HEWLETT
PACKARD

Large Deviations for Departures from a Shared Buffer

Neil O'Connell
Basic Research Institute in the
Mathematical Sciences
HP Laboratories Bristol
HPL-BRIMS-96-03
February, 1996

In this paper we describe how the joint large deviation properties of traffic streams are altered when the traffic passes through a shared buffer according to a FCFS service policy with stochastic service capacity. We also consider the stationary case, proving large deviation principles for the state of the system in equilibrium and for departures from an equilibrium system.

LARGE DEVIATIONS FOR DEPARTURES FROM A SHARED BUFFER

Neil O'Connell, BRIMS, Hewlett-Packard Labs, Bristol

BRIMS Technical Report HPL-BRIMS-96-003 [1]

AMS Subject Classifications: 60F10, 60K25, 60K30.

Abstract

In this paper we describe how the joint large deviation properties of traffic streams are altered when the traffic passes through a shared buffer according to a FCFS service policy with stochastic service capacity. We also consider the stationary case, proving large deviation principles for the state of the system in equilibrium and for departures from an equilibrium system.

---

[1] This is a revised version of 'Large deviations in queueing networks', DIAS Technical Report DIAS-APG-9413, and has been submitted to *J. Appl. Prob.*

1 Introduction

Consider a single server queue with arrivals process X_n and service process C_n: for each integer time n, X_n denotes the amount of work arriving at the queue and C_n denotes the amount of work that can be serviced; the queue length at time n is defined recursively by the Lindley equation

$$Q_n = (Q_{n-1} + X_n - C_n)^+. \tag{1}$$

For each n set

$$A_n = \sum_{k=1}^{n} X_k, \qquad B_n = \sum_{k=1}^{n} C_k, \qquad W_n = A_n - B_n, \tag{2}$$

with the convention that $A_0 = B_0 = W_0 = 0$. If X and C are stationary processes and $EX_1 < EC_1$, then Q is stationary and

$$Q_0 \stackrel{d}{=} \sup_n W_n. \tag{3}$$

The identity (3) can be used to deduce the asymptotic behaviour of the queue-length distribution from the large deviation properties of W. More precisely, if the sequence W_n/n satisfies the large deviation principle (LDP) with good rate functions I, then, under mild hypotheses on I [2, 9, 11, 16], the tails of the queue-length distribution satisfy the order relation

$$\log P(Q_0 > b) \sim -\delta b \tag{4}$$

for large b, where

$$\delta = \inf_{w>0} I(w)/w. \tag{5}$$

As this is such a general result, it may be useful for real applications: in particular, it provides a basis for predicting overflow probabilities at a single buffered resource [6, 10].

It also suggests the possibility of a kind of network calculus at the level of rare events. For example, given an arbitrary network with several inputs, it may be possible to estimate the probability of overflow at a given (buffered) node in terms of the large deviation properties of the inputs. The obvious starting point is to ask how the large deviation properties of traffic are altered when the traffic passes through a buffer, possibly sharing that buffer with other traffic. In a recent paper, de Veciana, Courcoubetis and Walrand [9] give a partial answer to this question. Suppose we have two independent, ergodic arrival processes X^1 and X^2 sharing an (initially empty) deterministic buffer according to a work conserving policy with service rate c. Suppose also that the corresponding partial sums satisfy the LDP with respective rate functions I_1 and I_2. Then, under certain conditions [9, Corollary 3.2], if D_n^1 denotes the total departures upto time n corresponding to the first traffic stream, the sequence D_n^1/n satisfies the LDP with good rate function given by I_1 on the interval $[EX_1^1, c - EX_1^2]$. In this paper we consider a more general problem. For a single-server queue with an arbitrary number of inputs, served according to a FCFS policy with stochastic service rate, we prove that, under suitable hypotheses, the collection of associated departure processes satisfy the LDP with rate function given by the solution of a finite-dimensional optimisation problem. We also consider a stationary version of the problem, where the queue is assumed to be initially in equilibrium. In this case we prove an LDP for the state of the system in equilibrium, and an LDP for the departure processes. The latter results were motivated by an observation of Chang and Zajic [5] that (in the one-dimensional case) the LDP for a stationary departure process is not necessarily the same as the LDP for departures from an initially empty queue.

For related work, see [1, 4, 3, 13, 14, 17, 19, 20] and references therein. The methods used in this paper can be applied, in principle, to most network configurations. The main tools used are (i) the description of interesting objects as functionals of the inputs and configuration, (ii) the contraction principle, to deduce LDP's for these objects, and (iii) convex analysis, to simplify the corresponding rate functions. In view of (iii), this simplification is far easier if the inputs are assumed to satisfy a sample-path LDP with 'linear point-to-point geodesics' (see the hypothesis (H2) below); unfortunately, this assumption is not generally satisfied by the corresponding departure processes [15], and so the recursive application of these simplifying results is not possible in general.

The format of the paper is as follows. Departures from an initially empty buffer are considered in Section 2 and some equilibrium results are presented in Section 3; Section 4 is devoted to proofs.

We will adopt the following convention throughout the paper: if $\boldsymbol{x}$ is a vector-valued object, denote by x^i the components of $\boldsymbol{x}$ and by x the sum of the components of $\boldsymbol{x}$.

