Many-Sources Large Deviations for Max-Weight Scheduling

Somsak Kittipiyakul* and Tara Javidi*
ECE Department
University of California, San Diego
La Jolla, CA 92093
Email: {skittipi, tjavidi}@ucsd.edu

Vijay G. Subramanian†
Hamilton Institute
NUIM, Maynooth
Co. Kildare, Ireland
Email: Vijay.Subramanian@nuim.ie

***Abstract*—In this paper, we establish a many-sources large deviations principle (LDP) for the stationary workload of a multi-queue single-server system with simplex capacity, operated under a stabilizing and non-idling maximum-weight scheduling policy. Assuming a many-sources sample path LDP for the arrival processes, we establish an LDP for the workload process by employing Garcia's extended contraction principle that is applicable to quasi-continuous mappings. The LDP result can be used to calculate asymptotic buffer overflow probabilities accounting for the multiplexing gain, when the arrival process is an average of i.i.d. processes. We express the rate function for the stationary workloads in term of the rate functions of the finite-horizon workloads when the arrival processes have i.i.d. increments.**

I. Introduction

In this paper, we consider a single-server multi-class discrete-time queueing system where the server is allocated to queues according to a maximum weight scheduler, which is known to be stabilizing [1]. We provide a refined analysis of the statistical performance of this policy under stochastic arrivals. In particular, with K independent queues we seek to derive the probability of buffer overflow. Specifically, for a given finite value B, we consider the transient behavior, i.e., quantities such $P(W_{0,T} \geq B\mathbf{1}_K)$ where $W_{0,T} \in \mathbb{R}_+^K$ is the 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, as well as the stationary behavior, i.e., the similar probabilistic quantities as before for the limiting workload vector as $T \to \infty$. Like many recent papers on analysis of scheduling algorithms [2]–[8], our work considers logarithmic asymptotics to the probabilities by analyzing a large-deviation approximation to the problem. The present paper is closely related to [6], where the buffer overflow probability for the workload processes of a single-server multi-queue queueing system under max-weight policies and general compact and convex capacity regions was established. While [6] addresses the large-buffer scaling regime, this paper establishes similar results for a classical multi-class single-server (simplex capacity region) system under a "many-sources" asymptotic regime (see [7]–[14]).

In a many-sources asymptotic regime, one considers a sequence of queueing systems indexed by the number of the (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 $L \to \infty$. The motivation to consider many-sources scaling includes the following considerations: 1) practical interest in real applications when there are large number of flows to each user or node. This asymptote usually gives a more refined approximation to the probabilistic quantities of interest by incorporating the impact of the multiplexing gain [9]–[12], [15]–[17]; and 2) a cross-layer optimization for the optimal duration of the finite code blocks when the transmission channel is operated at high-SNR regime (see [18]).

Given a sample path large deviation principle for the arrival processes (in the space of real-valued sequences with the scaled uniform topology), we derive a large deviations principle for the workload. In particular, we first show that the workload is a quasi-continuous map of the arrival process. The first contribution of the paper is, thus, obtained based on a recent extension of the contraction principle by J. Garcia [19]. More precisely, we use Garcia's extended contraction principle together with an assumed sample path large deviations principle (LDP) (see Definition 2) for the arrival process to establish an LDP for the workload at any given time t as well as the stationary workload. The LDP results (Theorems 1 and 2) directly imply that the probability of buffer overflow has an exponential tail whose decay rate is dictated by a good rate function whose form is 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. When the arrival process has i.i.d. increments, we provide a simplified form for the rate function.

The outline of the paper is as follows. The problem formulation is given in Section II. Section III provides

*This work was supported in part by the Center for Wireless Communications, UCSD and UC Discovery Grant No. Com04-10176, ARO-MURI Grant No. W911NF-04-1-0224, NSF CAREER Award No. CNS-0533035, and AFOSR Grant No. FA9550-05-01-0430.

†This work was supported by SFI (Science Foundation of Ireland) grant 07/IN.1/I901.

background and preliminary results on the large deviations principle. The main results of the paper, which are the LDPs of the workloads, are given in Section IV. Section V gives simplified expressions of the rate functions. We conclude in Section VI with a discussion of future work.

II. Problem Formulation

We consider a discrete-time queueing system with K independent queues and one server with capacity c (bits per timeslot). For every queue $k \in \mathcal{K} := \{1, \ldots, K\}$ we assume that work (in bits) arrives into the queue given by a sequence $(A_t^k, t \in \mathbb{N})$ where $A_t^k \in \mathbb{R}_+$ is the work brought in at time $-t$. For $0 \leq m_1 \leq m_2$ integers, we define $A^k(m_1, m_2] := \sum_{t=m_1+1}^{m_2} A_t^k$ as the total amount of work to arrive for user k from timeslot $-m_2$ and until timeslot $-m_1 - 1$. We also write $A^k|_{(m_1,m_2]}$ to denote the finite sequence of arrivals A^k restricted to $\{-m_2, \ldots, -m_1 - 1\}$.

We assume a maximum-weight server allocation policy where the weights are functions of the unfinished workloads, and under which we are interested in the statistical properties of the unfinished workload in queue k at time t. Let $W_t^k \in \mathbb{R}_+$ be the unfinished workload (queue length) of queue k at the beginning of time $-t$ and R_t^k be the amount of service allocated to queue k during time $(-t, -t+1]$. Let $W_t := (W_t^k, k \in \mathcal{K})$ be the corresponding workload vector and $R_t := (R_t^k, k \in \mathcal{K})$ be the rate vector. One can define a simplex rate region $\mathcal{R}$,

$$\mathcal{R} := \left\{ \mathbf{r} = (r^1, \ldots, r^K) \in \mathbb{R}_+^K : \sum_{k=1}^{K} r^k \leq c \right\}, \quad (1)$$

as the set of server's operating points, i.e., $R_t \in \mathcal{R}$. At the beginning of timeslot $-t$, the rate vector $R_t \in \mathcal{R}$ is selected by a work-conserving max-weight scheduler H in response to the current workload W_t; that is, $R_t = H(W_t)$ where the scheduler H serves c bits from the queue k^* which has the largest workload W_t^k when the workload of the longest queue is at least c. In case of a tie, the scheduler chooses the queue with the lowest index. To make the scheduler non-idling, we assume the scheduler splits the service when the unfinished workload in each queue is less than c. That is, we assume that the scheduler assigns $H(\mathbf{x}) = \text{Proj}_{\mathcal{R}}(\mathbf{x})$ when $\mathbf{x} \in [0, c)^K$, where $\text{Proj}_B(\mathbf{b})$ is the projection of vector $\mathbf{b}$ on the set B. Specifically, for $\mathbf{x} \in \mathbb{R}_+^K$ we consider $H(\mathbf{x})$ to be given by

$$H(\mathbf{x}) := \begin{cases} \mathbf{e}(\mathbf{x}) & \text{if } \mathbf{x} \notin [0, c)^K; \\ \text{Proj}_{\mathcal{R}}(\mathbf{x}) & \text{if } \mathbf{x} \in [0, c)^K. \end{cases} \quad (2)$$

