Large Deviations for Queue Lengths at a Multi-Buffered Resource

Neil O'Connell
Basic Research Institute in the
Mathematical Sciences
HP Laboratories Bristol
HPL-BRIMS-96-10
February, 1996

In this paper we obtain the large deviation principle for scaled queue lengths at a multi-buffered resource, and simplify the corresponding variational problem in the case where the inputs are assumed to be independent.

1 Introduction and summary

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 $Q_n = (Q_n^1, Q_n^2)$.

The idea is to consider the queue lengths to be a function of the inputs and deduce the large deviation properties of the latter from those of the former. To do this, it is convenient to introduce some notation.

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

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

and its polygonal approximation by

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

For $\lambda \in \mathbb{R}^2$ set

$$\Lambda(\lambda) = \lim_{n \to \infty} \frac{1}{n} \log E e^{\lambda \cdot S_n(1)}, \tag{3}$$

whenever this limit exists. Write Λ^* for the convex dual of Λ. Denote by $\mathcal{A}$ the space of absolutely continuous functions ϕ on $[0,1]$ with $\phi(0) = 0$. Dembo and Zajic [2] establish very general conditions for which $\tilde{S}_n$ satisfies the LDP in $\mathcal{A}^2$ with respect to the uniform topology, with good convex rate function given by

$$I(\phi) = \int_0^1 \Lambda^*(\dot{\phi}) ds. \tag{4}$$

For such an LDP to hold in the *i.i.d.* case (see, for example, [3, Chapter 5]) it is sufficient that the scaled cumulant generating function (3) exists and is finite everywhere; this is a classical result, due to Varadhan and Mogulskii. If $\tilde{S}_n$ satisfies the LDP with rate function given by (4) and we can write

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

for some continuous function f, then we have by the contraction principle [2, Theorem 4.2.1] that the sequence Q_n/n satisfies the LDP in $\mathbb{R}_+^2$ with good rate function given by

$$J(a) = \inf\{I(\phi) : \ f(\phi) = a\}, \tag{5}$$

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

$$L(a) = \inf\{I(\phi) : \ f_1(\phi) = a\}, \tag{6}$$

where f_1 denotes the first component of f. The mapping f is formally defined as follows. For each t, define a mapping $\Pi_t : \mathcal{A}^2 \to \mathbb{R}_+^2$ by letting q be the solution to the pair of integral equations $(i \neq j)$

$$q_i(t) = \int_0^t \left[\dot{\phi}_i(s) - \left[p_i c 1_{(q_j>0)} + (c - \dot{\phi}_2) 1_{(q_j=0)}\right] 1_{(q_i>0)}\right] ds, \tag{7}$$

and setting $\Pi_t(\phi) = q(t)$. Existence and uniqueness of a solution can be verified by first considering piecewise linear arguments where solutions can be constructed recursively on intervals, then checking that Π_t is continuous on the space of piecewise linear functions with respect to the total variation norm topology and using the fact that piecewise linear functions are dense in $\mathcal{A}^2$ with respect to that topology. (By piecewise linear we mean linear at all but finitely many points; the fact that these are dense in $\mathcal{A}^2$ follows from the fact that simple functions are dense in $L_1[0,1]$, the space of integrable ($\mathbb{R}^2$-valued) functions on the unit interval, with respect to the norm

$$\|f\|_1 = \int_0^1 |f(s)| ds,$$

(see, for example, [8]) and $\mathcal{A}^2$ is isomorphic to $L_1[0,1]$, when equipped with the total variation norm

$$\|\phi\|_{TV} = \|\dot{\phi}\|_1.)$$

Now set $f = \Pi_1$ and note that $f(\tilde{S}_n) = Q_n/n$, as required. The above construction is useful for describing the dynamics of the system; it is not clear, however, that f is continuous. This can be checked using the following representation for the mapping f, which was recently obtained by Toomey [14]. If f_i denotes the i^{th} component of f, and $i \neq j$, then

$$f_i(\phi) = \inf_{0<t<1} \sup_{0<s<1} [w_i(\phi)(s,1) + w_j(\phi)(s,t)],$$

where

$$w_i(\phi)(u,v) = \phi_i(v) - \phi_i(u) - p_i c(v-u),$$

for $u < v$, and zero otherwise.

Our main result provides simple expressions for the rate functions J and L in 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 1 *In the above setting:*

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

$$J(a) = \inf\{\tau\Lambda^*(x,y) + \tau'\Lambda^*(x',y') : \ (\tau,\tau',x,x',y,y') \in E(a)\}, \quad (8)$$

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

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

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

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

This compliments and extends results of de Veciana and Kesidis [4] where the the tail asymptotics for the limiting distribution of Q_n^1 are obtained in the ergodic case; of Weber [15] on the large deviation principle for queue lengths in a similar system with state-dependent service; and of Ignatyul *et al* [10], Borovkov and Mogulskii [1], on the large deviation behaviour of random walks in a two-dimensional quadrant. See also Dupuis and Ellis [6, 7], and references therein, for related work.

