Queue Lengths and Departures at Single-Server Resources

Neil O'Connell
Basic Research Institute in the
Mathematical Sciences
HP Laboratories Bristol
HPL-BRIMS-96-04
February, 1996

In this paper I will review and illustrate some large deviation results for queues with interacting traffic, both for shared buffer and shared capacity models. These results are examples of a general scheme which can be applied to an endless 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.

QUEUE LENGTHS AND DEPARTURES AT SINGLE-SERVER RESOURCES

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

BRIMS Technical Report HPL-BRIMS-96-004 [1]

Abstract

In this paper I will review and illustrate some large deviation results for queues with interacting traffic, both for shared buffer and shared capacity models. These results are examples of a general scheme which can be applied to an endless 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.

[1] To appear in the Proceedings of the Royal Statistical Society Research Workshop on Stochastic Networks, Edinburgh, 1995.

1 Introduction

In this paper I will review and illustrate some large deviation results for queues with interacting traffic, both for shared buffer and shared capacity models. These results are examples of a general scheme which can be applied to an endless 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: I will begin by motivating such a scheme and briefly describing how it works.

We will 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 continuous function of the partial sums process corresponding to X. To make this more precise, for $t \geq 0$ set

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

and write $\tilde{S}_n$ for the polygonal approximation to S_n:

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

For $\mu \in \mathbb{R}^d$, denote by $\mathcal{A}_\mu(\mathbb{R}_+)$ the space of absolutely continuous paths $\phi : \mathbb{R}_+ \to \mathbb{R}^d$, with $\phi(0) = 0$ and limits

$$\lim_{t\to\infty} \frac{\phi(t)}{1+t} = \mu,$$

equipped with the topology induced by the norm

$$\|\phi\|_u = \sup_t \left|\frac{\phi(t)}{1+t}\right|. \tag{3}$$

Our supposition is that there exists a continuous function $f : \mathcal{A}_\mu(\mathbb{R}_+) \to \mathcal{X}$, for some Hausdorff topological space $\mathcal{X}$, such that $O_n = f(\tilde{S}_n)$, for each n. (Note that we are also implicitly assuming that $\tilde{S}_n \in \mathcal{A}_\mu(\mathbb{R}_+)$, for each n.)

For example, suppose $d = 1$ and X_k is the amount of work arriving at time $-k$ at a single-server queue with constant service capacity $c > 0$. Suppose also that the limit

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

exists almost surely and is less than c. The queue length at time zero is given by

$$Q_0 = \sup_{n \geq 0} \sum_{k=0}^{n} (X_k - c), \tag{4}$$

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

$$f(\phi) = \sup_{t>0} [\phi(t) - ct]. \tag{5}$$

It is easy to check that f is a continuous function.

Why is this a useful supposition? To answer this, we need to introduce some large deviation theory.

Let $\mathcal{X}$ be a Hausdorff topological space with Borel σ-algebra $\mathcal{B}$, and let μ_n be a sequence of probability measures on $(\mathcal{X}, \mathcal{B})$. We say that μ_n satisfies the *large deviation principle* (LDP) with rate function I, if for all $B \in \mathcal{B}$,

$$-\inf_{x \in B^\circ} I(x) \leq \liminf_n \frac{1}{n} \log \mu_n(B) \leq \limsup_n \frac{1}{n} \log \mu_n(B) \leq -\inf_{x \in \bar{B}} I(x); \tag{6}$$

if, for each n, Z_n is a realisation 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.