Above $\mathbf{e}(\mathbf{x})$ is defined as the K-dimensional vector whose elements are zeros except for the k^*th element which is c, where $k^* = \min\{k : k \in \arg\max_{i \in \mathcal{K}} x_i\}$. For example, when $K = 2$, the scheduler H in (2) becomes

$$H(\mathbf{x}) = \begin{cases} (c, 0), & \text{if } x^1 \geq x^2, x^1 \geq c \\ (0, c), & \text{if } x^1 < x^2, x^2 \geq c, \\ \text{Proj}_{\mathcal{R}}(\mathbf{x}), & \text{if } x^1 < c, x^2 < c. \end{cases} \quad (3)$$

For $t \in \mathbb{N}$, the dynamics of the workloads of queue $k \in \mathcal{K}$ is

$$W_{t-1}^k = [W_t^k - R_t^k]^+ + A_t^k, \quad (4)$$

where for $x \in \mathbb{R}$, $[x]^+ := \max\{0, x\}$. We assume that the arrival vector A_t happens any time in $(-t, -t+1)$ but cannot be served in that timeslot t.

In this paper, we are interested in the asymptotic probabilities of 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 $W_T \in \mathcal{R}$. The index T in $W_{0,T}$ reminds us of this initial condition.[1] The infinite-horizon workload, $\mathcal{W}$, is defined as $\mathcal{W} = \mathcal{W}(A) := \lim_{T \to \infty} W_{0,T}(A|_{(0,T]})$. We assume that the limit exists but may be infinite. It can be shown that $\mathcal{W}$ is the stationary workload when the system is stable. We will use the function G_T to mean $G_T(A|_{(0,T]}) = W_{0,T}(A|_{(0,T]})$ and the function G to mean $G(A) = \mathcal{W}(A)$. To aid in describing our results we further define $G_T^{\mathbf{a}}$ and $G^{\mathbf{a}}$ in the following way:

Definition 1: For a function $F : \mathcal{X} \mapsto \mathcal{Y}$ and $x \in \mathcal{X}$, we define

$$F^x := \{y \in \mathcal{Y} : (\exists x_n \to x) \text{ such that } F(x_n) \to y\}. \quad (5)$$

Note that $F^{(\cdot)}$ is a set-valued mapping. It is single-valued at x where F is continuous (i.e., $F^x = \{F(x)\}$).

We consider a sequence of queueing systems indexed by $L \in \mathbb{N}$ and will be interested in the behavior of the queueing system L as L becomes large. For each user $k \in \mathcal{K}$ and system indexed by L, we assume a stationary arrival process of work brought into the system given by a sequence $A^{k,L} := (A_t^{k,L}, t \in \mathbb{N})$ where $A_t^{k,L} \in \mathbb{R}_+$ is the work (in bits) brought in at time $-t$ into the queue of user k. The arrivals to different queues/users are mutually independent. We follow the many-sources scaling regime on the system with index L. The arrival process to each queue k is assumed to be an average of L i.i.d. processes, i.e., $A^{k,L} := \frac{1}{L} \sum_{i=1}^{L} A^{k,(i)}$, where each $A^{k,(i)}$ is an independent identically distributed copy of a stationary process A. We denote the mean arrival rate by $\mu := EA_1^{k,L} = EA_1$. Also let $A^L := (A^{k,L}, k \in \mathcal{K})$ be the sequence of arrival vectors.

A. Main Results

Assuming that the sequence of the arrival processes $\{A^L\}$ satisfies a many-sources sample path LDP with a continuous rate function (Assumptions 1 and 2, respectively, given in Section III), the main results of the paper are the following LDP's for the finite and infinite-horizon workloads. We also provide a simplification of the rate functions when the arrival processes have i.i.d. increments.

[1] The initial condition is normally taken to be the zero vector but the result remains valid even when the initial condition is within $\mathcal{R}$. With $W_T \in \mathcal{R}$, we always have the workload at time $-T+1$ be $W_{T-1} = [W_T - H(W_T)]^+ + A_T = A_T$ from the non-idling condition that we imposed on the server allocation mechanism.

Theorem 1: For $t \in \mathbb{N}$, the sequence of the finite-horizon workloads $\{W_{0,t}(A^L|_{(0,t]}) := G_t(A^L|_{(0,t]})\}$ satisfies an LDP on $\mathbb{R}_+^K$ with the rate function I_t, where for $\mathbf{b} \in \mathbb{R}_+^K$

$$I_t(\mathbf{b}) = \inf_{\mathbf{x} \in \mathbb{R}_+^{K \times t} : G_t^{\mathbf{x}} \ni \mathbf{b}} I^\sharp{}_t(\mathbf{x}) \tag{6}$$

Theorem 2: If $K\mu < c$, the sequence of infinite-horizon workloads $\{\mathcal{W}(A^L) := G(A^L)\}$ satisfies an LDP on $\mathbb{R}_+^K$ with rate function J, where for $\mathbf{b} \in \mathbb{R}_+^K$

$$J(\mathbf{b}) = \inf_{a \in \mathcal{D}_\mu^K : G^a \ni \mathbf{b}} I^\sharp(a). \tag{7}$$

In the above results, a denotes a sequence taking values in $\mathbb{R}_+^K$ and $\mathcal{D}_\mu^K$ is a special subset of sequences taking values in $\mathbb{R}_+^K$ which will be clarified in Section III-A .

III. Background and Assumptions

A. Topology for Sample Paths

Since a large deviations principle is defined with 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 topology as in [13] for our analysis. Let $\mathcal{D}$ denote the space of sample paths (non-negative discrete-time functions), i.e., $\mathcal{D} := \{x : \mathbb{N} \mapsto \mathbb{R}_+\}$, and let $\mathcal{D}^K$ be the K cartesian product of $\mathcal{D}$. Let $||\cdot||_u$ be the scaled uniform norm on $\mathcal{D}$, i.e., $||x||_u := \sup_{t \in \mathbb{N}} \left|\frac{x(0,t]}{t}\right|$ for all $x \in \mathcal{D}$ while for all $a = (a^k, k \in \mathcal{K}) \in \mathcal{D}^K$, where $a^k \in \mathcal{D}$, the scaled uniform norm of a is $||a||_u := \max_{k \in \mathcal{K}} ||a^k||_u$. Define a subspace $\mathcal{D}_\mu$ of $\mathcal{D}$ which contains all the arrival paths whose average arrival rate is equal to the expected rate μ, i.e., $\mathcal{D}_\mu := \left\{x \in \mathcal{D} : \lim_{t \to \infty} \frac{x(0,t]}{t} = \mu\right\}$ and $\mathcal{D}_\mu^K$ the K products of $\mathcal{D}_\mu$. Again, we equip $\mathcal{D}_\mu$ and $\mathcal{D}_\mu^K$ with the scaled uniform topology. For metric spaces like $\mathbb{R}_+^n$, $n \in \mathbb{N}$, we use the square uniform topology with the square metric ρ [20], where $\rho(\mathbf{x}, \mathbf{y}) := \max_{i \in \{1,\ldots,n\}} |x_i - y_i|$.