2 Departures from a shared buffer

Suppose we have d arrival streams $\boldsymbol{X} = (X^1, \ldots, X^d)$ sharing an infinite buffer, initially empty, according to a FCFS policy with stochastic service rate C: we will begin by making this statement precise. For the moment, the only assumption is that $X^1, \ldots, X^d$ and C are non-negative sequences of random variables, indexed by the positive integers. For each n, set

$$\boldsymbol{A}_n = \sum_{k=1}^{n} \boldsymbol{X}_k, \qquad B_n = \sum_{k=1}^{n} C_k. \tag{6}$$

The total amount of work in the queue at time n is given by the recursion ($Q_0 = 0$)

$$Q_n = (Q_{n-1} + X_n - C_n)^+, \tag{7}$$

and the total departures (amount of work serviced) up to time n is given by

$$D_n = A_n - Q_n, \tag{8}$$

or, equivalently,

$$D_n = \inf_{0 \leq k \leq n} (A_k - B_k) + B_n. \tag{9}$$

It remains to specify the quantities of interest, namely the amounts of work, $\boldsymbol{D}_n = (D_n^1, \ldots, D_n^d)$, serviced from each input stream by time n. To do this we set

$$T_n = \sup\{k \leq n : \; A_k \leq D_n\}, \tag{10}$$

$$\boldsymbol{D}_n = \boldsymbol{A}_{T_n}^i + (D_n - A_{T_n})\boldsymbol{X}_{T_n+1}/X_{T_n+1}. \tag{11}$$

Note that $D_n = D_n^1 + \cdots + D_n^d$, and so our notation is consistent.

In words, work is serviced in the order received and simultaneous arrivals from each source are thoroughly mixed in the queue.

To state the main result we set, for $0 \leq t \leq 1$,

$$\boldsymbol{S}_n(t) = \left(\frac{1}{n} \boldsymbol{A}_{[nt]}, \frac{1}{n} B_{[nt]} \right), \tag{12}$$

and

$$\boldsymbol{R}_n(t) = \frac{1}{n} \boldsymbol{D}_{[nt]}. \tag{13}$$

For each positive integer k denote by $\mathcal{L}^k$ the subspace of paths in $L_\infty([0,1])^k$ with non-decreasing components, by $\mathcal{C}^k \subset \mathcal{L}^k$ the subspace of continuous paths starting at zero, and by $\mathcal{A}^k \subset \mathcal{C}^k$ the set of those paths with absolutely continuous components. We will record the following hypotheses.

(H1) **LDP for arrivals and service processes.** *For all* $\gamma \in \mathbb{R}$,

$$\sup_k E e^{\gamma(X_k + C_k)} < \infty, \tag{14}$$

and the sequence S_n *satisfies the LDP in* $\mathcal{L}^{d+1}$ *with good rate function* I.

(H2) **Linear geodesics.** *For each* $\boldsymbol{\lambda} \in \mathbb{R}^{d+1}$, *the limit*

$$\Lambda(\boldsymbol{\lambda}) = \lim_{n\to\infty} \frac{1}{n} \log E \exp\left(\sum_{k=1}^{n} \boldsymbol{\lambda} \cdot \boldsymbol{X}_k \right) \tag{15}$$

exists as an extended real number and is finite in a neighbourhood of the origin. Note that Λ *is automatically convex. (H1) holds and the rate function* I *is given by*

$$I(\phi) = \begin{cases} \int_0^1 \Lambda^*(\dot{\phi}) ds & \phi \in \mathcal{A}^{d+1} \\ \infty & \textit{otherwise} \end{cases} \tag{16}$$

where Λ^* *is the Fenchel-Legendre transform of* Λ.

(H3) **Asymptotic independence of arrivals and service processes.** *(H2) holds, and* Λ^* *is of the form*

$$\Lambda^*(\boldsymbol{x}, c) = \Lambda_a^*(\boldsymbol{x}) + \Lambda_b^*(c), \tag{17}$$

for $(\boldsymbol{x}, c) \in \mathbb{R}^d \times \mathbb{R}$.

For fairly general conditions under which the hypothesis (H2) is satisfied, see [7].

Theorem 2.1 *If (H1) is satisfied, then the sequence* $\boldsymbol{R}_n$ *satisfies the LDP in* $\mathcal{L}^d$ *with good rate function given by*

$$I_d(\boldsymbol{\psi}) = \inf\{I(\boldsymbol{\phi}) : \ \boldsymbol{\Delta}(\boldsymbol{\phi}) = \boldsymbol{\psi}\}, \tag{18}$$

where $\boldsymbol{\Delta} : \mathcal{C}^{d+1} \to \mathcal{C}^d$ *is defined by*

$$\boldsymbol{A}(\boldsymbol{\phi}) = (\phi^1, \ldots, \phi^d), \tag{19}$$

$$D(\boldsymbol{\phi})(t) = \inf_{0\le\nu\le1} [A(\boldsymbol{\phi})(\nu t) - \phi^{d+1}(\nu t)] + \phi^{d+1}(t), \tag{20}$$

$$T(\boldsymbol{\phi})(t) = \inf\{r : \ A(\boldsymbol{\phi})(r) = D(\boldsymbol{\phi})(t)\}, \tag{21}$$

$$\boldsymbol{\Delta}(\boldsymbol{\phi}) = \boldsymbol{A}(\boldsymbol{\phi}) \circ T(\boldsymbol{\phi}). \tag{22}$$

Corollary 2.2 *If (H3) is satisfied, then D_n/n satisfies the LDP in $\mathbb{R}_+^d$ with good rate function given by*

$$\begin{aligned}\Lambda_d^*(z) \;=\; & \inf\{\beta\Lambda_a^*(x/\beta) + \sigma\Lambda_a^*\left(\frac{z-x}{\sigma}\right) + \beta\Lambda_b^*(c) + (1-\beta)\Lambda_b^*\left(\frac{z-x}{1-\beta}\right) : \\ & \beta, \sigma \in [0,1],\; c \in \mathbb{R},\; \beta+\sigma \leq 1,\;\; x \leq \beta c\}.\end{aligned}$$

We can deduce the large deviation properties of departures from a queue fed by a single arrivals stream by setting $d = 1$. This generalises results of de Veciana *et al.* [9, Theorem 3.1] and Chang *et al* [4], and sharpens a result of [12, Theorem 3]. For a closer look at the optimisation problem in Corollary 2.2, see [15].

