Linear geodesics are not generally preserved by a FIFO queue

A. J. Ganesh* and N. O'Connell†

May 30, 1996

Abstract

Large deviations techniques have proved to be very useful in characterizing the tail of queue length distributions, for queues with very general arrival and service processes. Extending these ideas to networks of queues requires the characterization of the large deviations behaviour at the output of a queue in terms of the behaviour of the input and service processes. We show in this paper that the output process does not, in general, possess a desirable property that is usually assumed for the input process. This property is the convexity of its large deviations rate function, or equivalently, the linear geodesic property of its sample paths. The lack of this property implies that a simple, inductive approach to characterizing the tail of queue length distributions in networks is not feasible in general.

Keywords : Queueing networks, large deviations, multiclass queues.

1 Preliminaries

Consider a discrete time queue with d arrival streams $\mathbf{X} = (X^1, \ldots, X^d)$ sharing an infinite buffer according to an FCFS policy with stochastic service

*Department of Computer Science, University of Edinburgh, JCMB, Kings Buildings, Edinburgh EH9 3JZ. Research supported by a fellowship from BP and the Royal Society of Edinburgh.

†BRIMS, Hewlett Packard Laboratories, Filton Road, Stoke Gifford, Bristol BS12 6QZ.

rate C. $\mathbf{X}_k$ denotes the number of arrivals of each type in time slot k, while C_k denotes the maximum number of customers of any type that can be served in this time slot. We suppose that the queue is initially empty. Define

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

Let $A_n = \sum_{j=1}^{d} \mathbf{A}_n^j$ denote the total number of arrivals, and D_n the total number of departures up to time n. Assuming that the queue is work-conserving, we have

$$D_n = \inf_{0 \le k \le n} (A_k - B_k) + B_n \tag{2}$$

It remains to specify the amount of work, $\mathbf{D}_n = (D_1, \ldots, D_n)$, serviced from each input stream by time n. To do this, we set,

$$T_n = \sup\{k \le n \ : \ A_k \le D_n\} \tag{3}$$

$$\mathbf{D}_n = \mathbf{A}_{T_n} + (D_n - A_{T_n})\mathbf{X}_{T_n+1}/X_{T_n+1} \tag{4}$$

This corresponds to the assumption that simultaneous arrivals from different sources are thoroughly mixed in the queue.

The large deviations behaviour of the input and service processes has been used by a number of authors to derive asymptotics on the tail of the queue length, see, for example. In order to extend such an approach to networks of queues, it is desirable to obtain an expression for the large deviations behaviour of the departure process in terms of those of the arrival and service processes. This has been done by N. O'Connell [4]. We begin by reviewing this result.

Define $\mathbf{S}_n(t) = (\mathbf{A}_{[nt]}/n, B_{[nt]}/n)$, $\mathbf{R}_n(t) = \mathbf{D}_{[nt]}/n$. For each positive integer k, let $\mathcal{L}^k$ denote the subspace of paths in $L_\infty([0,1]^k)$ with non-decreasing components, $\mathcal{C}^k \subset \mathcal{L}^k$ the set of continuous paths starting at zero, and $\mathcal{A}^k \subset \mathcal{C}^k$ the set of those paths with absolutely continuous components. The following assumptions are made in [4].

(A1) For all $\gamma \in I\!R$, $\sup_k E[\exp \gamma(X_k + C_k)] < \infty$.

(A2) For each $\lambda \in I\!R^{d+1}$, the limit

$$\Lambda(\lambda) = \lim_{n\to\infty} \frac{1}{n} \log E[\exp(\lambda \cdot \mathbf{S}_n(1))] \tag{5}$$

exists as an extended real number and is finite in a neighborhood of the origin. The sequence $\mathbf{S}_n$ satisfies the large deviation principle (LDP) in $\mathcal{L}^{d+1}$ with good rate function I given by

$$I(\phi) = \begin{cases} \int_0^1 \Lambda^*(\dot{\phi})ds, & \text{if } \phi \in \mathcal{A}^{d+1} \\ \infty, & \text{otherwise} \end{cases} \tag{6}$$

where Λ^* is the convex conjugate of Λ.

