Stochastics and Stochastics Reports, 2002 Vol. 73 (1–2), pp. 25–35

Taylor & Francis
Taylor & Francis Group

A LARGE DEVIATION PRINCIPLE WITH QUEUEING APPLICATIONS

A.J. GANESH[a,†,*] and NEIL O'CONNELL[b]

[a]*Microsoft Research, 7 JJ Thomson Ave., Cambridge CB3 OFB, UK*
[b]*BRIMS, Hewlett-Packard Labs, Filton Road, Bristol BS12 6QZ, UK*

(Revised 11 May 1999; In final form 22 July 2001)

In this paper, we present a large deviation principle for partial sums processes indexed by the half line, which is particularly suited to queueing applications. The large deviation principle is established in a topology that is finer than the topology of uniform convergence on compacts and in which the queueing map is continuous. Consequently, a large deviation principle for steady-state queue lengths can be obtained immediately via the contraction principle.

Keywords: Large deviations; Queues; Inverse contraction principle

Keywords: 60F10; 60K25

The main result in this paper provides a new tool for looking at large deviations for queueing systems in equilibrium. Equilibrium systems have generally been treated on a case-by-case basis, with much work and/or additional hypotheses necessary to prove large deviation principles (see, for example, Refs. [1,6,7,16,18,23]). We provide a simple sufficient condition for the usual sample path LDP (as in Mogulskii's theorem) to be strengthened to a topology for which the reflection mappings appearing in many queueing applications are continuous and the contraction principle can be applied. A step in this direction was made by Dobrushin and Pechersky [13], who introduce a finer topology (a gauge topology)

*Corresponding author.
[†]Research partially carried out while the author was at BRIMS, Hewlett-Packard Labs.

ISSN 1045-1129 print/ISSN 1029-0346 online
DOI: 10.1080/1045112021000004199

which allows one to treat the single server queue with constant service rate, and prove the LDP in this topology for a class of Markov jump processes. However, this does not easily extend to more complicated network configurations, or even to the single server queue with stochastic service rate. The main result in this paper can be (and has been) applied to some quite complicated multidimensional systems with interacting traffic [21,22,24].

The context in which the need for our main result arises is a general scheme which can be applied to a variety of network problems where the goal is to establish probability approximations for aspects of a system (such as queue lengths) under very general ergodicity and mixing assumptions about the network inputs.

Suppose that the inputs to a network can be represented by a sequence of random variables (X_k) in $\mathbb{R}^d$, and that the (sequence of) objects of interest, (O_n), can be expressed as a function of the partial sums process corresponding to X. To make this more precise, for $t \geq 0$ set

$$S_n(t) = \frac{1}{n} \sum_{k=1}^{[nt]} X_k, \tag{1}$$

where $[x]$ denotes the integer part of x, and write $\tilde{S}_n$ for the polygonal approximation to S_n:

$$\tilde{S}_n(t) = S_n(t) + \left(t - \frac{[nt]}{n}\right)\left(S_n\left(\frac{[nt]+1}{n}\right) - S_n\left(\frac{[nt]}{n}\right)\right). \tag{2}$$

Denote by $C(\mathbb{R}_+)$ the space of continuous functions on $\mathbb{R}_+$. Then $\tilde{S}_n \in C^d(\mathbb{R}_+)$ and our supposition is that there exists a function $f : C^d(\mathbb{R}_+) \to \mathcal{X}$, for some space $\mathcal{X}$, such that $O_n = f(\tilde{S}_n)$, for each n.

For example, suppose $d = 1$ and X_k is the difference between the amount of work arriving at time $-k$ at a single-server queue and the available service capacity at that time. Suppose also that the limit,

$$\mu := \lim_{n\to\infty} \sum_{k=1}^{n} X_k/n$$

exists almost surely and is less than 0. Then the queue length at time zero is given by

$$Q_0 = \sup_{n\geq 0} \sum_{k=0}^{n} X_k, \tag{3}$$

or, equivalently, $Q_0/n = f(\tilde{S}_n)$, where $f : C(\mathbb{R}_+) \to \mathbb{R} \bigcup \{\infty\}$ is defined by

$$f(\phi) = \sup_{t\geq 0} \phi(t). \tag{4}$$

If the sequence X_k is stationary and ergodic, then Q_0 represents the equilibrium queue-length distribution. In this example, $O_n = Q_0/n$.