B. Large Deviations Principle

The following definition of a large deviations principle is taken from [13]. For an excellent full introduction to the theory, definitions, and tools, see [21] and for queueing applications, see [14].

Definition 2 (Large deviations principle): A sequence of random variables X^L in a Hausdorff space $\mathcal{X}$ with σ-algebra $\mathcal{B}$ is said to satisfy a large deviations principle (LDP) with good rate function I if, for any $B \in \mathcal{B}$,

$$\begin{aligned} -\inf_{x \in B^o} I(x) &\le \liminf_{L \to \infty} \frac{1}{L} \log P(X^L \in B) \\ &\le \limsup_{L \to \infty} \frac{1}{L} \log P(X^L \in B) \le -\inf_{x \in \bar{B}} I(x), \end{aligned} \tag{8}$$

where B^o and $\bar{B}$ are the interior and the closure of B, respectively, and the rate function $I : \mathcal{X} \mapsto \mathbb{R}_+ \cup \{\infty\}$ has compact level sets, where the level sets are defined as $\{x : I(x) \le \alpha\}$, for $\alpha \in \mathbb{R}$.

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

We are interested in finding an LDP for the sequence of the workloads $\mathcal{W}(A^L)$ and $W_{0,T}(A^L|_{(0,T]})$, assuming the following sample path LDP of the arrival processes A^L.

C. 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 $\mathcal{D}_\mu^K$ equipped with the scaled uniform topology with rate function $I^\sharp$, where the rate function $I^\sharp$ is given as

$$I^\sharp(a) := \sup_{t \in \mathbb{N}} I^\sharp{}_t(a|_{(0,t]}) = \lim_{t \to \infty} I^\sharp{}_t(a|_{(0,t]}) \tag{9}$$

for $a \in \mathcal{D}_\mu^K$, where for $\mathbf{x} = (\mathbf{x}^k \in \mathbb{R}_+^t, k \in \mathcal{K}) \in \mathbb{R}_+^{Kt}$,

$$I^\sharp{}_t(\mathbf{x}) := \sum_{k=1}^K \Lambda_t^*(\mathbf{x}^k), \tag{10}$$

and Λ_t^* is the convex conjugate or Fenchel-Legendre transform of Λ_t:

$$\Lambda_t^*(\mathbf{y}) := \sup_{\theta \in \mathbb{R}^t} \theta \cdot \mathbf{y} - \Lambda_t(\theta), \qquad \text{for } \mathbf{y} \in \mathbb{R}^t, \tag{11}$$

$$\Lambda_t(\theta) := \log E \exp\left(\theta \cdot A|_{(0,t]}\right), \qquad \text{for } \theta \in \mathbb{R}^t. \tag{12}$$

Remark 1: Assumption 1 implies that the sequence $\{A^L\}$ also satisfies an LDP on $\mathcal{D}^K$ equipped with the scaled uniform topology, with rate function $I^\sharp$ where $I^\sharp(a) = \infty$ for $a \in \mathcal{D}^K / \mathcal{D}_\mu^K$ [14]. It is shown in [14, Lemma 7.8] that under Assumption 1, $\Lambda_t^*(\cdot)$ is non-negative, Λ_t^* is convex, and $\Lambda_t^*(\mu \mathbf{1}_t) = 0$, where $\mathbf{1}_n$ is the vector of all ones in $\mathbb{R}^n$. Hence, $I^\sharp{}_t(\mu \mathbf{1}_{Kt}) = 0$ and $I^\sharp{}_t$ is convex.

In this paper, we also assume the following continuity condition on the rate function $I^\sharp$ in (9):

Assumption 2: $I^\sharp$ is continuous on its effective domain defined as $\{x \in \mathcal{D}^K : I^\sharp(x) < \infty\}$.

Remark 2: As shown in [13] and [14], the above many-sources sample path LDP (Assumption 1) holds when the underlying arrival process A satisfies mild regularity conditions. This implies that several standard stationary processes used for traffic modeling, such as i.i.d. increment processes, Markov-modulated, a general class of Gaussian, and fractional Brownian processes (for long-range dependent or heavy-tailed traffic), satisfy Assumptions 1 and 2.

D. Garcia's Extended Contraction Principle

The contraction principle (see [21, 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, the usual contraction principle fails to provide sufficient structure. Instead, we will utilize the following powerful extension of the contraction principle

for quasi-continuous transformations on metric spaces, given by Garcia [19]. First, let us provide the definition of quasi-continuity on metric spaces:

Fact 1: [19, Theorem 3.2] If $\mathcal{X}, \mathcal{Y}$ are complete metric spaces, a function $F : \mathcal{X} \mapsto \mathcal{Y}$ is quasi-continuous if and only if for each $x \in \mathcal{X}$, 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.

Remark 3: 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. But if $F(0) = 1/2$, then F is not quasi-continuous. From this example, we can infer that our scheduling function H is quasi-continuous.

Remark 4: An interesting property is that if F is a continuous function and G is a quasi-continuous function, then $F \circ G$ is quasi-continuous but $G \circ F$ is not necessarily quasi-continuous [19].

Fact 2 (Garcia's Extended Contraction Principle): Assume $\Omega \xrightarrow{X^L} \mathcal{X} \xrightarrow{F} \mathcal{Y}$, $\mathcal{X}, \mathcal{Y}$ are metric spaces, and $\{X^L\}$ satisfies a large deviation principle with good rate function $I^\sharp$. If at every x with $I^\sharp(x) < \infty$, F is quasi-continuous and $I^\sharp$ is continuous, then $\{F(X^L)\}$ satisfies the LDP with rate function given by

$$I(y) = \inf \left\{ I^\sharp(x) : y \in F^x \right\}. \tag{13}$$

Hence, given Assumption 2, the LDP's for the sequences of finite- and infinite-horizon workloads would follow as a direct consequence of the quasi-continuity of the mappings G_t and G. The quasi-continuity of the workload mappings is inherited from the quasi-continuity of the scheduler H.

IV. Analysis: LDP's for Workloads

In this section, we present the main result of the paper: LDP's for the sequences of the finite- and infinite-horizon workloads. We first establish an LDP 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 $\{W^L_{0,t} := G_t(A^L|_{(0,t]})\}$. The approach is to first show that the mapping $G_t : \mathbb{R}^{K \times t}_+ \mapsto \mathbb{R}^K_+$ is quasi-continuous, then use Garcia's extended contraction principle to obtain an LDP for the finite-horizon workloads from the LDP assumption for $\{A^L|_{(0,t]}\}$.

Lemma 1: For $t \in \mathbb{N}$, G_t is quasi-continuous on $\mathbb{R}^{K \times t}_+$ with respect to the uniform topology.