3 Equilibrium results

In the last section we assumed that the buffer was initially empty. In the case of a single input ($d = 1$), Chang and Zajic [5] prove a stationary version of Corollary 2.2 and make the important observation that the rate function for the departures in the stationary case is generally different from the 'transient' case when the service rate is stochastic (otherwise it is the same); the difference stems from the fact a large (positive) deviation in the departures can be encouraged by starting with a very long queue. Recall that the tail decay of the queue length distribution is determined by Λ_a^* and Λ_b^*. Their result states that, under additional mixing hypotheses, if there is just one arrivals process, the rate function for the *stationary* departure process is given by

$$\bar{\Lambda}_d^*(z) = \delta z - \sup_{x \leq z}[\delta x - \Lambda_a^*(x)] + \Lambda_b^*(z \vee \mu_b), \tag{23}$$

for $z \geq \Lambda_a'(0)$, where

$$\delta = \inf_{w,c}[\Lambda_a^*(w+c) + \Lambda_b^*(c)]/w. \tag{24}$$

The additional mixing hypothesis is assumed because the queue-length at time zero is not necessarily independent of subsequent service and arrivals. One might therefore expect a similar result to hold for departures from a shared buffer when the system is assumed to be initially in equilibrium; note, however, that to describe the state of the system in equilibrium requires more than just a single queue-length, or even d queue-lengths.

We will begin by setting up a stationary version of the system described in the previous section. Suppose $\{(\boldsymbol{X}_k, C_k) : k \in \mathbb{Z}\}$ is an ergodic, stationary sequence in $\mathbb{R}_+^d \times \mathbb{R}_+$. It is convenient to define cumulative arrivals and service on intervals: set

$$\boldsymbol{A}_{k,l} = \sum_{j=k+1}^{l} \boldsymbol{X}_j, \qquad\qquad B_{k,l} = \sum_{j=k+1}^{l} C_j; \tag{25}$$

we will write $\boldsymbol{A}_n$ for $\boldsymbol{A}_{0,n}$ and B_n for $B_{0,n}$. As before, for $0 \leq t \leq 1$ we set

$$\boldsymbol{S}_n(t) = \left(\frac{1}{n}\boldsymbol{A}_{[nt]}, \frac{1}{n}B_{[nt]}\right), \tag{26}$$

and for $0 \leq r \leq L/n$, set $\boldsymbol{O}_n(r) = \boldsymbol{N}([rn])/n$. We will continue to refer to the hypotheses recorded in the previous section. The following hypothesis will guarantee stability, and facilitate the proof of Theorem 3.1 below.

(H4) **Stability.** *(H2) holds, and* $\mu = \nabla\Lambda(0)$ *exists with* $A(\bar{\mu}) < \mu^{d+1}$.

The (total) amount of work in the queue at time $n \in \mathbb{Z}$ is given by

$$Q_n = \sup_{k \geq 0}(A_{n-k,n} - B_{n-k,n}). \tag{27}$$

Lemma 3.1 *If (H4) holds, then* $P(Q_n < \infty) = 1$ *for all* n.

The hypothesis (H4) will be in force throughout this section. Note that Q satisfies the Lindley equations (1). The (total) departures during the interval $(k, l]$ is given by

$$D_{k,l} = A_{k,l} + Q_k - Q_l. \tag{28}$$

Set

$$K = \inf\{k \geq 0 : Q_{-k} = 0\}, \tag{29}$$

and note that $Q_0 = A_{-K,0} - B_{-K,0}$.

Just as in the previous set-up, we need to specify the quantities of interest, and this requires an assumption about how service is distributed between inputs. Let

$$L = -\sup\{k \leq 0 : A_{-K,k} \leq D_{-K,0}\}$$

and define the departures $\boldsymbol{D}_{-K,0} = (D^1_{-K,0}, \ldots, D^d_{-K,0})$ from the respective inputs on the interval $(-K, 0]$ to be

$$\boldsymbol{D}_{-K,0} = \boldsymbol{A}_{-K,-L} + \epsilon,$$

where

$$\epsilon = (D_{-K,0} - A_{-K,-L})X_{1-L}/X_{1-L}.$$

To justify our notation, note that $D^1_{-K,0} + \cdots + D^d_{-K,0} = D_{-K,0}$.

To describe the state of the system at time 0, we consider the following $\mathbb{R}^d_+$-valued process: set $\boldsymbol{N}(L) = \boldsymbol{A}_{-L,0} - \epsilon$, and for $k = 1, \ldots, L-1$ set $\boldsymbol{N}(k) = \boldsymbol{A}_{-k,0}$. Note that $N^i(k)$ is the amount of work of type i that's been waiting in the queue for at most k units of time; $N(L) = Q_0$, and $N^i(L)$ is the amount of work of type i in the queue at time 0. For clarity we will write $\boldsymbol{Q}_0$ for $\boldsymbol{N}(L)$; $\boldsymbol{Q}_n$, the respective amounts of work in the queue at any other time n, is defined similarly. Write $\tilde{\boldsymbol{O}}_n$ for the polygonal approximation (see, for example, (38)) to $\{\boldsymbol{A}_{-nt,0}/n,\ t \geq 0\}$ on the interval $[0, \tilde{L}]$, where $\tilde{L} = L - \epsilon/X_{1-L}$. Note that $\tilde{\boldsymbol{O}}_n(\tilde{L}) = \boldsymbol{Q}_0/n$. The first result of this section is an LDP for $\tilde{\boldsymbol{O}}_n$.

We will begin our preparations by extending the range of path spaces: for each positive integer k denote by $\mathcal{L}^k_\tau$ the subspace of paths in $L_\infty([0,\tau])^k$ with non-decreasing components, by $\mathcal{C}^k_\tau \subset \mathcal{L}^k_\tau$ the subspace of continuous paths starting at zero, and by $\mathcal{A}^k_\tau \subset \mathcal{C}^k_\tau$ the set of those paths with absolutely continuous components. Finally, we set

$$\mathcal{B}_k = \{\boldsymbol{\theta} \in \mathcal{C}^k_\tau :\ \tau > 0\}, \tag{30}$$

and equip $\mathcal{B}_k$ with the topology defined by the metric

$$d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \sup_{0 \leq t \leq \tau_1 \wedge \tau_2} \sum_{i=1}^{k} |\theta^i_1(t) - \theta^i_2(t)| + \sum_{i=1}^{k} |\theta^i_1(\tau_1) - \theta^i_2(\tau_2)|, \tag{31}$$

for $\boldsymbol{\theta}_1 \in \mathcal{C}^k_{\tau_1}$, $\boldsymbol{\theta}_2 \in \mathcal{C}^k_{\tau_2}$.