The idea is to deduce a large deviation principle (see below) for O_n from one which can generally be assumed for $\tilde{S}_n$. This can be done using the *contraction principle*, which we will now describe.

Let $\mathscr{X}$ be a Hausdorff topological space with Borel σ-algebra $\mathscr{B}$, and let μ_n be a sequence of probability measures on $(\mathscr{X}, \mathscr{B})$. We say that μ_n satisfies the *large deviation principle* (LDP) with rate function I, if $I : \mathscr{X} \to \mathbb{R} \bigcup \{\infty\}$ is non-negative and lower semicontinuous, and for all $B \in \mathscr{B}$,

$$-\inf_{x \in B^0} I(x) \leq \liminf_n \frac{1}{n} \log \mu_n(B) \leq \limsup_n \frac{1}{n} \log \mu_n(B) \leq -\inf_{x \in \bar{B}} I(x); \qquad (5)$$

here, B^0 and $\bar{B}$ denote the interior and closure of B, respectively. If, for each n, Z_n is a realization of μ_n, it is sometimes convenient to say that the sequence Z_n satisfies the LDP. A rate function is *good* if its level sets are compact. The contraction principle states that if Z_n satisfies the LDP in a Hausdorff topological space $\mathscr{X}$ with good rate function I, and f is a continuous mapping from $\mathscr{X}$ into another Hausdorff topological space $\mathscr{Y}$, then the sequence $f(Z_n)$ satisfies the LDP in $\mathscr{Y}$ with good rate function given by

$$J(y) = \inf\{I(x) : f(x) = y\}.$$

Now consider the partial sums process $\tilde{S}_n$. Denote by $\tilde{S}_n[0,1]$ the restriction of $\tilde{S}_n$ to the unit interval, by $C[0,1]$ the space of continuous functions on [0,1], equipped with the uniform topology, and by $\mathscr{A}[0,1]$ the subspace of absolutely continuous functions ϕ on [0,1] with $\phi(0) = 0$. Dembo and Zajic [9] establish quite general conditions for which $\tilde{S}_n[0,1]$ satisfies the LDP in $\mathscr{A}[0,1]$ with good convex rate function given by

$$I_1(\phi) = \begin{cases} \int_0^1 \Lambda^*(\dot{\phi})\, \mathrm{d}s & \phi \in \mathscr{A}[0,1] \\ \infty & \text{otherwise,} \end{cases} \qquad (6)$$

where Λ^* is the Fenchel–Legendre transform of the scaled cumulant generating function

$$\Lambda(\lambda) = \lim_{n \to \infty} \frac{1}{n} \log E\, \mathrm{e}^{n\lambda \cdot S_n(1)}, \qquad (7)$$

which is assumed to exist for each $\lambda \in \mathbb{R}^d$ as an extended real number. For such an LDP to hold in the *i.i.d.* case, it is sufficient that the moment generating function $E\, \mathrm{e}^{\lambda \cdot X_1}$ exists and is finite everywhere; this is a classical result, due to Varadhan [25] and Mogulskii [19]. This LDP is usually extended to the space

$C(\mathbb{R}_+)$, (of continuous functions on $\mathbb{R}_+$), via the Dawson Gärtner theorem for projective limits. However, the projective limit topology (the topology of uniform convergence on compact intervals) is not strong enough for many applications; in particular, the function f defined by Eq. (4) is not continuous in this topology on any supporting subspace, and so the contraction principle does not apply. To see this, consider

$$\phi_n(t) = \begin{cases} t/n, & 0 \le t \le n, \\ 1, & t \ge n. \end{cases}$$

$\phi_n \to \phi \equiv 0$, uniformly on compact sets, but $f(\phi_n) = 1$ for all n whereas $f(\phi) = 0$.

