The Annals of Applied Probability
2001, Vol. 11, No. 2, 379–404

SAMPLE PATH LARGE DEVIATIONS FOR QUEUES WITH MANY INPUTS[1]

BY DAMON J. WISCHIK

Statistical Laboratory Cambridge

This paper presents a large deviations principle for the average of real-valued processes indexed by the positive integers, one which is particularly suited to queueing systems with many traffic flows. Examples are given of how it may be applied to standard queues with finite and infinite buffers, to priority queues and to finding most likely paths to overflow.

1. Introduction. Consider a queue fed by several different input processes. Many quantities of interest in queueing theory, such as the amount of work in the queue, can be expressed as functions of the sequence of variables $(x_t)_{t\in\mathbb{N}}$, where x_t is the total amount of work received t time steps ago.

The sequence (x_t) will typically live in a space on which the quantity of interest is a continuous function. For example, let $\mathscr{X}_\mu$ be the space of real-valued sequences $\mathbf{x} = (x_t)$ for which $t^{-1}\sum_{i=1}^{t} x_i < \mu$ eventually. Then the amount of work Q in a queue with an infinite buffer and fixed service rate $C > \mu$ is given by

$$Q(\mathbf{x}) = \left[\sup_{t>0}\left(\sum_{i=1}^{t} x_i - Ct\right)\right]^+.$$

There is a simple topology on $\mathscr{X}_\mu$, which we call the uniform topology, that makes Q continuous.

The principal result of this paper is a large deviations principle (LDP) for a sequence of random processes $\mathbf{X}^L$ in $\mathscr{X}_\mu$ equipped with the uniform topology. This can be used to understand the large deviations behavior of a wide range of queueing systems. Consider a sequence of queueing systems, in which the Lth system has input $\mathbf{X}^L$. We will use the contraction principle to deduce, from the LDP for $\mathbf{X}^L$, LDP's for various quantities such as $Q(\mathbf{X}^L)$.

In this paper we will be motivated by one particular limiting regime, in which $\mathbf{X}^L$ is the average of L processes. This is known in queueing theory as the *many sources asymptotic*, and was described in an early paper of Weiss (1986). It is well suited to modern telecommunications networks, in which a switch may have hundreds of different input flows. Another limiting regime which has been widely studied is the *large buffer asymptotic*, in which $\mathbf{X}^L$ is a speeded-up version of a base process $\mathbf{X}$. We will see that large

Received June 1998; revised February 2000.

[1] This work was carried out under a Research Studentship from the UK Engineering and Physical Sciences Research Council.

AMS 2000 *subject classifications*. Primary 60K25; secondary 60F10, 60G17, 60K30.

Key words and phrases. Effective bandwidth, sample path large deviations, many sources, priority queues, paths to overflow.

deviations in this regime can often be found as a special case of the many sources regime.

The rest of this paper is in two parts. In Section 2, the sample path large deviations principle for $\mathbf{X}^L$ is established. O'Connell (1997a) has proved a sample path large deviations principle for the large buffer regime, and the proof given here for the many sources regime is similar. We also give several examples of processes satisfying the sample path LDP, including fractional Brownian motion.

In Section 3, the sample path LDP is used together with the contraction principle to study large deviations in three different queueing problems: standard queues with finite and infinite buffers, likely paths to overflow and priority queues. There are many other possible applications; for example, it is used by Wischik (1999) in studying the output of a queue. Several authors have used this approach to study large deviations under the large buffer regime; we will see that under the many sources regime, large deviations often possess a richer structure.

2. Large deviations for averages of processes.

We will be concerned with the set $\mathscr{X}$ of real-valued processes indexed by the natural numbers $\{1, 2, \ldots\}$. Throughout this paper, t will represent a natural number. Denote a process in $\mathscr{X}$ by $\mathbf{x}(0, \infty)$, and its truncation to the set $\{s+1 \cdots t\}$ by $\mathbf{x}(s, t]$ for $s < t$. When the meaning is unambiguous, $\mathbf{x}(0, \infty)$ and $\mathbf{x}(0, t]$ may be written as $\mathbf{x}$. Let $\mathbf{1}$ be the constant process taking value 1 at each time step. Denote by x_t the value of the process at time t, and by $x(s, t]$ the cumulative process $x(s, t] = \sum_{i=s+1}^{t} x_i$, with $x(t, t] = 0$.

Consider a sequence of random processes $(\mathbf{X}^L)_{L\in\mathbb{N}}$, where each $\mathbf{X}^L$ takes values in $\mathscr{X}$. We will prove results about the limit of the $\mathbf{X}^L$: the principal result of this section is a sample path large deviations principle for the $\mathbf{X}^L$. It will be helpful to think of $\mathbf{X}^L$ as the average of L independent, identically distributed processes. However, despite the title of this section, we will (for the moment) assume only that the $\mathbf{X}^L$ take values in $\mathscr{X}$.

It should be explained here what is meant by a large deviations principle. For a full introduction to the theory, and details of the tools and definitions we will be using, see Dembo and Zeitouni (1993). A sequence of random variables X^L in a Hausdorff space $\mathscr{X}$ with σ-algebra $\mathscr{B}$ is said to satisfy a large deviations principle (LDP) with good rate function I if, for any $B \in \mathscr{B}$,

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

where $I: \mathscr{X} \to \mathbb{R}^+ \cup \{\infty\}$ has compact level sets. If X is a process, this is called a sample path LDP. The left- and right-hand sides of this inequality are referred to as the large deviations lower and upper bounds. To avoid measure-theoretic complications, we will assume throughout this paper that $\mathscr{B}$

contains the Borel σ-algebra, and thus that all the open and closed sets we will be interested in are measurable.

We want to find a sample path LDP in a space appropriate for queueing applications. This will be done in four steps. The first step is to find an LDP for finite truncations of the process. If $\mathbf{X}^L$ is the average of L processes, a finite truncation is just the average of L vectors, and there are standard tools for dealing with this. The next step is to extend the LDP to the entire process. This is done by taking projective limits, again a standard step. The third step takes most of the work. Many queueing functions of interest are not continuous with respect to the projective limit topology, so we need to strengthen the LDP to a more appropriate topology. O'Connell (1997a) has introduced a suitable topology: that given by the *uniform norm*

$$\|\mathbf{x}\| = \sup_{t>0} \left| \frac{x(0,t]}{t} \right|. \tag{1}$$

As well as choosing this finer topology, we need to restrict the LDP by incorporating a notion of stability; this is the final step.

We will find conditions under which $\mathbf{X}^L$ satisfies an LDP, with the uniform topology, and with good rate function

$$\mathbf{I}(\mathbf{x}) = \sup_{t>0} \sup_{\boldsymbol{\theta}\in\mathbb{R}^t} \boldsymbol{\theta}\cdot\mathbf{x}(0,t] - \boldsymbol{\Lambda}_t(\boldsymbol{\theta}), \tag{2}$$

where $\boldsymbol{\Lambda}_t(\boldsymbol{\theta})$ is the moment generating function

$$\lim_{L\to\infty} \frac{1}{L} \log \mathbb{E} \exp\bigl(L\boldsymbol{\theta}\cdot\mathbf{X}^L(0,t]\bigr).$$

An LDP *for truncated sequences.* The following lemma establishes an LDP for any finite truncation of the process. It is a direct restatement of the Gärtner–Ellis theorem for the average of vectors in $\mathbb{R}^t$ [see Dembo and Zeitouni (1993), Theorem 2.3.6].

ASSUMPTION 1 (Finite-time regularity). *Define the logarithmic moment generating function $\boldsymbol{\Lambda}_t^L(\boldsymbol{\theta})$ for $\boldsymbol{\theta}\in\mathbb{R}^t$ by*

$$\boldsymbol{\Lambda}_t^L(\boldsymbol{\theta}) = \frac{1}{L} \log \mathbb{E} \exp\bigl(L\boldsymbol{\theta}\cdot\mathbf{X}^L(0,t]\bigr).$$

Assume that for each t and $\boldsymbol{\theta}$ the limiting moment generating function

$$\boldsymbol{\Lambda}_t(\boldsymbol{\theta}) = \lim_{L\to\infty} \boldsymbol{\Lambda}_t^L(\boldsymbol{\theta})$$

exists as an extended real number, and that the origin belongs to the interior of the effective domain of $\boldsymbol{\Lambda}_t$. Assume further that $\boldsymbol{\Lambda}_t$ is an essentially smooth, lower semicontinuous function.

LEMMA 1. *Under Assumption* 1, *for any fixed t, the sequence* $\mathbf{X}^L(0,t]$ *satisfies an* LDP *with good rate function*

$$\Lambda_t^*\big(\mathbf{x}(0,t]\big) = \sup_{\boldsymbol{\theta}\in\mathbb{R}^t} \boldsymbol{\theta}\cdot\mathbf{x}(0,t] - \Lambda_t(\boldsymbol{\theta}).$$

EXAMPLE 1 (Many sources). Let $\mathbf{X}^L$ be the average of L independent copies of the process $\mathbf{X}$. Then

$$\Lambda_t(\boldsymbol{\theta}) = \Lambda_t^L(\boldsymbol{\theta}) = \log\mathbb{E}\exp(\boldsymbol{\theta}\cdot\mathbf{X}(0,\ t]).$$

This example should be borne in mind, because it is the motivation behind all the following results.

The projective limit. Now we extend the LDP from finite truncations $\mathbf{X}(0,t]$ to the full process $\mathbf{X}(0,\infty)$. We need a little more care than this in stating the result, because the definition of the large deviations principle relies on open and closed sets and there are several useful topologies on the space of processes $\mathscr{X}$. We will use the topology of projective limits (i.e., the topology of pointwise convergence of sequences). The following lemma is a direct application of the Dawson–Gärtner theorem for projective limits [see Dembo and Zeitouni (1993), Theorem 4.6.1].

LEMMA 2. *Under Assumption* 1, *the sequence* $\mathbf{X}^L$ *satisfies an* LDP *in* $\mathscr{X}$ *under the topology of pointwise convergence, with good rate function*

$$\mathbf{I}(\mathbf{x}) = \sup_t \Lambda_t^*\big(\mathbf{x}(0,t]\big). \tag{3}$$

Strengthening the topology. The topology of pointwise convergence is not directly useful for many queueing applications. For example, if x_t is the amount of work arriving at a queue at time $-t$, and the queue is served at constant rate C, then the queue size at time 0 is

$$Q(\mathbf{x}) = \sup_{t\geq 0} x(0,t] - Ct$$

and this function is not continuous with respect to the topology of pointwise convergence. To see this, set $x_t^L = C$ for $t < L$, $x_L^L = C+1$ and $x_t^L = 0$ for $t > L$. Then $\mathbf{x}^L$ converges pointwise to the constant process of rate C, for which $Q = 0$, but $Q(\mathbf{x}^L) = 1 \not\to 0$. The answer is to show that the LDP holds in a finer topology.