The nature of the optimisation in (a) can be deduced from the random walk results in [10, 1]. There exist angles $0 < \gamma_1 < \gamma_2 < \pi/2$ such that if the argument of a is less than γ_1, $E_1(a)$ dominates ('the first queue starts building up first'), and if the argument of a is greater than γ_2, $E_2(a)$ dominates ('the second queue starts building up first'); otherwise, $\tau' = 0$ for the optimal path, which can be interpreted as 'both queues start building up at the same time'. The critical values γ_1 and γ_2 satisfy a complicated differential equation, involving functionals of Λ^*.

Theorem 1 can also be extended to the equilibrium case, where the LDP holds with rate functions given by the expressions in Theorem 1 without the restrictions $\tau + \tau' \leq 1$ for case (a) and $\tau \leq 1$ for cases (b) and (c). To check continuity in this case, one can use the stronger topology introduced in [13], for sample paths indexed by the half-line, to take care of the fact that the queue lengths in equilibrium depend on the entire history of the input processes, just as in [11], where equilibrium results for departures from a shared buffer are obtained. (The usual projective limit topologies are generally not strong enough to prove continuity, even in the case of a single input; see Dobrushin and Pechersky [5] or O'Connell [13] for further discussion on this point.)

The results of this paper have been applied to answer some basic questions on optimal resource allocation in [12]. For example, it is demonstrated there that for the above model, with two identical Gaussian sources and a fixed (large) amount of total buffer space to allocate between the inputs, the (asymptotic) overall frequency of overflow is minimised when top priority is given to either of the streams, and more buffer space allocated to the other.

2 Proof of Theorem

A fact that we will use repeatedly throughout the proof is that for any convex function g on $\mathbb{R}^d$, and $\phi \in \mathcal{A}^d$,

$$\int_s^t g(\dot{\phi})ds \geq (t-s)g\left(\frac{\phi(t)-\phi(s)}{t-s}\right). \tag{9}$$

(See, for example, [9, Theorem 4.1].)

(a) We wish to simplify the expression

$$J(a) = \inf\left\{\int_0^1 \Lambda^*(\dot{\phi})ds : \ f(\phi) = a\right\}.$$

To do this, consider a particular $\phi \in \mathcal{A}^2$ with $f(\phi) = a$ and set $q(t) = \Pi_t(\phi)$. Note that $q(1) = a$. The reader may find it easier to interpret the arguments that follow if one thinks of $q(t)$ as the vector of queue lengths at time t, and $\phi(t)$ as the vector of total arrivals up to time t. Consider the last time buffer i was empty:

$$\beta_i = \sup\{t \geq 0 : \ q_i(t) = 0\}.$$

First suppose $\beta_1 \leq \beta_2$. Set

$$\sigma_1 = \sup\{t \leq \beta_1 : \ q_2(t) = 0\},$$

$$\sigma_2 = \inf\{t \geq \beta_1 : \ q_2(t) = 0\}.$$

Define a new path $\phi^{(1)} \in \mathcal{A}^2$ by its derivatives on $[0,1]$:

$$\dot{\phi}_1^{(1)}(s) = \begin{cases} \mu_1 & 0 \leq s < \beta_1 \\ \dot{\phi}_1(s) & \beta_1 \leq s \leq 1 \end{cases}$$

$$\dot{\phi}_2^{(1)}(s) = \begin{cases} \mu_2 & 0 \leq s < \beta_1 \\ p_2 c & \beta_1 \leq s < \sigma_2 \\ \dot{\phi}_2(s) & \sigma_2 \leq s \leq 1 \end{cases}$$

Then $f(\phi^{(1)}) = f(\phi) = a$ and $I(\phi^{(1)}) \leq I(\phi)$. To see the latter, note that since q_2 is non-negative on $[\sigma_1, \sigma_2]$ we have

$$\phi_2(\sigma_2) - \phi_2(\sigma_1) \geq (\sigma_2 - \sigma_1) p_2 c;$$

we also have $\mu_2 \leq p_2 c$ by hypothesis, so by (9),

$$\begin{aligned} \int_{\sigma_1}^{\sigma_2} \Lambda_2^*(\dot{\phi}_2) &\geq (\sigma_2 - \sigma_1)\Lambda_2^*\left(\frac{\phi_2(\sigma_2) - \phi_2(\sigma_1)}{\sigma_2 - \sigma_1}\right) \\ &\geq (\sigma_2 - \sigma_1)\Lambda_2^*(p_2 c) \\ &\geq (\beta_1 - \sigma_1)\Lambda_2^*(\mu_2) + (\sigma_2 - \beta_1)\Lambda_2^*(p_2 c). \end{aligned}$$

The next step is to define $\phi^{(2)}$ by linearly interpolating $\phi^{(1)}$ on the intervals $[\beta_1, \beta_2]$ and $[\beta_2, 1]$:

$$\dot{\phi}^{(2)}(s) = \begin{cases} \mu & 0 \leq s < \beta_1 \\ \frac{\phi^{(1)}(\beta_2) - \phi^{(1)}(\beta_1)}{\beta_2 - \beta_1} & \beta_1 \leq s < \beta_2 \\ \frac{\phi^{(1)}(1) - \phi^{(1)}(\beta_2)}{1 - \beta_2} & \beta_2 \leq s \leq 1 \end{cases}$$