Proof: See Appendix. The idea of the proof relies on 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 $s \in (0, t-1]$. ■

Now, as already discussed, the proof of Theorem 1 is complete. We refer to the corresponding rate function, I_t, as the finite-horizon rate function. Next, we discuss the LDP for the infinite-horizon workloads.

B. LDP for Infinite-Horizon Workloads

In this section, we establish an LDP of the sequence of the infinite-horizon workloads $\{\mathcal{W}^L = G(A^L)\}$ where $A^L \in \mathcal{D}^K$. Similar to the last section, we first show that the mapping G is quasi-continuous on $\mathcal{D}^K_\mu$ when $K\mu < c$, and then use Garcia's extended contraction principle to establish the desired LDP.

Lemma 2: If $K\mu < c$, the mapping G is quasi-continuous on $\mathcal{D}^K_\mu$ with respect to the scaled uniform topology.

Proof: See Appendix. The main idea is to use the fact that the sum (over all queues) workload process behaves like that of a single queue. ■

Again, the above lemma and Garcia's extended contraction principle to the sequence of $\{A^L\}$ immediately give the LDP for the sequence of the infinite-horizon workload in Theorem 2. Recall that the set $\mathcal{D}^K_\mu$ contains all arrival sample paths a such that $I^\sharp(a) < \infty$ and $E[a^k_t] = \mu$ for all $k \in \mathcal{K}$ and $t \in \mathbb{N}$.

Let us now consider the problem of calculating the rate function. Eqn. (7) suggests that the rate function J, where $J(\mathbf{b}) = \inf_{a \in \mathcal{D}^K_\mu : G^a \ni \mathbf{b}} I^\sharp(a)$, could be interpreted as the minimum-cost solution among all paths $a \in \mathcal{D}^K_\mu$ such that $\mathbf{b} \in G^a$, where the cost of the path a is $I^\sharp(a)$ and convex. Hence, the problem of finding the rate functions is a deterministic optimal control problem like those in [4], [6].

The expressions for the rate functions I_t and J in (6) and (7) are of little use in their current forms, as their computation is far from straight forward. In the next section, we simplify the rate functions when the arrival processes are limited to having i.i.d. increments.

V. 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. In this case, the cost of a sample path $a \in \mathcal{D}^K$, which is $I^\sharp(a)$, 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 [18]). For these i.i.d. increment arrival processes, it is easy to show that for $\mathbf{x} \in \mathbb{R}^t_+$, $\Lambda^*_t(\mathbf{x}) = \sum_{i=1}^t \Lambda^*(x_i)$, where Λ^* is the Fenchel-Legendre transform of Λ and $\Lambda(\theta) = \log E \exp(\theta A_1)$ [14]. Hence, for a finite vector $\mathbf{a} = (a^k_i, k \in \mathcal{K}, i \in (0,t]) \in \mathbb{R}^{K \times t}_+$, the cost $I^\sharp{}_t(\mathbf{a})$ in (10) can be written as

$$I^\sharp{}_t(\mathbf{a}) = \sum_{i=1}^t \mathcal{X}^A(\mathbf{a}_i), \tag{14}$$

where we define $\mathcal{X}^A(\mathbf{x}) := \sum_{k=1}^K \Lambda^*(x^k)$, for $\mathbf{x} \in \mathbb{R}^K_+$, as the per-timeslot cost of a K-dimensional sample path. Next, we simplify the rate functions for the infinite-horizon and finite-horizon workloads, respectively.

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 3: For i.i.d. increment arrivals and $K\mu < c$, the infinite-horizon rate function J is simplified as

$$J(\mathbf{b}) = \inf_{t\geq 1} I_t(\mathbf{b}). \tag{15}$$

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 2, for $a \in \mathcal{D}_\mu^K$ where $K\mu < c$, we can find $t := s^*(a)$ such that $W_t(a) \in \mathcal{R}$. Hence, for a such that $\mathbf{b} \in G^a$, one can reduce the cost of the path by setting $a_v = \mu$ for all $v > t$ while keeping $G^a \ni \mathbf{b}$. This is because $\mathcal{X}^A(\mu \mathbf{1}_K) = 0$ and implies that $I^\sharp(a) = I^\sharp{}_t(a|_{(0,t]})$. On the other hand, since $W_t(a) \in \mathcal{R}$, we can write $\mathbf{b} \in G_t^{a|_{(0,t]}}$. All of these imply that

$$\begin{aligned} J(\mathbf{b}) &= \inf_{a\in\mathcal{D}_\mu^K : G^a \ni \mathbf{b}} I^\sharp(a) \\ &= \inf_{t\geq 1} \inf_{\mathbf{x}\in\mathbb{R}_+^{Kt} : G_t^{\mathbf{x}} \ni \mathbf{b}} I^\sharp{}_t(\mathbf{x}) = \inf_{t\geq 1} I_t(\mathbf{b}), \end{aligned}$$

by the definition of $I_t(\mathbf{b})$ in (6). ■

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

B. Finite-Horizon Rate Function

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

Lemma 4: For $t \in \mathbb{N}$, the finite-horizon rate function I_t is simplified as

$$I_t(\mathbf{b}) = \min_{u\in(0,t]} \inf_{\mathbf{x}\in\mathbb{A}(u,\mathbf{b})} I^\sharp{}_u(\mathbf{x}) \tag{16}$$

for $\mathbf{b} \in \mathbb{R}_+^K$, where

$$\begin{aligned} \mathbb{A}(u,\mathbf{b}) := \{a \in \mathbb{R}_+^{K\times u} : \mathbf{b} \in G_u^a, \\ G_{u-v}(a|_{(v,u]}) \notin \mathcal{R}, \forall v \in [1, u-1]\}. \end{aligned} \tag{17}$$

Proof: This follows the idea from the proof of Lemma 3. Let $t \in \mathbb{N}$. For $a \in \mathbb{R}_+^{K\times t}$ such that $\mathbf{b} \in G_t^a$, we let $u = \min\left\{t, \min\{s \in [1, t-1] : W_s = G_{t-s}(a|_{(s,t]}) \in \mathcal{R}\}\right\}$. In other words, $-u$ is the last time the workload vector is inside the capacity region $\mathcal{R}$ before time 0. By definition of I_t, we already know that the workload vector starts initially inside $\mathcal{R}$ at time $-t$. With this definition of u, we have $W_v \notin \mathcal{R}$ for all $v \in [1, u-1]$. We can find another path $\tilde{a} \in \mathbb{R}_+^{Kt}$ with a reduced cost while keeping the workloads at time $-u+1$ to 0 (i.e., W_{u-1} to W_0) intact by setting $\tilde{a}_v = \mu \mathbf{1}_K, \forall v \in (u, t]$ and $\tilde{a}_v = a_v$ otherwise. Since $\mathcal{X}^A(\mu\mathbf{1}_K) = 0$, we have $I^\sharp{}_t(a) \geq I^\sharp{}_t(\tilde{a}) = I^\sharp{}_u(a|_{(0,u]})$ and yet $\mathbf{b} \in G_u^{\tilde{a}|_{(0,u]}} = G_t^{\tilde{a}}$. Since by definition $W_v = G_{u-v}(a|_{(v,u]})$ for $v \in [1, u-1]$, we have