The uniform topology (1) defined above allows one to analyze a wide range of queueing problems. The idea is that it controls what happens over very large timescales. Under an additional assumption on the large timescale behavior of the process $\mathbf{X}^L$, we can show that the sample path LDP of Lemma 2 can be extended to this topology.

The results in Section 3 do not actually need a topology as strong as the uniform topology. The only properties of the topology they use are that it is stronger than the projective limit topology, and that it makes the queue size function continuous. There are weaker topologies that have these two properties, such as the *weak queue topology* used in Wischik (1999), defined by the metric

$$d(\mathbf{x}, \mathbf{y}) = |Q(\mathbf{x}) - Q(\mathbf{y})| + \sum_{t=1}^{\infty} \frac{1 \wedge |x_t - y_t|}{2^t}.$$

But the uniform topology is easier to work with, so we will use it in what follows.

ASSUMPTION 2 (Large timescale characteristics). *A* scaling function *is a function v: $\mathbb{N} \to \mathbb{R}$ for which $v(t)/\log t \to \infty$. For some scaling function v, define the scaled cumulant moment generating function*

$$\Lambda_t^L(\theta) = \frac{1}{v(t)} \boldsymbol{\Lambda}_t^L\Big(\frac{\mathbf{1}\theta v(t)}{t}\Big)$$

for $\theta \in \mathbb{R}$. From Assumption 1, for each t there is an open neighborhood of the origin in which the limit

$$\Lambda_t(\theta) = \lim_{L\to\infty} \Lambda_t^L(\theta)$$

exists. Assume that there is an open neighborhood of the origin in which these limits and the limit

$$\Lambda(\theta) = \lim_{t\to\infty} \Lambda_t(\theta)$$

exist uniformly in θ.

We also know from Assumption 1 *that for θ in some open neighborhood of the origin, the limit $\Lambda_t^L(\theta) - \Lambda_t(\theta) \to 0$ is uniform as $L \to \infty$. Assume that for θ in some open neighborhood of the origin, the limit*

$$\sqrt{\frac{v(t)}{\log t}}\Big(\Lambda_t^L(\theta) - \Lambda_t(\theta)\Big) \to 0 \tag{4}$$

is uniform in θ as $t, L \to \infty$.

THEOREM 3 (Sample path LDP for process averages). *Suppose $\mathbf{X}^L$ satisfies Assumptions* 1 *and* 2. *Then it satisfies an* LDP *in the space of real-valued sequences $\mathcal{X}$ equipped with the uniform topology* (1), *with good rate function* $\mathbf{I}$ *given in* (3).

EXAMPLE 2 (Many sources). In the case of Example 1, when $\mathbf{X}^L$ is the average of L independent processes with common distribution $\mathbf{X}$, the uniformity of the limit (4) is guaranteed, since $\boldsymbol{\Lambda}_t^L = \boldsymbol{\Lambda}_t$.

PROOF OF THEOREM 3. The processes $\mathbf{X}^L$ take values in the set $\mathscr{X}$ of real-valued sequences. Write $(\mathscr{X}, p)$ for $\mathscr{X}$ equipped with the projective limit topology, and $(\mathscr{X}, \|\cdot\|)$ for $\mathscr{X}$ equipped with the uniform topology. The identity map from $(\mathscr{X}, \|\cdot\|)$ to $(\mathscr{X}, p)$ is continuous; we assume as usual that all open and closed sets in $(\mathscr{X}, \|\cdot\|)$ are measurable; and we know that $\mathbf{X}^L$ satisfies an LDP in $(\mathscr{X}, p)$ with rate function $\mathbf{I}$. So, by the inverse contraction principle [see Dembo and Zeitouni (1993), Theorem 4.2.4], if $\mathbf{X}^L$ is exponentially tight in $(\mathscr{X}, \|\cdot\|)$, then it satisfies an LDP in $(\mathscr{X}, \|\cdot\|)$ with the same rate function.

It remains to show that $\mathbf{X}^L$ is exponentially tight in $(\mathscr{X}, \|\cdot\|)$: in other words, that there exist compact sets K_α in $(\mathscr{X}, \|\cdot\|)$ such that

$$\lim_{\alpha\to\infty} \limsup_{L\to\infty} \frac{1}{L} \log \mathbb{P}(\mathbf{X}^L \notin K_\alpha) = -\infty.$$

Choose the sets K_α as follows. For each t, let $\mu_t = \Lambda_t'(0)$, let $d_t = \sqrt{\log t / v(t)}$, let

$$K_\alpha(t) = \left\{ \mathbf{x} \in \mathscr{X} : \frac{x(0,t]}{t} \in [\mu_t - \alpha d_t, \mu_t + \alpha d_t] \right\}$$

and choose

$$K_\alpha = \bigcap_{t\in\mathbb{N}} K_\alpha(t).$$

Exponential tightness with these K_α will be shown in the following two lemmas. □

LEMMA 4. *The sets K_α are compact in the uniform topology.*

PROOF. Because we are working in a metric space, it suffices to show that the sets K_α are sequentially compact. So, let $\mathbf{x}^k$ be a sequence of processes. Since the T-dimensional truncation of $\bigcap_{t\le T} K_\alpha(t)$ is compact in $\mathbb{R}^T$, the intersection K_α is compact under the projective topology. That is, there is a subsequence $\mathbf{x}^{j(k)}$ which converges pointwise, say to $\mathbf{x}$. It remains to show that $\mathbf{x}^j \to \mathbf{x}$ under the uniform topology.

Given any ε, since $d_t \to 0$ as $t \to \infty$, we can find t_0 such that, for $t \ge t_0$, $2d_t\alpha < \varepsilon$. And since $\mathbf{x}$ and all the $\mathbf{x}^j$ are in K_α,

$$\sup_{t\ge t_0} \left| \frac{x^j(0,t]}{t} - \frac{x(0,t]}{t} \right| < \varepsilon.$$

Also, since the $\mathbf{x}^j$ converge pointwise, there exists a j_0 such that, for $j \ge j_0$,

$$\sup_{t< t_0} \left| \frac{x^j(0,t]}{t} - \frac{x(0,t]}{t} \right| < \varepsilon.$$

Putting these two together gives the result. □

LEMMA 5.

$$\lim_{\alpha\to\infty}\limsup_{L\to\infty}\frac{1}{L}\log\mathbb{P}(\mathbf{X}^L\notin K_\alpha)=-\infty.$$

PROOF. First, note that if

$$\lim_{\alpha\to\infty}\limsup_{L\to\infty}L^{-1}\log y_\alpha^L=-\infty,$$

and the same is true of z_α^L, then it is also true of $y_\alpha^L+z_\alpha^L$, by the principle of the largest term.

Also, note that

$$\mathbb{P}(\mathbf{X}^L\notin K_\alpha)\le\sum_t\mathbb{P}\big(X^L(0,t]/t>\mu_t+\alpha d_t\big)+\sum_t\mathbb{P}\big(X^L(0,t]/t<\mu_t-\alpha d_t\big).$$

We will adopt the strategy of breaking the infinite sums up into several parts: several finite-timescale parts and a long-timescale infinite part. Finite-timescale parts are easy to deal with individually, and we can control the behavior of $\mathbf{X}^L$ over long timescales. This strategy is also at the core of proofs for related large deviations results, proved directly by Courcoubetis and Weber (1996) and Botvich and Duffield (1995).

First, fix t and consider $\limsup_L L^{-1}\log\mathbb{P}(X^L(0,t]/t>\mu_t+\alpha d_t)$. By Chernoff's bound,

$$\mathbb{P}\big(X^L(0,t]/t>\mu_t+\alpha d_t\big)\le\exp\big[-Lv(t)\big(\theta(\mu_t+\alpha d_t)-\Lambda_t^L(\theta)\big]\big)$$

for any $\theta>0$. So the expression we are interested in is bounded above by $\limsup_L -v(t)(\theta(\mu_t+\alpha d_t)-\Lambda_t^L(\theta))$. Choosing any θ for which $\Lambda_t(\theta)$ is finite, it is clear that this quantity tends to $-\infty$ as $\alpha\to\infty$.

Now for the remaining terms. We have assumed that the limits $\Lambda_t^L(\theta)\to\Lambda_t(\theta)$ and $\Lambda_t(\theta)\to\Lambda(\theta)$ exist uniformly in θ in an open neighborhood of the origin. Since Λ_t^L is a cumulant moment generating function, it has a power series expansion, and so the coefficients in the power series also converge. Let $\Lambda_t^L(\theta)=\theta\mu_t^L+\frac{1}{2}\theta^2s_t^L+O(\theta^3)$, and denote the coefficients of Λ_t and Λ by dropping the superscripts and subscripts appropriately.

For fixed t_0, consider the remaining terms

$$\lim_{\alpha\to\infty}\limsup_{L\to\infty}\frac{1}{L}\log\sum_{t\ge t_0}\exp\big[-Lv(t)\big(\theta(\mu_t+\alpha d_t)-\Lambda_t^L(\theta)\big)\big].\tag{5}$$

Assume for the moment that $s>0$, and pick θ depending on L and t: $\theta_t^L=(d_t+\varepsilon_t^L)/s_t^L$, where $\varepsilon_t^L=\mu_t-\mu_t^L$. This gives as the typical exponent

$$-Lv(t)\left[\left\{\frac{\big(d_t+\varepsilon_t^L\big)^2}{2s_t^L}+O\big(d_t+\varepsilon_t^L\big)^3\right\}+\frac{\alpha-1}{s_t^L}d_t\big(d_t+\varepsilon_t^L\big)\right].$$

Because of our assumption on the uniformity of convergence (4), there exists a t_0 and L_0 such that, for $t\ge t_0$ and $L\ge L_0$, θ_t^L is positive; and because $d_t\to 0$, the term in brackets $\{\cdot\}$ is also positive. (If $s = 0$, pick $\theta_t^L = d_t + \varepsilon_t^L$; then the same conclusion holds.)