A useful tool in large deviation theory is the *contraction principle*. This states that if Z_n satisfies the LDP in a Hausdorff topological space $\mathcal{X}$ with good rate function I, and f is a continuous mapping from $\mathcal{X}$ into another Hausdorff topological space $\mathcal{Y}$, then the sequence $f(Z_n)$ satisfies the LDP in $\mathcal{Y}$ 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 $\mathcal{A}[0,1]$ the subspace of absolutely continuous functions on $[0,1]$ with $\phi(0) = 0$. Dembo and Zajic (1995) establish quite general conditions for which $\tilde{S}_n[0,1]$ satisfies the LDP in $\mathcal{A}[0,1]$ with good convex rate function given by

$$I(\phi) = \begin{cases} \int_0^1 \Lambda^*(\dot{\phi}) ds & \phi \in \mathcal{A}[0,1] \\ \infty & \text{otherwise,} \end{cases} \tag{7}$$

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

$$\Lambda(\lambda) = \lim_{n\to\infty} \frac{1}{n} \log E e^{n\lambda \cdot S_n(1)}, \tag{8}$$

which is assumed to exist for each $\lambda \in \mathbb{R}^{d+1}$ 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 $Ee^{\lambda \cdot X_1}$ exists and is finite everywhere; this is a classical result, due to Varadhan (1966) and Mogulskii (1976). This 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 (5) is not continuous in this topology on any supporting subspace, and so the contraction principle does not apply. This has motivated the consideration of stronger topologies by Dobrushin and Pechersky (1995) and O'Connell (1996). In the latter it is proved that if the LDP holds in $C[0,1]$ and Λ is differentiable at the origin with $\nabla\Lambda(0) = \mu$, then the LDP holds in the space $\mathcal{A}_\mu(\mathbb{R}_+)$ with the topology induced by the norm (3), and with good rate function given by

$$I(\phi) = \int_0^\infty \Lambda^*(\dot{\phi}) ds.$$

As we remarked earlier, the function f defined by (5) is continuous in this topology, provided $\mu < c$.

Getting back to our network problem we see that under very general conditions on the input process, if the objects of interest can be written as $O_n = f(\tilde{S}_n)$, for some continuous f, we have an LDP for O_n with rate function given by

$$J(y) = \inf\left\{ \int_0^\infty \Lambda^*(\dot{\phi}) ds : \ f(\phi) = y \right\}. \tag{9}$$

This will provide probability approximations for O_n. However, for it to be useful, we must first simplify the rate function J (as it stands, it is an infinite-dimensional optimisation problem). This is where we use the convexity of Λ^*: combined with Jensen's inequality it allows us to restrict our consideration to a set of piecewise linear paths that depends on f and the problem becomes finite-dimensional.

To illustrate this, consider the single-server queue with arrivals process (X_k) and constant capacity $c > 0$: if $\tilde{S}_n$ satisfies the LDP in $\mathcal{A}_\mu(\mathbb{R}_+)$ with

good convex rate function given by

$$I(\phi) = \int_0^\infty \Lambda^*(\dot{\phi})ds,$$

then the normalised queue length at time zero, Q_0/n, satisfies the LDP in $\mathbb{R}_+$ with good rate function

$$\begin{aligned} J(q) &= \inf\left\{\int_0^\infty \Lambda^*(\dot{\phi})ds : \sup_{t>0}[\phi(t) - ct] = q\right\} \\ &= \inf_{\tau>0}\inf\left\{\int_0^\tau \Lambda^*(\dot{\phi})ds : \phi(\tau) - c\tau = q\right\} \\ &= \inf_{\tau>0} \tau\Lambda^*(c + q/\tau). \end{aligned}$$

This fact has previously been demonstrated by several authors, under similar conditions (Chang, 1994; de Veciana et al, 1993; Duffield and O'Connell, 1995; Glynn and Whitt, 1994).

Finally, why is all this potentially useful? Because it is very general, and rate functions can (in principle) be estimated from real traffic observations: see, for example, Courcoubetis et al (1994) or Duffield et al (1995) for more about the estimation problem.

The outline of the paper is as follows. In Section 2, we present the LDP for departures of traffic streams from an initially empty shared buffer with stochastic service capacity; in Section 3 we present an equilibrium version of this result, along with an LDP for the state of the system in equilibrium. In Section 4 we consider a system with dedicated buffers, served with weighted priority by a single server; in Section 5 we consider the problem of optimal resource allocation in such a system, and present some surprising results.