Theorem 3.1 *If (H4) holds, then $\tilde{\boldsymbol{O}}_n$ satisfies the LDP in $\mathcal{B}_d$ with good rate function given by*

$$\begin{aligned} K(\boldsymbol{\theta}) &= \inf\{\rho\Lambda^*(\boldsymbol{x}, c) + \int_0^\beta \Lambda^*(\dot{\boldsymbol{\theta}}, \dot{\phi}) : \\ &\qquad \boldsymbol{x} \in \mathbb{R}^d_+,\ \beta, \rho, c \in \mathbb{R}_+,\ \phi \in \mathcal{A}_\beta,\ \tau(x - c) = \phi(\beta)\}. \end{aligned}$$

Corollary 3.2 *If (H4) is satisfied then $\boldsymbol{O}_n(\tilde{L})$ (or equivalently, $\boldsymbol{Q}_0/n$) satisfies the LDP in $\mathbb{R}^d_+$ with good rate function*

$$\begin{aligned} L(\boldsymbol{q}) &= \inf\{\rho\Lambda^*(\boldsymbol{x}, c) + \beta\Lambda^*(\boldsymbol{q}/\beta, \rho(x - c)/\beta) : \\ &\qquad \boldsymbol{x} \in \mathbb{R}^d_+,\ \beta, \rho, c \in \mathbb{R}_+\}. \end{aligned}$$

To state the LDP for the departures from an equilibrium system we define the cumulative departures from respective inputs upto time n, by

$$\boldsymbol{D}_n = \boldsymbol{Q}_0 + \boldsymbol{A}_n - \boldsymbol{Q}_n, \tag{32}$$

and set $\boldsymbol{R}_n(t) = \boldsymbol{D}_{[nt]}/n$.

Theorem 3.3 *If (H3) and (H4) are satisfied, and the queue is initially in equilibrium, then $\boldsymbol{R}_n$ satisfies the LDP in $\mathcal{L}^d$ with a good rate function (the expression is complicated), and $\boldsymbol{R}_n(1)$ (or equivalently, $\boldsymbol{D}_n/n$) satisfies the LDP in $\mathbb{R}^d_+$ with good rate function given by*

$$\begin{aligned} \tilde{\Lambda}^*_d(\boldsymbol{z}) \;=\; & \inf\{L(\boldsymbol{q}) + \beta_1\Lambda^*\left(\frac{\boldsymbol{z}_1 - \boldsymbol{q}}{\beta_1}, \frac{z_1}{\beta_1}\right) + \beta_2\Lambda^*(\boldsymbol{z}_2/\beta_2, c_2) \\ & +\tau\Lambda^*_a\left(\frac{\boldsymbol{z} - \boldsymbol{z}_1 - \boldsymbol{z}_2}{\tau}\right) + (1 - \beta_1 - \beta_2)\Lambda^*_b\left(\frac{z - z_1 - z_2}{1 - \beta_1 - \beta_2}\right) : \\ & \quad \boldsymbol{q}, \boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathbb{R}^d_+,\ c_2, \beta_1, \beta_2, \tau \in \mathbb{R}_+,\ \beta_1 + \beta_2 + \tau \le 1, \\ & \quad \beta_2 c_2 \ge z_2\}. \end{aligned}$$

4 Proofs

Proof of Theorem 2.1. Throughout the proof, $\|\cdot\|$ denotes the supremum norm on $\mathcal{L}$ and for $\phi \in \mathcal{L}^k$, with a slight abuse of notation we write $\|\phi\| = \sum_i \|\phi^i\|$; this metric generates the uniform product topology in $\mathcal{L}^k$. The maps $\boldsymbol{A} : \mathcal{C}^{d+1} \to \mathcal{C}^d$, $D : \mathcal{C}^{d+1} \to \mathcal{C}$, $T : \mathcal{C}^{d+1} \to \mathcal{C}$ and $\boldsymbol{\Delta} : \mathcal{C}^{d+1} \to \mathcal{C}^d$ are defined (as in the statement of the theorem) by

$$\boldsymbol{A}(\phi) = (\phi^1, \ldots, \phi^d) \tag{33}$$

$$D(\phi)(t) = \inf_{0 \le \nu \le 1}[A(\phi)(\nu t) - \phi^{d+1}(\nu t)] + \phi^{d+1}(t), \tag{34}$$

$$T(\phi)(t) = \inf\{r :\ A(\phi)(r) = D(\phi)(t)\}, \tag{35}$$

$$\boldsymbol{\Delta}(\phi) = \boldsymbol{A}(\phi) \circ T(\phi). \tag{36}$$

Lemma 4.1 *The mapping $\boldsymbol{\Delta} : \mathcal{C}^{d+1} \to \mathcal{C}^d$ is continuous.*

Proof. Suppose $\phi, \tilde{\phi} \in \mathcal{C}^{d+1}$ and $\|\phi - \tilde{\phi}\| \le \epsilon$ for some $\epsilon > 0$. Then $\|D(\phi) - D(\tilde{\phi})\| \le 2\epsilon$. For brevity, fix t and set $\tau = T(\phi)(t)$, $\tilde{\tau} = T(\tilde{\phi})(t)$,

$d = D(\phi)(t)$, $\tilde{d} = D(\tilde{\phi})(t)$. Since ϕ^i is non-decreasing for each i,

$$\begin{aligned}\sum_{i=1}^{d} |\phi^i(\bar{r}) - \phi^i(r)| &= |A(\phi)(\bar{r}) - A(\phi)(r)| \\ &\leq |A(\phi)(\bar{r}) - A(\tilde{\phi})(\bar{r})| + |d - \tilde{d}| \\ &\leq 3\epsilon.\end{aligned}$$