So the typical exponent in (5) is bounded above by

$$-Lv(t)\left[\frac{\alpha-1}{s_t^L}d_t\big(d_t+\varepsilon_t^L\big)\right]$$

for sufficiently large t and L. Indeed, for sufficiently large t and L we can bound it by $-Lv(t)\kappa(\alpha-1)d_t^2$ for some constant $\kappa > 0$. Therefore, by our choice of d_t, for t_0 sufficiently large, expression (5) is bounded above by

$$\lim_{\alpha\to\infty}\limsup_{L\to\infty}\frac{(\alpha-1)\kappa}{L}\log\sum_{t\geq t_0}t^{-L}.$$

It is easy to check that this is equal to $-\infty$. □

Stability. We have achieved the goal of a sample path LDP for averages of processes. But it is still not directly useful for queueing applications, because the queue size function is still not continuous, even with respect to the finer topology. The problem is that there is no notion of stability. If the mean arrival rate is higher than the service rate, the queue will be unstable. Mathematically speaking, the queue size function is only continuous on the subspace of processes for which the mean arrival rate is less than the service rate. The following theorem shows that the sample path LDP holds in this restricted space. First, we must define the mean rate of the arrivals.

DEFINITION 3 (Mean rate). Define the *mean rate* of the $\mathbf{X}^L$ to be the derivative $\Lambda'(0)$.

We will also explain here what we mean by stationarity. We do not need this definition immediately, but as it crops up again and again, it will be useful to give it now.

DEFINITION 4 (Stationarity). Say that $\mathbf{X}^L$ is *stationary* if the limiting moment generating functions $\mathbf{\Lambda}_t$ correspond to a stationary process. Note that if $\mathbf{X}^L$ is stationary, then the mean rate is equal to $t^{-1}\mathbf{\Lambda}_t'(\theta\mathbf{1})$ for all t, where the derivative is taken at $\theta = 0$.

THEOREM 6. *Under Assumptions* 1 *and* 2, *the* LDP *of Theorem* 3 *holds on the space $\mathscr{X}_\mu$, which has the uniform topology and is given by*

$$\mathscr{X}_\mu = \left\{\mathbf{x}\in\mathscr{X} : \frac{x(0,t]}{t}\leq\mu \quad \textit{eventually}\right\}$$

for any μ greater than the mean rate of the $\mathbf{X}^L$.

PROOF. By Dembo and Zeitouni [(1993), Lemma 4.1.5] it suffices to show that $\{\mathbf{x} : \mathbf{I}(\mathbf{x}) < \infty\} \subset \mathscr{X}_\mu$, and for L sufficiently large, $\mathbb{P}(\mathbf{X}^L \in \mathscr{X}_\mu) = 1$.

Recall that $\mathbf{I}(\mathbf{x}) = \sup_{\mathbf{t}} \mathbf{\Lambda}^*_{\mathbf{t}}(\mathbf{x}(\mathbf{0}, \mathbf{t}])$. Let $\mu = \Lambda'(0) + \varepsilon$, and pick $\theta > 0$ such that $\Lambda(\theta) < \theta(\mu - \frac{1}{2}\varepsilon)$. Now if $x(0,t]/t > \mu$, then, for sufficiently large t,

$$\mathbf{\Lambda}^*_t(\mathbf{x}(0,t]) = \sup_{\boldsymbol{\theta}} \boldsymbol{\theta} \cdot \mathbf{x}(0,t] - \mathbf{\Lambda}_t(\boldsymbol{\theta}) \geq \theta v(t)\left(\frac{x(0,t]}{t} - \left(\mu - \frac{1}{2}\varepsilon\right)\right) \geq \frac{1}{2}\theta v(t)\varepsilon.$$

So if $\mathbf{x} \notin \mathscr{X}_\mu$, then this inequality holds for infinitely many t, and since $v(t)$ is unbounded, $\mathbf{I}(\mathbf{x}) = \infty$.

Second, since $\Lambda^L_t(\theta) \to \Lambda_t(\theta)$ uniformly for t sufficiently large and $\Lambda_t(\theta) \to \Lambda(\theta)$, there exists $\theta > 0$ such that, for L and t sufficiently large, $\Lambda^L_t(\theta) < \theta(\mu - \frac{1}{2}\varepsilon)$. Then, by Chebyshev's inequality,

$$\sum_{t=1}^{\infty} \mathbb{P}\left(\frac{X^L(0,t]}{t} > \mu\right) \leq \sum_{t=1}^{\infty} \exp\left[-Lv(t)\left(\theta\mu - \Lambda^L_t(\theta)\right)\right],$$

which is finite for L sufficiently large. So, by the Borel–Cantelli lemma, $\mathbb{P}(\mathbf{X}^L \in \mathscr{X}_\mu) = 1$. □

This result will be used to study the large deviations behavior of a variety of queueing systems. Some of the systems can easily be studied directly. But the indirect route, via the sample path LDP, can give more insight. It also means there is less additional work for each different application.

2.1. *Examples.* We have already given the example of the many sources asymptotic, in which $\mathbf{X}^L$ is the average of L independent processes. We now give three more examples. The first shows how large buffer results can be obtained from the same theorems (though they usually turn out to have a less rich structure).

EXAMPLE 3 (Large buffer). Given a base process $\mathbf{X}$, let $X^L(0,t] = f(L)^{-1} \times X(0, f(L)t]$. This is the *large buffer asymptotic* regime. For a variety of processes $\mathbf{X}$, it is possible to choose a normalizing function $f(L)$ such that Assumption 1 is satisfied. Often, the normalizing function is just $f(L) = L$, and the limit $\mathbf{\Lambda}_t$ has the simple linear form $\mathbf{\Lambda}_t(\boldsymbol{\theta}) = \sum_{i=1}^t \mathbf{\Lambda}_1(\theta_i)$. For an account of conditions under which this occurs, see Dembo and Zajic (1995). In Example 5, the normalizing function is not linear and $\mathbf{\Lambda}_t$ has a more complicated form.

Suppose for now that $\mathbf{\Lambda}_t$ has the simple linear form: this gives as the rate function $\mathbf{I}(\mathbf{x}) = \sum_t \mathbf{\Lambda}^*_1(x_t)$. Then Assumption 2 is satisfied. To see this, choose $v(t) = t$, so that $\Lambda(\theta) = \mathbf{\Lambda}_1(\theta)$. Since $\Lambda^L_t(\theta)$ is given by

$$\Lambda^L_t(\theta) = \frac{1}{Lt} \log \mathbb{E} \exp\left(\theta X(0, Lt]\right),$$

and we have assumed that this converges as $L \to \infty$, we can by choosing t and L sufficiently large make $\Lambda^L_t(\theta) - \Lambda_t(\theta)$ arbitrarily small. Thus the limit (4) is uniform as $t, L \to \infty$. O'Connell (1997a) describes sample path large deviations under the large buffer asymptotic in more detail.

The next two examples are of fractional Brownian motion, a process with long-range dependence, by which we mean that the sum of covariance coefficients $\sum_{i=0}^{\infty}\mathrm{Cov}(X_0, X_i)$ is infinite. This makes it both appealing as a model for Internet traffic, since this phenomenon has been observed empirically by Leland, Taqqu, Willinger and Wilson (1994) and others, and also a problem for the standard large buffer asymptotic. But under the many sources asymptotic, it looks just like any other process.

EXAMPLE 4 (Fractional Brownian motion with many sources). As an illustration of the many sources asymptotic, let $\mathbf{X}^L$ be the average of L independent copies of the process $\mathbf{X}$, defined by $X(0, t] = \lambda t + \sigma Z_t$, where Z_t is a fractional Brownian motion with Hurst parameter H. Then $\mathbf{\Lambda}_t(\boldsymbol{\theta}) = \lambda\boldsymbol{\theta}\cdot\mathbf{1} + \frac{1}{2}\sigma^2\boldsymbol{\theta}\cdot S_t\boldsymbol{\theta}$, where the $t \times t$ matrix S_t is given by $(S_t)_{ij} = \frac{1}{2}(|j-i-1|^{2H} + |j-i+1|^{2H} - 2|j-i|^{2H})$, and so $\mathbf{\Lambda}_t(\theta\mathbf{1}) = \lambda\theta t + \frac{1}{2}\sigma^2\theta^2 t^{2H}$.

To check that Assumption 2 is satisfied, choose the scaling function $v(t) = t^{2(1-H)}$, so that $\Lambda_t^L(\theta) = \lambda\theta + \frac{1}{2}\sigma^2\theta^2$. This does not depend on L or t, so it is also equal to $\Lambda_t(\theta)$ and $\Lambda(\theta)$.

EXAMPLE 5 (Fractional Brownian motion with large buffer). To contrast the many sources and the large buffer asymptotic, consider the large buffer version of fractional Brownian motion. Let $\mathbf{X}$ be a fractional Brownian motion with Hurst parameter H, as in the previous example. Choose the scaling $X^L(0, t] = f(L)^{-1}X(0, f(L)t]$ with $f(L) = L^{1/2(1-H)}$. This gives $\mathbf{\Lambda}_t^L(\theta\mathbf{1}) = \mathbf{\Lambda}_t(\theta\mathbf{1}) = \lambda\theta t + \frac{1}{2}\sigma^2\theta^2 t^{2H}$, the same expression as before. This is not linear in t, so $\mathbf{\Lambda}_t(\boldsymbol{\theta})$ does not have the simple linear form described in Example 3.

For Assumption 2, as with any large buffer example the limit (4) is uniform for any scaling function v, and as in Example 4 we can choose $v(t) = t^{2(1-H)}$.

Applying the results in Section 3 to the LDP we obtain from this, we can rederive a result of Duffield and O'Connell (1995) for the workload in a queue fed by a single fractional Brownian motion source.

We shall revisit these examples in the next section, to see what they tell us about large deviations of queue size.

3. Large deviations for queues. In this section, the sample path LDP is applied to study large deviations in several queueing problems: standard queues with finite and infinite buffers, likely paths to overflow and priority queues.

The common approach will be to take the sample path LDP and then apply the contraction principle to find an LDP for the quantity of interest. The contraction principle says that if $\mathbf{X}^L$ satisfies the sample path LDP in $\mathscr{X}_\mu$, and if f is a continuous function on $\mathscr{X}_\mu$, then $f(\mathbf{X}^L)$ satisfies an LDP with good rate function $I(y) = \inf\{\mathbf{I}(\mathbf{x}) : \mathbf{x} \in \mathscr{X}_\mu, f(\mathbf{x}) = y\}$. See Dembo and Zeitouni (1993), Theorem 4.2.1, for a proof of the contraction principle.

First, though, we relate the abstract setting of the last section to queueing models and describe the limiting regime.

3.1. *Queueing model.* Consider a sequence of queues, indexed by L, in which the Lth queue has service rate C and buffer size B. Let X_t^L be the total amount of work arriving at the Lth queue at time $-t$. (Depending on the context, $\mathbf{X}^L$ will variously be called an input process, a source or a traffic flow.)