$$\begin{aligned} I_t(\mathbf{b}) &= \inf_{\mathbf{x}\in\mathbb{R}_+^{Kt} : G_t^{\mathbf{x}} \ni \mathbf{b}} I^\sharp{}_t(\mathbf{x}) \\ &= \min_{u\in(0,t]} \inf_{\mathbf{x}\in\mathbb{R}_+^{Ku} : \mathbf{b}\in G^{\mathbf{x}}, G_{u-v}(\mathbf{x}|_{(v,u]})\notin\mathcal{R}} I^\sharp{}_u(\mathbf{x}) \\ &= \min_{u\in(0,t]} \inf_{\mathbf{x}\in\mathbb{A}(u,\mathbf{b})} I^\sharp{}_u(\mathbf{x}), \end{aligned}$$

where $\mathbb{A}(u, \mathbf{b})$ is defined as in (17). ■

Remark 5: The above lemma reduces the set of feasible sample paths to the set $\mathbb{A}(u, \mathbf{b})$ for $u \in (0, t]$. It is interesting to note the property of the sample paths in this set. For any $\mathbf{x} \in \mathbb{A}(u, \mathbf{b})$, we have $\hat{W}_0(\mathbf{x}) = \hat{\mathbf{x}}(0, u] - c(u-1) = \hat{\mathbf{b}}$, recalling that the $\hat{}$ notation is the sum over queues. There is no wastage of service capacity over the $u-1$ timeslots because $\forall v \in [1, u-1]$, $W_v = G_{u-v}(\mathbf{x}|_{(v,u]}) \notin \mathcal{R}$ and hence $\hat{W}_v > c$. That is, any sample path $\mathbf{x} \in \mathbb{A}(u, \mathbf{b})$ has its sum of the arrivals over time $(0, u]$ and queues equal to $\hat{\mathbf{x}}(0, u] = \hat{\mathbf{b}} + c(u-1)$.

In addition, an immediate implication of Lemma 4 is that we can rewrite J in (7) as

$$\begin{aligned} J(\mathbf{b}) &= \inf_{t\geq 1} I_t(\mathbf{b}) \\ &= \inf_{t\geq 1} \min_{u\in(0,t]} \inf_{\mathbf{x}\in\mathbb{A}(u,\mathbf{b})} I^\sharp{}_u(\mathbf{x}) \\ &= \inf_{t\geq 1} \inf_{\mathbf{x}\in\mathbb{A}(t,\mathbf{b})} I^\sharp{}_t(\mathbf{x}). \end{aligned} \tag{18}$$

If we denote t^* as the optimizer of the last equation, then t^* is called the *critical timescale* (see [13]). It can be interpreted that t^* is the length of time which the buffers are most likely to take to fill from "empty" level (more precisely, anywhere within $\mathcal{R}$) to a given level $\mathbf{b}$.

Note that for fixed $u \in \mathbb{N}$, $\inf_{\mathbf{x}\in\mathbb{A}(u,\mathbf{b})} I^\sharp{}_u(\mathbf{x})$ is a optimization problem, with a convex cost function $I^\sharp{}_u(\cdot)$ and a set of mixed discrete and continuous feasible solutions $\mathbb{A}(u, \mathbf{b})$. This problem is difficult to solve analytically. However, in [22] we employ the additivity and convexity of the rate function $I^\sharp{}_t$ to further provide some simplified bounds of the rate functions.

VI. Conclusion

In this paper, we have established a many-sources LDP for the stationary (infinite-horizon) workload for multi-queue single-server system with simplex capacity, operated under the maximum-weight scheduling with the arrival processes assumed to satisfy a 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. We gave the associated rate functions and the expression of the infinite-horizon rate function in term of the finite-horizon ones, when the arrivals processes have i.i.d. increments.

Note that the quasi-continuity of the finite-horizon workload mapping and hence the LDP for the sequence of finite-horizon workload processes is valid even when the rate region

is MAC or any convex and compact set. The main difficulty in establishing LDP for the infinite-horizon workload is in showing the quasi-continuity of the infinite-horizon workload mapping. This is an interesting area of future research.

APPENDIX

Here we prove Lemmas 1 and 2. We first show the proof of Lemma 1 which uses the following fact:

Fact 3: Assume $\mathcal{X} \overset{F,G}{\to} \mathcal{Y}$, $\mathcal{X}, \mathcal{Y}$ are metric spaces, and $x \in \mathcal{X}$. If F is quasi-continuous at x and G is continuous at x, then $F + G$ is quasi-continuous at x.

Lemma 1: For $t \in \mathbb{N}$, G_t is quasi-continuous on $\mathbb{R}_+^{K\times t}$ with respect to the uniform topology.

Proof: Using our queueing equation we first observe the following recursive relation between G_t and G_{t-1} for any $t \in \{2, 3, \ldots\}$ and $\mathbf{x} = \mathbf{x}_{(0,t]} \in \mathbb{R}_+^{K\times t}$:

$$G_t(\mathbf{x}_{(0,t]}) = [G_{t-1}(\mathbf{x}_{(1,t]}) - H(G_{t-1}(\mathbf{x}_{(1,t]}))]^+ + \mathbf{x}_1, \quad (19)$$

where we used the fact that $W_0(\mathbf{x}_{(0,t]}) = G_t(\mathbf{x}_{(0,t]})$, and $W_1(\mathbf{x}_{(1,t]}) = G_{t-1}(\mathbf{x}_{(1,t]})$ when the initial backlog at time $-t$ is small, i.e., $W_t \in \mathcal{R}$.

Equation (19) says that $G_t(\mathbf{x}_{(0,t]})$ depends linearly on $\mathbf{x}_1$. This implies the following simple but consequential observations:

Observation 1: If G_t is strictly quasi-continuous at $\mathbf{x}_{(0,t]}$, then it is strictly quasi-continuous at $\tilde{\mathbf{x}}_{(0,t]} := (\tilde{\mathbf{x}}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t)$ for any $\tilde{\mathbf{x}}_1 \in \mathbb{R}_+^K$. If G_t is continuous at $\mathbf{x}_{(0,t]}$, then it is also continuous at $\tilde{\mathbf{x}}_{(0,t]}$.

Observation 2: If $G_t(\mathbf{x}^n_{(0,t]}) \to G_t(\mathbf{x}_{(0,t]})$ for a sequence $\{\mathbf{x}^n_{(0,t]}\}$ such that $\mathbf{x}^n_{(0,t]} \to \mathbf{x}_{(0,t]}$, then for any sequence $\{\tilde{\mathbf{x}}^n_{(0,t]} = (\tilde{\mathbf{x}}^n_1, \mathbf{x}^n_2, \ldots, \mathbf{x}^n_t)\}$ where $\tilde{\mathbf{x}}^n_1 \to \mathbf{x}_1$, we also have $G_t(\tilde{\mathbf{x}}^n_{(0,t]}) \to G_t(\mathbf{x}_{(0,t]})$.