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

2 Departures from a shared buffer

Suppose we have d arrival streams $X = (X^1, \ldots, X^d)$ 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

$$A_n = \sum_{k=1}^{n} X_k, \qquad B_n = \sum_{k=1}^{n} C_k. \tag{10}$$

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

$$Q_n = (Q_{n-1} + X_n - C_n)^+, \tag{11}$$

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

$$D_n^{(t)} = \hat{A}_n - Q_n, \tag{12}$$

or, equivalently,

$$D_n^{(t)} = \inf_{0 \le k \le n} (\hat{A}_k - B_k) + B_n. \tag{13}$$

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

$$T_n = \sup\{k \le n : \ \hat{A}_k \le D_n^{(t)}\}, \tag{14}$$

$$D_n^i = A_{T_n}^i + (D_n^{(t)} - \hat{A}_{T_n}) X_{T_n+1}^i / \hat{X}_{T_n+1}. \tag{15}$$

Note that $D_n^{(t)} = \hat{D}_n = D_n^1 + \cdots + D_n^d$. In words, work is serviced in the order received and simultaneous arrivals from each source are thoroughly mixed in the queue.

For $0 \le t \le 1$, set

$$S_n(t) = \left(\frac{1}{n} A_{[nt]}, \frac{1}{n} B_{[nt]} \right), \tag{16}$$

and write $\tilde{S}_n$ for the polygonal approximation to S_n. The following is a slight modification of Corollary 2.2 in (O'Connell, 1994).

Theorem 2.1 *Suppose the sequence of partial sums $\tilde{S}_n$ satisfies the LDP in $\mathcal{A}_\mu(\mathbb{R}_+)$ with good convex rate function given by*

$$I(\phi) = \int_0^1 \Lambda^*(\dot{\phi}) ds, \tag{17}$$

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

$$\Lambda(\lambda) = \lim_{n \to \infty} \frac{1}{n} \log E e^{n\lambda \cdot S_n(1)}, \tag{18}$$

which is assumed to exist for each $\lambda \in \mathbb{R}^{d+1}$ *as an extended real number, and* $\nabla\Lambda(0) = \mu$. *Suppose also that* Λ^* *is of the form*

$$\Lambda^*(x,c) = \Lambda_a^*(x) + \Lambda_b^*(c), \tag{19}$$

for $(x,c) \in \mathbb{R}^d \times \mathbb{R}$. *Then* D_n/n *satisfies the LDP in* $\mathbb{R}_+^d$ *with good rate function given by*

$$\begin{aligned}\Lambda_d^*(z) \;=\; & \inf\{\beta\Lambda_a^*(x/\beta) + \sigma\Lambda_a^*\left(\frac{z-x}{\sigma}\right) + \beta\Lambda_b^*(c) + (1-\beta)\Lambda_b^*\left(\frac{\hat{z}-\hat{x}}{1-\beta}\right) : \\ & \beta,\sigma \in [0,1],\ c \in \mathbb{R},\ \beta+\sigma \le 1,\ \ x \in \mathbb{R}_+^d,\ \ \hat{x} \le \beta c\}.\end{aligned} \tag{20}$$

De Veciana, Walrand and Courcoubetis (1994) showed that under similar hypotheses, with the arrivals assumed to be independent ($\Lambda_a^*(z) = \Lambda_1^*(z^1) + \cdots + \Lambda_d^*(z^d)$, say) and service assumed to be constant ($C_n = c$, say), the sequence of departures corresponding to the first stream (D_n^1/n) satisfies the LDP in $\mathbb{R}_+$ with rate function $\Lambda_{D^1}^*$ which is equal to Λ_1^* on the interval $[\mu_1, c - \mu_2 - \cdots - \mu_d]$, where $\mu_i = \Lambda_i'(0)$: a full description of $\Lambda_{D^1}^*$ can be obtained from (20) by taking an infimum over the second and subsequent variables.