(A3) The arrival and service processes are asymptotically independent in the sense that

$$\Lambda^*(\mathbf{x}, c) = \Lambda_a^*(\mathbf{x}) + \Lambda_b^*(c) \tag{7}$$

Under the above assumptions, it was shown in [4] that the sequence $\mathbf{R}_n$ satisfies the LDP in $\mathcal{L}^d$ with good rate function

$$I_d(\psi) = \inf\{I(\phi) : \Delta(\phi) = \psi\} \tag{8}$$

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

$$\mathbf{A}(\phi) = (\phi^1, \ldots, \phi^d) \tag{9}$$

$$D(\phi)(t) = \inf_{0\leq\nu\leq 1}\left[A(\phi)(\nu t) - \phi^{d+1}(\nu t)\right] + \phi^{d+1}(t) \tag{10}$$

$$T(\phi)(t) = \inf\{r : A(\phi)(r) = D(\phi)(t)\} \tag{11}$$

$$\Delta(\phi) = \mathbf{A}(\phi) \circ T(\phi) \tag{12}$$

It is immediate from the contraction principle that $\mathbf{D}_n/n$ satisfies the LDP in $I\!R^d$ with good rate function

$$\Lambda_d^*(\mathbf{z}) = \inf\{I_d(\psi) : \psi(0) = 0, \psi(1) = \mathbf{z}\}, \tag{13}$$

for I_d as in (8). From this, it was derived in [4] that

$$\begin{aligned} \Lambda_d^*(\mathbf{z}) \quad = \quad & \inf\{\beta\Lambda_a^*(\mathbf{x}/\beta) + \sigma\Lambda_a^*\left(\frac{\mathbf{z}-\mathbf{x}}{\sigma}\right) + \beta\Lambda_b^*(c) + (1-\beta)\Lambda_b^*\left(\frac{z-x}{1-\beta}\right) : \\ & \beta, \sigma \in [0,1],\ c \in I\!R,\ \beta+\sigma \leq 1,\ x \leq \beta c\}. \end{aligned} \tag{14}$$

The last result has the interpretation that the most likely path of the arrival and service processes which results in the departure process having mean rate $\mathbf{z}$ on the interval $[0, n]$ is as follows. The arrival process has rate

$\mathbf{x}/\beta$ on the interval $[0, \beta n]$ and rate $(\mathbf{z}-\mathbf{x})/\sigma$ on $[\beta n, (\beta+\sigma)n]$. The service rate during $[0, \beta n]$ is c, which is greater than the aggregate arrival rate during this period. So the queue is empty during $[0, \beta n]$. The queue is non-empty throughout $[\beta n, n]$, during which period the service rate, at $(z-x)/(1-\beta)$, is no larger than the total arrival rate, which is $(z-x)/\sigma$. The rigorous statement underlying this intuition, proved in [4], is the following:

$$\Lambda_d^*(\mathbf{z}) = I_d(\psi), \tag{15}$$

where $\psi(t)$, $0 \le t \le 1$ is specified by,

$$\psi(0) = 0,\ \dot{\psi}(t) = \begin{cases} \frac{\mathbf{x}}{\beta}, & 0 < t < \beta, \\ \frac{\mathbf{z}-\mathbf{x}}{1-\beta}, & \beta < t < 1. \end{cases} \tag{16}$$

Here β, $\mathbf{x}$ are those achieving the infimum in (14), and I_d is as defined in (8).

The result in (14) applies to a queue started empty. A similar but more involved expression was derived for a queue in equilibrium, which we shall deal with later. Under the above assumptions, we can derive an expression for the asymptotics of the queue length distribution. The problem of extending this derivation to an arbitrary queue in a feed-forward queueing network remains open. The arrival process into any queue in such a network is an aggregate of the departure processes from its predecessors (or splittings thereof) and possibly of an external arrival process. Therefore, the result above suggests that we approach this problem using the LDP for the departure process. This would work if the departure process also satisfied assumptions (A1)-(A3). We show below that there are situations where the departure process fails to satisfy (A2). Note that, due to the convexity of Λ^*, (A2) corresponds to the assumption of linear geodesics. We shall give two counterexamples to show that departures from an empty queue need not have linear geodesics.

