DIAS-STP-94-13

Large deviations in queueing networks

Neil O'Connell*

May 11, 1994 (revised)

Abstract

The main result of this paper describes how the large deviation properties of a traffic stream are altered when the traffic passes through a buffer, possibly sharing that buffer with cross traffic. We also consider the effect of priority service policies. The heuristics behind the approach suggest a general method for determining the large deviation properties of traffic at any point in an arbitrary network, which we illustrate with an example.

1 Introduction

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

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

For each $n \in Z_+$ set

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

with the convention that $A_0 = S_0 = W_0 = 0$. If X and C are stationary processes and $EX_1 < EC_1$, then Q is stationary and

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

The identity (3) can be used to deduce the asymptotic behaviour of the queue-length distribution from the large deviation properties of A and S. More precisely, if A and S are independent and the sequences A_n/n and S_n/n satisfy the large deviation principle (LDP) with respective good rate functions I_A and I_S, then W_n/n satisfies the LDP with rate function given by

$$I_W(w) = \inf_y [I_A(w+y) + I_S(y)], \tag{4}$$

and, under mild hypotheses on I_W [1, 5, 8, 11], the tails of the queue-length distribution satisfy the order relation

$$\log P(Q_0 > b) \sim -\delta b \tag{5}$$

*Dublin Institute for Advanced Studies, 10 Burlington Road, Dublin 4. Research supported by grants from EOLAS and Mentec Computer Systems Ltd. under the Higher Education-Industry Cooperation Scheme.

DIAS-STP-9413

for large b, where

$$\delta = \inf_{c>0} I_W(c)/c. \tag{6}$$

As this is such a general result, it may be useful for real applications: in particular, it provides a basis for predicting overflow probabilities at a single buffered resource [4, 7].

It also suggests the possibility of a kind of network calculus at the level of rare events. For example, given an arbitrary network with several inputs, it may be possible to estimate the probability of overflow at a given (buffered) node in terms of the large deviation properties of the inputs. The obvious starting point is to ask how the large deviation properties of traffic are altered when the traffic passes through a buffer, possibly sharing that buffer with other traffic. In a recent paper, de Veciana, Courcoubetis and Walrand [5] give a partial answer to this question. Suppose we have two independent, ergodic arrival processes X^1 and X^2 sharing a deterministic buffer according to a work conserving policy with service rate c. Suppose also that the corresponding partial sums satisfy the LDP with respective rate functions I_1 and I_2. Then, under certain conditions [5, Corollary 3.2], if D_n^1 denotes the total departures upto time n corresponding to the first traffic stream, the sequence D_n^1/n satisfies the LDP with good rate function I_{D^1} given by I_1 on the interval $[0, c - EX_1^2]$. The main result of this paper provides a full description of I_{D^1}, using sample path large deviation theory, when the service policy is FCFS (first come, first served). We also consider more general service policies and describe how our results can be extended to more complicated networks.

2 Sample path large deviations in R^d

Denote by $D_{[0,1]}(R^d)$ the space of right continuous paths $[0,1] \to R^d$ having left limits equipped with the uniform topology, and by $\mathcal{AC}^0$ the set of paths that are absolutely continuous and start at 0. Let X_k be a sequence of random variables and set

$$S_n(t) = \frac{1}{n}\sum_{k=1}^{[nt]} X_k \tag{7}$$

for $t \in [0,1]$. Dembo and Zajic [10] establish very general conditions for which

(SP) *the sequence of partial sums processes $S_n(\cdot)$ satisfy the large deviation principle on $D_{[0,1]}(R^d)$ with good convex rate function given by*

$$I(\phi) = \begin{cases} \int_0^1 \Lambda^*(\dot{\phi})ds & \phi \in \mathcal{AC}^0 \\ \infty & \textit{otherwise} \end{cases} \tag{8}$$

where Λ^* is the Fenchel-Legendre transform of

$$\Lambda(\lambda) = \lim_{n\to\infty} \frac{1}{n} \log Ee^{\langle \lambda, S_n(1)\rangle}. \tag{9}$$

(By convention, all infimums over empty sets are infinite.) In particular, (SP) holds for bounded ψ-mixing stationary processes and hypercontractive Markov chains.

To illustrate how easy it is to use the property (SP) we consider some examples. For the sake of clarity, assume that I is continuous. To begin with, the one dimensional LDP is a trivial consequence of (SP). Indeed, for all open sets G,

$$-\log P(S_n(1) \in G) \sim \inf_{\phi(1)\in G} \int_0^1 \Lambda^*(\dot{\phi})ds. \tag{10}$$

But since Λ is convex we have by Jensen's inequality that for any $\phi \in \mathcal{AC}^0$,

$$\int_0^1 \Lambda^*(\dot{\phi})ds \geq \Lambda^*(\phi(1)), \tag{11}$$

and so (10) becomes

$$-\log P(S_n(1) \in G) \sim \inf_{x \in G} \Lambda^*(x). \tag{12}$$

Intuitively, the path that minimises the RHS of (10) is the 'most likely path' of S_n on $\{S_n(1) \in G\}$; in this case the most likely path is a straight line. We can also use (SP) to better understand the relation (5):

$$\begin{aligned} -\log P(\sup_t S_n(t) > 1) &\sim \inf_{\|\phi\|>1} \int_0^1 \Lambda^*(\dot{\phi})ds && (13)\\ &= \inf_{0\leq\tau\leq 1} \inf_{\phi(\tau)>1} \int_0^1 \Lambda^*(\dot{\phi})ds && (14)\\ &= \inf_{0\leq\tau\leq 1} \tau\Lambda^*(1/\tau). && (15) \end{aligned}$$

Note that the partial sum process can be defined on any finite interval and the property (SP) will hold on that interval if it holds on $[0,1]$; we can therefore deduce (5) from (15) by localisation over successively longer time intervals.