The lack of continuity was observed by Dobrushin and Pechersky [13], who introduce a finer topology (a gauge topology) which allows one to treat the single server queue with constant service rate, and prove the LDP in this topology for a class of Markov jump processes. In this topology, the restriction of the mapping in Eq. (4) to a certain subspace of paths ϕ with limits

$$\lim_{t\to\infty} \phi(t)/t = \mu < 0,$$

is continuous. However, this does not easily extend to more complicated network configurations, or even to the single server queue with (stochastic) time-varying capacity.

We consider the set of paths

$$\mathscr{Y} = \bigcap_{j=1}^{d} \left\{ \phi \in \mathscr{C}^d(\mathbb{R}_+) : \lim_{t\to\infty} \frac{\phi^j(t)}{1+t} \text{ exists} \right\},$$

where $\phi^j(t)$ denotes the jth component of $\phi(t)$, and equip $\mathscr{Y}$ with the norm

$$\|\phi\|_u = \sup_j \sup_t \left| \frac{\phi^j(t)}{1+t} \right|.$$

Note that $\mathscr{Y}$ can be identified with the Polish space $C^d(\mathbb{R}_+^*)$ of continuous functions on the extended (and compactified) real line, equipped with the supremum norm, via the bijective mapping $\phi(t) \mapsto \phi(t)/(1+t)$. In particular, $\mathscr{Y}$ is a Polish space. We prove the following.

THEOREM 1 *Suppose that for each $\theta \in \mathbb{R}^d$, the limit*

$$\Lambda(\theta) = \lim_{n\to\infty} \frac{1}{n} \log E\, \mathrm{e}^{n\theta\cdot\tilde{S}_n(1)}, \tag{8}$$

exists as an extended real number, and the sequence $\tilde{S}_n[0,1]$ satisfies the LDP in

$C^d[0,1]$ *with good rate function given by*

$$I_1(\phi) = \begin{cases} \int_0^1 \Lambda^*(\dot{\phi})\,\mathrm{d}s & \phi \in \mathscr{A}^d[0,1], \\ \infty & \textit{otherwise} \end{cases}$$

where Λ^* *is the convex dual of* Λ. *If* Λ *is differentiable at the origin, then* $\tilde{S}_n$ *satisfies the LDP in* $\mathscr{Y}$ *with good rate function*

$$I_\infty(\phi) = \begin{cases} \int_0^\infty \Lambda^*(\dot{\phi})\,\mathrm{d}s & \phi \in \mathscr{A}^d(\mathbb{R}_+) \cap \mathscr{Y}, \\ \infty & \textit{otherwise} \end{cases}$$

Here, $\mathscr{A}^d[0,1]$ (resp. $\mathscr{A}^d(\mathbb{R}_+)$) denotes the space of absolutely continuous functions, ϕ, from [0,1] (resp. $\mathscr{R}_+$) to $\mathscr{R}^d$ with $\phi(0) = 0$. Although, our topology on $\mathscr{Y}$ is quite different from the gauge topology introduced by Dobrushin and Pechersky [13], conceptually, it is quite similar: the idea is to get some kind of uniform control over the sample average. We have also used some ideas from their paper in the proof of Theorem 1 below, in order to construct compact sets that support most of the measure. Eichelsbacher and O'Connell [15] extend the sample path LDP to an even finer topology on $\mathscr{Y}$ than considered here, under the additional assumption that X_k is an *iid* sequence. We remark also that Deuschel and Stroock [11] prove a version of Schilder's theorem in the space $\mathscr{Y}$, using essentially Gaussian techniques.