2 Counterexamples

2.1 Single customer class

Consider a queue with a stochastic server and a single class of customers (so $d = 1$). Then, (14) simplifies to

$$\begin{aligned} \Lambda_d^*(z) &= \inf\{\beta\Lambda_a^*(x/\beta) + \sigma\Lambda_a^*\left(\frac{z-x}{\sigma}\right) + \beta\Lambda_b^*(c) + (1-\beta)\Lambda_b^*\left(\frac{z-x}{1-\beta}\right) : \\ &\qquad \beta, \sigma \in [0,1],\ c \in I\!R,\ \beta + \sigma \le 1,\ x \le \beta c\}. \end{aligned} \tag{17}$$

where now x and z are scalars.

Let $EX = \Lambda_a'(0)$ denote the mean number of arrivals, and $EC = \Lambda_b'(0)$ the mean number of services in each time slot. The following properties of Λ_a^*, Λ_b^* are well-known, see [3] for instance. Λ_a^* (respectively, Λ_b^*) is non-negative, and zero only at EX (respectively EC). Both Λ_a^* and Λ_b^* are convex, and finite on a non-empty interval, in the interior of which they are analytic. We assume that the queue is stable, namely $EX < EC$. Suppose that for some $\alpha \in [0, EX]$,

$$\Lambda_a^*(\alpha) = \Lambda_b^*(\alpha) \tag{18}$$

and also that these functions are finite in a neighborhood of α. Without loss of generality, we can take α to be the largest number in $[0, EX]$ for which (18) holds. Then, since $\Lambda_a^*(EX) = 0 < \Lambda_b^*(EX)$, we have $\Lambda_a^*(x) < \Lambda_b^*(x)$ for all $x \in (\alpha, EX]$, and consequently that $(\Lambda_a^*)'(\alpha) < (\Lambda_b^*)'(\alpha)$. It follows from this that

$$\exists\, \epsilon > 0\ :\quad \Lambda_b^*(x) < \Lambda_a^*(x) < +\infty \quad \forall\, x \in [\alpha - \epsilon, \alpha], \tag{19}$$

and also that

$$\exists\, 0 < x_1 < \alpha < x_2 < EX :\ \frac{x_1 + x_2}{2} = \alpha,\ \frac{\Lambda_b^*(x_1) + \Lambda_a^*(x_2)}{2} < \Lambda_a^*(\alpha). \tag{20}$$

We shall show that the departure process conditioned to have mean rate α does not have a linear geodesic. Let $\phi(t) = (\phi^1(t), \phi^2(t))$, be defined on $[0,1]$ by $\phi(0) = 0$ and

$$\dot\phi^1(t) = \begin{cases} x_2, & 0 < t < \frac{1}{2}, \\ EX, & \frac{1}{2} < t < 1. \end{cases} \qquad \dot\phi^2(t) = \begin{cases} EC, & 0 < t < \frac{1}{2}, \\ x_1, & \frac{1}{2} < t < 1, \end{cases} \tag{21}$$

where x_1, x_2 are as in (20). Then, since $EX < EC$, we have from (9)-(12) that

$$\Delta(\phi)(t) = \begin{cases} x_2 t, & 0 \le t \le \frac{1}{2}, \\ \frac{1}{2}x_2 + (t - \frac{1}{2})x_1, & \frac{1}{2} \le t \le 1. \end{cases} \tag{22}$$

and in particular that $\Delta(\phi)(1) = \alpha$. Therefore, by (8) and (13),

$$\begin{aligned} \Lambda_d^*(\alpha) &\le I(\phi) \\ &= \int_0^1 \Lambda_a^*(\dot\phi^1(s))ds + \int_0^1 \Lambda_b^*(\dot\phi^2(s))ds \\ &= \int_0^{1/2} \Lambda_a^*(x_2)ds + \int_{1/2}^1 \Lambda_a^*(EX)ds + \int_0^{1/2} \Lambda_b^*(EC)ds + \int_{1/2}^1 \Lambda_b^*(x_1)ds \\ &= \frac{1}{2}\left[\Lambda_a^*(x_2) + \Lambda_b^*(x_1)\right] \\ &< \Lambda_a^*(\alpha) \end{aligned} \tag{23}$$