Theorem 2.1 also generalises the one-dimensional results of De Veciana, Walrand and Courcoubetis (1994), Chang *et al* (1994) and Duffield and O'Connell (1994).

A natural question to ask, if one were hoping to consider the departure process as an arrival process at a subsequent queue and iterate the results, is whether the departure process satisfies the same hypotheses as the arrival process. The answer is that this is not generally the case (Ganesh and O'Connell, 1996).

3 Equilibrium results for a shared buffer

In the last section we assumed that the buffer was initially empty. In the case of a single input ($d = 1$), Chang and Zajic (1995) prove a stationary version of Theorem 2.1 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. In this section we present the LDP 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 $\{(X_k, C_k) : k \in \mathbb{Z}\}$ is a stationary, ergodic sequence in $\mathbb{R}_+^d \times \mathbb{R}_+$, with $E\hat{X}_1 < EC_1$ for stability. It is convenient to define cumulative arrivals and service on intervals: set

$$A_{k,l} = \sum_{j=k+1}^{l} X_j, \qquad B_{k,l} = \sum_{j=k+1}^{l} C_j; \tag{21}$$

we will write A_n for $A_{0,n}$ and B_n for $B_{0,n}$. As before, for $0 \le t \le 1$ we set

$$S_n(t) = \left(\frac{1}{n}A_{[nt]}, \frac{1}{n}B_{[nt]}\right), \tag{22}$$

and write $\tilde{S}_n$ for the polygonal approximation to S_n. The (total) amount of work in the queue at time $n \in \mathbb{Z}$ is given by

$$Q_n^{(t)} = \sup_{k \ge 0}(\hat{A}_{n-k,n} - B_{n-k,n}). \tag{23}$$

The (total) departures during the interval $(k, l]$ is given by

$$D_{k,l}^{(t)} = \hat{A}_{k,l} + Q_k^{(t)} - Q_l^{(t)}. \tag{24}$$

Set

$$K = \inf\{k \ge 0 : \ Q_{-k}^{(t)} = 0\}, \tag{25}$$

and note that $Q_0^{(t)} = \hat{A}_{-K,0} - B_{-K,0}$.

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

$$L = -\sup\{k \le 0 : \ \hat{A}_{-K,k} \le D_{-K,0}^{(t)}\}$$

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

$$D_{-K,0}^i = A_{-K,-L}^i + \epsilon^i,$$

where

$$\epsilon^i = (D_{-K,0}^{(t)} - \hat{A}_{-K,-L})X_{1-L}^i / \hat{X}_{1-L}.$$

Note that $\hat{D}_{-K,0} = D^1_{-K,0} + \cdots + D^d_{-K,0} = D^{(t)}_{-K,0}$.

To describe the state of the system at time 0, we consider the following $\mathbb{R}^d_+$-valued process: set $N(L) = A_{-L,0} - \epsilon$, and for $k = 1, \ldots, L-1$ set $N(k) = A_{-k,0}$. 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; $\hat{N}(L) = Q_0$, and $N^i(L)$ is the amount of work of type i in the queue at time 0. For clarity we will write Q_0 for $N(L)$; Q_n, the respective amounts of work in the queue at any other time n, is defined similarly. Write $\tilde{O}_n$ for the polygonal approximation to $\{A_{-nt,0}/n,\ t \geq 0\}$ on the interval $[0, \tilde{L}]$, where $n\tilde{L} = L - \epsilon/X_{1-L}$. Note that $\tilde{O}_n(\tilde{L}) = Q_0/n$. To state the LDP for $\tilde{O}_n$, we need to define a suitable path space. 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; now set

$$\mathcal{B}_k = \{\theta \in \mathcal{C}^k_\tau :\ \tau > 0\}, \tag{26}$$

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

$$d(\theta_1, \theta_2) = \sup_{0 \leq t \leq \tau_1 \wedge \tau_2} \sum_{i=1}^k |\theta^i_1(t) - \theta^i_2(t)| + \sum_{i=1}^k |\theta^i_1(\tau_1) - \theta^i_2(\tau_2)|, \tag{27}$$

for $\theta_1 \in \mathcal{C}^k_{\tau_1}$, $\theta_2 \in \mathcal{C}^k_{\tau_2}$.