Using the recursive relation in (19), we prove this lemma by induction on $t \in \mathbb{N}$. For $t = 1$, $G_1(\mathbf{a}_1) = \mathbf{a}_1$, hence G_1 is continuous on $\mathbb{R}_+^K$. Assuming that G_t is quasi-continuous on $\mathbb{R}_+^{Kt}$, we want to show that G_{t+1} is quasi-continuous on $\mathbb{R}_+^{K(t+1)}$. Using the fact that the $[\cdot]^+$ function is continuous, Remark 4, and Fact 3, it suffices to show that the function $F_t := G_t - H \circ G_t$, is quasi-continuous on $\mathbb{R}_+^{Kt}$ to show that G_{t+1} is quasi-continuous. In particular, for any arrival sample path $\mathbf{a} = \mathbf{a}_{(0,t]} \in \mathbb{R}_+^{K\times t}$, we need to show that F_t is quasi-continuous at $\mathbf{a}_{(0,t]}$ with respect to the uniform topology. It suffices to show that it is possible to select a sequence $\hat{\mathbf{a}}^n \to \mathbf{a}$ for which

$$G_t(\hat{\mathbf{a}}^n_{(0,t]}) \quad \to G_t(\mathbf{a}_{(0,t]}), \quad (20)$$

$$H \circ G_t(\hat{\mathbf{a}}^n_{(0,t]}) \quad \to H \circ G_t(\mathbf{a}_{(0,t]}), \quad (21)$$

such that both $G_t(\cdot)$ and $H \circ G_t(\cdot)$ are continuous at every $\hat{\mathbf{a}}^n_{(0,t]}$.

We show this by first noting that the induction hypothesis, i.e., quasi-continuity of G_t, and the definition of quasi-continuity ensure that there exists a sequence $\{\mathbf{a}^n_{(0,t]}\}$ such that $\mathbf{a}^n_{(0,t]} \to \mathbf{a}$, in the uniform topology, such that $G_t(\mathbf{a}^n_{(0,t]}) \to G_t(\mathbf{a}_{(0,t]})$, and $G_t(\cdot)$ is continuous at $\mathbf{a}^n_{(0,t]}$ for all n. We will construct the desired sequence $\{\hat{\mathbf{a}}^n_{(0,t]}\}$ from this sequence $\{\mathbf{a}^n_{(0,t]}\}$. We proceed by considering the following two cases, depending on the value of $\mathbf{a}_1$.

Case 1: $\mathbf{a}_1 > \mathbf{0}$, i.e., every component of the $\mathbf{a}_1 \in \mathbb{R}^K$ is positive. Let $\epsilon > 0$ be the smallest component of $\mathbf{a}_1$. Since $H(\cdot)$ is quasi-continuous, it is possible to choose a sequence of (workload) vectors $\{w^n\}$ such that $w^n \to G_t(\mathbf{a}_{(0,t]})$, and H is continuous at w^n for all n. Now, we define

$$\begin{aligned}
\tilde{\mathbf{a}}^n_1 &:= w^n - [F_{t-1}(\mathbf{a}^n_{(1,t]})]^+ && (22)\\
&= w^n - G_t(\mathbf{a}^n_{(0,t]}) + \mathbf{a}^n_1 && (23)\\
&= \big(w^n - G_t(\mathbf{a}_{(0,t]})\big) + \Big(G_t(\mathbf{a}_{(0,t]}) - G_t(\mathbf{a}^n_{(0,t]})\Big)\\
&\quad + (\mathbf{a}^n_1 - \mathbf{a}_1) + \mathbf{a}_1. && (24)
\end{aligned}$$

It is clear from the last equality that $\tilde{\mathbf{a}}^n_1 \to \mathbf{a}_1$ with respect to uniform topology. We need to ensure that $\tilde{\mathbf{a}}^n_1 \geq 0$ since negative quantities are involved in the definition. We do this by using the facts that every component of $\mathbf{a}_1 \in \mathbb{R}^K$ is greater or equal to $\epsilon > 0$, and that $w^n \to G_t(\mathbf{a}_{(0,t]})$, $G_t(\mathbf{a}^n_{(0,t]}) \to G_t(\mathbf{a}_{(0,t]})$, and $\mathbf{a}^n_1 \to \mathbf{a}_1$. These facts imply that there exists an n_ϵ such that for all $n > n_\epsilon$ we have $||w^n - G_t(\mathbf{a}_{(0,t]})|| < \epsilon/3$, $||G_t(\mathbf{a}_{(0,t]}) - G_t(\mathbf{a}^n_{(0,t]})|| < \epsilon/3$ and $||\mathbf{a}^n_1 - \mathbf{a}_1|| < \epsilon/3$ (with the L_1 norm) which then together with (24) imply that, for the sequence $\tilde{\mathbf{a}}^{m+n_\epsilon}_1$, we always have non-negativity of all components. Hence, we construct a new sequence $\{\hat{\mathbf{a}}^n_{(0,t]}\}$ where $\hat{\mathbf{a}}^n_1 = \tilde{\mathbf{a}}^{n+n_\epsilon}_1$ and $\hat{\mathbf{a}}^n_{(1,t]} = \mathbf{a}^{n+n_\epsilon}_{(1,t]}$.

This new sequence $\hat{\mathbf{a}}^n_{(0,t]}$ is the sequence we are after because using the induction hypothesis together with Observations 1 and 2, we have that $G_t(\hat{\mathbf{a}}^n) \to G_t(\mathbf{a})$, and G_t is continuous at $\hat{\mathbf{a}}^n$ for all n. Furthermore, by construction

$$\begin{aligned}
G_t(\hat{\mathbf{a}}^n) &= \hat{\mathbf{a}}^n_1 + [F_{t-1}(\hat{\mathbf{a}}^n_{(1,t]})]^+\\
&= \tilde{\mathbf{a}}^{n+n_\epsilon}_1 + [F_{t-1}(\mathbf{a}^{n+n_\epsilon}_{(1,t]})]^+\\
&= w^{n+n_\epsilon}. && (25)
\end{aligned}$$

Hence, we have shown that there exists a sequence $\hat{\mathbf{a}}^n_{(0,t]}$ satisfying (20) and (21). In addition, the continuity of $H \circ G_t$ at $\hat{\mathbf{a}}^n_{(0,t]}$ for all n is a direct consequence of continuity of G_t at $\hat{\mathbf{a}}^n_{(0,t]}$ and continuity of H at w^{n+n_ϵ}, which is equal to $G_t(\hat{\mathbf{a}}^n)$, for all n.

