \) where \( \hat{a}_1^n = a_1^n + \epsilon, \) \( \hat{a}_1^n = a_1^n + \epsilon \).

This new sequence \( \hat{a}^n_{i0} \) is the sequence we are after because using the induction hypothesis together with Observations \(1\) and \(2\) we have that \( G^{wc}(\hat{a}^n_{i0}) \rightarrow G^{wc}(a_{i0}) \), and \( G^{wc} \) is continuous at \( a^n_{i0} \) for all \( n \). Furthermore, by construction

\[
G^{wc}(\hat{a}^n_{i0}) = a^n_{i0} + [F_{t-1}(\hat{a}^n_{10})]^+ = a_1^n + [F_{t-1}(a^n_{10})]^+ = w^n + a_1^n.
\]

Hence, we have shown that there exists a sequence \( \hat{a}^n_{i0} \) satisfying \((22)\) and \((23)\). In addition, the continuity of \( H^{wc} \circ G^{wc} \) at \( \hat{a}^n_{i0} \) for all \( n \) is a direct consequence of continuity of \( G^{wc} \) at \( a^n_{i0} \) and continuity of \( H^{wc} \) at \( w^n \), which is equal to \( G^{wc}(\hat{a}^n_{i0}) \), for all \( n \).

**Case 2:** \( a_1 \geq 0 \). Let \( K_1 := \{k \in K : a_k^1 = 0\} \) and let \( K_2 := \text{arg \, max}_{k \in K \setminus K_1} W^k_{0}(a_{i0})C^k \). Without loss of generality, by permuting the user labels we can assume that the first \( K := K_1 \cup K_2 \) components of \( a_1 \) are either in \( K_1 \) or in \( K_2 \); thus, the rest of the \( K - K \) components are both positive and not part of the scheduling decision made at time \( 0 \) with arrival sequence \( a_{i0} \). Now consider the sequence \( a_m^n := [1/mC]K + a_1 \) where \( [1/mC]K \) is short-hand for a vector with \( 1/(mC) \) in the first \( K \) components and \( 0 \) in the remaining coefficients; by construction \( a_m^n \) converges to \( a_1 \) such that for every \( m \) every component of \( a_m^n \) is positive. We construct a sequence \( \{a_{m0}^{n}\} \) with this \( a_{m0}^{n} \) and \( a_{10}^{n}(1,1) = a_{10}(1,1) \). It is obvious that \( G^{wc}(a_{m0}^{n}(1,1)) \rightarrow G^{wc}(a_{i0}(1,1)) \) since

\[
G^{wc}(a_{m0}^{n}(1,1)) = a_{m0}^{n} + [F_{t-1}(a_{m0}^{n}(1,1))]^+ = a_{10}^{n} + [F_{t-1}(a_{10}(1,1))]^+ = [1/mC]_K + G^{wc}(a_{i0}(1,1)).
\]

When \( G^{wc}(a_{i0}(1,1)) \notin \bigcap_{k=1}^{K} C^k \), by construction we have

\[
H^{wc} \circ G^{wc}(a_{i0}(1,1)) = H^{wc}(G^{wc}(a_{i0}(1,1))) = H(a_{i0}(1,1)) = H \circ G^{wc}(a_{i0}(1,1)),
\]
where the function $H(\cdot)$ is the regular max-weighted scheduling function (with $\mathcal{R}_s$). On the other hand, if $G_{\mu,c}^m(a_{(0,t)}) \in \prod_{k=1}^{K} [0, C_k)$, then the continuity of Proj$_{\mathcal{R}_s}(\cdot)$ yields $H_{\mu,c} \circ G_{\mu,c}^m(a_{(0,t)}) \to H_{\mu,c} \circ G_{\mu,c}^m(a_{(0,t)})$.