To illustrate how Theorem 1 can be applied, we will continue to work through the single-server queue example: suppose $d = 1$ and consider the function f defined by Eq. (4). Recall that $f(\tilde{S}_n)$ is equal in distribution to the normalized queue length at a single-server queue. If $\Lambda'(0) = \mu$, say, then a corollary of Theorem 1 is that the LDP holds in the subspace

$$\mathscr{Y}^\mu = \left\{ \phi \in \mathscr{Y} : \quad \lim_{t\to\infty} \frac{\phi(t)}{1+t} = \mu \right\}.$$

If $\mu < 0$, then the restriction of f to $\mathscr{Y}^\mu$ is finite and continuous. To see this, note that $\mathscr{Y}^\mu$ is a metric space and so we can check continuity using sequences. Take $\phi_n \to \phi$ in $\mathscr{Y}^\mu$. Then, for any $\varepsilon > 0$ there exists N such that for all $n > N$ we have $\phi_n(t) \le \phi(t) + \varepsilon(1+t)$ for all $t \ge 0$. Since $\phi \in \mathscr{Y}^\mu$ we also have $\phi(t) < (\mu + \varepsilon)(1+t)$ for all t sufficiently large ($t > T$ say). Thus, for all $n > N$ and $t > T$, $\phi_n(t) < (\mu + 2\varepsilon)(1+t)$ which, in particular, is negative for small ε. Now, since $\phi(0) = \phi_n(0) = 0$ for all n, we can completely ignore what happens outside the interval $[0, T]$. The suprema are attained on this interval; moreover, $\phi_n \to \phi$ in $\mathscr{Y}^\mu$ implies that $\phi_n \to \phi$ uniformly on $[0, T]$, and so the supremum converges

on this interval. Hence, f is continuous on $\mathscr{Y}^{\mu}$. We can therefore apply the contraction principle and Jensen's inequality to get that the sequence $Q_0/n = f(\tilde{S}_n)$ satisfies the LDP in $\mathbb{R}_+$ with rate function given by

$$J(q) = \inf\left\{\int_0^\infty \Lambda^*(\dot{\phi})\,\mathrm{d}s : \sup_{t>0}\phi(t) = q\right\}$$

$$= \inf_{\tau>0}\inf\left\{\int_0^\tau \Lambda^*(\dot{\phi})\,\mathrm{d}s : \phi(\tau) = q\right\} = \inf_{\tau>0}\tau\Lambda^*(q/\tau).$$

This fact has previously been demonstrated by several authors [5,12,14,17], under similar conditions. The *i.i.d.* case is due to Cramér [8] and Borovkov [4]. The advantage of our approach is that the existence of an LDP is established by continuity which, using the above topology, is quite accessible, and the rate function is easier to compute.

Finally, we remark that this result is not specifically designed for the single-server queue. It is widely applicable, and ideally suited to problems where reflection mappings exist. It has been used, for example, to obtain comprehensive equilibrium large deviations results for a multiclass FIFO queue [21] and can also be applied to systems with dedicated buffers [20,22] (the latter corresponds to the random walk in a quadrant, and is the subject of many recent papers: see, for example, Bertsimas *et al.* [2,3]).

PROOF OF THEOREM 1 We have from the assumptions of the theorem and the Dawson–Gärtner theorem for projective limits ([10], Theorem 4.6.1) that the sequence $\tilde{S}_n$ satisfies the LDP on $\mathscr{C}^d(\mathbb{R}_+)$, equipped with the topology of uniform convergence on compacts, with the good rate function

$$I(\phi) = \begin{cases} \int_0^\infty \Lambda^*(\dot{\phi})\,\mathrm{d}s & \phi \in \mathscr{A}^d(\mathbb{R}_+), \\ \infty & \text{otherwise} \end{cases}$$

We first show that $\tilde{S}_n$ satisfies the LDP on $\mathscr{Y}$ with the same rate function by showing that $\mathscr{D}_I \subset \mathscr{Y}$ and $P(\tilde{S}_n \in \mathscr{Y}) = 1$.