Theorem 3.1 *Under the hypotheses of Theorem 2.1, $\tilde{O}_n$ satisfies the LDP in $\mathcal{B}_d$ with good rate function given by*

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

Corollary 3.2 *Under the hypotheses of Theorem 2.1, Q_0/n satisfies the LDP in $\mathbb{R}^d_+$ with good rate function*

$$L(q) = \inf\{\rho\Lambda^*(x,c) + \beta\Lambda^*(q/\beta, \rho(\hat{x} - c)/\beta) :\ x \in \mathbb{R}^d_+,\ \beta, \rho, c \in \mathbb{R}_+\}. \tag{28}$$

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

$$D_n = Q_0 + A_n - Q_n. \tag{29}$$

Theorem 3.3 *Under the hypotheses of Theorem 2.1, D_n/n satisfies the LDP in $\mathbb{R}_+^d$ with good rate function given by*

$$\begin{aligned}\tilde{\Lambda}_d^*(z) &= \inf\{L(q) + \beta_1\Lambda^*\left(\frac{z_1 - q}{\beta_1}, \frac{\hat{z}_1}{\beta_1}\right) + \beta_2\Lambda^*(z_2/\beta_2, c_2) \\ &\quad +\tau\Lambda_a^*\left(\frac{z - z_1 - z_2}{\tau}\right) + (1 - \beta_1 - \beta_2)\Lambda_b^*\left(\frac{\hat{z} - \hat{z}_1 - \hat{z}_2}{1 - \beta_1 - \beta_2}\right) : \\ &\quad q, z_1, z_2 \in \mathbb{R}_+^d,\ c_2, \beta_1, \beta_2, \tau \in \mathbb{R}_+,\ \beta_1 + \beta_2 + \tau \leq 1, \\ &\quad \beta_2 c_2 \geq \hat{z}_2\}.\end{aligned}$$

Proofs of the above results can be found in (O'Connell, 1994).

Again, a natural question to ask here is whether the departure process satisfies the same hypotheses assumed to hold for the arrivals; again, generally not is the answer (Ganesh and O'Connell, 1996). There is, however, one situation where it is the case, namely if the arrivals processes are independent Poisson processes and the service times are exponential; then the departure processes are also independent Poisson processes.

4 Queue lengths at a system with dedicated buffers and shared service capacity

Consider a single-server queue with two inputs (X_n^1) and (X_n^2) and constant service capacity c shared between the inputs according to a weighted priority scheme with weights $p_1 + p_2 = 1$. To be more precise, X_n^1 and X_n^2 are sequences of non-negative random variables and, starting with an empty system, the respective queue lengths at time n are defined recursively by the equations

$$\begin{aligned}Q_n^1 &= (Q_{n-1}^1 + X_n^1 - \max(c - Q_{n-1}^2 - X_n^2, p_1 c))^+ \\ Q_n^2 &= (Q_{n-1}^2 + X_n^2 - \max(c - Q_{n-1}^1 - X_n^1, p_2 c))^+\end{aligned}$$

with $Q_0^1 = Q_0^2 = 0$. We will write $Q_n = (Q_n^1, Q_n^2)$.