Since for each $m$ we have that $a^{m}_{\infty}$ has all elements strictly positive, we can use the construction from Case 1 but with $a^{m}_{(0,t)}$ in place of $a_{(0,t)}$. In particular, for each $m$, we can now generate a sequence $(\tilde{a}^{m,n}_{\cdot})$ such that $\tilde{a}^{m,n}_{t} \to a^{m}_{t}$ as $n \to +\infty$, $\tilde{a}^{m,n}_{(1,t)} = a^{m}_{(1,t)}$, and by using Observations 1 and 2 the following hold

\begin{align}
G_{\mu,c}^m(\tilde{a}^{m,n}_{(0,t)}) & \to G_{\mu,c}^m(a^{m}_{(0,t)}), \quad (25) \\
H_{\mu,c} \circ G_{\mu,c}^m(\tilde{a}^{m,n}_{(0,t)}) & \to H_{\mu,c} \circ G_{\mu,c}^m(a^{m}_{(0,t)}), \quad (26)
\end{align}

with both $G_{\mu,c}^m(\cdot)$ and $H_{\mu,c} \circ G_{\mu,c}^m(\cdot)$ being continuous at $\tilde{a}^{m,n}_{(0,t)}$ for all $m$, $n$.

Now we define the sequence $\tilde{a}^{m} = \tilde{a}^{m,n}$ as the sequence we are after. By construction, we have $\tilde{a}^{m}_{(0,t)} = a_{(0,t)}$, and both $G_{\mu,c}^m(\cdot)$ and $H_{\mu,c} \circ G_{\mu,c}^m(\cdot)$ continuous at all $\tilde{a}^{m}_{(0,t)}$. Since $G_{\mu,c}^m(\tilde{a}^{m}_{(0,t)}) \to G_{\mu,c}^m(a_{(0,t)})$ and $H_{\mu,c} \circ G_{\mu,c}^m(\tilde{a}^{m}_{(0,t)}) \to H_{\mu,c} \circ G_{\mu,c}^m(a_{(0,t)})$, it follows from (25) and (26) that $G_{\mu,c}^m(\tilde{a}^{m}_{(0,t)}) \to G_{\mu,c}^m(a_{(0,t)})$ and $H_{\mu,c} \circ G_{\mu,c}^m(\tilde{a}^{m}_{(0,t)}) \to H_{\mu,c} \circ G_{\mu,c}^m(a_{(0,t)})$.

Remark 9: The proof above can be carried out for every rate-region in the class that we are interested in. The argument presented in Case 1 would remain exactly the same but the argument presented in Case 2 would have to be modified to account for a further characterization of $\arg\max_{R \in \mathcal{R}} < W_0(a_{(0,t)}), R >$, especially when it is not a singleton. The components of $a_{(0,t)}$ will need to be adjusted in such a manner so as not to perturb $\arg\max_{R \in \mathcal{R}} < W_0(\tilde{a}^{m}_{(0,t)}), R >$ for the adjusted sequence $\tilde{a}^{m}_{(0,t)}$. As the case of a non-singleton $\arg\max_{R \in \mathcal{R}} < W_0(\tilde{a}^{m}_{(0,t)}), R >$ will correspond to a specific set of values of $W_0(a_{(0,t)})$ (a cone) such that the boundary of the rate-region and a hyper-plane intersect at more than one point, we will need to use the normal corresponding to the hyper-plane in constructing the appropriate perturbation. Thus, on a case-by-case basis the same argument can be carried out for every rate-region.

Finally, we prove Lemma 3.

**Lemma 3** For $t \in \mathbb{N}$, $G_t(\cdot)$ is quasi-continuous on $\mathbb{R}^{K\times 1}$.  

**Proof:** Consider the queuing equation, i.e.,

$$W_{T-1} = W_T - R_T^+ + a_T \quad T \in \mathbb{N},$$

where $R_T^+ \in H(W_T)$, where $H(\cdot)$ is the maximal monotone correspondence defined in (20). Assume that we start the system at (fixed) time $T \in \mathbb{N}$ with workload vector $W_T \in \mathbb{R}^{K}$; we will often assume that $W_T = 0$. First, for $t \in \mathbb{N}$ we define a correspondence $G_t(W_t, a_{(0,t)}) : R_{(0,t)}^{\infty} \to \mathbb{R}^{K}$ that represents all possible workload vectors at time $0$ that can be achieved from the inputs $(W_t, a_{(0,t)})$. This results from the successive application of the queuing equation where we use all possible values in $H(\cdot)$ based upon the workload vector that results at each step. Our goal is show that $G_t(\cdot)$ admits a quasi-continuous selection which we shall call $G_t(\cdot)$ for $t > 1$ it suffices to establish