There are several ways in which we can interpret this, depending on what $\mathbf{X}^L$ represents, though none of the results in this paper relies on a particular interpretation. Here are two possibilities, corresponding to examples from the previous section.

The first example is the one we have in mind throughout this paper: when the total input flow is the aggregate of many independent flows. This sort of scaling is well suited to modern telecommunications networks, in which a switch may have hundreds of inputs but only a small amount of buffer space per input.

EXAMPLE 6 (Many sources). In the many sources asymptotic, $\mathbf{X}^L$ is the average of L independent identically distributed flows. So the Lth queue can be thought of as multiplexing together L different flows, with its resources growing in proportion: it has service rate LC and buffer size LB.

The next example has been much more widely studied. For Markov-modulated fluid sources and for many others, the probability of loss decays exponentially in buffer size, so a good way to reduce loss is to make the buffers larger; and it is natural to study asymptotic regimes in which the buffer size increases. The observation that this is largely inaccurate when there are many input flows or when the sources exhibit long-range dependence [see Choudhury, Lucantoni and Whitt (1994) and Leland, Taqqu, Willinger and Wilson (1994) for example] has prompted some of the work on the many sources asymptotic.

EXAMPLE 7 (Large buffer). In the large buffer asymptotic, described in Example 3, $\mathbf{X}^L$ is a speeded up version of a base process: $X^L(0,t] = f(L)^{-1}\times X(0, f(L)t]$. So the Lth queue can be thought of as having a single input $\mathbf{X}$ and fixed service rate C, but increasing buffer size $f(L)B$.

Several authors, including O'Connell (1996, 1998), Paschalidis (1996) and Puhalskii and Whitt (1998), have used the contraction principle approach to study the large deviations behavior of various queueing systems under this asymptotic.

3.2. *Buffer size in a queue.* In this section, we look at a standard queue with a constant service rate. Some of the following results have previously been proved directly, but it is instructive to see the techniques used in deriving them from the sample path LDP, as these same techniques will be used in the following sections.

Consider a queue with constant service rate C fed with input process $\mathbf{x}$. The amount of work in the queue at time $-u$ may be defined to be

$\lim_{t\to\infty} Q_u\ (\mathbf{x}(0,t])$, where $Q_u(\mathbf{x}(0,t])$ is given by the Lindley recursion

$$Q_{s-1} = \left(Q_s + x_s - C\right)^+, \qquad Q_t = 0.$$

If the input is a stationary process, the stationary queue size may be written as

$$Q(\mathbf{x}) = \sup_t x(0,t] - Ct.$$

Lemma 13 shows that this function is continuous on $\mathscr{X}_\mu$ for any $\mu < C$. By the contraction principle, this immediately gives Corollary 7: an LDP for workload in queues with infinite buffers, which when simplified duplicates the results of Botvich and Duffield (1995) for linear scaling functions $v(t)$, of Duffield (1996) for general scaling functions and of Simonian and Guibert (1995) for the special case of Markov-modulated fluid sources. The estimate which this LDP provides can be refined with the Bahadur–Rao improvement, as described by Likhanov and Mazumdar (1999), but for the purposes of this paper we will stick with large deviations.

COROLLARY 7. *Under Assumptions* 1 *and* 2, *if* $\mathbf{X}^L$ *has mean rate less than* C, *then* $Q(\mathbf{X}^L)$ *satisfies an* LDP *with good rate function*

$$I(b) = \inf_{\mathbf{x}\in\mathscr{X}_C\,:\,Q(\mathbf{x})=b} \mathbf{I}(\mathbf{x}).$$

PROOF. The only point to note is that the infimum is taken over $\mathscr{X}_C$. But it might as well have been taken over $\mathscr{X}_\mu$ for any μ greater than the mean rate and less than C, since the rate function will be infinite on $\mathscr{X}_C\backslash\mathscr{X}_\mu$ by Theorem 6. □

We can do the same thing for queues with finite buffers. The queue size $\overline{Q}$ in a queue with a finite buffer B is defined similarly to Q, except that it cannot fill to greater than B and any excess work is discarded. More precisely, define the queue size at time $-u$ to be $\lim_{t\to\infty} \overline{Q}_u(\mathbf{x}(0,t])$, where $\overline{Q}_u(\mathbf{x}(0,t])$ is given by the Lindley recursion

$$\overline{Q}_{s-1} = \left(\overline{Q}_s + x_s - C\right)^+ \wedge B, \qquad \overline{Q}_t = 0.$$

Lemma 13 also shows that $\overline{Q}$ is a continuous function of the input process, and so we obtain Corollary 8: an LDP for workloads in queues with finite buffers.

COROLLARY 8. *Under Assumptions* 1 *and* 2, *if* $\mathbf{X}^L$ *has mean rate less than* C, *then* $\overline{Q}(\mathbf{X}^L)$ *satisfies an* LDP *with good rate function*

$$\overline{I}(b) = \inf_{\mathbf{x}\in\mathscr{X}_C\,:\,\overline{Q}(\mathbf{x})=b} \mathbf{I}(\mathbf{x}).$$

These expressions for the rate functions are not very informative, and so Theorem 9 gives a more manageable expression for $I(b)$. In fact, if the process is stationary, then, for $b \leq B$, $\overline{I}(b)$ and $I(b)$ are identical [and for $b > B$, $\overline{I}(b) = \infty$]; this is shown in Theorem 10. The proofs of these theorems are deferred to the end of this section.

THEOREM 9. *Under Assumptions* 1 *and* 2, *if the mean rate is less than* C *and if furthermore* $\mathbf{\Lambda}_t'(\theta\mathbf{1}) < Ct$ *at* $\theta = 0$ *for all* t, *then* $I(b)$ *is increasing in* b *and is given by*

$$I(b) = \inf_{\mathbf{x}\in\mathscr{X}_C\,:\,Q(\mathbf{x})=b} \mathbf{I}(\mathbf{x}) \tag{6}$$

$$= \inf_t \inf_{\mathbf{x}\in\mathbb{R}^t\,:\,x(0,t]=b+Ct} \mathbf{\Lambda}_t^*\big(\mathbf{x}(0,t]\big) \tag{7}$$

$$= \inf_t \sup_\theta \theta(b+Ct) - \mathbf{\Lambda}_t(\theta\mathbf{1}). \tag{8}$$

THEOREM 10. *If* $I(b)$ *is finite, then the optimal timescale* $\hat{t}$ *and the optimizing path* $\hat{\mathbf{x}}(0,\hat{t}\,]$ *are both attained; and if the optimal spacescale* $\hat{\theta}$ *is attained, then*

$$\hat{\mathbf{x}}(0,\hat{t}\,] = \nabla\mathbf{\Lambda}_{\hat{t}}(\hat{\theta}\mathbf{1}).$$

For a queue with a finite buffer B *and stationary input whose mean rate is less than* C, *if* $b \leq B$, *then* $\overline{I}(b) = I(b)$ *and the same path* $\hat{\mathbf{x}}$ *is optimal.*

Note that Theorem 9 does not assume stationarity. The condition that $\mathbf{\Lambda}_t'(\theta\mathbf{1}) < Ct$ at $\theta = 0$ for all t is implied by stationarity, but allows some additional cases such as appear in Wischik (1999).

The optimal $\hat{\theta}$ and $\hat{t}$ appearing in Theorem 10 are called the *operating point* of the switch, or the *critical spacescale* and *timescale*. Courcoubetis, Siris and Stamoulis (1999) give a detailed account, with simulation results, of how they are affected by the traffic mix and the queue parameters under the many sources asymptotic regime.

Examples. To illustrate the different forms that this rate function can take, we will go back to the two examples of Section 3.1—the many sources asymptotic and the large buffer asymptotic—paying particular attention to the interpretation of the critical timescale.

EXAMPLE 8 (Fractional Brownian motion with many sources). As in Example 6, consider a sequence of queues indexed by L in which the Lth queue Q^L is fed by an aggregate $L\mathbf{X}^L$ of L independent inputs and has service rate LC, and suppose the event of interest is that the queue size reaches Lb. As in Example 4, let each source be a fractional Brownian motion with mean rate λ and Hurst parameter H. We can calculate the critical spacescale and timescale:

$$\hat{\theta} = \frac{b + (C-\lambda)\hat{t}}{\sigma^2 \hat{t}^{2H}}$$

and

$$\hat{t} = \frac{b}{C-\lambda}\frac{H}{1-H}$$

(or rather, $\hat{t}$ is an integer close to this value; but we will ignore this minor complication). This gives rate function

$$I(b) = \frac{1}{2\sigma^2} b^{2(1-H)}(C-\lambda)^{2H}\left(\frac{H}{1-H}\right)^{2(1-H)}\frac{1}{H^2}$$

and large deviations approximation

$$\log \mathbb{P}\Big(Q^L(L\mathbf{X}^L) = Lb\Big) \approx -LI(b) \quad \text{for large } L.$$

Under the large buffer asymptotic, the rate function is exactly the same, but it has a very different interpretation, as we now illustrate.

EXAMPLE 9 (Fractional Brownian motion with large buffer). Instead of a sequence of queues, we will consider a single queue with fixed service rate C and fed by a single input flow $\mathbf{X}$, as in Example 7. Let the input flow again be a fractional Brownian motion, and consider the event that the queue size reaches $f(L)b$, where $f(L) = L^{1/2(1-H)}$.

As we saw in Example 5, the moment generating function $\mathbf{\Lambda}_t$ is exactly the same as for the many sources asymptotic, and so the rate function $I(b)$ is the same, too. This similarity disguises the fact that the results have very different interpretations. To see this, note that b is just a scaling factor so we may as well set $b = 1$, and let $\beta = f(L)$. Then the large deviations approximation amounts to

$$\log \mathbb{P}\Big(Q(\mathbf{X}) = \beta\Big) \approx -\beta^{2(1-H)} I(1) \quad \text{for large } \beta.$$

Notice that when $H = \frac{1}{2}$ the decay is exponential in β: many other sources including Markov-modulated fluid sources share this exponential decay. But when $H > \frac{1}{2}$ the source has long-range dependence and the decay is less than exponential, which means that increasing the buffer size does not give as much of a reduction in loss probability. This phenomenon was observed in real network traffic by Leland, Taqqu, Willinger and Wilson (1994), and it has stimulated much interest in long-range-dependent traffic models. But as we saw in the last example, it makes no difference to the many sources approximation whether $H = \frac{1}{2}$ or $H > \frac{1}{2}$.

There are some noteworthy differences between the many sources and large buffer asymptotics as regards the critical timescale $\hat{t}$ identified in Theorem 9. We point out these differences in the next example.