The first equality above follows from (6) and (7), the second from the definition of ϕ in (21), and the last from the fact that $\Lambda_a^*(EX) = \Lambda_b^*(EC) = 0$, see [3] for example. The last inequality above holds because of (20). Notice that the departure process $\Delta(\phi)$ in (22), corresponding to the arrival process ϕ^1 and service process ϕ^2, is not linear but has different slopes x_2 and x_1 in two different periods of equal length.

Next, let $\psi(t)$ be linear on $[0,1]$ with $\psi(0) = 0$ and $\psi(1) = \alpha$, so that $\dot\psi(t) = \alpha$ for all $t \in (0,1)$. Consider any $\phi \in \mathcal{A}^2$ such that $\psi = \Delta(\phi)$. That is, (ϕ^1, ϕ^2) is any pair of arrival and service processes (excluding those whose rate function is $+\infty$), corresponding to which ψ is the departure process. Then, by (9)-(12),

$$\psi(t) = \inf_{0 \le s \le t}\left[\phi^1(s) - \phi^2(s)\right] + \phi^2(t), \tag{24}$$

from which it is clear that $\psi(t) \le \phi^1(t)$. In particular, $\phi^1(1) \ge \alpha$. If $\phi^1(1) = \alpha$, then, by (6),

$$\begin{aligned} I(\phi) &\ge \int_0^1 \Lambda_a^*(\dot\phi^1(s))ds \\ &\ge \Lambda_a^*\left(\int_0^1 \dot\phi^1(s)ds\right) \\ &= \Lambda_a^*(\alpha). \end{aligned} \tag{25}$$

The first inequality is due to the non-negativity of Λ_a^*, Λ_b^*, the second holds because of Jensen's inequality and the convexity of Λ_a^*, while the equality is because $\phi^1(1)$ was assumed to be α. If $\phi^1(1) > \alpha$, define

$$\tau = \sup\{t \in [0,1] : \phi^1(t) \leq \alpha t\} \tag{26}$$

and note that $\tau < 1$. Hence, by continuity of ϕ, $\phi^1(\tau) = \alpha\tau$.

Lemma 1 *Suppose that $\psi(t) = \alpha t$ for all $t \in [0,1]$, where ψ is defined by (24), and that $\phi \in \mathcal{A}^2$. Then, with τ given by (26),*

$$\phi^2(t) - \phi^2(\tau) = \alpha(t-\tau) \quad \forall\, t \in [\tau, 1].$$

Proof : As noted above, $\phi^1(\tau) = \alpha\tau = \psi(\tau)$, the latter equality holding by assumption regarding ψ. From this, we see that the infimum in (24) corresponding to $t = \tau$ is achieved at $s = \tau$. Consequently, (24) implies that

$$\psi(t) = \inf_{\tau \leq s \leq t} \left[\phi^1(s) - \phi^2(s)\right] + \phi^2(t) \quad \forall\, t \geq \tau. \tag{27}$$

If the infimum above is achieved at τ for all $t \in [\tau, 1]$, then, for all t in this interval, $\phi^2(t) - \phi^2(\tau) = \psi(t) - \psi(\tau) = \alpha(t - \tau)$, and so the lemma is established. Otherwise, because ϕ is absolutely continuous, one of the following must hold:

$$\exists\, \epsilon > 0 : \dot{\phi}^1(s) - \dot{\phi}^2(s) < 0 \quad \forall\, s \in (\tau, \tau+\epsilon), \tag{28}$$

or

$$T \triangleq \inf\{s > \tau : \phi^1(s) - \phi^2(s) < \phi^1(\tau) - \phi^2(\tau)\} \in (\tau, 1). \tag{29}$$