Case 2: $\mathbf{a}_1 \geq \mathbf{0}$. Without loss of generality by permuting the user labels, we can assume that the first k components of $\mathbf{a}_1$ are 0 while the rest of the $K-k$ components are positive. Now the sequence $\mathbf{a}^m_1$ with $1/m$ in the first k components and the non-zero values of $\mathbf{a}_1$ in the remaining coefficients converges to $\mathbf{a}_1$ such that for every m every component of $\mathbf{a}^m_1$ is positive. We construct a sequence $\{\mathbf{a}^m_{(0,t]}\}$ with this $\mathbf{a}^m_1$ and $\mathbf{a}^m_{(1,t]} = \mathbf{a}_{(1,t]}$. For ease of exposition we denote the vector with $1/m$ in the first k positions and 0s in the remaining $K-k$ positions by $[1/m]_k$. It is obvious that $G_t(\mathbf{a}^m_{(0,t]}) \to G_t(\mathbf{a}_{(0,t]})$ since

$$\begin{aligned}
G_t(\mathbf{a}^m_{(0,t]}) &= \mathbf{a}^m_1 + [F_{t-1}(\mathbf{a}^m_{(1,t]})]^+\\
&= \mathbf{a}^m_1 + [F_{t-1}(\mathbf{a}_{(1,t]})]^+ = [1/m]_k + G_t(\mathbf{a}_{(0,t]}).
\end{aligned}$$

When $G_t(\mathbf{a}_{(0,t]}) \notin [0,C)^K$, for m large enough,[2] we have

$$H \circ G_t(\mathbf{a}^m_{(0,t]}) = \mathbf{e}\left(G_t(\mathbf{a}^m_{(0,t]})\right) = \mathbf{e}\left(G_t(\mathbf{a}_{(0,t]})\right) = H \circ G_t(\mathbf{a}_{(0,t]}),$$

where the function $\mathbf{e}$ is defined in the definition of H in (2). On the other hand, if $G_t(\mathbf{a}_{(0,t]}) \in [0,C)^K$, then the continuity of $\text{Proj}_{\mathcal{R}}(\cdot)$ yields $H \circ G_t(\mathbf{a}^m_{(0,t]}) \to H \circ G_t(\mathbf{a}_{(0,t]})$.

Since for each m we have that $\mathbf{a}^m_1$ has all elements strictly positive, we can use the similar construction as in Case 1 but with $\mathbf{a}^m_{(0,t]}$ in place of $\mathbf{a}_{(0,t]}$. In particular, for each m, we can now generate a sequence $\{\tilde{\mathbf{a}}^{m,n}_{(0,t]}\}$ such that $\tilde{\mathbf{a}}^{m,n}_1 \to \mathbf{a}^m_1$ as $n \to +\infty$, $\tilde{\mathbf{a}}^{m,n}_{(1,t]} = \mathbf{a}^n_{(1,t]}$, and by using Observations 1 and 2, the following hold

$$G_t(\tilde{\mathbf{a}}^{m,n}_{(0,t]}) \to G_t(\mathbf{a}^m_{(0,t]}), \tag{26}$$

$$H \circ G_t(\tilde{\mathbf{a}}^{m,n}_{(0,t]}) \to H \circ G_t(\mathbf{a}^m_{(0,t]}), \tag{27}$$

with both $G_t(\cdot)$ and $H \circ G_t(\cdot)$ being continuous at $\tilde{\mathbf{a}}^{m,n}_{(0,t]}$ for all n.

Now we define the sequence $\hat{\mathbf{a}}^m_{(0,t]} = \tilde{\mathbf{a}}^{m,m}_{(0,t]}$ as the sequence we are after. By construction, we have $\hat{\mathbf{a}}^m_{(0,t]} \to \mathbf{a}_{(0,t]}$ and both $G_t(\cdot)$ and $H \circ G_t(\cdot)$ continuous at all $\hat{\mathbf{a}}^m_{(0,t]}$. Since $G_t(\mathbf{a}^m_{(0,t]}) \to G_t(\mathbf{a}_{(0,t]})$ and $H \circ G_t(\mathbf{a}^m_{(0,t]}) \to H \circ G_t(\mathbf{a}_{(0,t]})$, it follows from (26) and (27) that $G_t(\hat{\mathbf{a}}^m_{(0,t]}) \to G_t(\mathbf{a}_{(0,t]})$ and $H \circ G_t(\hat{\mathbf{a}}^m_{(0,t]}) \to H \circ G_t(\mathbf{a}_{(0,t]})$. ■

Next, we prove Lemma 2:

Lemma 2: If $K\mu < c$, the mapping G is quasi-continuous on $\mathcal{D}^K_\mu$ with respect to the scaled uniform topology.

Proof: The proof follows the concept in [13]. Let $K\mu < c$ and $A \in \mathcal{D}^K_\mu$. Consider any sequence $\{A^n\}$ such that $A^n \to A$. The main step of the proof is based on the following claim:

Claim 1: There exists a $s^* = s^*(A) < \infty$ and n'_0 such that, when $n > n'_0$, the workloads at time $-s^*$ of the arrival sample paths A^n and A stay within the rate region $\mathcal{R}$, i.e., $W_{s^*}(A^n) \in \mathcal{R}$ and $W_{s^*}(A) \in \mathcal{R}$.

With this claim and by the definition of G_{s^*}, the workloads at time zero for A^n and A are $G(A^n) = G_{s^*}(A^n|_{(0,s^*]})$ and $G(A) = G_{s^*}(A|_{(0,s^*]})$, respectively, when $n > n'_0$. In other words, we have transformed the infinite-horizon workload into the finite-horizon workload whose mapping is already known to be quasi-continuous by Lemma 1. The proof is now complete since G_{s^*} is quasi-continuous on $\mathbb{R}^{K\times s^*}_+$ and $A^n|_{(0,s^*]} \to A|_{(0,s^*]}$.

What is left is to show Claim 1. To do this, we look at the sum arrival processes and the sum workload processes and follow the proof in [13], [14] for the (aggregate) single-queue scenario. Given the definition of H and the simplex capacity region $\mathcal{R}$, the queue dynamics for the sum workload is that of a single queue whose arrivals are the sum of the arrivals, i.e.,

$$\hat{W}_{t-1} = [\hat{W}_t - c]^+ + \hat{A}_t, \tag{28}$$

[2]E.g., m being greater than the reciprocal of the maximum positive component of $G_t(\mathbf{a}_{(0,t]})$.

where we define the hat $(\hat{\cdot})$ notation to mean the sum over all users, i.e. $\hat{A}_t = \sum_{k=1}^K A^k_t$ and $\hat{W}_t = \sum_{k=1}^K W^k_t$. Recursion of the queue dynamics (28) and letting $T \to \infty$ where $W_T \in \mathcal{R}$, gives the standard expression for the stationary sum workload [14]:

$$\hat{W}_0(A) = \sup_{t\in\mathbb{N}} \hat{A}(0,t] - c(t-1). \tag{29}$$

To prove the claim we use the fact that the rate region $\mathcal{R}$ is simplex, hence $\hat{W}_s \le c \Leftrightarrow W_s \in \mathcal{R}$. That is, it suffices to show that there are a n'_0 and a finite s such that, for $n \ge n'_0$, $\hat{W}_s(A) \le c$ and $\hat{W}_s(A^n) \le c$.