$$G_{t-1}(W_t, a_{(1,t)}) - H(G_{t-1}(W_t, a_{(1,t)})) := \{ x \in \mathbb{R}^K : x = y - z \text{ for some } y \in G_{t-1}(W_t, a_{(1,t)}) \text{ and } z \in H(G_{t-1}(W_t, a_{(1,t)})) \}$$

admits a quasi-continuous selection which we shall call (with an abuse of notation) $G_{t-1}(W_t, a_{(1,t)}) - H(G_{t-1}(W_t, a_{(1,t)}))$. Since $[\cdot]^+$ is a continuous function, we have $[\cdot]^+ \circ G_{t-1}(W_t, a_{(1,t)}) - H([\cdot]^+ \circ G_{t-1}(W_t, a_{(1,t)}))$ being a quasi-continuous function. Finally, by properties of projections and by the definition of quasi-continuity, we get the quasi-continuity of $G_t(W_t, a_{(0,t)})$. We should add a note of caution here that even though for notational convenience we write

$$G_t(W_t, a_{(0,t)}) = [G_{t-1}(W_t, a_{(1,t)}) - H(G_{t-1}(W_t, a_{(1,t)}))]^+ + a_1,$$

it need not be the case that the quasi-continuous selection that we obtain for $G_t$ be related using the above queuing equation to the quasi-continuous selection for $G_{t-1}$. Additionally, we may not even use the quasi-continuous selection $H(\cdot)$. Therefore, the property highlighted in Remark 8 need not hold.

The proof will once again use mathematical induction where our induction step will assume that $G_{t-1}(\cdot)$ is osc. Since $W_0 \leq W_t + a_1(0, t)$ for any $W_0 \in G_t(W_t, a_{(0,t)})$, it follows that $G_t(W_t, a_{(0,t)})$ is osc. Now using [56, Ex. 12.8b, pg. 536] we know that $H(\cdot)$ is both lb and osc, and therefore by [56, Thm. 5.19, pg. 160] it is also osc. Then using [56, Prop. 5.52b, pp. 184–185] we have $H(G_{t-1}(\cdot))$ being osc. Therefore it also follows that $G_{t-1}(\cdot) - H(G_{t-1}(\cdot))$ is also osc. Once this has been demonstrated the induction step is very easy as $G_t(\cdot)$ is obtained from $G_{t-1}(\cdot) - H(G_{t-1}(\cdot))$ by continuous transformations, as mentioned above. This same method also allows us to establish the initial step of the induction procedure; note that we will be dealing with $H(W_t)$ and $W_t - H(W_t)$ in this case.

Since $G_{t-1}(\cdot) - H(G_{t-1}(\cdot))$ is lb, using [56, Ex. 12.8b, pg. 536] we have $G_{t-1}(\cdot) - H(G_{t-1}(\cdot))$ also being osc. Finally, we get a quasi-continuous selection by [60, Thm. 2.2] and [61, Thm. 3.4] since $G_{t-1}(\cdot) - H(G_{t-1}(\cdot))$ is osc and takes compact values. The required result then follows by setting $G_t(W_t, a_{(0,t)}) = G_t(0, a_{(0,t)})$.

**APPENDIX C**

**PROOF OF LEMMAs 5, 6**

**Lemma 5** Consider an arrival process $a \in D_{\mu}^{K}$. There exists a $s^* = s^*(a) \times \infty$ such that the workloads at time $-s^*$ under $a$ lies within the rate region $\mathcal{R}_s$, i.e., $G_{\mu,c}(a_{(-s^*,:)} \in \mathcal{R}_s$. Furthermore, for any sequence of arrival processes $\{a^n \in D_{\mu}^{K}\}$ converging to $a$ (in scaled uniform topology), the workloads at time $-s^*$ under $a^n$, when $n$ is large enough, also lie within the rate region $\mathcal{R}_s$, i.e., $\exists n_0$ such that $G_{\mu,c}(a^n_{(-s^*,:)} \in \mathcal{R}_s$ for $n > n_0$.