EXAMPLE 10 (Timescales). In the many sources asymptotic, the timescale $\hat{t}$ is easy to interpret: it is the length of time which the buffer is most likely to take to fill from empty to a given level Lb. In the large buffer asymptotic, $\hat{t}$ has a slightly different interpretation. It is a scaling parameter which relates the buffer level $f(L)b$ to the time taken to reach that level, $f(L)\hat{t}$.

In the latter case, the time taken to fill the buffer tends to ∞ and so the rate function $I(b)$ depends only on the infinite-time characteristics of the source. For Markov-modulated fluid sources [and many other sources which satisfy conditions described by Dembo and Zajic (1995)], it is appropriate to take $f(L) = L$ and so $\mathbf{\Lambda}_t(\theta\mathbf{1}) = t\lim_{L\to\infty} L^{-1}\log\mathbb{E}\exp(\theta X(0, L])$. Then the rate function $I(b)$ simplifies to $I(b) = \sup_\theta \theta b$, where the supremum is taken over all θ such that $\Lambda_1(\theta) \le C$.

By contrast, under the many sources asymptotic, the rate function depends on the characteristics of the source $\log\mathbb{E}\exp(\theta X(0, t])$ over all timescales t.

More LDP's. There are actually three more useful LDP's, which are easily confused with Corollaries 7 and 8. The first gives the probability that a queue with an infinite buffer is nonempty. At first sight, we can find this from Corollary 7: just consider the event $b > 0$. But the upper bound we get is useless, because it involves the closure of this set—which is $b \ge 0$, the entire space. So for a better bound, we can go back to the sample path LDP and look at the closure of the set of sample paths for which $Q(\mathbf{x}) > 0$, now not the entire space. The same technique can be used for the events that a queue with a finite buffer is nonempty or overflows. The infinite buffer result has been proved by Botvich and Duffield (1995), and the finite buffer results have been proved by Courcoubetis and Weber (1996). The proof of Corollary 11 is deferred to the end of this section. The proof of Corollary 12 is similar, and is omitted.

COROLLARY 11. *Under the assumptions of Theorem* 9, *the event* $\{Q > 0\}$ *has large deviations lower bound* $-I(0^+)$ *and upper bound* $-I^+(0)$. *If, in addition,* $B > 0$, *then the event* $\{\overline{Q} > 0\}$ *has the same large deviations bounds. Here,* $I(b^+) = \lim_{a\downarrow b} I(a)$ *and* $I^+(0)$ *is given by*

$$I^+(0) = \sup_\theta \theta C - \mathbf{\Lambda}_1(\theta\mathbf{1}).$$

COROLLARY 12. *Under Assumptions* 1 *and* 2, *if* $\mathbf{X}^L$ *is stationary and has mean rate less than C, then the event that* $\overline{Q}$ *overflows has large deviations lower bound* $-I(B^+)$ *and upper bound* $-I(B)$ [*or* $-I^+(0)$ *if* $B = 0$].

Note that in Corollary 11 $\mathbf{X}^L$ need not be stationary, and so $Q(\mathbf{X}^L)$ may not be, either. This corollary, like Theorem 9, makes the weaker assumption that $\mathbf{\Lambda}_t'(\theta\mathbf{1}) < Ct$ at $\theta = 0$ for all t. On the other hand, our results for queues with positive but finite buffers—Theorem 10 and Corollary 12—do require stationarity.

Proofs. The rest of this section is given over to proofs.

LEMMA 13. *The queue size functions Q and $\overline{Q}$ are continuous on $\mathscr{X}_\mu$ if $\mu < C$.*

PROOF. Consider a sequence of processes $\mathbf{x}^k \to \mathbf{x}$ in $\mathscr{X}_\mu$ under the uniform topology. That is, given ε, there is a k_0 such that, for $k \geq k_0$,

$$\sup_t \left| \frac{x^k(0,t]}{t} - \frac{x(0,t]}{t} \right| < \varepsilon.$$

And since $\mathbf{x} \in \mathscr{X}_\mu$, there is a t_0 such that, for $t \geq t_0$,

$$x(0,t]/t < \mu.$$

Then for $k \geq k_0$ and $t \geq t_0$, choosing $\varepsilon = C - \mu$,

$$x^k(0,t]/t < C$$

and the same holds for $\mathbf{x}$. So the expression for queue size Q simplifies: for $k \geq k_0$, $Q(\mathbf{x}^k) = Q(\mathbf{x}^k(0,t_0])$, and the same holds for $\mathbf{x}$. Thus for $k \geq k_0$,

$$|Q(\mathbf{x}^k) - Q(\mathbf{x})| = \left| \sup_{t \leq t_0} \Bigl(x^k(0,t] - Ct \Bigr) - \sup_{t \leq t_0} \Bigl(x(0,t] - Ct \Bigr) \right|,$$

which tends to 0 as $k \to \infty$.

Now for $\overline{Q}$. Since $Q(\mathbf{x}) = Q(\mathbf{x}(0,t_0])$, the infinite-buffer queue must empty at some time in $[-t_0, 0]$. For suppose it does not. Let $s \leq t_0$ be the last time at which the queue, started from empty at $-t_0$, is empty; then $Q(\mathbf{x}(0,t_0]) = Q(\mathbf{x}(0,s]) = x(0,s] - Cs$. But $Q(\mathbf{x}) = q + x(0,s] - Cs$, where $q > 0$ is the queue size at time $-s$, leading to a contradiction.

So Q empties at some time in $[-t_0, 0]$. So, too, must $\overline{Q}$, because $\overline{Q} \leq Q$. In other words, $\overline{Q}(\mathbf{x}) = \overline{Q}(\mathbf{x}(0,t_0])$. The same holds for $\mathbf{x}^k$ for k sufficiently large, and so we deduce that $\overline{Q}$ is also continuous. □

PROOF OF THEOREM 9. If $b = 0$, then (7) and (8) take the value 0 at $t = 0$. Now consider the sample path given by $\mathbf{x}(0,t] = \nabla\mathbf{\Lambda}_t(\mathbf{0})$. Since $\mathbf{\Lambda}'_t(\theta\mathbf{1}) < Ct$ at $\theta = 0$ for all t, $x(0,t] < Ct$ for all t, and so $Q(\mathbf{x}) = 0$. And $\mathbf{x}$ has rate $\mathbf{I}(\mathbf{x}) = 0$, so (6) also takes the value 0. So we restrict our attention to the case $b > 0$.

Note that because $b + Ct$ is greater than $\mathbf{\Lambda}'_t(\theta\mathbf{1})$ at $\theta = 0$, we may take the supremum only over $\theta \geq 0$; thus (8) is increasing in b.

First, (7) = (8). Fix t. Then $\mathbf{X}^L(0,t] \cdot \mathbf{1}$ is just a real-valued random variable, and from Assumption 1 it satisfies an LDP with good rate function given by the expression in (8). Another way of finding this is by contracting from the sample path LDP for $\mathbf{X}^L(0,t]$, which gives as rate function the expression in (7). By the uniqueness of the rate function, these are equal.

Next, (6) $\geq$ (7). It will be helpful to introduce some new notation. For a finite process $\mathbf{x}$ and an infinite process $\mathbf{y}$, write $\mathbf{x} :: \mathbf{y}$ for the concatenation of the two. And recall that we may replace $\mathscr{X}_C$ in (6) with $\mathscr{X}_\mu$ for any μ greater than the mean arrival rate and less than C, because by Theorem 6 the sample path rate function is infinite on $\mathscr{X}_C\backslash\mathscr{X}_\mu$.

Suppose that (6) is finite (otherwise the inequality is trivial). The sample path rate function $\mathbf{I}$ is good, so an optimal path $\hat{\mathbf{x}}$ is attained. Now $Q(\hat{\mathbf{x}}) = \sup_t \hat{x}(0,t] - Ct = b$, and this supremum must be attained since otherwise there is a sequence t_n for which $\hat{x}(0,t_n]/t_n \to C$, which cannot happen in $\mathscr{X}_\mu$. So $\hat{\mathbf{x}} = \hat{\mathbf{x}}(0,t] :: \hat{\mathbf{y}}$ for some $\hat{\mathbf{y}}$, with $\hat{x}(0,\hat{t}\,] = b + C\hat{t}$ and $Q(\hat{\mathbf{y}}) = 0$. Clearly, $\boldsymbol{\Lambda}^*_t(\mathbf{x}(0,t])$ is increasing in t for any $\mathbf{x}$, so

$$\mathbf{I}(\hat{\mathbf{x}}) = \sup_s \boldsymbol{\Lambda}^*_{\hat{t}+s}\Big(\hat{\mathbf{x}} :: \hat{\mathbf{y}}(0,s]\Big) \ge \boldsymbol{\Lambda}^*_{\hat{t}}\Big(\hat{\mathbf{x}}(0,\hat{t}\,]\Big).$$

Taking the infimum over t and $\mathbf{x}(0,t]$ gives the result.

Finally, (6) $\le$ (7). Assume that (7) is finite (since otherwise the inequality is trivial). For a given t, an optimal $\hat{\mathbf{x}}(0,\hat{t}\,]$ is attained by goodness of the rate function $\boldsymbol{\Lambda}^*_t$. And an optimal $\hat{t}$ is also attained. For suppose not, and take a sequence $t_n \to \infty$ and $\mathbf{x}^n(0,t_n]$, with $x^n(0,t_n]/t_n \to C$ and $\boldsymbol{\Lambda}^*_{t_n}(\mathbf{x}^n)$ bounded above by K say. By the contraction principle and the goodness of the rate function $\mathbf{I}$, we can extend $\mathbf{x}^n(0,t_n]$ to $\mathbf{x}^n(0,\infty)$, with $\mathbf{I}(\mathbf{x}^n) < K$. Since $\mathbf{I}$ is good it has compact level sets, so the $\mathbf{x}^n$ have a convergent subsequence, say $\mathbf{x}^k \to \mathbf{x}$, also with $\mathbf{I}(\mathbf{x}) < K$. But then $x(0,t_k]/t_k \to C$ also, and so $\mathbf{I}(\mathbf{x}) = \infty$, giving a contradiction.