Since $A^n \to A$ under the scaled uniform topology, for any given $\epsilon > 0$, there exists a n_0 such that for $n \ge n_0$, $\max_{k\in K}\sup_{t\in\mathbb{N}} |\frac{A^{n,k}(0,t]}{t} - \frac{A^k(0,t]}{t}| < \epsilon$. Hence, $\sup_t |\frac{\hat{A}^n(0,t]}{t} - \frac{\hat{A}(0,t]}{t}| < K\epsilon$. Since $A \in \mathcal{D}^K_\mu$, there is a $t_0 < \infty$ such that for $t > t_0$ and $k \in K$, $\frac{A^k(0,t]}{t} \le \mu+\epsilon$. Therefore, it follows that $\frac{\hat{A}(0,t]}{t} \le K\mu + K\epsilon$ for $t > t_0$. Since $K\mu < c$, we choose $\epsilon = (c - K\mu)/4K$. We now have that for all $n \ge n_0$ and $t \ge t_0$, $\frac{\hat{A}^n(0,t]}{t} < K(\mu+2\epsilon) = (c+K\mu)/2 < c$, and we also have that $\frac{\hat{A}(0,t]}{t} \le K(\mu+\epsilon) = (c+3K\mu)/4 < c$. In other words, for all $n \ge n_0$, the workload at time zero is a function of only the arrivals within time $(0,t_0]$ and hence,

$$\hat{W}_0(A) = \sup_{1\le t\le t_0} \hat{A}(0,t] - c(t-1),$$

$$\hat{W}_0(A^n) = \sup_{1\le t\le t_0} \hat{A}^n(0,t] - c(t-1).$$

Let $s \le t_0$ and $s^n \le t_0$ be the minimum values of the optimizing t's in the above equations, respectively. It can be shown as in [14, Lemma 5.4] that $\hat{W}_s(A) \le c$ and $\hat{W}_{s^n}(A^n) \le c$ (and in addition, $\hat{W}_v(A) > c$ and $\hat{W}_{v^n}(A^n) > c$ for all $v \in (0,s)$ and $v^n \in (0,s^n)$).

Next we show that there exists n_1 such that for $n \ge n_1$, $s^n = s$. This is not difficult because it is known that $\hat{W}_0$ is continuous on $\mathcal{D}_{K\mu}$ [13, Lemma 13]. Since $\hat{A}^n \to \hat{A}$ on $\mathcal{D}_{K\mu}$, we have $\hat{W}_0(A^n) \to \hat{W}_0(A)$ and $s^n \to s$. Since $s^n, s \in \mathbb{N}$, there exists a n_1 such that $s^n = s$ for $n \ge n_1$. The claim is now proved by taking $n'_0 = \max(n_1, n_0)$. ■

References

[1] M. Andrews, A. Stolyar, K. Kumaran, R. Vijayakumar, K. Ramanan, and P. Whiting, "Scheduling in a queuing system with asynchronously varying service rates," *Probab. Eng. Inf. Sci.*, vol. 18, pp. 191–217, 2004.

[2] D. Bertsimas, I. Paschalidis, and J. Tsitsiklis, "Asymptotic buffer overflow probabilities in multiclass multiplexers: an optimal control approach," *IEEE Trans. Autom. Control*, vol. 43, no. 3, pp. 315–335, 1998.

[3] A. L. Stolyar and K. Ramanan, "Largest weighted delay first scheduling: Large deviations and optimality," *Annals of Applied Probabilities*, vol. 11, no. 1, pp. 1–48, Feb. 2001.

[4] S. Shakkottai, "Effective Capacity and QoS for wireless scheduling," *IEEE Trans. Autom. Control*, vol. 53, no. 3, pp. 749–761, Apr. 2008.

[5] L. Ying, R. Srikant, A. Eryilmaz, and G. Dullerud, "A Large Deviations analysis of scheduling in wireless networks," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 5088–5098, Nov. 2006.

[6] V. G. Subramanian, "Large deviations of max-weight scheduling policies of convex rate regions," in *2008 ITA*, 2008.

[7] C.-W. Yang, A. Wierman, S. Shakkottai, and M. Harchol-Balter, "Tail asymptotics for policies favoring short jobs in a many-flows regime," *SIGMETRICS Perform. Eval. Rev.*, vol. 34, no. 1, pp. 97–108, 2006.
[8] S. Shakkottai and R. Srikant, "Many-sources delay asymptotics with applications to priority queues," *Queueing Systems Theory and Applications (QUESTA)*, vol. 39, pp. 183–200, Oct. 2001.
[9] A. Weiss, "A new technique for analyzing large traffic systems," *Advances in Applied Probability*, vol. 18, pp. 506–532, 1986.
[10] D. D. Botvich and N. G. Duffield, "Large deviations, the shape of the loss curve, and economies of scale in large multiplexers," *Queueing System*, vol. 20, pp. 293–320, 1995.
[11] C. Courcoubetis and R. Weber, "Buffer overflow asymptotics for a buffer handling many traffic sources," *Journal of Applied Probability*, vol. 33, pp. 886–903, 1996.
[12] A. Simonian and J. Guibert, "Large deviations approximation for fluid queues fed by a large number of on/off sources," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 1017–1027, Aug. 1995.
[13] D. J. Wischik, "Sample path large deviations for queues with many inputs," *Ann. Appl. Probab.*, 2001.
[14] A. Ganesh, N. O'Connell, and D. Wischik, *Big Queues*. Springer-Verlag, 2004.
[15] E. Buffet and N. G. Duffield, "Exponential upper bounds via martingales for multiplexers with markovian arrivals," *J. Appl. Prob.*, vol. 31, pp. 1049–1060, 1994.
[16] N. G. Duffield, "Exponential bounds for queues with markovian arrivals," *Queueing Systems*, vol. 17, pp. 413–430, 1994.
[17] G. L. Choudhury, D. M. Lucantoni, and W. Whitt, "Squeezing the most out of atm," *IEEE Trans. Comm.*, vol. 44, pp. 203–217, Feb 1996.
[18] S. Kittipiyakul, P. Elia, and T. Javidi, "High-SNR analysis of outage-limited communications with bursty and delay-limited information," *submitted for publication*, 2007.
[19] J. Garcia, "An extension of the Contraction Principle," *Journal of Theoretical Probability*, vol. 17, no. 2, pp. 403–434, Apr. 2004.
[20] J. Munkres, *Topology*, 2nd ed. Prentice Hall, 2000.
[21] A. Dembo and O. Zeitouni, *Large Deviations techniques and applications*, 2nd ed. Springer, 1998.
[22] S. Kittipiyakul, "Cross-layer optimization for transmission of delay-sensitive and bursty traffic in wireless systems," Ph.D. dissertation, University of California at San Diego, 2008.