By considering $\tilde{S}_n(t) - t\nabla\Lambda(0)$ we can, without loss of generality, assume that $\nabla\Lambda(0) = 0$. Let ϕ belong to the domain of I. Then ϕ is absolutely continuous, and we have from the non-negativity and convexity of Λ^*, and Jensen's inequality, that

$$t\Lambda^*(\phi(t)/t) \le I(\phi).$$

Since this holds for all t, we must have $\Lambda^*(\phi(t)/t) \to 0$ as $t \to \infty$. Now, by the assumption that Λ is differentiable at the origin, Λ^* has a unique zero at $\nabla\Lambda(0) = 0$. Hence, $\phi(t)/t \to 0$ as $t \to \infty$, and so $\phi \in \mathscr{Y}$. Moreover, by the assumption that

$\tilde{S}_n[0,1]$ satisfies the LDP in $C^d[0,1]$ and the contraction principle, $\tilde{S}_n(1) = (1/n)\sum_{k=1}^n X_k$ satisfies the LDP in $\mathbb{R}$ with rate function Λ^*. As already noted, Λ^* has a unique zero at $E[X_1] = 0$. Hence, for any $\varepsilon > 0$, there is a $\delta > 0$ such that $P(|\tilde{S}_n(1)| > \varepsilon) < \mathrm{e}^{-n\delta}$ for all n sufficiently large. Thus, by the Borel–Cantelli Lemma, $\tilde{S}_n(1) = (1/n)\sum_{k=1}^n X_k \to 0$ as $n \to \infty$, from which it is immediate that, for every fixed n,

$$\lim_{t\to\infty} \frac{\tilde{S}_n(t)}{t} = 0 \ \ a.s.$$

Hence, $P(\tilde{S}_n \in \mathscr{Y}) = 1$.

We now have by ([10], Lemma 4.1.5) that $\tilde{S}_n$ satisfies the LDP in $\mathscr{Y}$ when equipped with the topology of uniform convergence on compact intervals, and that the rate function is I_∞ (the restriction of I to $\mathscr{Y}$). To strengthen this to the topology induced by the norm $\|\cdot\|_u$, we appeal to the inverse contraction principle ([10], Corollary 4.2.6), by which it suffices to prove exponential tightness of the sequence $\tilde{S}_n$ in the space $(\mathscr{Y}, \|\cdot\|_u)$.

By the Dawson–Gärtner theorem and the assumption that I_1 is good, I_∞ is a good rate function in the topology of uniform convergence on compact intervals. Hence, the set

$$K_\alpha = \{\phi \in \mathscr{C}^d(\mathbb{R}_+) : I_\infty(\phi) \le \alpha\}$$

is compact in this topology, and

$$\limsup_n \frac{1}{n} \log P\{\tilde{S}_n \notin K_\alpha\} \le -\alpha. \tag{9}$$

By assumption, Λ is finite in a neighbourhood of the origin, so we can find $\theta_0 > 0$ such that $|\Lambda_j(\theta)| < \infty$ for all $|\theta| \le \theta_0$ and $j = 1, \ldots, d$. Here,

$$\Lambda_j(\theta) := \lim_{n\to\infty} \frac{1}{n} \log E\, \mathrm{e}^{n\theta \tilde{S}_n^j(1)},$$

where $\tilde{S}_n^j$ is the jth component of $\tilde{S}_n$; in other words, $\Lambda_j(\theta) = \Lambda(\theta \mathrm{e}_j)$ where e_j denotes the jth unit vector. We now define the following quantities:

$$\delta(n) = \sup_{m\ge n} \sup_{|\theta|<\theta_0} \max_{1\le j\le d} \left| \frac{1}{m} \log E\, \mathrm{e}^{m\theta \tilde{S}_m^j(1)} - \Lambda_j(\theta) \right|,$$

$$\theta(n) = \min\left\{ \frac{\theta_0}{2}, \sqrt{\delta(n)}, n^{-1/4} \right\},$$

$$d_\alpha(n) = (\alpha+1)\left[\max_{1\le j\le d} \frac{|\Lambda_j(\theta(n))| + |\Lambda_j(-\theta(n))| + \delta(n)}{\theta(n)} + n^{-1/4} \right].$$

It is not hard to see that $\delta(n)$, $\theta(n)$ and $d_\alpha(n)$ all decrease to zero as n increases to infinity (for $d_\alpha(n)$, we use the assumption that $\nabla\Lambda_j(0) = 0$).