3 Departures from a shared buffer

We begin with a slight variation of the problem considered by de Veciana *et al.* [5]. We have two independent arrival streams X_n^1 and X_n^2 sharing a deterministic buffer according to a FCFS policy with stochastic service rate C_n. (C is assumed to be independent of X^1 and X^2.) Set

$$A_n^i = \sum_{k=1}^n X_k^i, \qquad S_n = \sum_{k=1}^n C_k. \tag{16}$$

Denote by D^1 the cumulative departure process corresponding to the first arrival stream. If we start with an empty buffer, then the total departures from the buffer upto time n is given by

$$D_n = \inf_{k\leq n}[A_k^1 + A_k^2 - S_k] + S_n. \tag{17}$$

The key fact is that if we set

$$T_n = \inf\{k \leq n : A_k^1 + A_k^2 \geq D_n\}, \tag{18}$$

then $D_{T_n}^1$ is 'close' to A_n^1; in the proof of Theorem 3.1 below we restrict our attention to continuous paths and the above becomes equality. For $i = 1, 2$ set

$$\Lambda_i(\theta) = \lim_{n\to\infty} \frac{1}{n} \log Ee^{\theta A_n^i}, \qquad \Lambda_S(\theta) = \lim_{n\to\infty} \frac{1}{n} \log Ee^{\theta S_n} \tag{19}$$

whenever these limits exist. For convenience we will use a slight variant of the hypothesis (SP) to state our theorem, namely to restrict attention to non-decreasing sample paths. Denote by $\mathcal{D}$ the subspace of non-decreasing paths in $D_{[0,1]}(R)$.

Theorem 3.1 *If the partial sums processes $A^i_{[n\cdot]}/n$ and $S_{[n\cdot]}/n$ satisfy the LDP in $\mathcal{D}$ with respective rate functions given by*

$$I_i(\phi) = \begin{cases} \int_0^1 \Lambda_i^*(\dot\phi)ds & \phi \in \mathcal{AC}^0 \cap \mathcal{D} \\ \infty & \text{otherwise} \end{cases} \tag{20}$$

for $i = 1, 2$ and

$$I_S(\phi) = \begin{cases} \int_0^1 \Lambda_S^*(\dot\phi)ds & \phi \in \mathcal{AC}^0 \cap \mathcal{D} \\ \infty & \text{otherwise,} \end{cases} \tag{21}$$

then so does the sequence of processes $D^1_{[n\cdot]}/n$, with rate function given by

$$I_{D^1}(\phi) = \begin{cases} \int_0^1 \Lambda_{D^1}^*(\dot\phi)ds & \phi \in \mathcal{AC}^0 \cap \mathcal{D} \\ \infty & \text{otherwise} \end{cases} \tag{22}$$

where

$$\Lambda^*_{D^1}(z) = \inf_{x\beta(x,y,c)=z} [\beta(x,y,c)[\Lambda_1^*(x) + \Lambda_2^*(y)] + \Lambda_S^*(c)] \tag{23}$$

and

$$\beta(x,y,c) = \frac{c}{x+y} \wedge 1. \tag{24}$$

Proof. For $\phi_1, \phi_2, \phi_S \in \mathcal{AC}^0 \cap \mathcal{D}$, $0 \le t \le 1$, define (by analogy with the stopping time T_n above)

$$\tau(t) = \inf\left\{ r : \ \phi_1(r) + \phi_2(r) = \inf_{0\le\nu\le 1} [\phi_1(\nu t) + \phi_2(\nu t) - \phi_S(\nu t)] + \phi_S(t) \right\}, \tag{25}$$

and observe that

$$\dot\tau(t) = \beta\left(\dot\phi_1(\tau(t)), \dot\phi_2(\tau(t)), \dot\phi_S(t)\right). \tag{26}$$