**Proof:** Consider the normalized sum arrivals and the normalized sum workloads, and follow the proof in [24].
where recall that the hat (·) notation means the normalized sum over all users, i.e., $\hat{a}_t = \sum_{k=1}^{K} a_t^k / C^k$ and $\hat{W}_t = \sum_{k=1}^{K} W_t^k / C^k$. Recursion of the queue dynamics [27] and letting $T \to \infty$ where $W_T \in R_s$, gives the standard expression for the infinite-horizon sum workload [26], namely,

$$\hat{W}_0 := \hat{G}(a) = \sup_{t \in T} \hat{a}(0, t] - (t - 1), \quad (28)$$

where $\hat{G}$ represents the infinite-horizon normalized sum workload mapping; in other words, for all $s$, $\hat{W}_s = \hat{G}(a)(s, \infty)$ represent the normalized sum workload at time $s$, under arrival sequence $a$.

To prove the lemma we use the fact that the rate region $R_s$ is simplex, hence $\hat{W}_s \leq 1 \Leftrightarrow \hat{W}_s \in R_s$. Thus, it suffices to show that there is a finite $n_0$ and a finite $s$ such that $\hat{G}(a)(s, \infty) \leq 1$, and for $n \geq n_0$, $G(a)(s, m, \infty) \leq 1$.

Since $a \in D_\mu$, there is a $t_0 < \infty$ such that for all $\epsilon > 0$, $t > t_0$ and $k \in K$, $\hat{a}_t^{(0, t]} \leq C^k \mu^k$. Since $\mu$ is int$(R_s)$, we choose $\epsilon = (1 - \hat{\mu})/4K$. We now have that for all $t \geq t_0$, $\hat{a}_t^{(0, t]} \leq \hat{\mu} + C^k = (1 + 3\hat{\mu})/4 < 1$. In other words, the workload at time zero is a function of only the arrivals within time $(0, t_0]$ and hence,

$$\hat{W}_0(a) = \sup_{1 \leq t \leq t_0} \hat{a}(0, t] - (t - 1).$$

Let $s^* \leq t_0 < \infty$ be the minimum values of the optimizing $t$'s in the above equation. It can be shown [26, Lemma 5,4] that

$$\hat{G}(a)(s, \infty) \leq 1. \quad (29)$$

It is known that $\hat{G}$ is continuous on $D_\mu$ [24, Lemma 13] when $\hat{\mu} < 1$. However, this together with continuity of shift operator implies that for all $\{a^n\}$ such that $a^n$ converges to $a$ in scaled uniform topology,

$$\hat{G}(a^n)(u, \infty) \to \hat{G}(a)(u, \infty) \quad \text{for all } u \in (0, s^*]. \quad (29)$$

In particular, [29] implies that there exists $n_0$ such that for all $n \geq n_0$, the normalized sum workload under arrival sequence $a^n$ at time $u = s^*$ is no more than 1 packets, i.e.,

$$\hat{G}(a^n)(s^*, \infty) < 1 \quad \text{for all } n \geq n_0. \quad (29)$$

However, since the rate region is a simplex, the workload vectors at time $s^*$, under $a$ and $a^n$ lie in the rate region $R_s$. Hence, we have the assertion of the lemma.

Next, we prove Lemma 6.

Lemma 6 Let $\mu \in \text{int}(R_s)$, $a$ be an arrival sequence with rate $\mu$ with $I^*(a) < +\infty$, and $W = G^{\text{wc}}(a)$ be its corresponding steady state workload. For any $W \in ^*G^{\text{wc}}$ there exists a sequence of arrivals $\{a^n \in D_\mu^k\}$ such that $a^n$ converges to $a$ in the scaled uniform norm topology, $G^{\text{wc}}(a^n) \to G^{\text{wc}}(a)$, $G^{\text{wc}}$ is continuous at $a^n$, and $I^*(a^n) \to I^*(a)$.