In the former case, the infimum in (27) corresponding to $t = \tau + \epsilon$ is achieved at $s = \tau + \epsilon$, and so

$$\psi(\tau+\epsilon) - \psi(\tau) = \phi^1(\tau+\epsilon) - \phi^1(\tau) > \alpha\epsilon, \tag{30}$$

where the inequality follows from the definition of τ in (26). In the latter case, we see from the continuity of ϕ that the infimum in (27) corresponding to $t = T$ is achieved at $s = T$. So $\psi(T) = \phi^1(T)$ and

$$\psi(T) - \psi(\tau) = \phi^1(T) - \phi^1(\tau) > \alpha(T - \tau), \tag{31}$$

where the inequality follows from (26). Now, both (30) and (31) contradict the assumption that $\psi(t) = \alpha t$ for all $t \in [0,1]$. Therefore, neither (28) nor (29) can hold, implying that the infimum in (27) must be achieved at τ for all $t \in [\tau, 1]$. This completes the proof of the lemma.

□

From the above lemma and (26), we obtain using (6) that

$$\begin{aligned} I(\phi) &\geq \int_0^\tau \Lambda_a^*(\dot{\phi}^1(s))ds + \int_\tau^1 \Lambda_b^*(\dot{\phi}^2(s))ds \\ &\geq \tau\Lambda_a^*(\alpha) + (1-\tau)\Lambda_b^*(\alpha) \\ &= \Lambda_a^*(\alpha). \end{aligned} \tag{32}$$

The first inequality is due to the non-negativity of Λ_a^*, Λ_b^* and the second due to their convexity and Jensen's inequality. The equality follows from the definition of α in (18).

Let ψ be given by $\psi(t) = \alpha t$, $t \in [0,1]$. Since either (25) or (32) applies to any $\phi \in \mathcal{A}^2$ for which $\Delta\phi = \psi$, observe from (8) that $I_d(\psi) \geq \Lambda_a^*(\alpha)$. Therefore, by (23), ψ does not achieve the infimum in (13) corresponding to $z = \alpha$. In other words, the departure process with constant rate α is not the most likely to achieve an average departure rate α; this is achieved by a process with a nonlinear path. This implies that I_d cannot be expressed in the form

$$I_d(\xi) = \int_0^1 \Lambda^*(\dot{\xi})ds$$

for any convex function Λ^* and so the departure process does not satisfy assumption (A2).

The above conclusion applies to a queue started empty. We now consider a queue in stationarity. Let Q_0 denote the queue length at time 0. It is shown in [4] that the scaled queue lengths Q_0/n satisfy an LDP in $\mathbb{R}$ with rate function L, which is explicitly computed. For our purposes, it is enough to note that $L(0) = 0$ and that $L(q) \geq 0$ for all $q > 0$. Suppose the scaled initial queue length is q, and that the scaled process of arrivals and services is described by $\phi = (\phi^1, \phi^2)$. Then, the scaled departure process up to time t is given by

$$D(q,\phi)(t) = \phi^2(t) \wedge \inf_{0\leq\nu\leq 1}[q + \phi^1(\nu t) - \phi^2(\nu t) + \phi^2(t)], \tag{33}$$

where $x \wedge y$ denotes $\min\{x, y\}$. Notice that since $\phi(0) = \mathbf{0}$, we recover (10) for the departure process from an empty queue by substituting $q = 0$. We shall show that for any $q > 0$ and $\phi \in \mathcal{L}^2$ such that $D(q, \phi(t)) = \alpha t$ for all

$t \in [0,1]$, we have $L(q) + I(\phi) \geq \Lambda^*(\alpha)$. This will enable us to conclude that the departure process does not have linear geodesics, even in equilibrium.