It's not hard to check that the mapping $T : \mathcal{D}^3 \to \mathcal{D}$ defined by

$$T(\phi_1, \phi_2, \phi_S) = \phi_1 \circ \tau \tag{27}$$

is continuous, so for any open $B \subset \mathcal{D}$ the set $T^{-1}(B)$ is either open or empty in $\mathcal{D}^3$; as it is clearly non-empty we have

$$\begin{aligned}
&-\liminf_{n\to\infty} \frac{1}{n} \log P(D^1_{[n\cdot]}/n \in B) && (28)\\
&\le \inf_{\phi_1,\phi_2,\phi_S \in T^{-1}(B)} [I_1(\phi_1) + I_2(\phi_2) + I_S(\phi_S)] && (29)\\
&= \inf_{\phi_1,\phi_2,\phi_S \in \mathcal{AC}^0 \cap T^{-1}(B)} \int_0^1 [\Lambda_1^*(\dot\phi_1) + \Lambda_2^*(\dot\phi_2) + \Lambda_S^*(\dot\phi_S)]dt && (30)\\
&= \inf_{\phi_1,\phi_2,\phi_S \in \mathcal{AC}^0 \cap T^{-1}(B)} \left[\int_0^1 [\Lambda_1^*(\dot\phi_1(\tau(r))) + \Lambda_2^*(\dot\phi_2(\tau(r)))]\dot\tau(r)dr + \int_0^1 \Lambda_S^*(\dot\phi_S)dt \right. && (31)\\
&\qquad \left. + \int_{\tau(1)}^1 [\Lambda_1^*(\dot\phi_1) + \Lambda_2^*(\dot\phi_2)]dt \right] \\
&= \inf_{\phi_1,\phi_2,\phi_S \in \mathcal{AC}^0 \cap T^{-1}(B)} \int_0^1 [[\Lambda_1^*(\dot\phi_1(\tau(t))) + \Lambda_2^*(\dot\phi_2(\tau(t)))]\beta(\dot\phi_1(\tau(t)), \dot\phi_2(\tau(t)), \dot\phi_S(t)) && (32)\\
&\qquad + \Lambda_S^*(\dot\phi_S(t))]dt \\
&= \inf_{\phi \in B \cap \mathcal{AC}^0} \int_0^1 \inf_{x\beta(x,y,c)=\dot\phi} [\beta(x,y,c)[\Lambda_1^*(x) + \Lambda_2^*(y)] + \Lambda_S^*(c)]\, dt. && (33)
\end{aligned}$$

To justify the last step note that the 'most likely paths' ϕ_1, ϕ_2, ϕ_S implicitly defined on $[0, \tau(1)]$ and $[0,1]$ by their derivatives are absolutely continuous (and increasing) if, and only if, ϕ is; this follows from the fact their derivatives at each time s, as defined, are an integrable function of $\dot{\phi}(s)$.

The proof of the corresponding upper bound is identical. □

Corollary 3.2 *Under the hypotheses of Theorem 3.1, the sequence D_n^1/n satisfies the LDP with rate function given by (23).*

Note that if $C_n = c$ for all n,

$$\Lambda^*_{D^1}(z) = \begin{cases} [\Lambda_1^*(z) + \Lambda_2^*((c-z) \wedge \Lambda_2'(0))] \wedge \inf_{x \geq z} \frac{z}{x} \left[\Lambda_1^*(x) + \Lambda_2^*\left(\frac{c-z}{z}x\right)\right] & z \in [0,c], \\ \infty & \text{otherwise;} \end{cases} \tag{34}$$

the infimum is taken to be infinite at $z = 0$. If Λ_i^* is convex, then $\Lambda_i^*(x)/x$ is non-decreasing for $x \geq \Lambda_i'(0)$. To see this, let X be a random variable with

$$P(X = x) = z/x = 1 - P(X = \Lambda_i'(0)), \tag{35}$$

where $x \geq z \geq \Lambda_i'(0)$. Then

$$EX = z + (1 - z/x)\Lambda_i'(0) \geq z, \tag{36}$$

and by Jensen's inequality,

$$z\Lambda_i^*(x)/x = E\Lambda_i^*(X) \geq \Lambda_i^*(EX) = \Lambda_i^*(z + (1 - z/x)\Lambda_i'(0) \geq \Lambda_i^*(z). \tag{37}$$

It therefore follows from (34) that if Λ_1^* and Λ_2^* are convex, and $\Lambda_1'(0) + \Lambda_2'(0) < c$ (in other words the queue is stable) then $\Lambda^*_{D^1} = \Lambda_1^*$ on the interval $\mathcal{D}_1^\circ \cap [\Lambda_1'(0), c - \Lambda_2'(0)]$, where $\mathcal{D}_1^\circ$ is the interior of the effective domain of Λ_1^*; this is in agreement with the result of de Veciana *et al.* [5, Corollary 3.2]. Furthermore, $\Lambda^*_{D^1} \geq \Lambda_1^*$.

The heuristics behind Corollary 3.2 are as follows. (These heuristics are only valid when the input and service rate functions are convex.) Suppose x, y and c are the minimisers in (23). On the event $\{D_n/n \approx x\}$ (by $\approx$ we mean 'contained in a neighbourhood of size $1/n$'), the most likely paths for the cumulative arrivals are straight lines upto time $\beta(x,y,c)n$ with respective slopes x and y, and straight lines thereafter with respective slopes $\Lambda_1'(0)$, $\Lambda_2'(0)$; the most likely path for the cumulative service process is straight with slope c. An intuition for these heuristics makes it clear how one would perform a calculus of rate functions on more complicated networks. We will demonstrate this with an example in §6.