Proof: Lemma 5 implies that for any given arrival sequence $a \in D_\mu^k$, there exists a $s^*$ such that

$$G^{\text{wc}}(a) = G^{\text{wc}}_s(a)(0, s^*].$$

However, $G^{\text{wc}}_s$ is quasi-continuous on $R_s$. This implies that there exists a sequence of finite arrivals $\{a^n\}$ such that

1) $\hat{a}_n(0, s^*) \rightarrow a(0, s^*)$;
2) $G^{\text{wc}}(a)$ is continuous at $a^n(0, s^*)$; and
3) $G^{\text{wc}}(a)(0, s^*) \rightarrow G^{\text{wc}}(a)(0, s^*)$.

Now construct the sequence of arrivals $\{a^n\}$ via concatenation of $a(0, s^*)$ and $\hat{a}_n(0, s^*)$. It is immediate that $a^n \rightarrow a$. Appealing to Lemma 5 for $n$ large enough (greater than $n_0$) we have

1) convergence of $G^{\text{wc}}(a^n)$ to $G^{\text{wc}}(a)$ since

$$G^{\text{wc}}(a^n) = G^{\text{wc}}_s(a^n)(0, s^*) \to G^{\text{wc}}_s(a)(0, s^*) \to G^{\text{wc}}(a); \quad (28)$$

2) continuity of $G^{\text{wc}}(\cdot)$ at $a^n$. For any sequence converging to $a^n$ in $D_\mu^k$ by appealing to Lemma 5, we know that far enough along every sequence only the arrivals in $(0, s^*)$ matter. Now using the fact that projection is continuous on $D_\mu^k$, we get the result from the continuity of $G^{\text{wc}}_s$ at $a^n(0, s^*)$.

This establishes the quasi continuity of function $G^{\text{wc}}$. Lastly, Assumptions 3 and 2 ensure that $I^*(a^n) \rightarrow I^*(a)$.

Finally, we prove Lemma 7.

Lemma 7 If $\mu \in \text{int}(R_s)$, the mapping $G^{\text{wc}}(\cdot)$ is almost compact on $D_\mu^k$ with respect to the scaled uniform norm topology.

Proof: This follows almost exactly along the same lines as the proof of Lemma 6. For any $a^n \rightarrow a$ we proved the existence of a $n_0$ such that for $n \geq n_0$ such that the workload vectors only depended on the arrivals within time $(0, t_0]$. Thus, the proof of almost compactness simply follows from Lemma 4. Note that we have used the fact that the projection operator is continuous on $D_\mu^k$.

APPENDIX D

PROOF OF LEMMA 10

Next we prove Lemma 10 which gives the bounds on $I_t$.

Lemma 10 For $K = 2$, $b \in R_+^2$, $I_t(b)$ can be bounded as

$$I_t(b) \geq \min_{u \in (0, t]} u \sum_{k=1}^{K} \Lambda^*_t \left( \frac{1}{u} \text{Proj}_{X_u(b)}(0) \right)^{k}$$

and when $b \notin [0, 1)^2$,

$$I_t(b) \leq \min_{u \in (0, t]} u \sum_{k=1}^{K} \Lambda^*_t \left( \frac{1}{u} (b^k + (u - 1)H(b)) \right)^{k},$$

where we recall that the convex set $X(u, b) \subseteq R_+^2$ is defined as

$$X(u, b) := \{b + v : v \in R_+^2 \text{ and } v_1 + v_2 = (u - 1)\}.$$
is easy to see that both bounds turn to be \( I_1(b) \). Owing to this we can also assume that \( u > 1 \) since the terms corresponding to \( u = 1 \) in both bounds evaluate to \( I_1(b) \).