For each n define a path $S_n : [0,1] \to \mathbb{R}_+^2$ by

$$S_n(t) = \left(\frac{1}{n}\sum_{k=1}^{[nt]} X_k^1, \frac{1}{n}\sum_{k=1}^{[nt]} X_k^2\right), \tag{30}$$

and denote its polygonal approximation by $\tilde{S}_n$. For $\lambda \in \mathbb{R}^2$ set

$$\Lambda(\lambda) = \lim_{n\to\infty} \frac{1}{n} \log E e^{\lambda \cdot S_n(1)}, \tag{31}$$

whenever this limit exists. Write Λ^* for the convex dual of Λ. Assuming $\tilde{S}_n$ satisfies the LDP in $\mathcal{A}_\mu(\mathbb{R}_+)$, where $\mu = \nabla\Lambda(0)$, with good rate function given by

$$I(\phi) = \int_0^\infty \Lambda^*(\dot{\phi})ds,$$

and we can write

$$Q_n/n = \Pi(\tilde{S}_n),$$

for some continuous function Π, we have by the contraction principle that the sequence Q_n/n satisfies the LDP in $\mathbb{R}_+^2$ with good rate function given by

$$J(q) = \inf\{I(\phi) : \ \Pi(\phi) = q\}, \tag{32}$$

and the first queue length Q_n^1/n satisfies the LDP in $\mathbb{R}_+$ with good rate function given by

$$L(q) = \inf\{I(\phi) : \ \Pi(\phi)^1 = q\}. \tag{33}$$

The mapping Π is formally defined in (O'Connell, 1994a), where the following simplifications of J and L are obtained for the case where the inputs are assumed to be independent:

$$\Lambda(\lambda) = \Lambda_1(\lambda_1) + \Lambda_2(\lambda_2).$$

Set $\mu_i = \Lambda_i'(0)$.

Theorem 4.1 *In the above setting:*

(a) If $\mu_i \le p_i c$ $(i = 1, 2)$,

$$J(a) = \inf\{\tau\Lambda^*(x, y) + \tau'\Lambda^*(x', y') : \ (\tau, \tau', x, x', y, y') \in E(a)\}, \tag{34}$$

where $E(a) = E_1(a) \cup E_2(a)$ and

$$\begin{aligned} E_1(a) \ = \ & \{(\tau, \tau', x, x', y, y') \in \mathbb{R}_+^6 : \ \tau + \tau' \le 1, \ y \le p_2 c, \\ & \tau(x + y - c) + \tau'(x' - p_1 c) = a_1, \ \tau'(y' - p_2 c) = a_2\}, \end{aligned}$$

$$\begin{aligned} E_2(a) \ = \ & \{(\tau, \tau', x, x', y, y') \in \mathbb{R}_+^6 : \ \tau + \tau' \le 1, \ x \le p_1 c, \\ & \tau(x + y - c) + \tau'(y' - p_2 c) = a_2, \ \tau'(x' - p_1 c) = a_1\}. \end{aligned}$$

(b) If $\mu_1 + \mu_2 \le c$ and $\mu_2 \le p_2 c$ then

$$L(a) = \inf\{\tau\Lambda^*(c - y + a/\tau, y) : \ 0 \le \tau \le 1, \ y \le p_2 c\}.$$

(c) If $\mu_1 + \mu_2 \leq c$ *and* $\mu_2 \geq p_2 c$ *then*

$$L(a) = \inf_{0 \leq \tau \leq 1} \tau \Lambda_1^*(p_1 c + a/\tau).$$

This compliments and extends results of de Veciana and Kesidis (1993) and Bertsimas, Paschilidis and Tsitilidis (1995), where the the tail asymptotics for the limiting distribution of Q_n^1 are obtained in the ergodic case; of Weber (1995) on the large deviation principle for queue lengths in a similar system with state-dependent service; and of Ignatyul *et al* (1993), Borovkov and Mogulskii (1995), on the large deviation behaviour of random walks in a two-dimensional quadrant. See also (Dupuis and Ellis, 1994) and references therein, for related work.

This can also be extended to the equilibrium case, where the LDP holds with rate functions given by the expressions in Theorem 4.1 without the restrictions $\tau + \tau' \leq 1$ for case (a) and $\tau \leq 1$ for cases (b) and (c).