Thus,

$$\begin{aligned}\sum_{i=1}^{d} |\phi^i(r) - \tilde{\phi}^i(\bar{r})| &\leq \sum_{i=1}^{d} |\phi^i(r) - \phi^i(\bar{r}) + \phi^i(\bar{r}) - \tilde{\phi}^i(\bar{r})| \\ &\leq \sum_{i=1}^{d} |\phi^i(\bar{r}) - \phi^i(r)| + \epsilon \\ &\leq 4\epsilon.\end{aligned}$$

Since this holds for all t we have

$$\|\Delta(\phi) - \Delta(\tilde{\phi})\| \leq 4\epsilon, \tag{37}$$

and the result follows. □

For each t, set

$$\tilde{S}_n(t) = S_n(t) + \left(t - \frac{[nt]}{n}\right)\left(S_n\left(\frac{[nt]+1}{n}\right) - S_n\left(\frac{[nt]}{n}\right)\right), \tag{38}$$

and

$$\tilde{\boldsymbol{R}}_n(t) = \boldsymbol{R}_n(t) + \left(t - \frac{[nt]}{n}\right)\left(\boldsymbol{R}_n\left(\frac{[nt]+1}{n}\right) - \boldsymbol{R}_n\left(\frac{[nt]}{n}\right)\right). \tag{39}$$

Note that $\Delta(\tilde{S}_n) = \tilde{\boldsymbol{R}}_n$.

Recall that S_n and $\tilde{S}_n$ are called *exponentially equivalent* if, for all $\delta > 0$,

$$\limsup_{n\to\infty} \frac{1}{n} \log P(\|S_n - \tilde{S}_n\| > \delta) = -\infty. \tag{40}$$

Lemma 4.2 *If (H1) is satisfied then* S_n *and* $\tilde{S}_n$ *(*$\boldsymbol{R}_n$ *and* $\tilde{\boldsymbol{R}}_n$*) are exponentially equivalent in* $\mathcal{L}^{d+1}$ *(*$\mathcal{L}^d$*).*

Proof. Observe that

$$\|S_n - \tilde{S}_n\| = \sup_{k\leq n}(X_k + C_k)/n; \tag{41}$$

this implies that for all $\gamma > 0$,

$$\begin{aligned} P(\|S_n - \tilde{S}_n\| > \delta) &\leq n \sup_{k\leq n} P(X_k + C_k > \delta n) \\ &\leq n e^{-\gamma\delta n} \sup_k E e^{\gamma(X_k+C_k)}, \end{aligned}$$

and now (40) follows from (H1) by considering first the normalised logarithmic limit as $n \to \infty$ and then letting $\gamma \to \infty$. The exponential equivalence of $\boldsymbol{R}_n$ and $\tilde{\boldsymbol{R}}_n$ follows, by a similar argument, from the fact that

$$\|\boldsymbol{R}_n - \tilde{\boldsymbol{R}}_n\| \leq \sup_{k\leq n}(C_k)/n. \tag{42}$$

□

Lemma 4.3 *If (H1) is satisfied then $\tilde{S}_n$ satisfies the LDP in C^{d+1}.*

Proof. By the previous lemma, $\tilde{S}_n$ satisfies the LDP in $\mathcal{L}^{d+1}$. The LDP in C^{d+1} follows from the fact that the domain of I is contained in C^{d+1} and $\tilde{S}_n \in C^{d+1}$ almost surely for all n (see, for example, [8, Lemma 4.1.5]). □

We can now combine the above lemmas and apply the contraction principle [8, Theorem 4.2.1] to get that $\tilde{\boldsymbol{R}}_n$ satisfies the LDP in C^d with good rate function

$$I_d(\psi) = \inf\{I(\phi) : \phi \in C^{d+1}, \ \Delta(\phi) = b\psi\}. \tag{43}$$

Now since C^d is a closed subspace of $\mathcal{L}^d$, $\tilde{\boldsymbol{R}}_n$ also satisfies the LDP in $\mathcal{L}^d$, and the result follows from the exponential equivalence of $\boldsymbol{R}_n$ and $\tilde{\boldsymbol{R}}_n$ in $\mathcal{L}^d$. □

Proof of Corollary 2.2. We begin with a trivial application of the contraction principle to get that $\boldsymbol{D}_n/n$ satisfies the LDP in $\mathbb{R}^d_+$ with good rate function given by

$$\begin{aligned} J(z) &= \inf\{I(\phi) : \phi(T(\phi)(1)) = z\} \\ &= \inf\left\{\int_0^1 \Lambda^*(\dot{\phi}) : \phi(T(\phi)(1)) = z\right\}. \end{aligned}$$

To simplify this, consider a path $\phi \in \mathcal{A}^{d+1}$ with $\phi(T(\phi)(1)) = z$: we will show (and the statement of the theorem will follow) that there exists a path $\phi_1 \in \mathcal{A}^{d+1}$ with

$$\phi_1(T(\phi_1)(1)) = z, \tag{44}$$

and

$$I(\phi) \geq I(\phi_1) = \beta\Lambda_a^*(x/\beta) + (\tau-\beta)\Lambda_a^*\left(\frac{z-x}{\tau-\beta}\right) + \beta\Lambda_b^*(c) + (1-\beta)\Lambda_b^*(c'), \tag{45}$$

for some $0 \leq \beta \leq \tau \leq 1$, $x \in \mathbb{R}^d$, $c, c' \in \mathbb{R}$ with $x \leq \beta c$ and $z - x = (1-\beta)c'$. Set

$$\beta = \sup\{t: \ A(\phi)(\tau) - A(\phi)(t) = \phi^{d+1}(1) - \phi^{d+1}(t)\}, \quad \tau = T(\phi)(1), \tag{46}$$

and recall that by the hypothesis (H2) there exists $\mu \in \mathbb{R}^{d+1}$ with $\Lambda^*(\mu) = 0$. Define ϕ_1 by its derivatives on $[0,1]$:

$$\dot{\phi}_1^i(s) = \begin{cases} \phi^i(\beta)/\beta & 0 \leq s < \beta \\ \{\phi^i(\tau) - \phi^i(\beta)\}/(\tau-\beta) & \beta \leq s < \tau \\ \mu^i & \tau \leq s \leq 1 \end{cases} \tag{47}$$

for $i = 1, \ldots, d$, and

$$\dot{\phi}_1^{d+1} = \begin{cases} \phi^{d+1}(\beta)/\beta & 0 \leq s < \beta \\ \{\phi^{d+1}(1) - \phi^{d+1}(\beta)\}/(1-\beta) & \beta \leq s \leq 1. \end{cases} \tag{48}$$