We first show the lowerbound \([18]\). As we have noted earlier that for \( a_{i}(0,t) \in \Lambda(u,b) \), the \([-]^{+} \) function can be removed from the queue dynamics. Hence, we have \( a(0,u) = b + \sum_{i=1}^{u-1} H(W_i) \), where \( W_i \in \mathbb{R}_s \) and \( W_i \notin \mathbb{R}_s \) for all \( i \in (0,u-1) \). Using this and the fact that \( H(W_0) \in \{ v \in \mathbb{R}^2_+ \mid v^1 + v^2 = 1 \} \), for all \( i \in (0,u-1) \), we have \( a(0,u) \geq \mathbb{X}(u,b) \) is defined above. Now, given any point \( d \in \mathbb{X}(u,b) \), the constant-speed linear path with increments of \( d/u \) is the minimum-cost path among all the paths with the same destination (using Property 1). In addition, among all the paths to destinations in \( \mathbb{X}(u,b) \), the closest constant-speed linear paths \( a^{*}(0,u) \) to the equal line is the minimum-cost path (using Property 2). Since the closest point in \( \mathbb{X}(u,b) \) to the equal line is \( \text{Proj}(X_{u,b}(0)) \), we have \( a^{*}(0,u) = \text{Proj}(X_{u,b}(0)) \), \( i \in (0,u) \). Since the set of paths with destination in \( \mathbb{X}(u,b) \) includes all paths in \( \Lambda(u,b) \), from [7] we have the lowerbound \([18]\):

\[
I_1(b) = \min_{u \in (0,t)} \inf_{X \in \Lambda(u,b)} \sum_{k=1}^{K} \sum_{i=1}^{u} \Lambda_i^{+}(x_i^{(k)})
\]

To show the upperbound \([19]\), we only need to show that the constant-speed linear path \( a|_{(0,u)} = (a_i = \frac{1}{u}(b + (u - 1)H(b)) \), \( i \in (0,u) \), is in \( \Lambda(u,b) \), when \( b \notin [0,1)^2 \). Without loss of generality, we consider only when \( b^1 \geq b^2 \) and \( b^1 \geq 1 \). In this case, we set \( H(b) = (1,0) \) and the queue dynamics gives

\[
W_i = \frac{u-i}{u}(b + (u - 1)(1,0)) - (u - 1 - i)(1,0),
\]

for all \( i \in (0,u-1) \). Since \( b^1 \geq 1 \), we have \( W_i^1 \geq 1 \) and \( W_i^2 \geq W_i^2 \), and hence we can once again set \( H(W_i) = (1,0) \) for all \( i \in (0,u-1) \). Hence, \( a|_{(0,u)} \in \Lambda(u,b) \).

### References


Vijay G. Subramanian received the B.Tech. degree from the Department of Electrical Engineering, Indian Institute of Technology, Madras, India, in 1993, the M.S. degree in electronic communication engineering from the Indian Institute of Science, Bangalore, India, in 1995, and the Ph.D. degree in electrical engineering from the University of Illinois at Urbana-Champaign, Urbana, in 1999.

From 1999 to 2006, he was with the Networks Business, Motorola, Arlington Heights, IL. From May 2006 to Nov. 2010, he was a Research Fellow at the Hamilton Institute, NUIM, Ireland. Since Nov. 2010 he is a Senior Research Associate, Dept of EECS, Northwestern University. His research interests include information theory, communications, communication networks, wireless networks, queuing theory, and applied probability and stochastic processes.

Tara Javidi studied electrical engineering at Sharif University of Technology, Tehran, Iran, from 1992 to 1996. She received the M.S. degrees in electrical engineering (systems), and in applied mathematics (stochastics) from the University of Michigan, Ann Arbor, in 1998 and 1999, respectively. She received the Ph.D. degree in electrical engineering and computer science from the University of Michigan, Ann Arbor, in 2002.

From 2002 to 2004, she was an Assistant Professor at the Electrical Engineering Department, University of Washington, Seattle. She joined University of California, San Diego, in 2005, where she is currently an Associate Professor of Electrical and Computer Engineering. She was a Barbour Scholar during 19992000 academic year and received an NSF CAREER Award in 2004. Her research interests are in communication networks, stochastic resource allocation, stochastic control theory, and wireless communications.

Somsak Kittipiyakul received the S.B. degree and M.Eng. degrees in electrical engineering and computer science from Massachusetts Institute of Technology, Cambridge, in 1996. He received the Ph.D. degree in electrical and computer engineering from the University of California, San Diego, in 2008.

He is currently a lecturer at the School of Information, Computer and Communication Technology, Sirindhorn International Institute of Technology, Thammasat University, Thailand. His research interests include resource allocation and scheduling in wireless networks and systems.