Then $I(\phi^{(2)}) \leq I(\phi^{(1)})$, again by (9), and $f(\phi^{(2)}) = a$. The case $\beta_1 > \beta_2$ is treated similarly. We have thus shown that for any ϕ with $f(\phi) = a$, there exists a less expensive path $\phi^{(2)}$ (that is, $I(\phi^{(2)}) \leq I(\phi)$) with the same outcome $f(\phi^{(2)}) = a$. Also, there exist times τ and τ' with $\tau + \tau' \leq 1$ (these are

$$\tau = \min(\beta_1, \beta_2)$$

and

$$\tau' = \max(\beta_1, \beta_2) - \tau)$$

with the property that for some $\tau + \tau' < 1$ one of the components of $\phi^{(2)}$ is linear on the intervals $[0, \tau]$ and $[\tau, 1]$ and the other is linear on the intervals $[0, \tau]$, $[\tau, \tau']$ and $[\tau', 1]$; finally, the cost $\phi^{(2)}$ on the interval $[0, \tau]$ is zero. We can therefore restrict the class of paths in the optimisation to paths with the above properties, leading to the expression for J given in the statement of the theorem.

(b) We simplify L in exactly the same manner. Consider a particular $\phi \in \mathcal{A}^2$ with $f_1(\phi) = a$ and denote by q the corresponding solution to the integral equations (7); note that $q_1(1) = a$. Set

$$\beta_1 = \sup\{t \geq 0 : \ q_1(t) = 0\},$$

$$\sigma_1 = \sup\{t \leq \beta_1 : \ q_2(t) = 0\},$$

$$\sigma_2 = \inf\{t \geq \beta_1 : \ q_2(t) = 0\}.$$

Then the path $\phi^{(2)}$, defined as in part (a), has $I(\phi^{(2)}) \leq I(\phi^{(1)})$ and $f_1(\phi^{(2)}) = a$.

(c) As before, consider a particular $\phi \in \mathcal{A}^2$ with $f_1(\phi) = a$ and denote by q the corresponding solution to the integral equations (7); set

$$\beta_1 = \sup\{t \geq 0 : \ q_1(t) = 0\}.$$

In this case we define a new path $\phi^{(1)}$ by

$$\dot{\phi}_1^{(1)}(s) = \begin{cases} \mu_1 & 0 \leq s < \beta_1 \\ \frac{\phi_1(1) - \phi_1(\beta_1)}{1 - \beta_1} & \beta_1 \leq s \leq 1 \end{cases}$$

$$\dot{\phi}_2^{(1)}(s) = \mu_2, \qquad 0 \leq s < 1.$$

Then

$$f_1(\phi^{(1)}) = \phi_1(1) - \phi_1(\beta_1) - (1-\beta_1)p_1 c \geq a$$

and $I(\phi^{(1)}) \leq I(\phi)$. Now define $\phi^{(2)}$ by

$$\dot{\phi}_1^{(2)}(s) = \begin{cases} \mu_1 & 0 \leq s < \beta_1 \\ \frac{a}{1-\beta_1} + p_1 c & \beta_1 \leq s \leq 1 \end{cases}$$

and $\phi_2^{(2)} = \phi_2^{(1)}$. Note that $f_1(\phi^{(2)}) = a$. Since $\mu_1 \leq p_1 c$, by hypothesis,

$$\Lambda_1^*\left(\frac{a}{1-\beta_1} + p_1 c\right) \leq \Lambda_1^*\left(\frac{\phi_1(1) - \phi_1(\beta_1)}{1-\beta_1}\right),$$

from which it follows that $I(\phi^{(2)}) \leq I(\phi^{(1)})$, and the proof is complete. □

Acknowledgements. Part of this work was carried out while the author was at the Dublin Institute for Advanced Studies, supported by grants from EOLAS and Mentec Computer Systems Limited, under the Higher Education-Industry Cooperation Scheme, and at Trinity College Dublin.

References

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

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

[3] Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications.* Jones and Bartlett, 1993.

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

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

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

[7] Paul Dupuis and Richard S. Ellis. 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.

[8] Avner Friedman. *Foundations of Modern Analysis.* Dover, 1982.

[9] Jack Gunson. Inequalities in mathematical physics. In: *Inequalities*, W. Norrie Everitt, ed. Marcel Dekker Inc., 1991.

[10] I.A. Ignatyuk, V. Malyshev and V.V. Scherbakov. Boundary effects in large deviation problems. Preprint.

[11] Neil O'Connell. Large deviations for departures from a shared buffer. Submitted to *J. Appl. Prob.*

[12] Neil O'Connell. Queue lengths and departures at single-server resources. To appear in the *Proceedings of the RSS Workshop on Stochastic Networks*, Edinburgh, 1995.

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

[14] Fergal Toomey. Large deviations for shared resource systems: reflection mapping approach. DIAS Technical Report.

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