By the contraction principle and the goodness of the rate function, we can extend $\hat{\mathbf{x}}(0,\hat{t}\,]$ to $\hat{\mathbf{x}} = \hat{\mathbf{x}}(0,\infty)$, where $\mathbf{I}(\hat{\mathbf{x}}(0,\hat{t}\,]) = \mathbf{I}(\hat{\mathbf{x}})$. If $Q(\hat{\mathbf{x}}) = b$ the inequality is proved. So suppose $Q(\hat{\mathbf{x}}) = b' > b$. Then there is some $s > \hat{t}$ with $\hat{x}(0,s] = b'$. But then

$$\inf_t \inf_{\mathbf{x}\,:\,x(0,t]=b+Ct} \boldsymbol{\Lambda}^*_t(\mathbf{x}) \ge \inf_{s>\hat{t}} \inf_{\mathbf{x}\,:\,x(0,s]=b'+Cs} \boldsymbol{\Lambda}^*_s(\mathbf{x}) \ge \inf_{s>\hat{t}} \inf_{\mathbf{x}\,:\,x(0,s]=b+Cs} \boldsymbol{\Lambda}^*_s(\mathbf{x}),$$

where the last inequality is because for fixed t, (8) is increasing in b. The inequalities must then both be equalities. We can repeatedly apply this argument until we find an optimal $\hat{\mathbf{x}}$ such that $Q(\hat{\mathbf{x}}) = b$. For otherwise, as in the previous paragraph, there are arbitrarily large optimal $\hat{t}$, leading to a contradiction. □

PROOF OF THEOREM 10. First, we prove that $\overline{I}(b) = I(b)$. If $I(b)$ is infinite, then $\overline{I}(b)$ must certainly be infinite, as any path which makes $\overline{Q}(\mathbf{x}) = b$ makes $Q(\mathbf{x}) \ge b$. So suppose $I(b)$ is finite, and let the optimizing path in Theorem 9 be $\hat{\mathbf{x}}(0,\hat{t}\,]$. We may assume that this path never causes the buffer to exceed level b. For suppose that under $\hat{\mathbf{x}}$ the buffer reaches level $b' > b$ at time $-s$. Consider the truncated process $\tilde{\mathbf{x}}(0,s] = \mathbf{x}(\hat{t}-s,\hat{t}\,]$. By stationarity, $\boldsymbol{\Lambda}^*_{\hat{t}}(\hat{\mathbf{x}}) \ge \boldsymbol{\Lambda}^*_s(\tilde{\mathbf{x}})$. And

$$\boldsymbol{\Lambda}^*_s(\tilde{\mathbf{x}}) \ge \inf_{\mathbf{x}\in\mathbb{R}^s\,:\,x(0,s]=b'+cs} \boldsymbol{\Lambda}^*_s(\mathbf{x}) \ge \inf_{\mathbf{x}\in\mathbb{R}^s\,:\,x(0,s]=b+cs} \boldsymbol{\Lambda}^*_s(\mathbf{x}),$$

where the second inequality follows because (8) is increasing in b. Because the optimal path does not cause the buffer to exceed level b, it is also optimal for the finite buffer case, and so $\overline{I}(b) = I(b)$.

Now fix t and suppose that $\hat{\theta}$ is optimal in (8). By Assumption 1, $\mathbf{\Lambda}_t$ must be differentiable at $\hat{\theta}\mathbf{1}$. Set $\hat{\mathbf{x}} = \nabla\mathbf{\Lambda}_t(\hat{\theta}\mathbf{1})$. Differentiating (8) gives $\hat{\mathbf{x}} \cdot \mathbf{1} = b + Ct$. But by Dembo and Zeitouni [(1993), Lemma 2.3.9], $\mathbf{\Lambda}_t^*(\hat{\mathbf{x}})$ is equal to (8), and so $\hat{\mathbf{x}}$ is optimal. □

PROOF OF COROLLARY 11. Let F be the event that $Q > 0$. For the large deviations lower bound, we will prove that $\inf_{\mathbf{x}\in F} \mathbf{I}(\mathbf{x}) = \lim_{b\downarrow 0} I(b)$, and for the large deviations upper bound,

$$\inf_{\mathbf{x}\in \overline{F}} \mathbf{I}(\mathbf{x}) = \inf_{t>0} \inf_{\mathbf{x}\,:\,x(0,\,t]=Ct} \mathbf{I}(\mathbf{x}). \tag{9}$$

This reduces to

$$\inf_{t>0} \sup_{\theta} \theta Ct - \mathbf{\Lambda}_t(\theta\mathbf{1})$$

as in Theorem 9. By convexity, $\mathbf{\Lambda}_t(\theta\mathbf{1}) \le t\mathbf{\Lambda}_1(\theta\mathbf{1})$, so the optimum is attained at $t = 1$ and we are left with $I^+(0)$.

Since $F = \cup_{b>0}\{Q = b\}$, $\inf_{\mathbf{x}\in F} \mathbf{I}(\mathbf{x}) = \inf_{b>0} I(b)$. But because $I(b)$ is increasing, this is $\lim_{b\downarrow 0} I(b)$.

LHS $\le$ *RHS in* (9) Suppose $x(0, t] = Ct$ for some $t > 0$. For $\varepsilon > 0$, let $\mathbf{x}^\varepsilon = (x_1 + \varepsilon, x_2, \ldots)$. Then $Q(\mathbf{x}^\varepsilon) > 0$ so $\mathbf{x}^\varepsilon \in F$. But as $\varepsilon \to 0$, $\mathbf{x}^\varepsilon \to \mathbf{x}$, so $\mathbf{x} \in \overline{F}$. Thus $\{\mathbf{x} : \exists t > 0, x(0, t] = Ct\} \subset \overline{F}$. Taking the infimum of $\mathbf{I}$ over these sets gives the result.

LHS $\ge$ *RHS in* (9) Let $\mathbf{x} \in \overline{F}$. Then there exist $\mathbf{x}^n \to \mathbf{x}$ in F, and $Q(\mathbf{x}^n) \to Q(\mathbf{x})$ by Lemma 13. If $Q(\mathbf{x}) > 0$, then

$$\mathbf{I}(\mathbf{x}) \ge \inf_{b>0} I(b) \ge \inf_{t>0} \sup_{\theta} \theta Ct - \mathbf{\Lambda}_t(\theta\mathbf{1})$$

because the optimal $\hat{t}$ in (8) must be strictly positive for $b > 0$.

So suppose $Q(\mathbf{x}^n) \to 0$. As in Lemma 13, there exist an n_0 and t_0 such that, for $n \ge n_0$,

$$Q(\mathbf{x}^n) = \sup_{t\le t_0} x^n(0, t] - Ct.$$

And because $Q(\mathbf{x}^n) > 0$, the supremum must be attained at $t > 0$. Some t must be repeated infinitely often as $n \to \infty$; for that t, $x(0, t] = Ct$. Taking the infimum over such $\mathbf{x}$ gives the result.

Now for $\{\overline{Q} > 0\}$. If $\overline{Q}(\mathbf{x}) > 0$, then $Q(\mathbf{x}) > 0$ also, so the same upper bound works. And as for $Q > 0$, the lower bound is straightforward. □

3.3. *Paths to overflow.* The expression for the rate function in Corollary 7 tells us more than just the probability that the queue size reaches a certain level. It tells us *how* the queue reaches that level. Because the rate function $\mathbf{I}$ is good, the infimum in

$$I(b) = \inf_{\mathbf{x}\in C\,:\,Q(\mathbf{x})=b} \mathbf{I}(\mathbf{x})$$

is attained. And Theorems 9 and 10 tell us what that sample path looks like: $\hat{\mathbf{x}}$ is the path most likely to make the queue fill from empty to level b, and it takes time $\hat{t}$ to do so. Furthermore, the sample path LDP tells us the likelihood of any deviation from this path.

The problem of most likely paths to overflow under the many sources asymptotic has been studied before using direct methods. Weiss (1986) solves it for two-state Markov-modulated fluid sources, and Mandjes and Ridder (1999) solve it for general Markov sources and for periodic sources. The advantage of our sample path LDP method is that it can be applied very easily to general input processes.

EXAMPLE 11 (Markov-modulated fluid source). Let $\mathbf{X}^L$ be the average of L independent sources distributed like $\mathbf{X}$, where $\mathbf{X}$ is a Markov chain which produces an amount of work h each time step while in the on state and no work while in the off state, and which flips from on to off with probability p and from off to on with probability q. If θ and t are the critical space- and timescales, then the most likely path to overflow is given by

$$x_s = \nabla \mathbf{\Lambda}_t(\theta \mathbf{1}) = \frac{\mathbb{E}\big(X_s e^{\theta X(0,t]}\big)}{\mathbb{E}\big(e^{\theta X(0,t]}\big)}.$$

To calculate this, first define

$$A_t = \mathbb{E}\Big(e^{\theta X(0,t]}\Big| X_0 = \text{on}\Big)$$

and

$$B_t = \mathbb{E}\Big(e^{\theta X(0,t]}\Big| X_0 = \text{off}\Big).$$

We can find expressions for A_t and B_t by conditioning on X_1:

$$\begin{pmatrix} A_t \\ B_t \end{pmatrix} = \begin{pmatrix} (1-p)e^{\theta h} & p \\ q e^{\theta h} & 1-q \end{pmatrix}^t \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

We can now calculate

$$\begin{aligned}
&\mathbb{E}\Big(X_s \exp(\theta X(0,\ t])\Big)\\
&\quad = \mathbb{E}\big[X_s \mathbb{E}\big(\exp(\theta X(0,\ s-1])\big| X_s\big) \exp(\theta X_s) \mathbb{E}\big(\exp(\theta X(s,\ t])\big| X_s\big)\big]\\
&\quad = \frac{q}{p+q} h A_{s-1} \exp(\theta h) A_{t-s}.
\end{aligned}$$

The first equality follows from the Markov property, and the second equality follows from reversibility. This gives

$$x_s = \frac{q h e^{\theta h} A_{t-s} A_{s-1}}{q A_t + p B_t}.$$

If $p + q < 1$, the path to overflow $s \mapsto x_s$ is concave over $s \in (0, t]$: the sources start slowly, then conspire to produce lots of work in the middle of the critical time period, then slow down again at the end. (If $p+q>1$, it is convex.)

Multistate Markov models exhibit more varied behavior.

EXAMPLE 12 (Gaussian sources). Suppose $\mathbf{X}^L$ is the average of L independent Gaussian processes, each with mean λ and covariance structure $\mathrm{Cov}(X_0, X_i) = \gamma_i$. It is easy to work out the optimal path: $\nabla\boldsymbol{\Lambda}_t(\theta\mathbf{1}) = \lambda\mathbf{1} + \theta V\mathbf{1}$, where $V_{ij} = \gamma_{|i-j|}$.