We can deduce the large deviation properties of departures from a queue fed by a single arrivals stream by setting $X_n^2 = 0$ for all n. This generalises results of de Veciana *et al.* [5, Theorem 3.1] and Chang *et al* [2], and sharpens a result of [9, Theorem 3].

Corollary 3.3 *If we have just one arrivals process X_n^1 in a queue with stochastic service rate C_n and if the hypotheses of Theorem 3.1 are satisfied, the normalised total departures D_n/n satisfy the LDP with rate function given by*

$$\Lambda_D^*(z) = [\Lambda_1^*(z) + \Lambda_S^*(z \vee \Lambda_S'(0))] \wedge \left[\inf_{x \geq z} z\Lambda_1^*(x)/x + \Lambda_S^*(z)\right]. \tag{38}$$

In fact, the sequence of processes $D_{[n\cdot]}/n$ satisfy the LDP in $\mathcal{D}$ with rate function given by (23) where $\Lambda^*_{D^1}$ is replaced by Λ_D^* defined above. If Λ_1^* is convex, (38) becomes

$$\Lambda_D^*(z) = [\Lambda_1^*(z) + \Lambda_S^*(z \vee \Lambda_S'(0))]. \tag{39}$$

For constant service rate c, we have

$$\Lambda_D^*(z) = \begin{cases} \Lambda_1^*(z) & 0 \le z < c, \\ \inf_{x \ge c} c\Lambda_1^*(x)/x & z = c, \\ \infty & \textit{otherwise}. \end{cases} \tag{40}$$

Again, if Λ_1^* is convex, this simplifies to

$$\Lambda_D^*(z) = \begin{cases} \Lambda_1^*(z) & 0 \le z \le c, \\ \infty & \textit{otherwise}. \end{cases} \tag{41}$$

Throughout this section we have assumed that the buffer is initially empty. If the arrivals were assumed to be stationary and the queue assumed to be stable, one could prove a stationary version of Theorem 3.1 where the queue is assumed to be initially in equilibrium. Chang and Zajic [3] prove a stationary version of Corollary 3.3 and make the important observation that the rate function for the departures in the stationary case is generally different from the above when the service is stochastic (otherwise it is the same); the difference stems from the fact a large (positive) deviation in the departures can be encouraged by starting with a very long queue. Recall that the tail decay of the queue length distribution is determined by Λ_1^* and Λ_S^*. Their result states that, under additional mixing hypotheses, if there is just one arrivals process, the rate function for the *stationary* departure process is given by

$$\tilde{\Lambda}_D^*(z) = \delta z - \sup_{x \le z}[\delta x - \Lambda_1^*(x)] + \Lambda_S^*(z \vee \Lambda_S'(0)), \tag{42}$$

for $z \ge \Lambda_1'(0)$, where

$$\delta = \inf_{w,c}[\Lambda_1^*(w+c) + \Lambda_S^*(c)]/w. \tag{43}$$

The additional mixing hypothesis is required because the queue-length at time zero is not independent of subsequent service and arrivals.

4 Priority traffic

The heuristics which were used above to justify Corollary 3.2 intuitively can also be applied to more complicated service policies. We illustrate this with an example. Suppose we have two independent cumulative arrival processes A^1 and A^2 sharing a deterministic buffer according to policy with total capacity c that prioritises the first traffic stream with weight $0 < p < 1$. In other words, if there is traffic from both streams in the buffer the first stream is served at rate pc and the second at rate $(1-p)c$; spare capacity is open to traffic from either stream. This kind of service scheme is known as *generalised processor sharing*. In a recent paper, de Veciana and Kesidis [6] provide large deviation approximations for the tails of the queue length corresponding to the first traffic stream. Here we consider the departures.

It is clear from the heuristics that under the hypotheses of Theorem 3.1 the most likely paths of A^1 and A^2 on $\{D_n^1/n \approx z\}$ will be straight upto time

$$T_n = \sup\{k : D_n^1 = A_k^1\} \tag{44}$$

with respective slopes x and y, say, and straight thereafter with slopes $\Lambda_1'(0)$ and $\Lambda_2'(0)$. Note that this implies the most likely path for D^1 will be straight with slope z, and $T_n/n \approx z/x$ with high probability. If $x > pc$ and $y > (1-p)c$ then $z = pc$; if $x > pc$ and $y < (1-p)c$ then $z = (c-y) \wedge x$; otherwise $z = x$. Putting all this

together we expect the normalised departures corresponding to the first stream, D_n^1/n, to satisfy the LDP with rate function given by

$$\Lambda^*_{D^1}(z) = \inf_{x\beta(x,y)=z} \frac{z}{x}[\Lambda_1^*(x) + \Lambda_2^*(y)], \tag{45}$$

where

$$\beta(x,y) = \begin{cases} pc/x & x > pc,\ y > (1-p)c \\ \frac{c-y}{x} \wedge 1 & x > pc,\ y \le (1-p)c \\ 1 & \text{otherwise.} \end{cases} \tag{46}$$