That ϕ_1 satisfies (44) is easy to check and the inequality in (45) follows from Jensen's inequality. □

Terminology. Troughout the remainder of this paper we will refer to the modifications in (47) and (48) as 'straightening'. For example, we say that ϕ_1^{d+1} is obtained from ϕ^{d+1} by straightening ϕ^{d+1} on the intervals $[0,\beta]$ and $[\beta, 1]$.

Proof of Theorem 3.1.

For $t \geq 0$ set

$$\boldsymbol{P}_n(t) = \left(\sum_{k=0}^{[nt]} \boldsymbol{X}_{-k}, \sum_{k=0}^{[nt]} C_{-k}\right). \tag{49}$$

To apply the contraction principle here, we need to work in a stonger topology than the usual projective limit topology (note that $\tilde{\boldsymbol{O}}_n$ is a function of

P_n on the entire half line). Denote by $\mathcal{A}_\mu$ the space of absolutely continuous functions $\phi : \mathbb{R}_+ \to \mathbb{R}^{d+1}$ with non-decreasing components, $\phi(0) = 0$ and

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

equipped with the topology induced by the norm

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

In [18] it is shown that if (H4) is satisfied then $\tilde{P}_n$ satisfies the LDP in $\mathcal{A}_\mu$ with good rate function given by

$$I_\infty(\varphi) = \int_0^\infty \Lambda^*(\dot{\varphi})ds.$$

Define mappings $\sigma : \mathcal{A}_\mu \to \mathbb{R}_+$, $D : \mathcal{A}_\mu \to \mathbb{R}_+$, $\beta : \mathcal{A}_\mu \to \mathbb{R}_+$ and $\Theta : \mathcal{A}_\mu \to \mathcal{B}_d$ by

$$\sigma(\varphi) = \inf\{s \geq 0 : \ A(\varphi)(s) - \phi^{d+1}(s) = \sup_t [A(\varphi)(t) - \phi^{d+1}(t)]\}, \quad (50)$$

$$D(\varphi) = \inf_{s\leq\sigma(\varphi)} [A(\varphi)(s) - \phi^{d+1}(s)] + \phi^{d+1}(\sigma(\varphi)), \quad (51)$$

$$\beta(\varphi) = \inf\{s \geq 0 : \ A(\varphi)(\sigma(\varphi)) - A(\varphi)(s) = D(\varphi)\}, \quad (52)$$

and for $0 \leq \tau \leq \beta(\varphi)$,

$$\theta(\varphi)(\tau) = A(\varphi)(\tau). \quad (53)$$

Note that $\tilde{L} = n\beta(\tilde{P}_n)$ and so $\theta(\tilde{P}_n) = \tilde{O}_n$. The following lemma will allow us to apply the contraction principle to obtain the LDP for $\tilde{O}_n$.

Lemma 4.4 *The mapping* $\Theta : \mathcal{A}_\mu \to \mathcal{B}_d$ *is continuous.*

Proof. Note that for $\varphi, \tilde{\varphi} \in \mathcal{A}_\mu$, $\beta(\varphi)$ and $\beta(\tilde{\varphi})$ are finite (since $A(\mu) < \mu^{d+1}$), and we have

$$A(\varphi)(\beta(\varphi)) = \sup_t [A(\varphi)(t) - \phi^{d+1}(t)], \quad (54)$$

and

$$A(\tilde{\varphi})(\beta(\tilde{\varphi})) = \sup_t [A(\tilde{\varphi})(t) - \tilde{\varphi}^{d+1}(t)]. \quad (55)$$

Thus, if $\|\varphi - \tilde{\varphi}\|_u < \epsilon$ for some $\epsilon > 0$, there exists $t(\varphi) > 0$, independent of $\tilde{\varphi}$, such that

$$|A(\varphi)(\beta(\varphi)) - A(\tilde{\varphi})(\beta(\tilde{\varphi}))| < 2\epsilon(1 + t). \quad (56)$$

(The value of t is chosen so that

$$\left|\frac{\phi(s)}{1+s} - \mu\right| < \epsilon,$$

for all $s > t$.) Now, since the paths $A(\bar{\varphi})$ and $A(\varphi)$ are non-decreasing, we have

$$\begin{aligned}\sum_i |A^i(\bar{\varphi})(\beta(\bar{\varphi})) - A^i(\bar{\varphi})(\beta(\varphi))| &= |A(\bar{\varphi})(\beta(\bar{\varphi})) - A(\bar{\varphi})(\beta(\varphi))| \\ &\leq |A(\bar{\varphi})(\beta(\bar{\varphi})) - A(\varphi)(\beta(\varphi))| \\ &\quad +|A(\bar{\varphi})(\beta(\varphi)) - A(\varphi)(\beta(\varphi))| \\ &< 2\epsilon(2 + t(\varphi) + \beta(\varphi)),\end{aligned}$$

and so we have

$$\begin{aligned}\sum_i |A^i(\varphi)(\beta(\varphi)) - A^i(\bar{\varphi})(\beta(\bar{\varphi}))| &\leq |A(\varphi)(\beta(\varphi)) - A(\bar{\varphi})(\beta(\varphi))| \\ &\quad +|A(\bar{\varphi})(\beta(\varphi)) - A(\bar{\varphi})(\beta(\bar{\varphi}))| \\ &< 2\epsilon(3 + t(\varphi) + 2\beta(\varphi)).\end{aligned}$$