Let ψ be linear on $[0,1]$ with $\psi(t) = \alpha t$ for all $t \in [0,1]$. Fix $q > 0$ and let $\phi \in \mathcal{C}^2$ be such that $D(q,\phi) = \psi$. Then, by (33), either $\phi^2(t) = \alpha t$ for all $t \in [0,1]$, or $q + \phi^1(s) - \phi^2(s) < 0$ for some $s \in [0,1]$. In the former case, we have by (6), (7) and the non-negativity of the Λ^* that

$$I(\phi) \geq \int_0^1 \Lambda_b^*(\dot{\phi}^2(s))ds = \Lambda_b^*(\alpha). \tag{34}$$

In the latter case, we have by the continuity of ϕ that

$$\tau \triangleq \inf\{s \in [0,1] : q + \phi^1(s) - \phi^2(s) < 0\} \ \in \ [0,1).$$

It follows that $D(q, \phi(t)) = \phi^2(t)$ for all $t \in [0,\tau]$, whereas, for $t \in [\tau, 1]$,

$$D(q,\phi)(t) - D(q,\phi)(\tau) = \phi^2(t) - \phi^2(\tau) + \inf_{\tau \leq s \leq t} [q + \phi^1(s) - \phi^2(s)],$$

because $q + \phi^1(s) - \phi^2(s)$ takes its minimum value on $[0,\tau]$ at τ, and this value is zero. Hence, we can rewrite the above as

$$\begin{aligned} &D(q,\phi)(t) - D(q,\phi)(\tau) \\ &= \inf_{\tau \leq s \leq t} \Big[\Big(\phi^1(s) - \phi^1(\tau)\Big) - \Big(\phi^2(s) - \phi^2(\tau)\Big)\Big] + \phi^2(t) - \phi^2(\tau). \end{aligned} \tag{35}$$

Define $\tilde{\phi}(t) = \phi(t) - \phi(\tau)$, $t \in [\tau, 1]$. Then, we have from above that

$$\psi(t) = \begin{cases} \phi^2(t), & \text{if } t \in [0,\tau], \\ \inf_{\tau \leq s \leq t} \Big[\tilde{\phi}^1(s) - \tilde{\phi}^2(s)\Big] + \tilde{\phi}^2(t) + \phi^2(\tau), & \text{if } t \in (\tau, 1]. \end{cases} \tag{36}$$

Comparing this with (24), we see that the departure process on $[\tau, 1]$ is identical to that from an empty queue with arrival and service processes given by $\tilde{\phi}$. This is not surprising because the queue does, in fact, become empty at time τ by definition of τ. Since $\tilde{\phi}$, restricted to $[\tau, 1]$, is merely a shifted version of ϕ on this interval, $I(\phi) = I(\tilde{\phi})$ for ϕ restricted to this interval. Therefore,

$$I(\phi) \geq \int_0^\tau \Lambda_b^*(\dot{\phi}^2(s))ds + I(\tilde{\phi}).$$

Now, since $\psi(t) = \alpha t$, $\dot{\phi}^2(s) = \alpha$ for all $s \in [0, \tau]$. Also, by the same derivation as leads to (25) and (32), we have $I(\tilde{\phi}) \geq (1-\tau)\Lambda_a^*(\alpha)$. Finally, since $\Lambda_a^*(\alpha) = \Lambda_b^*(\alpha)$ by definition of α, and $L(q) \geq 0$ for all q, we get

$$L(q) + I(\phi) \geq \Lambda_a^*(\alpha).$$

This holds for all initial queue lengths $q \geq 0$, and arrival and service processes ϕ, that result in a linear departure process $\psi(t) = \alpha t$. Note that (23) continues to hold for departures in equilibrium because it was derived for departures from an empty queue, and we have $L(0) = 0$, see [4]. Therefore,

$$\Lambda_d^*(\alpha) < \inf\{L(q) + I(\phi) : D(q,\phi)(t) = \alpha t \ \ \forall\, t \in [0,1]\}, \tag{37}$$

which implies that, conditional on a mean departure rate of α, the most likely path is not linear. Thus, even in equilibrium, the departure process does not necessarily have linear geodesics.