Consider the set

$$D_\alpha = \left\{ \phi \in K_\alpha : \max_{1\le j\le d} \left|\frac{\phi^j(t)}{1+t}\right| \le d_\alpha([t]) \ \text{ for all } t \ge 1 \right\},$$

where $[t]$ denotes the integer part of t. Now $\phi \equiv 0$ belongs to D_α because $\Lambda^*(0) = 0$ as a consequence of the assumption that $\nabla\Lambda(0) = 0$. Thus, D_α is non-empty. The exponential tightness of $\tilde{S}_n$ in $(\mathscr{Y}, \|\cdot\|_u)$ will be established by the following two lemmas. □

LEMMA 1 *For each $\alpha > 0$, D_α is compact in $(\mathscr{Y}, \|\cdot\|_u)$.*

PROOF Let ϕ_n be a sequence in D_α. Since K_α is compact in $\mathscr{Y}$ equipped with the topology of uniform convergence on compact intervals, there exists a subsequence $n(k)$ such that $\phi_{n(k)}$ converges to some $\phi \in K_\alpha$ in this topology. It follows that, for each $T > 0$, and for each j,

$$\limsup_{k\to\infty} \sup_{t\le T} \left|\frac{\phi^j_{n(k)}(t)}{1+t} - \frac{\phi^j(t)}{1+t}\right| = 0, \ \ j = 1, \ldots, d.$$

Note that this implies, for each t and j, that

$$\left|\frac{\phi^j(t)}{1+t}\right| \le d_\alpha([t]),$$

and so $\phi \in D_\alpha$. Now for each $\varepsilon > 0$, there is a finite T such that $d_\alpha(T) \le \varepsilon$. Hence, for k sufficiently large,

$$\begin{aligned} \|\phi_{n(k)} - \phi\|_u &\le \sup_j \left\{ \sup_{t\le T} \left|\frac{\phi^j_{n(k)}(t)}{1+t} - \frac{\phi^j(t)}{1+t}\right| + \sup_{t>T} \left|\frac{\phi^j_{n(k)}(t)}{1+t} - \frac{\phi^j(t)}{1+t}\right| \right\} \\ &\le \varepsilon + 2d_\alpha(T) \le 3\varepsilon. \end{aligned}$$

The set D_α is therefore sequentially compact, and hence compact, in the metric space $(\mathscr{Y}, \|\cdot\|_u)$. □

LEMMA 2 *Under the hypotheses of Theorem 1,*

$$\lim_{\alpha\to\infty} \limsup_n \frac{1}{n} \log P(\tilde{S}_n \notin D_\alpha) = -\infty.$$

PROOF By Chernoff's inequality,

$$P\bigcup_{t\geq 1}\left\{\tilde{S}_n^j(t) > (1+t)d_\alpha(t)\right\} \leq P\bigcup_{k=1}^{\infty}\bigcup_{i=0}^{n-1}\left\{\tilde{S}_n^j(k+(i/n)) > (1+k)d_\alpha(k+1)\right\}$$

$$\leq \sum_{k=1}^{\infty}\sum_{i=0}^{n-1} E\exp\left[n\theta(k+1)\tilde{S}_n^j(k+(i/n)) - n\theta(k+1)(k+(i/n))d_\alpha(k+1)\right]$$

$$= \sum_{k=1}^{\infty}\sum_{i=0}^{n-1} E\exp\left[(nk+i)\theta(k+1)\tilde{S}_{nk+i}^j(1) - (nk+i)\theta(k+1)d_\alpha(k+1)\right]$$

$$\leq \sum_{k=1}^{\infty}\sum_{i=0}^{n-1} \exp(nk+i)[\Lambda_j(\theta(k+1)) + \delta(nk+i) - \theta(k+1)d_\alpha(k+1)],$$

where the last inequality is valid for n sufficiently large, when it follows from the inequality,

$$E\left[\mathrm{e}^{n\theta\tilde{S}_n^j(1)}\right] \leq \mathrm{e}^{n[\Lambda_j(\theta)+\delta(n)]},$$

which holds for all $n \geq 1$ and all $|\theta| < \theta_0$ by definition of $\delta(n)$. We note that $0 \leq \theta(k) < \theta_0/2$ for all $k \in \mathbb{N}$ by definition. Now,