5 Resource allocation

Suppose we have two buffers of sizes an and $(1-a)n$ (n is large and $0 < a < 1$) and service capacity c per unit time distributed between the buffers with respective priority weights p and $1-p$. The two input streams are independent, and are characterised by their rate functions Λ_1^* and Λ_2^*. How should we allocate service capacity and buffer space—that is, how should we choose p and a—in order to minimise the overall frequency of buffer-overflow? Well, we can approximate the overall frequency of overflow by

$$P(Q^1 > an \text{ or } Q^2 > (1-a)n),$$

where Q^1 and Q^2 are the queue lengths at an infinite buffer version of the system in equilibrium. Applying the principle of the largest term and an equilibrium version of Theorem 4.1, we have

$$P(Q^1 > an \text{ or } Q^2 > (1-a)n) \approx e^{-\delta(a,p)n},$$

where $\delta(a,p) = [a\delta_1(p)] \wedge [(1-a)\delta_2(p)]$ and

$$\delta_1(p) = \inf_{\tau \geq 0, x_2 \geq 0} \tau[\Lambda_1^*(1/\tau + (pc) \vee (c - x_2)) + \Lambda_2^*(x_2)],$$

$$\delta_2(p) = \inf_{\tau \geq 0, x_1 \geq 0} \tau[\Lambda_2^*(1/\tau + ((1-p)c) \vee (c - x_1)) + \Lambda_1^*(x_1)].$$

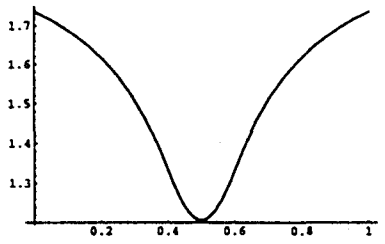

Figure 1: Plot of $\delta(a^*(p), p)$ against p for the parameter values $\mu_1 = \mu_2 = 0.4$, $\sigma_1^2 = \sigma_2^2 = 0.1$, $c = 1$. The optimal policy has $p = 1$ and $a = 0.13$.

The problem of minimising the overall frequency of overflow is thus approximately equivalent to the problem of maximising $\delta(a, p)$ with respect to a and p. For fixed p, this is achieved by setting

$$a = a^*(p) = \frac{\delta_1(p)}{\delta_1(p) + \delta_2(p)},$$

yielding

$$\delta(a^*(p), p) = \frac{\delta_1(p)\delta_2(p)}{\delta_1(p) + \delta_2(p)};$$

thus, to maximise $\delta(a, p)$ we should choose $p = p^*$ to minimise

$$\frac{\delta_1(p)\delta_2(p)}{\delta_1(p) + \delta_2(p)}$$

and set $a^* = a^*(p^*)$.

For example, suppose $\Lambda_i^*(x) = (x - \mu_i)^2/2\sigma_i^2$, $\mu_1 + \mu_2 < c$. Then, after some straightforward calculations, we get

$$\delta_2(p) = \begin{cases} \frac{2(c-\mu_1-\mu_2)}{\sigma_1^2+\sigma_2^2} & pc > \frac{2\sigma_1^2(c-\mu_1-\mu_2)}{\sigma_1^2+\sigma_2^2} + \mu_1, \\ 2\frac{(1-p)c-\mu_2}{\sigma_2^2} & pc \le \mu_1 \\ \frac{(1-p)c-\mu_2}{\sigma_2^2} + \frac{1}{\sigma_2}\sqrt{\left(\frac{(1-p)c-\mu_2}{\sigma_2}\right)^2 + \left(\frac{pc-\mu_1}{\sigma_1}\right)^2} & \text{otherwise} \end{cases}$$

and a similar expression for $\delta_1(p)$. Figures 1–3 are plots of $\delta(a^*(p), p)$

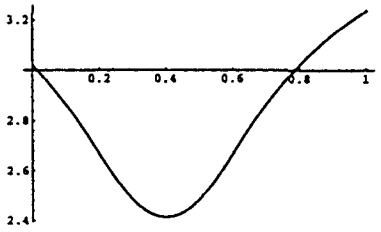

Figure 2: Plot of $\delta(a^*(p), p)$ against p for the parameter values $\mu_1 = 0.2$, $\mu_2 = 0.4$, $\sigma_1^2 = \sigma_2^2 = 0.1$, $c = 1$. The optimal policy has $p = 1$ and $a = 0.075$.