Consider the earlier fractional Brownian motion example, Example 4. For this process, $\gamma_i = \frac{1}{2}\sigma^2[(i-1)^{2H} - 2i^{2H} + (i+1)^{2H}]$, and so the most likely path to overflow is given by

$$x_i = \lambda + \tfrac{1}{2}\theta\sigma^2\Big(i^{2H} - (i-1)^{2H} + (t-i+1)^{2H} - (t-i)^{2H}\Big).$$

If $H > \frac{1}{2}$, the source exhibits long-range dependence, and the most likely input path $\mathbf{x}$ leading to overflow is concave; whereas if $H < \frac{1}{2}$, the path to overflow is convex.

Now let $\mathbf{X}$ be a single-step autoregressive process: $X_t = \lambda + a(X_{t-1} - \lambda) + (1-a^2)\varepsilon_t$, where $\varepsilon_t \sim N(0,\sigma^2)$ and $|a| < 1$. Then $\gamma_t = \sigma^2 a^t$, and the most likely path to overflow is

$$x_i = \lambda + \theta\sigma^2\left(1 + \frac{1-a^i}{1-a} + \frac{1-a^{t-i+1}}{1-a}\right).$$

If $a > 0$, then the path to overflow is concave; whereas if $a < 0$, it starts and finishes at a high rate and in between it oscillates.

EXAMPLE 13 (Large buffer). By contrast, in the large buffer asymptotic it is often the case that the buffer is most likely to fill up at a constant rate. Suppose that the base process $\mathbf{X}$ leads to a limiting moment generating function $\boldsymbol{\Lambda}_t$ with the simple linear form $\boldsymbol{\Lambda}_t(\theta) = \sum \boldsymbol{\Lambda}_1(\theta_i)$. Then $\boldsymbol{\Lambda}_t^*(\mathbf{x}(0,t]) = \sum \boldsymbol{\Lambda}_1(x_i)$, and because $\boldsymbol{\Lambda}_1$ is convex, the most likely path $\mathbf{x}$ to overflow is constant, and so the queue fills up at a steady rate.

3.4. *Priority queues.* The sample path LDP for the average of processes can be applied to a wide variety of queueing models. We have seen in the last two sections how it gives overflow probabilities and sample paths to overflow for a standard queue. As a further illustration of the power of the technique, in this section we look at another queueing discipline: the priority queue. This has been studied under the large buffer regime by Berger and Whitt (1998), and related queueing models have been studied by Kulkarni, Gün and Chimento (1995) and O'Connell (1998).

Consider a sequence of priority queues, indexed by L. The Lth queue has two inputs, $L\mathbf{X}^L$ and $L\mathbf{Y}^L$, and service rate LC. Think of $\mathbf{X}^L$ and $\mathbf{Y}^L$ as the averages of L processes. The two streams are assumed to be independent. The first stream $\mathbf{X}^L$ has high priority; the second stream $\mathbf{Y}^L$ has low priority. Let Q^L and R^L be respectively the stationary amounts of high-priority and low-priority work waiting to be served.

Kelly (1996) notes that the amount of high-priority traffic Q is exactly the amount of work in a standard queue with service rate C and only the high-priority input $\mathbf{X}$, and that the total amount of work $Q+R$ is the amount of work in a standard queue with service rate C and the aggregate input $\mathbf{X}+\mathbf{Y}$. Therefore, results from Section 3.2 can be applied directly to find the high-priority loss probability and the aggregate loss probability. But this leaves some open questions, such as how much low-priority work there is in the queue. Such questions can be answered with methods very similar to those of Section 3.2.

THEOREM 14. *Suppose that* $\mathbf{X}^L$ *and* $\mathbf{Y}^L$ *satisfy Assumptions* 1 *and* 2 *with limiting moment generating functions* $\mathbf{\Lambda}_t$ *and* $\mathbf{M}_t$, *respectively. Suppose also that the sum of the mean arrival rates for* $\mathbf{X}^L$ *and* $\mathbf{Y}^L$ *is strictly less than* C. *Then the pair* (Q^L, R^L) *satisfies an* LDP *with good rate function*

$$I(q,r) = \inf_{\substack{\mathbf{x}\in\mathscr{X}_C,\, \mathbf{y}\in\mathscr{X}_C:\\ Q(\mathbf{x})=q,\, R(\mathbf{x},\mathbf{y})=r}} \sup_t \mathbf{\Lambda}_t^*\big(\mathbf{x}(0,t]\big) + \sup_t \mathbf{M}_t^*\big(\mathbf{y}(0,t]\big). \tag{10}$$

This is bounded below by

$$\inf_t \inf_{s\le t} \sup_{\theta,\phi} \theta(q+Cs) + \phi\big(r+C(t-s)\big) - \mathbf{\Lambda}_t\big(\theta\mathbf{1}(0,s] + \phi\mathbf{1}(s,t]\big) - \mathbf{M}_t(\phi\mathbf{1}). \tag{11}$$

Let $I(\cdot, r) = \inf_{q\ge 0} I(q,r)$. *This is bounded below by*

$$\inf_t \sup_\theta \theta(r+Ct) - \mathbf{\Lambda}_t(\theta\mathbf{1}) - \mathbf{M}_t(\theta\mathbf{1}). \tag{12}$$

PROOF. Let $\mathbf{I}_X(\mathbf{x}) = \sup_t \mathbf{\Lambda}_t^*(\mathbf{x})$, and define $\mathbf{I}_Y$ similarly. Because $\mathbf{X}^L$ and $\mathbf{Y}^L$ are independent, the pair $(\mathbf{X}^L, \mathbf{Y}^L)$ satisfies an LDP with good rate function $\mathbf{I}(\mathbf{x},\mathbf{y}) = \mathbf{I}_X(\mathbf{x}) + \mathbf{I}_Y(\mathbf{y})$. Let λ and μ be the mean rates for $\mathbf{X}^L$ and $\mathbf{Y}^L$. Since $\lambda+\mu < C$, we can pick an $\varepsilon > 0$ such that $\lambda+\mu+2\varepsilon < C$: then by Theorem 6, $(\mathbf{X}^L, \mathbf{Y}^L)$ satisfies the LDP on $(\mathscr{X}_{\lambda+\varepsilon}, \mathscr{X}_{\mu+\varepsilon})$, and the rate function $\mathbf{I}$ is infinite outside this space. So if we can show that (Q, R) is continuous on this space, then using the contraction principle we can deduce (10).

Now Q depends only on the high-priority process: it is defined as though there were no other inputs to the queue. So, by Lemma 13, it is continuous on $\mathscr{X}_{\lambda+\varepsilon}$. Also, $Q+R$ is the aggregate workload, and does not depend on the structure of the queue: so, again by Lemma 13, $Q+R$ is continuous on $\mathscr{X}_{\lambda+\varepsilon} \times \mathscr{X}_{\mu+\varepsilon}$. Thus (Q, R) is continuous.

The bound on the rate function $I(q,r)$ may be obtained by noting a few properties of the optimal paths to overflow. The optimal paths must be attained, because the rate function is good. As in Theorem 9, there must be a last time $-t$ at which the high-priority and low-priority queues are both empty. And there must be a last time $-s \ge -t$ at which the high-priority queue is last empty. Because $Q(\mathbf{x}) = q$, it must be that $x(0,s] = q + Cs$. And because

$R(\mathbf{x},\mathbf{y})=r$, it must be that $x(s,t]+y(0,t]=r+C(t-s)$. So

$$I(q,r)\geq \inf_t \inf_{s\leq t} \inf_{\substack{\mathbf{x},\mathbf{y}\in\mathbb{R}^t:\\ x(0,s]=q+Cs,\\ x(s,t]+y(0,t]=r+C(t-s)}} \mathbf{\Lambda}_t^*(\mathbf{x})+\mathbf{M}_t(\mathbf{y}). \tag{13}$$

Now fix s and t. As in Theorem 9, the pair $(X^L(0,s], X^L(s,t]+Y^L(0,t])$ is just an $\mathbb{R}^2$-valued random variable, and by Assumption 1 it satisfies an LDP with a good rate function which simplifies to the expression in (11). Another way of finding this LDP is by contracting from the sample path LDP for $(\mathbf{X}^L(0,t],\mathbf{Y}^L(0,t])$ which gives as rate function the expression in (13). By the uniqueness of the rate function, these are equal.

We can obtain the lower bound on $I(\cdot,r)$ in a similar way, by noting that if $R(\mathbf{x},\mathbf{y})=r$ then there exists a last time $-t$ at which both queues were empty, and since then $x(0,t]+y(0,t]\geq r+Ct$. The argument of the previous paragraph can be applied to paths for which $x(0,t]+y(0,t]=q+r+Ct$. The resulting expression is increasing in q [it is a special case of (8) which is increasing in b], and setting $q=0$ yields the result. □

To help interpret this result, we will give an alternative description in terms of the service seen by the low-priority stream. A sensible first guess would be that the service is a random amount, the service rate C less a random amount of high-priority work. More thought would throw up various complications about queue sizes and leftover workloads. In fact, both of these cases arise, and a system can switch from one to the other as its parameters change. We will give an example to illustrate this transition.

But first, to make these statements precise we will introduce the idea of effective bandwidths. They are described in more detail by Kelly (1996). Consider a single queue with many independent inputs, as in Section 3.2. The overflow probability depends on the moment generating function $\mathbf{\Lambda}_t(\theta\mathbf{1})$. Suppose the critical space- and timescales are $\hat{\theta}$ and $\hat{t}$, and consider replacing a small proportion of the inputs by constant rate inputs, producing $(\hat{\theta}\hat{t})^{-1}\mathbf{\Lambda}_{\hat{t}}(\hat{\theta}\mathbf{1})$ units of work every time step. Locally, at $(\hat{\theta},\hat{t})$, these new inputs have the same moment generating function as the original inputs, and so the operation of the queue is not affected by the replacement. For this reason, $\lambda(\theta,t)=(\theta t)^{-1}\mathbf{\Lambda}_t(\theta\mathbf{1})$ is called the *effective bandwidth* of a source.

We can use this idea to describe the service seen by the low-priority stream. Consider a single queue fed by a process with effective bandwidth $\mu(\theta,t)$, but where the service is an independent stochastic process $\widetilde{\mathbf{C}}$ with effective bandwidth $\widetilde{C}(\theta,t)$. As above, if the critical space- and timescales are $\hat{\theta}$ and $\hat{t}$, replacing a small part of the service with constant service of rate $\widetilde{C}(\hat{\theta},\hat{t})$ does not affect the operation of the queue, and so we will call $\widetilde{C}(\theta,t)$ the *effective service rate*. Before we use this idea to describe the service seen by the low-priority stream, we had better check that it actually exists; that is, that the appropriate cumulant moment generating functions converge.