$$d_\alpha(k)\theta(k) \geq (\alpha+1)(|\Lambda_j(\theta(k))| + \delta(k) + \theta(k)k^{-1/4}) \quad \forall\, k \geq 1$$

by definition of d_α. Moreover, $\delta(k)$ is a decreasing, non-negative function of k and $\theta(k) \geq \min\{\theta_0/2, k^{-1/4}\}$ by definition, so we get

$$\Lambda_j(\theta(k+1)) + \delta(nk+i) - \theta(k+1)d_\alpha(k+1)$$

$$\leq -\alpha(|\Lambda_j(\theta(k+1))| + \delta(k+1) + \theta(k+1)(k+1)^{-1/4})$$

$$\leq -\alpha\min\{\theta_0/2, (k+1)^{-1/4}\}(k+1)^{-1/4} \leq -\alpha\min\{\theta_0/2, 1\}(k+1)^{-1/2}$$

$$\leq -\frac{\alpha}{2}\min\{\theta_0/2, 1\}\left(k+\frac{i}{n}\right)^{-1/2} \quad \forall\, k \geq 1,\ 0 \leq i \leq n-1.$$

Consequently, defining $\tilde{\alpha} = \alpha\min\{\theta_0/2, 1\}$, we get

$$P\bigcup_{t\geq 1}\left\{\tilde{S}_n^j(t) > (1+t)d_\alpha(t)\right\} \leq \sum_{k=1}^{\infty}\sum_{i=0}^{n-1}\exp\left[-\frac{\tilde{\alpha}}{2}(nk+i)\left(k+\frac{i}{n}\right)^{-1/2}\right]$$

$$\leq \sum_{j=n}^{\infty}\exp\left[-\frac{\tilde{\alpha}}{2}\sqrt{n}\sqrt{j}\right] \leq D\exp\,-\frac{\tilde{\alpha}}{4}\sqrt{n(n-1)}$$

for some finite constant D that remains bounded as α and n increase to infinity. Here, we have used the inequality

$$\sum_{k\geq k_0} e^{-\rho\sqrt{k}} \leq \frac{4\mathrm{e}^{-1}+2}{\rho^2}\exp\left(-\frac{\rho}{2}\sqrt{k_0-1}\right),$$

obtained as follows:

$$\sum_{k\geq k_0}\mathrm{e}^{-\rho\sqrt{k}} \leq \int_{k_0-1}^{\infty}\mathrm{e}^{-\rho\sqrt{x}}\,\mathrm{d}x = \frac{2}{\rho^2}\int_{\rho\sqrt{k_0-1}}^{\infty} z\mathrm{e}^{-z}\,\mathrm{d}z = \frac{2}{\rho^2}[\rho\sqrt{k_0-1}+1]\mathrm{e}^{-\rho\sqrt{k_0-1}}$$

$$\leq \frac{4\mathrm{e}^{-1}+2}{\rho^2}\exp\left(-\frac{\rho}{2}\sqrt{k_0-1}\right)$$

To obtain the last inequality above, we used the fact that $x\mathrm{e}^{-x} \leq e^{-1}$ and $\mathrm{e}^{-x} \leq 1$ for all $x \geq 0$. Likewise, we can show that

$$P\bigcup_{t\geq 1}\left\{\tilde{S}_n^j(t) < -(1+t)d_\alpha(t)\right\} \leq D\exp\,-\frac{\tilde{\alpha}}{4}\sqrt{n(n-1)}.$$

Therefore,

$$\limsup_n \frac{1}{n}\log P\bigcup_{t>1}\left\{\left|\tilde{S}_n^j(t)\right| > (1+t)d_\alpha(t)\right\} \leq -\frac{\alpha}{4}\min\left\{\frac{\theta_0}{2},1\right\}. \tag{10}$$

The statement of the lemma can now be obtained from Eqs. (9) and (10), via the principle of the largest term. □

This concludes the proof of the theorem.

References

[1] Bertsimas, D., Paschalidis, I. and Tsitsiklis, J.N. (1998) "On the large deviation behaviour of acyclic networks of G/G/1 queues", *Ann. Appl. Probab.* **8**(4).