We end this subsection with some comments about the scope and implications of the above results. A careful look at the proof shows that the result relied on α being less than EX and on the rate functions of the arrival and service processes intersecting at α. If the service process is deterministic, the latter cannot happen, and in this case it can be shown that the departure process has linear geodesics. This makes it possible to analyze networks of deterministic server queues, as in Chang [2]. Likewise, if we consider only $\alpha > EX$, then too it can be shown that the departure process conditioned on having mean rate α is linear. Since we are typically interested in the problem of queue lengths exceeding some large threshold, and since in well-designed networks this requires departure rates exceeding their mean, we are usually only interested in the rate function of departures for $\alpha > EX$. Since we have linear geodesics in this region, the study of networks of queues using a recursive approach is again feasible. Such an approach has been taken in Bertsimas *et al.*, [1]. We shall next show that neither of these features comes to our rescue when dealing with multi-class queues. In this case, the joint departure process can have non-linear geodesics even if the server is deterministic, and even if we consider departures whose aggregate rate exceeds their mean.

2.2 Two customer classes

Consider a queue multiplexing two customer classes, and served deterministically at rate c. Suppose that customers from the first class arrive deterministically at rate a, while those of the second have a stochastic arrival process satisfying assumptions (A1)-(A3) with the rate function Λ_2^*. We assume that the mean aggregate arrival rate is strictly less than the service rate, c. Note that the two arrival streams are trivially independent, as are the arrival and service processes. We denote the large deviations rate function of the first arrival process by Λ_1^* and that of the service process by Λ_b^*. So

$$\Lambda_1^*(x) = \begin{cases} 0, & \text{if } x = a, \\ +\infty, & \text{else.} \end{cases} \qquad \Lambda_b^*(x) = \begin{cases} 0, & \text{if } x = c, \\ +\infty, & \text{else.} \end{cases} \tag{38}$$

For some $\epsilon > 0$ and $b < c + \epsilon - a$, let $\mathbf{z} = (a - \epsilon, b)$ and consider the departure process conditioned to have mean rate $\mathbf{z}$. We shall show that this departure process does not have linear geodesics.

Let $\psi \in \mathcal{A}^2$ be linear with $\psi(0) = 0$ and $\psi(1) = \mathbf{z}$, so $\dot{\psi} = \mathbf{z}$. We show that there is no $\phi \in \mathcal{A}^3$ with $I(\phi) < +\infty$ such that $\Delta(\phi) = \psi$. In other words, there is no process of arrivals and services whose rate function is finite, corresponding to which ψ is the departure process. Suppose otherwise. Let $\phi \in \mathcal{A}^3$ have $I(\phi) < +\infty$, so that

$$\phi^1(t) = at, \quad \phi^3(t) = ct, \tag{39}$$

and suppose that

$$\Delta(\phi)(t) = \psi(t) = ((a - \epsilon)t, bt), \tag{40}$$

where $\epsilon > 0$ and $a + b - \epsilon < c$. Observe from (9)-(12) that

$$\Delta(\phi)(t) = \left(\phi^1\Big(T(\phi)(t)\Big), \phi^2\Big(T(\phi)(t)\Big)\right). \tag{41}$$

Therefore, by (39) and (40), $T(\phi)(t) = (a - \epsilon)t/a$, and so $\phi^2\Big((a-\epsilon)t/a\Big) = bt$. In addition, by (11),

$$D(\phi)(t) = \phi^1(T(\phi)(t)) + \phi^2(T(\phi)(t)) = (a + b - \epsilon)t. \tag{42}$$

But, by (10),

$$D(\phi)(t) = \inf_{0 \le s \le t} [\phi^1(s) + \phi^2(s) - \phi^3(s)] + \phi^3(t),$$

and so, by (39) and the fact, noted above, that $\phi^2(s) = abs/(a-\epsilon)$, we get

$$\begin{aligned} D(\phi)(t) &= \inf_{0\le s\le t}[as + \frac{abs}{a-\epsilon} - cs] + ct \\ &= \begin{cases} ct, & \text{if } \frac{a}{a-\epsilon}(a+b-\epsilon) \ge c, \\ \frac{a}{a-\epsilon}(a+b-\epsilon)t, & \text{else.} \end{cases} \end{aligned}$$

Because of our assumption that $a+b-\epsilon < c$, we have $D(\phi)(t) > (a+b-\epsilon)t$ in either case above, contradicting (42). We have thus shown that, if $I(\phi) < +\infty$, then $\Delta(\phi) = \psi$ is impossible for $\psi(t) = \mathbf{z}t$ with $\mathbf{z} = (a-\epsilon, b)$. Therefore, by (8), $I_d(\psi) = +\infty$.