Also, for $r < \beta(bvphi) \wedge \beta(\bar{\varphi})$,

$$\sum_i |A^i(\varphi)(r) - A^i(\bar{\varphi})(r)| < \epsilon(1 + \beta(\varphi)).$$

It follows that $d(\boldsymbol{\theta}(\varphi), \boldsymbol{\theta}(\bar{\varphi})) < 2\epsilon(2 + t(\varphi) + \beta(\varphi))$, and the statement follows. □

We now have by the contraction principle that $\tilde{\boldsymbol{O}}_n$ satisfies the LDP in $\mathcal{B}_d$ with good rate function given by

$$K(\boldsymbol{\theta}) = \inf\{I_\infty(\varphi) : \boldsymbol{\theta}(\varphi) = \boldsymbol{\theta}\}. \tag{57}$$

This can now be simplified by noting that for any $\varphi \in \mathcal{A}_{\boldsymbol{\mu}}$ with $\boldsymbol{\theta}(\varphi) = \boldsymbol{\theta}$, there exists a path φ_1 with $\boldsymbol{\theta}(\varphi_1) = \boldsymbol{\theta}$ and

$$K(\varphi) \geq K(\varphi_1) = (\sigma - \beta)\Lambda^*(\boldsymbol{x}, c) + \int_0^\beta \Lambda^*(\dot{\boldsymbol{\theta}}, \dot{\psi})$$

where $\boldsymbol{x} \in \mathbb{R}_+^d$, $\sigma > \beta > 0, c, d > 0$, $\psi \in \mathcal{A}_\beta$, $(\sigma - \beta)(x - c) = \psi(\beta)$; the path is obtained by straightening φ (see remark after proof of Corollary 2.2) on the interval $[\beta(\varphi), \sigma(\varphi)]$, and setting

$$\varphi_1(s) = \boldsymbol{\mu}(s - \sigma(\varphi)) + \varphi(\sigma(\varphi))$$

on $[\sigma(\varphi), \infty]$. $\square$

Proof of Theorem 3.3.

By the arguments in (and in the notation of) the proofs of Theorems 2.1 and 3.1, $(\tilde{\boldsymbol{P}}_n, \tilde{S}_n)$ satisfies the LDP in $\mathcal{A}_\mu \times \mathcal{A}^{d+1}$ with good rate function given by

$$I_{\infty,1}(\varphi, \phi) = I_\infty(\varphi) + I(\phi). \tag{58}$$

For $0 \leq r \leq 1$ we define the mappings $\rho_r : \mathcal{A}_\mu \times \mathcal{A}^{d+1} \to \mathcal{A}_\mu$ by

$$\rho_r(\varphi, \phi) = \begin{cases} \phi(r-s) - \phi(r) & 0 \leq s \leq r \\ \varphi(s-r) + \phi(r) & s > r, \end{cases} \tag{59}$$

and $\tilde{\boldsymbol{\Delta}} : \mathcal{A}_\mu \times \mathcal{A}^{d+1} \to \mathcal{A}^d$ by

$$\tilde{\boldsymbol{\Delta}}(\varphi, \phi)(r) = \boldsymbol{A}(\varphi)(\beta(\varphi)) + \phi(r) - \boldsymbol{A}(\rho_r(\varphi, \phi))(\beta(\rho_r(\varphi, \phi))). \tag{60}$$

Note that $\tilde{\boldsymbol{\Delta}}(\tilde{\boldsymbol{P}}_n, \tilde{S}_n) = \tilde{\boldsymbol{R}}_n$, where $\tilde{\boldsymbol{R}}_n$ is the polygonal approximation to $\boldsymbol{R}_n$. The continuity of $\tilde{\boldsymbol{\Delta}}$, and hence the LDP for $\tilde{\boldsymbol{R}}_n$, follows from Lemma 4.4; just as in the proof of Theorem 2.1, we can deduce that the LDP holds for $\boldsymbol{R}_n$ in the space $\mathcal{L}^d$, with good rate function given by

$$\tilde{I}_d(\psi) = \inf\{I_\infty(\varphi) + I(\phi) : \ \tilde{\boldsymbol{\Delta}}(\varphi, \phi) = \psi\}. \tag{61}$$

As in the proof of Corollary 2.2, the contraction principle implies that $\boldsymbol{D}_n/n$ satisfies the LDP in $\mathbb{R}^d$ with good rate function given by

$$\tilde{J}(z) = \inf\{I_\infty(\varphi) + I(\phi) : \ \tilde{\boldsymbol{\Delta}}(\varphi, \phi)(1) = z\}. \tag{62}$$

To simplify this, consider a path (φ, ϕ) with $\tilde{\boldsymbol{\Delta}}(\varphi, \phi)(1) = z$ and set

$$\beta_1 = \inf\{r > 0 : \ \Theta(\rho_r(\varphi, \phi))(1) = 0\}, \tag{63}$$

$$\beta_2 = \sup\{r \leq 1 : \ \Theta(\rho_r(\varphi, \phi))(1) = 0\}, \tag{64}$$

$$\tau = \inf\{t \leq 1 : \ \Theta(\varphi)(1) + \boldsymbol{A}(\phi)(t) = \tilde{\Delta}(\varphi, \phi)(1)\}. \tag{65}$$