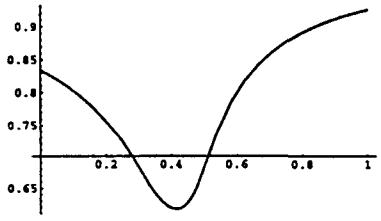

Figure 3: Plot of $\delta(a^*(p), p)$ against p for the parameter values $\mu_1 = \mu_2 = 0.4$, $\sigma_1^2 = 0.1$, $\sigma_2^2 = 0.3$, $c = 1$. The optimal policy has $p = 1$ and $a = 0.19$.

against p for different parameter values. To interpret these, recall that the optimal policy, if one wishes to minimise the overall frequency of overflow, is to choose p in order to maximise $\delta(a^*(p), p)$ (and take $a = a^*(p^*)$). In the case where the input streams have the same mean and variance (Figure 1), the optimal policy is to give top priority to either one of the streams, and about 87% of the buffer space to the other. This may seem surprising. Note however, that this is not a fair policy: the stream with top priority will typically experience shorter waiting times. In the cases where the first stream has a higher mean (Figure 2), or a higher variance (Figure 3), the optimal policy gives top priority to the second stream and more buffer space to the first.

References

[1] A.A. Borovkov and A.A. Mogulskii (1995). Large deviations for stationary Markov chains in a quarter plane. Preprint.

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

[3] Dimitris Bertsimas, Ioannis Ch. Paschalidis and John N. Tsitsiklis (1995). RSS Research Workshop in Stochastic Networks, Edinburgh, August 1995.

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

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

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

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

[8] 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.*

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

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

[11] G. de Veciana, C. Courcoubetis and J. Walrand (1993). Decoupling bandwidths for networks: a decomposition approach to resource management. Memorandum No. UCB/ERL M93/50, University of California.

[12] G. de Veciana and G. Kesidis (1993). Bandwidth allocation for multiple qualities of service using generalised processor sharing. Preprint.

[13] R.L. Dobrushin and E.A. Pechersky (1995). Large deviations for random processes with independent increments on infinite intervals. Preprint.

[14] N.G. Duffield, J.T. Lewis, Neil O'Connell, Raymond Russell and Fergal Toomey (1995). Entropy of ATM traffic streams: a tool for estimating QoS parameters. *IEEE Journal of Selected Areas in Communications* 13(6):981–990.

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

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

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

[18] Paul Dupuis and Richard S. Ellis (1994). Large deviation analysis of queueing systems. To appear in the Proceedings of the IMA workshop "Stochastic Networks, February 28 - March 4, 1994", F. Kelly and R. Williams, eds. Springer-Verlag.

[19] A. Ganesh and Neil O'Connell (1996). The linear geodesic property is not generally preserved by a FIFO queue. In preparation.

[20] Peter W. Glynn and Ward Whitt (1995). Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *J. Appl. Prob.*, to appear.

[21] I.A. Ignatyuk, V. Malyshev and V.V. Scherbakov (1993). Boundary effects in large deviation problems. Preprint.

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

[23] Neil O'Connell (1994). Large deviations for departures from a shared buffer. *J. Appl. Prob.*, to appear. (Revised version of 'Large deviations in queueing networks', DIAS Technical Report DIAS-APG-9413.)

[24] Neil O'Connell (1994a). Large deviations for queue lengths at a multi-buffered resource. *J. Appl. Prob.*, to appear.

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

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

[27] Richard Weber (1995). Estimation of overflow probabilities for state-dependent service of traffic streams with dedicated buffers. RSS Research Workshop in Stochastic Networks, Edinburgh, August 1995.