LEMMA 15 (Effective service). *Under the assumptions of Theorem* 14, *the service seen by the low-priority queue has an effective service rate.*

PROOF. O'Connell (1997b) shows that the departure map (which maps the aggregate input process to the aggregate departure process) is continuous under the uniform topology. Let $\mathbf{d}$ be the departure process from the high-priority queue. The service seen by the low-priority queue is $\widetilde{\mathbf{C}}$, where $\widetilde{C}_t = C - d_t$. Since the departure map is continuous, the service map is also continuous. Therefore the service process satisfies a large deviations principle, say with good rate function $\mathbf{J}$.

Applying Varadhan's integral lemma [Dembo and Zeitouni (1993), Theorem 4.3.1], and using the fact that the service process is bounded, we find that

$$\lim_{L\to\infty} \frac{1}{L} \log \mathbb{E} \exp(L\boldsymbol{\theta} \cdot \widetilde{\mathbf{C}}(0,\ t]) = \sup_{\mathbf{c}\in\mathbb{R}^t} \boldsymbol{\theta} \cdot \mathbf{c} - \mathbf{J}(\mathbf{c}).$$

In particular, the limit exists. □

We are now in a position to make precise the earlier claim about the service seen by the low-priority queue. The effective service rate is difficult to deal with analytically, but fortunately we can avoid doing so by using Theorem 14. The following corollary is a restatement of the bound (12). The terminology is due to Berger and Whitt (1998), who independently obtained the corresponding result for the large buffer asymptotic regime. As noted in Example 3, large buffer results can be deduced from a special case of the corresponding many sources results.

COROLLARY 16 (Empty buffer approximation). *The effective service rate seen by the low-priority queue is bounded below by the empty buffer approximation to the service rate, $\widetilde{C}(\theta, t) = C - \lambda(\theta, t)$, in the following sense:*

$$I(\cdot, r) \geq EI(r) = \inf_t \sup_\theta \theta\big(r + t\widetilde{C}(\theta, t)\big) - \theta t \mu(\theta, t),$$

where $\mu(\theta, t)$ is the effective bandwidth of the low-priority source.

This is just the usual rate function (8) for overflow in a single queue, but with the service rate C replaced by the effective service rate $\widetilde{C}$. It is called the *empty buffer approximation* because it is the rate function for the total workload reaching r—so if the most likely way for this to happen leaves the high-priority buffer empty, then $EI(r)$ will agree with $I(\cdot, r)$.

Berger and Whitt (1998) stress the point that this approximation gives a simple admission control region. But it is also interesting to consider the conditions under which the inequality is strict. When there is equality, the two queues operate essentially independently. But when the inequality is strict, the low-priority queue gets extra benefit from the sharing arrangement. Such an arrangement seems desirable from an engineering perspective. The following example illustrates how the queue and traffic parameters control whether or not there is extra benefit to the low-priority traffic.

EXAMPLE 14 (Phase transition in priority queues). It is often hard to simplify rate functions like $I(q, r)$ because the queue could overflow over any timescale. But for periodic processes, the queue can only overflow over timescales less than the period, so the calculations are easier.

Consider a sequence of priority queues indexed by L. Let the high-priority stream $\mathbf{X}^L$ be the average of L independent copies of a stationary periodic process of random phase, which produces 4 units of work every second time step. Let the low-priority stream $\mathbf{Y}^L$ be the average of L independent copies of the process that independently at each time step produces 1 unit of work with probability p and no work with probability $1 - p$. Let the service rate C be in the range $[3, 4)$.

These figures are chosen so that the entire queue empties every second time step, so that if it overflows it must do so in a single time step. This means that the only sample paths we need consider in (10) are those over a single time step. So

$$I(0, r) = \inf_{0 \le x \le C} \mathbf{\Lambda}_1^*(x) + \mathbf{M}_1^*(r + C - x),$$

$$I(q, r) = \mathbf{\Lambda}_1^*(q + C) + \mathbf{M}_1^*(r) \quad \text{for } q > 0.$$

Since $q + C$ is greater than the mean rate of $\mathbf{\Lambda}$, $\mathbf{\Lambda}_1^*(q + C) \ge \mathbf{\Lambda}_1^*(C)$. Taking $x = C$, we see that $I(0, r) \le I(q, r)$ for $q > 0$; and since $I(\cdot, r) = \inf_{q \ge 0} I(q, r)$, it must be that $I(\cdot, r) = I(0, r)$.

Now for the empty buffer approximation. Since $EI(r)$ is the rate function of the sample path most likely to give total queue size r,

$$EI(r) = \inf_{0 \le x \le C + r} \mathbf{\Lambda}_1^*(x) + \mathbf{M}_1^*(r + C - x).$$

Clearly, $I(\cdot, r) \ge EI(r)$. When is this inequality strict? Let $g(x) = \mathbf{\Lambda}_1^*(x) + \mathbf{M}_1^*(r + C - x)$. It is easy to calculate that, for $r < 1$,

$$g(x) = h(x/4 \mid 1/2) + h(r + C - x \mid p),$$

where $h(x|p) = x \log(x/p) + (1 - x)\log(1 - x)/(1 - p)$, and to show that $g(x)$ is convex. So $I(\cdot, r) > EI(r)$ if and only if $g'(C) < 0$, where

$$g'(C) = \frac{1}{4} \log \frac{C}{4 - C} - \log \frac{r}{1 - r} + \log \frac{p}{1 - p}.$$

In other words, there is extra benefit to the low-priority traffic when the service rate is small, or when the low-priority buffer is large, or when there is little low-priority work.

4. Conclusion. A sample path large deviations principle is an LDP factory: it makes it easy to study the large deviations in a wide range of queueing problems. Many LDP's have previously been found in this way, under the large buffer asymptotic regime. This paper presents a sample path LDP for the many sources asymptotic regime, and applies it to study three queueing problems. Existing results for standard queues have been refined, and new results have been presented for likely paths to overflow and for priority queues.

We have seen that the large buffer asymptotic can often be described as a special case of the many sources asymptotic. This means that large deviations of queueing systems under the many sources asymptotic, which depend on the characteristics of the traffic over all timescales, tend to have richer structure than those under the large buffer asymptotic, which depend only on the long-timescale characteristics of the traffic.

Acknowledgments. I am very grateful to Frank Kelly and to Neil O'Connell for many valuable discussions. I am also grateful for the thoughtful comments of the anonymous referee.

REFERENCES

BERGER, A. W. and WHITT, W. (1998). Effective bandwidths with priorities. *IEEE/ACM Trans. Networking* **6**.

BOTVICH, D. and DUFFIELD, N. (1995). Large deviations, the shape of the loss curve, and economies of scale in large multiplexers. *Queueing Systems* **20** 293–320.

CHOUDHURY, G. L., LUCANTONI, D. M. and WHITT, W. (1994). On the effectiveness of effective bandwidths for admission control in ATM networks. In *Proceedings of the 14th International Teletraffic Congress—ITC 14* 411–420. North Holland, Amsterdam.

COURCOUBETIS, C., SIRIS, V. A. and STAMOULIS G. D. (1999). Application of the many sources asymptotic and effective bandwidth to traffic engineering. *Telecommunication Systems*. To appear.

COURCOUBETIS, C. and WEBER, R. (1996). Buffer overflow asymptotics for a buffer handling many traffic sources. *J. Appl. Probab.* **33** 886–903.

DEMBO, A. and ZAJIC, T. (1995). Large deviations: from empirical mean and measure to partial sums process. *Stochastic Process. Appl.* **57** 191–224.

DEMBO, A. and ZEITOUNI, O. (1998). *Large Deviations Techniques and Applications*, 2nd ed. Springer, New York.

DUFFIELD, N. G. (1996). Economies of scale in queues with sources having power-law large deviation scalings. *J. Appl. Probab.* **33** 840–857.

DUFFIELD, N. G. and O'CONNELL, N. (1995). Large deviations and overflow probabilities for the general single-server queue, with applications. *Math. Proc. Cambridge Philos. Soc.* **118** 363–374.

KELLY, F. (1996). Notes on effective bandwidths. In *Stochastic Networks: Theory and Applications* (F. P. Kelly, S. Zachary and I. Ziedins, eds.) 141–168. Oxford Univ. Press.

KULKARNI, V. G., GÜN, L. and CHIMENTO, P. F. (1995). Effective bandwidth vectors for multiclass traffic multiplexed in a partitioned buffer. *IEEE J. Selected Areas in Communications* **13** 1039–1047.

LELAND, W. E., TAQQU, M. S., WILLINGER, W. and WILSON, D. V. (1994). On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Trans. Networking* **2** 1–15.

LIKHANOV, N. and MAZUMDAR, R. R. (1999). Cell loss asymptotics for buffers fed with a large number of independent stationary sources. *J. Appl. Probab.* To appear.

MANDJES, M. and RIDDER, A. (1999). Optimal trajectory to overflow in a queue fed by a large number of sources. *Queueing Systems* **31** 137–170.

O'CONNELL, N. (1996). Queue lengths and departures at single-server resources. In *Stochastic Networks: Theory and Applications* (F. P. Kelly, S. Zachary and I. Ziedins, eds.) Chap. 5. Oxford Univ. Press.

O'CONNELL, N. (1997a). A large deviation principle with queueing applications. Technical Report HPL-BRIMS-97-05, BRIMS, Hewlett Packard Labs, Bristol.

O'CONNELL, N. (1997b). Large deviations for departures from a shared buffer. *J. Appl. Probab.* **34** 753–766.

O'CONNELL, N. (1998). Large deviations for queue lenghts at a multi-buffered resource. *J. Appl. Probab.* **34** 240–245.

PASCHALIDIS, I. C. (1996). Large deviations in high speed communications networks. Ph.D. dissertation, MIT Laboratory for Information and Decision Systems, Cambridge, MA.

PUHALSKII, A. A. and WHITT, W. (1998). Functional large deviation principles for waiting and departure processes. *Probab. Engrg. Inform. Sci.* 479–507.

SIMONIAN, A. and GUIBERT, J. (1995). Large deviations approximation for fluid sources fed by a large number of on/off sources. *IEEE J. Selected Areas in Communications* **13** 1017–1027.

WEISS, A. (1986). A new technique for analyzing large traffic systems. *Adv. in Appl. Probab.* **18** 506–532.

WISCHIK, D. (1999). The output of a switch, or, effective bandwidths for networks. *Queueing Systems* **32** 383–396.

STATISTICAL LABORATORY
CENTRE FOR MATHEMATICAL SCIENCES
WILBERFORCE ROAD
CAMBRIDGE CB3 0WB
UNITED KINGDOM
E-MAIL: D.J.Wischik@statslab.cam.ac.uk