5 Applications

5.1 The effect of cross traffic on a deterministic flow

Consider a deterministic stream with rate d^{-1} sharing a deterministic buffer with an arbitrary cross stream A^2 according to a FCFS policy with service rate 1. Denote by Λ_1^* and Λ_2^* the rate functions corresponding to the two input streams, and by $\Lambda^*_{D^1}$ the rate function corresponding to the departures of the initially deterministic stream. Note that

$$\Lambda_1^*(x) = \infty 1_{x \neq d^{-1}}. \tag{47}$$

By Theorem 3.1 we have, assuming A^2 satisfies the sample path LDP hypothesis and $\Lambda_2'(0) < 1 - d^{-1}$,

$$\Lambda^*_{D^1}(z) = \begin{cases} zd\Lambda_2^*\left(\frac{1-z}{z}\frac{1}{d}\right) & 0 \le z < d^{-1}, \\ 0 & z = d^{-1}, \\ \infty & \text{otherwise.} \end{cases} \tag{48}$$

This example was considered by Kelly and Key [12] in the case where A^2 is Poisson and the queue is heavily loaded, using results of van den Berg and Resing [14] on the approximate distribution of the departure process; they consider the large deviation properties of the limiting departure process after the reference stream has passed through a long sequence of queues in tandem, sharing each queue with an independent Poisson cross stream, assuming each queue is heavily loaded. We will now apply our result to this example, and show that the heavy traffic and large deviation limits *do not commute.*

If the cross stream is Poisson with rate λ, then (48) becomes

$$\Lambda^*_{D^1}(z) = \begin{cases} zd\lambda - (1-z) + (1-z)\log\left(\frac{1-z}{zd\lambda}\right) & 0 \le z < d^{-1}, \\ 0 & z = d^{-1}, \\ \infty & \text{otherwise.} \end{cases} \tag{49}$$

In the heavy traffic limit ($\lambda \nearrow 1 - d^{-1}$) this becomes

$$\Lambda^*_{D^1}(z) = \begin{cases} z(d-1) - (1-z) + (1-z)\log\left(\frac{1-z}{z(d-1)}\right) & 0 \le z \le d^{-1}, \\ \infty & \text{otherwise.} \end{cases} \tag{50}$$

Compare this with the rate function corresponding to the heavy traffic approximation departure stream, a renewal process with interarrival time distribution $1 + P(d-1)$, where $P(\lambda)$ denotes a Poisson distribution with rate λ:

$$\tilde{\Lambda}^*_{D^1}(z) = \begin{cases} z(1-d^{-1}) - (1-z) + (1-z)\log\left(\frac{1-z}{z(1-d^{-1})}\right) & 0 \le z \le 1, \\ \infty & \text{otherwise.} \end{cases} \tag{51}$$

One can check that $\tilde{\Lambda}^*_{D^1}$ is uniformly greater than $\Lambda^*_{D^1}$ on $(0, d^{-1})$ and uniformly smaller on $[d^{-1}, \infty)$: in other words, conclusions based on $\tilde{\Lambda}^*_{D^1}$ regarding overflow probabilities in subsequent buffers are uniformly more pessimistic than those based on $\Lambda^*_{D^1}$.

5.2 Gaussian inputs

Suppose we have two Gaussian inputs with respective rate functions

$$\Lambda_i^*(x) = (x-\mu_i)^2/2\sigma_i^2, \tag{52}$$

$i = 1, 2$, sharing a deterministic buffer with service rate 1. Note that strictly speaking our results do not apply to this case, because the departure process is not well-defined for arrivals that can take negative values; however, assuming that the parameters are such that negative arrivals are unlikely, this is not a problem. Under the hypotheses of Theorem 3.1, the rate function corresponding to departures of the first stream is given by

$$\Lambda_{D^1}^*(z) = \begin{cases} \Lambda_1^*(z) & 0 \le z \le 1-\mu_2, \\ F(z,(1-z)\wedge\mu_2) \wedge \inf_{x\ge z} \frac{z}{x} F\left(x, \frac{1-z}{z}x\right) & 1-\mu_2 < z \le 1, \\ \infty & \text{otherwise}, \end{cases} \tag{53}$$

where

$$F(x,y) = \frac{(x-\mu_1)^2}{2\sigma_1^2} + \frac{(y-\mu_2)^2}{2\sigma_2^2}. \tag{54}$$

We can simplify (53) by observing that for each z, $x \mapsto zF(x,(1-z)x/z)$ is convex; it follows that if x_z denotes the point at which this function attains its minimum,

$$\Lambda_{D^1}^*(z) = \begin{cases} F(z,(1-z)\wedge\mu_2) \wedge \frac{z}{x_z \vee z} F\left(x_z \vee z, \frac{1-z}{z}(x_z \vee z)\right) & z \in [0,1], \\ \infty & \text{otherwise}; \end{cases} \tag{55}$$

the minimiser x_z is given by

$$x_z^2 = \frac{\mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2}{\sigma_2^2 + (1-z)^2\sigma_1^2/z^2}. \tag{56}$$

Figures 1–4 are plots of Λ_1^* and $\Lambda_{D^1}^*$ on the interval $(0,1)$ for various parameter values. Clearly the key parameter is the variance of the cross stream: a regular cross stream is more influencial than a bursty cross stream, and in all cases the large deviation properties of the first stream are improved (from the point of view of minimising overflow probabilities at subsequent buffers).