[2] Bertsimas, D., Paschalidis, I. and Tsitsiklis, J.N. (1998) "Asymptotic buffer overflow probabilities in multiclass multiplexers: an optimal control approach", *IEEE Trans. Autom. Control* **43**, 315–335.

[3] Bertsimas, D., Paschalidis, I. and Tsitsiklis, J.N. (1999) "Large deviations analysis of the generalized processor sharing policy", *Queueing Syst.* **32**, 319–349.

[4] Borovkov, A.A. (1976) Random Processes in Queueing Theory (Springer, Berlin).

[5] Chang, C.S. (1994) "Stability, queue length and delay of deterministic and stochastic queueing networks", *IEEE Trans. Autom. Control* **39**, 913–931.

[6] Chang, C.S. (1995) "Sample path large deviations and intree networks", *Queueing Syst.* **20**, 7–36.

[7] Chang, C.S. and Zajic, T. (1995) "Effective bandwidths of departure processes from queues with time varying capacities", *INFOCOM*.

[8] Cramér, H. (1954) "On some questions connected with mathematical risk", *Univ. Calif. Publ. Statist.* **2**, 99–125.

[9] Dembo, A. and Zajic, T. (1995) "Large deviations: from empirical mean and measure to partial sums process", *Stoch. Proc. Appl.* **57**, 191–224.

[10] Dembo, A. and Zeitouni, O. (1993) Large Deviations Techniques and Applications (Jones and Bartlett, Boston).
[11] Deuschel, J.-D. and Stroock, D.W. (1989) Large Deviations (Academic Press, NewYork).
[12] de Veciana, G., Courcoubetis, C. and Walrand, J. (1994) "Decoupling bandwidths for networks: a decomposition approach to resource management", *INFOCOM*.
[13] Dobrushin, R.L. and Pechersky, E.A. (1998) "Large deviations for random processes with independent increments on infinite intervals", *Probl. Inform. Transm.* **34**, 354–384.
[14] Duffield, N.G. and O'Connell, N. (1995) "Large deviations and overflow probabilities for the general single server queue, with applications", *Proc. Camb. Phil. Soc.* **118**(1).
[15] Eichelsbacher, P. and O'Connell, N. (1999) "Sample path large deviations in finer topologies", *Stochastics Stochastic Rep.* **67**, 231–254.
[16] Ganesh, A. and Anantharam, V. (1996) "Stationary tail probabilities in exponential server tandems with renewal arrivals", *Queueing Syst.* **22**, 203–247.
[17] Glynn, P.W. and Whitt, W. (1994) "Logarithmic asymptotics for steady-state tail probabilities in a single-server queue", *J. Appl. Prob.* **31A**, 131–156.
[18] Majewski, K. (1998) "Heavy traffic approximations of large deviations of feed-forward queueing networks", *Queueing Syst.* **28**, 125–155.
[19] Mogulskii, A.A. (1976) "Large deviations for trajectories of multi dimensional random walks", *Th. Prob. Appl.* **21**, 300–315.
[20] O'Connell, N. (1996) "Queue lengths and departures at single-server resources", In: Kelly, F.P., Zachary, S. and Ziedins, I., eds, Stochastic Networks: Theory and Applications (Clarendon Press, Oxford).
[21] O'Connell, N. (1997) "Large deviations for departures from a shared buffer", *J. Appl. Prob.* **34**, 753–766.
[22] O'Connell, N. (1998) "Large deviations for queue lengths at a multi-buffered resource", *J. Appl. Prob.* **35**, 240–245.
[23] Ramanan, K. and Dupuis, P. (1998) "Large deviation properties of data streams that share a buffer", *Annu. Appl. Probab.* **8**(4), 1.
[24] Toomey, F. (1998) "Bursty traffic and finite capacity queues", *Annu. Oper. Res.* **79**, 45–62.
[25] Varadhan, S.R. (1966) "Asymptotic probabilities and differential equations", *Comm. Pure Appl. Math.* **19**, 261–286.