If $\beta_1 \leq 1$, then $\beta_2 \leq \tau \leq 1$, and we can find a path (φ_1, ϕ_1) with $\tilde{\boldsymbol{\Delta}}(\varphi_1, \phi_1)(1) = z$ and

$$\begin{aligned} I_{\infty,1}(\varphi, \phi) &\geq I_{\infty,1}(\varphi_1, \phi_1) \\ &= \rho\Lambda^*(x, c) + \beta\Lambda^*(q/\beta, c') \\ &\quad + \beta_1\Lambda^*\left(\frac{z_1 - q}{\beta_1}, c_1\right) + \beta_2\Lambda^*(z_2/\beta_2, c_2) \\ &\quad + \tau\Lambda_a^*\left(\frac{z - z_1 - z_2}{\tau}\right) + (1 - \beta_1 - \beta_2)\Lambda_b^*\left(\frac{z - z_1 - z_2}{1 - \beta_1 - \beta_2}\right), \end{aligned}$$

for some $x, q, z_1, z_2 \in \mathbb{R}_+^d$, $\beta, \rho, c, c', c_1, c_2, \beta_1, \beta_2, \tau \in \mathbb{R}_+$, with $\rho(x-c) = \beta c'$, $\beta_1 + \beta_2 + \tau \leq 1$, and $\beta_1 c_1 = z_1$, $\beta_2 c_2 \geq z_2$. The path φ_1 is obtained by straightening φ on the intervals $[0, \beta(\varphi)]$, $[\beta(\varphi), \sigma(\varphi)]$ and setting

$$\varphi_1(s) = \mu(s - \sigma(\varphi)) + \varphi(\sigma(\varphi))$$

on $[\sigma(\varphi), \infty]$; ϕ_1 is obtained by straightening ϕ on $[0, \beta_1]$ and modifying ϕ on $[\beta_1, 1]$ in the same manner as in the proof of Corollary 2.2. The case $\beta_1 = \infty$ implies surplus queue-length at time 1, which can be removed by modifying φ in such a way that $\tilde{\Delta}(\varphi, \phi)(1)$ is unaltered and $I_{\infty,1}(\varphi, \phi)$ is reduced. $\square$

Acknowledgements. The author would like to thank the anonymous referee for constructive remarks and Gustavo de Veciana for helpful correspondence and suggestions; thanks also to Dmitri Botvich, Nick Duffield, John Lewis, Raymond Russell and Fergal Toomey for many enlightening discussions. Most of this research was carried out while the author was at the Dublin Institute for Advanced Studies, supported by grants from EOLAS and Mentec Computer Systems Ltd. under the Higher Education-Industry Cooperation Scheme, and at Trinity College Dublin.

References

[1] Dimitris Bertsimas, Ioannis Ch. Paschalidis and John N. Tsitsiklis. On the large deviations behaviour of acyclic networks of G/G/1 queues. LIDS Report: LIDS-P-2278, 1994.

[2] Cheng-Shang Chang. Stability, queue length and delay of deterministic and stochastic queueing networks. *IEEE Trans. on Automatic Control* 39:913–931, 1994.

[3] Cheng-Shang Chang. Approximations of ATM networks: effective bandwidths and traffic descriptors. Submitted.

[4] Cheng-Shang Chang, Philip Heidelberger, Sandeep Juneja and Perwez Shahabuddin. Effective Bandwidth and Fast Simulation of ATM Intree Networks. *Performance Evaluation* 20:45–66, 1994.

[5] C.-S. Chang and T. Zajic. Effective bandwidths of departure processes from queues with time varying capacities. INFOCOM, 1995.

[6] C. Courcoubetis, G. Kesidis, A. Ridder, J. Walrand and R. Weber. Admission control and routing in ATM networks using inferences from measured buffer occupancy. To appear in *IEEE Trans. Comm.*

[7] Amir Dembo and Tim Zajic. Large deviations: from empirical mean and measure to partial sums process. *Stoch. Proc. Appl.* 57:191–224, 1995.

[8] Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications.* Jones and Bartlett, London, 1992.

[9] G. de Veciana, C. Courcoubetis and J. Walrand. Decoupling bandwidths for networks: a decomposition approach to resource management. Memorandum No. UCB/ERL M93/50, University of California.

[10] N.G. Duffield, J.T. Lewis, Neil O'Connell, Raymond Russell and Fergal Toomey. Entropy of ATM traffic streams: a tool for estimating QoS parameters. Research Report DIAS-APG-9417. *IEEE Journal of Selected Areas in Communications*, special issue on Advances in the Fundamentals of Networking, 13(6):981–990.

[11] N.G. Duffield and Neil O'Connell. Large deviations and overflow probabilities for the general single server queue, with applications. *Proc. Camb. Phil. Soc.* 118(1), 1995.

[12] N.G. Duffield and Neil O'Connell. Large deviations for arrivals, departures, and overflow in some queues of interacting traffic. *Proceedings of the 11th IEE Teletraffic Symposium,* Cambridge, March 1994.

[13] Paul Dupuis and Richard S. Ellis. The large deviation principle for a general class of queueing systems, I. *Trans. Amer. Math. Soc.*, to appear.

[14] A. Ganesh and V. Anantharam. Stationary tail probabilities in exponential server tandems with renewal arrivals. In: *Stochastic Networks*, Frank P. Kelly and Ruth J. Williams, eds., Springer-Verlag, 1995.

[15] A. Ganesh and Neil O'Connell. The linear geodesic property is not generally preserved by a FIFO queue. BRIMS Technical Report HPL-BRIMS-96-006.

[16] Peter W. Glynn and Ward Whitt. Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *J. Appl. Prob.*, to appear.

[17] F.P. Kelly and P.B. Key. Dimensioning playout buffers from an ATM network. *Proceedings of the 11th IEE Teletraffic Symposium,* Cambridge, March 1994.

[18] Neil O'Connell (1996). Stronger topologies for sample path large deviations in Euclidean space. BRIMS Technical Report HPL-BRIMS-96-005.

[19] Shyam Parekh and Jean Walrand. A quick simulation method for excessive backlogs in networks of queues. *IEEE Trans. Aut. Contr.* 34:54–66, 1989.

[20] O. Yaron and M. Sidi. Performance and stability of communications networks via robust exponential bounds. *IEEE/ACM Trans. Networking,* 1(3):372–385, June 1993.