6 Towards a calculus on networks

In this section we describe how one could perform a calculus of rate functions on more complicated networks using the heuristics underlying Corollary 3.2; we will illustrate the method with an example. Consider the network represented in Figure 6. The input processes A^i are assumed to be independent, each having stationary increments and each satisfying the sample path large deviation hypothesis of Theorem 3.1 with convex rate functions (in other words the point-to-point geodesics are straight lines) which we denote by Λ_i^*. For simplicity we assume that all buffers are shared according to a FCFS policy with service rate 1, except for buffer d which has service rate 1/2 (otherwise there would be no possibility of overflow at this buffer), and that the system is stable.

Suppose we wish to determine the rate function corresponding to $\tilde{D}^2 + \tilde{D}^3$. We thus consider the event

$$\left\{(\tilde{D}_n^2 + \tilde{D}_n^3)/n \approx z\right\}. \tag{57}$$

On this event, the most likely paths of A^1 and D^2 are straight upto time $\beta_b n$ with respective slopes x_1 and y_2 say, where

$$\beta_b = \frac{1}{x_1 + y_2} \wedge 1, \tag{58}$$

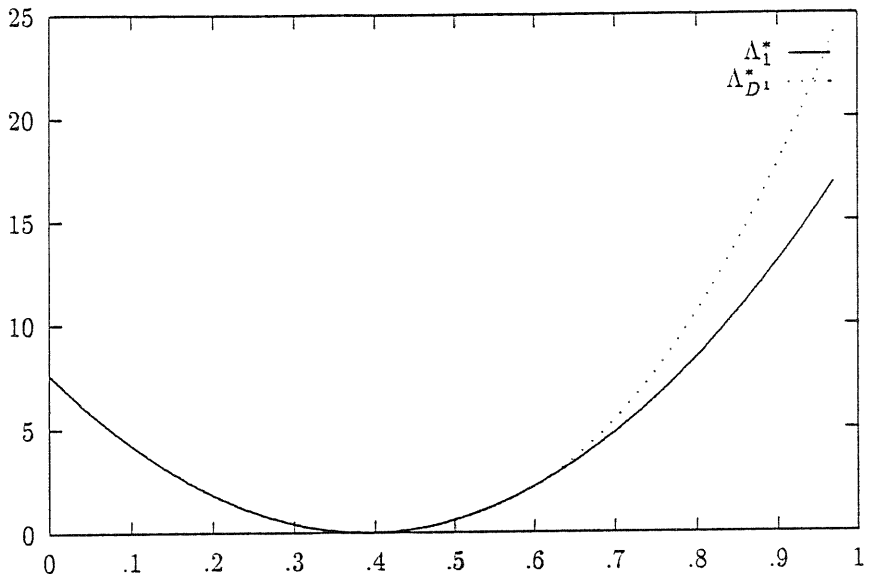

Figure 1: Plot of Λ_1^* and $\Lambda_{D^1}^*$ for $\mu_1 = \mu_2 = 0.4$ and $\sigma_1^2 = \sigma_2^2 = 0.1$.

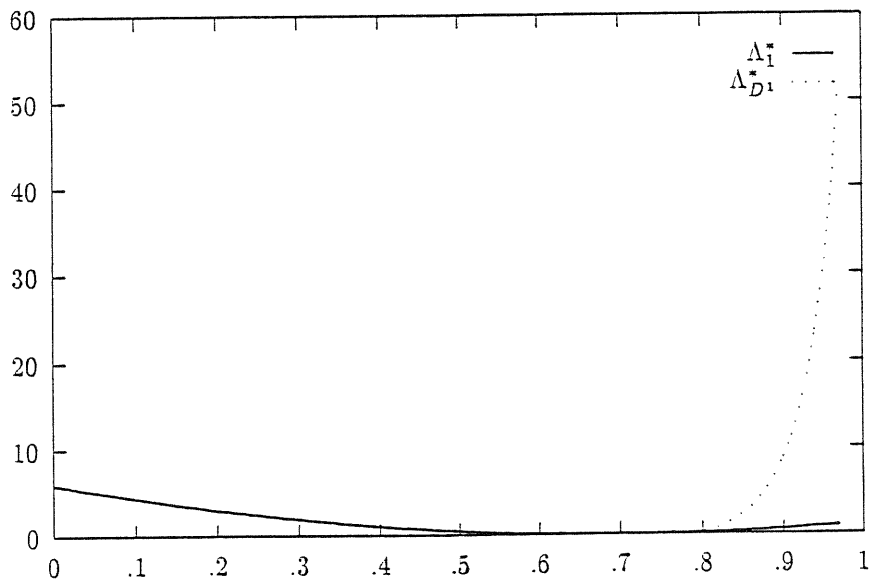

Figure 2: Plot of Λ_1^* and $\Lambda_{D^1}^*$ for $\mu_1 = 0.7$, $\mu_2 = 0.2$, $\sigma_1^2 = 0.2$ and $\sigma_2^2 = 0.01$.

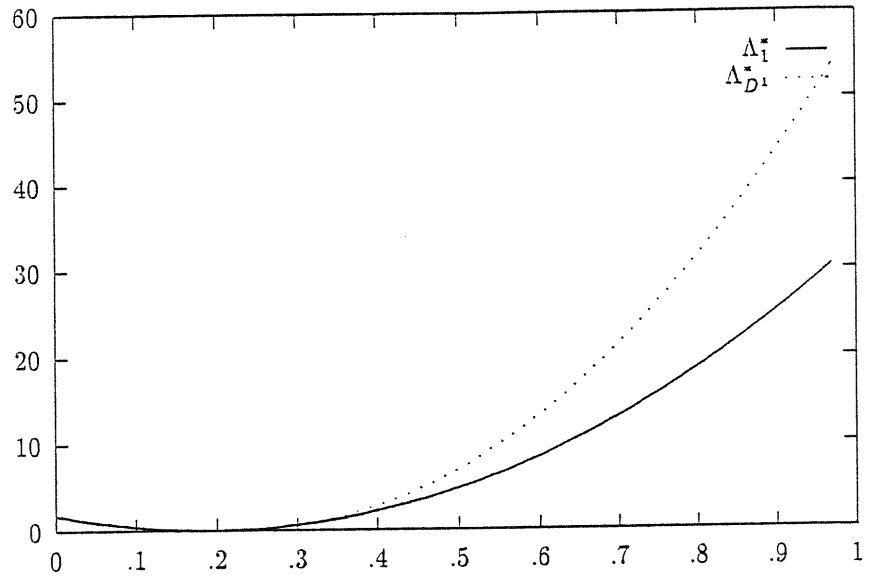

Figure 3: Plot of Λ_1^* and $\Lambda_{D^1}^*$ for $\mu_1 = 0.2$, $\mu_2 = 0.7$, $\sigma_1^2 = 0.1$ and $\sigma_2^2 = 0.1$.

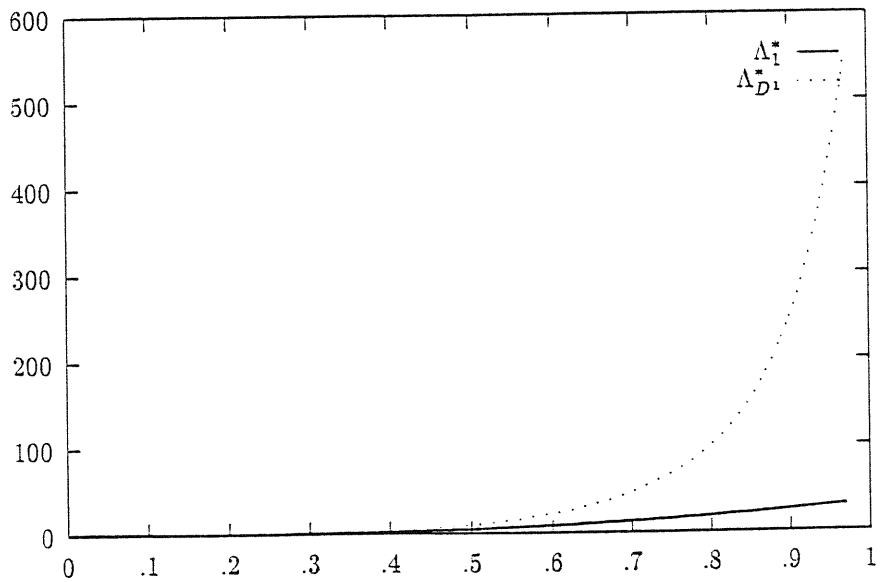

Figure 4: Plot of Λ_1^* and $\Lambda_{D^1}^*$ for $\mu_1 = 0.2$, $\mu_2 = 0.7$, $\sigma_1^2 = 0.1$ and $\sigma_2^2 = 0.01$.

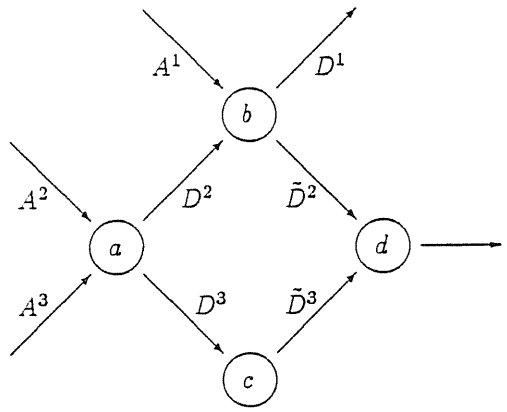

Figure 5: A network example.

following their mean slope thereafter; the most likely path of D^3 is straight with slope y_3, say, upto time $\beta_c n$, where

$$\beta_c = \frac{1}{y_3} \wedge 1, \tag{59}$$

and according to its mean thereafter. These are subject to the constraint

$$y_2\beta_b + y_3\beta_c = z. \tag{60}$$

We must now consider the most likely paths of A^2 and A^3 on the event

$$\left\{ D^2_{\beta_b n}/n \approx y_2,\ D^3_{\beta_c n}/n \approx y_3 \right\}. \tag{61}$$

Suppose $\beta_b < \beta_c$. By the usual Jensen-type arguments we expect A^1 and A^2 to be linear on the intervals $[0, \beta_a n]$ and $[\beta_a n, (\beta_a + \beta'_a)n]$, where $\beta_a n$ and $(\beta_a + \beta'_a)n$ are the times for which

$$A^2_{\beta_a n} = D^2_{\beta_b n} \tag{62}$$

and

$$A^3_{(\beta_a+\beta'_a)n} = D^3_{\beta_c n}. \tag{63}$$

Suppose A^i $(i = 2, 3)$ has consecutive slopes x_i and x'_i on these intervals, and its mean slope thereafter. Proceeding as before we find that

$$\beta_a = \left(\frac{1}{x_2 + x_3} \wedge 1 \right) \beta_b, \tag{64}$$

$$\beta'_a = \left(\frac{1}{x'_2 + x'_3} \wedge 1 \right) (\beta_c - \beta_b), \tag{65}$$

and we impose the constraints

$$x_2\beta_a = y_2, \qquad x_3\beta_a + x'_3\beta'_a = y_3. \tag{66}$$

Combining this with (60) we get

$$x_2\beta_a\beta_b + x_3\beta_a\beta_c + x_3'\beta_a'\beta_c = z. \tag{67}$$

Putting things together we expect the rate function corresponding to the input at d to be given by

$$\Lambda^*_{\tilde{D}^2+\tilde{D}^3}(z) = \inf_{E\cup F} \left\{\beta_b\Lambda_1^*(x_1) + \beta_a[\Lambda_2^*(x_2) + \Lambda_3^*(x_3)] + \beta_a'[\Lambda_2^*(x_2') + \Lambda_3^*(x_3')]\right\}, \tag{68}$$

where E is the set of $x_1, x_2, x_3, x_2', x_3' > 0$ and $\beta_a, \beta_a', \beta_b, \beta_c \in [0,1]$ which satisfy $\beta_b < \beta_c$,

$$\beta_a = \left(\frac{1}{x_2 + x_3} \wedge 1\right)\beta_b, \tag{69}$$

$$\beta_a' = \left(\frac{1}{x_2' + x_3'} \wedge 1\right)(\beta_c - \beta_b), \tag{70}$$

$$\beta_b = \frac{1}{x_1 + \beta_a x_2} \wedge 1, \tag{71}$$

$$\beta_c = \frac{1}{x_3\beta_a + x_3'\beta_a'} \wedge 1, \tag{72}$$

$$x_2\beta_a\beta_b + x_3\beta_a\beta_c + x_3'\beta_a'\beta_c = z, \tag{73}$$

and F is the set of $x_1, x_2, x_3, x_2', x_3' > 0$ and $\beta_a, \beta_a', \beta_b, \beta_c \in [0,1]$ which satisfy $\beta_b \geq \beta_c$,

$$\beta_a = \left(\frac{1}{x_2 + x_3} \wedge 1\right)\beta_c, \tag{74}$$

$$\beta_a' = \left(\frac{1}{x_2' + x_3'} \wedge 1\right)(\beta_b - \beta_c), \tag{75}$$

$$\beta_b = \frac{1}{x_2\beta_a + x_2'\beta_a'} \wedge 1, \tag{76}$$

$$\beta_c = \frac{1}{x_3\beta_a} \wedge 1, \tag{77}$$

$$x_2\beta_a\beta_b + x_2'\beta_a'\beta_b + x_3\beta_a\beta_c = z. \tag{78}$$

Recall that infimums over empty sets are infinite by convention.

It is tempting to now apply (5) to estimate the tail of the queue-length distribution at d, but strictly speaking this can only be done if we know the rate function of the *stationary* version of $\tilde{D}^2 + \tilde{D}^3$. The author is presently extending the results of this paper to the stationary case in [13].

Although the above approach may seem complicated, there is no inherent difficulty in writing a program to carry out the analysis on an arbitrary network and solve the optimisation problems that arise. The method can also be applied to networks with feedback, although in this case the solutions will be implicit.

Acknowledgements. The author would like to thank Gustavo de Veciana for helpful correspondence and suggestions; thanks also to Dmitri Botvich, Nick Duffield, John Lewis, Raymond Russell and Fergal Toomey for many enlightening discussions. The plots were prepared by Fergal Toomey.

References

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

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

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

[4] 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.*

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

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

[7] N.G. Duffield, J.T. Lewis, Neil O'Connell, Raymond Russell and Fergal Toomey. The entropy of an arrivals process: a tool for estimating QoS parameters of ATM traffic. *Proceedings of the 11th IEE Teletraffic Symposium*, Cambridge, March 1994.

[8] N.G. Duffield and Neil O'Connell. Large deviations and overflow probabilities for the general single server queue, with applications. To appear in *Proc. Camb. Phil. Soc.*

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

[10] Amir Dembo and Tim Zajic. Large deviations: from empirical mean and measure to partial sums process. Preprint.

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

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

[13] Neil O'Connell. Queue-lengths and departures at a single-server, multi-class queue. In preparation.

[14] J.L. van den Berg and J.A.C. Resing. The change of traffic characteristics in ATM networks. COST 242 document.