We now show that $\Lambda_d^*(\mathbf{z}) < +\infty$ for $\mathbf{z}$ as above. Since $\epsilon > 0$ was arbitrary, we assume without loss of generality that $a - \epsilon > 0$ and define

$$x_1 = a + 2(b-\epsilon) - c, \quad x_2 = \frac{a}{a-2\epsilon}(c - a + 2\epsilon). \tag{43}$$

Since the only requirement we imposed above was that $a+b-\epsilon < c$, it is clear that b and ϵ can be chosen so that $x_1 \ge 0$. Also, $x_2 > 0$ since it was assumed that c is larger than a. Let $\phi \in \mathcal{A}^3$ be defined by

$$\phi(0) = 0, \quad \dot\phi(t) = \begin{cases} (a, x_1, c), & 0 < t < 1/2, \\ (a, x_2, c), & 1/2 < t < 1. \end{cases} \tag{44}$$

Since x_1 and x_2 are non-negative, ϕ has non-decreasing components as required by the definition of $\mathcal{A}^3$. Note that

$$a + x_1 = 2(a+b-\epsilon) - c < c$$

by the assumption that $a+b-\epsilon < c$, whereas

$$a + x_2 = \frac{ac}{a-2\epsilon} > c.$$

In other words, the aggregate arrival rate $a + x_1$ is less than the service rate c during $[0, 1/2]$ whereas, at $a + x_2$, it is greater than c during $[1/2, 1]$. Therefore, the joint departure process $\Delta(\phi)$ is given by

$$\Delta(\phi)(0) = 0, \quad \frac{d}{dt}\Delta(\phi)(t) = \begin{cases} (a, x_1), & 0 < t < 1/2, \\ \left(\frac{a}{a+x_2}c, \frac{x_2}{a+x_2}c\right), & 1/2 < t < 1. \end{cases} \tag{45}$$

This is intuitively clear from the description of the queue, but can also be formally established using (9)-(12). Hence, we have from (43) that

$$\begin{aligned}\Delta(\phi)^1(1) &= \frac{1}{2}\left(a + \frac{ac}{a+x_2}\right) = a - \epsilon, \\ \Delta(\phi)^2(1) &= \frac{1}{2}\left(x_1 + \frac{x_2 c}{a+x_2}\right) = b.\end{aligned}$$

Therefore, by definition of $\mathbf{z}$, $\Delta\phi(1) = \mathbf{z}$. Furthermore, by (44), (6) and (38), we have

$$\Lambda^*(\phi) = \frac{1}{2}\left(\Lambda_2^*(x_1) + \Lambda_2^*(x_2)\right) \tag{46}$$

for x_1, x_2 as in (43). Therefore, $\Lambda^*(\phi)$ is finite if $\Lambda_2^*(x_1)$ and $\Lambda_2^*(x_2)$ are, as is true if, for instance, the second arrival process is Poisson. It now follows from (8) and (13) that $\Lambda_d^*(\mathbf{z}) < +\infty$. But we showed earlier that $I_d(\psi) = +\infty$ for ψ given by $\psi(t) = \mathbf{z}t$. Therefore, the departure process with linear path does not achieve the infimum in (13), implying that the departure process does not satisfy a large deviations principle with action functional that is the integral of a convex rate function. In other words, it is not true that

$$I_d(\xi) = \int_0^1 \Lambda^*(\xi(s))ds$$

for any convex function Λ^*. Consequently, the joint departure process does not satisfy assumption (A2), and so a recursive approach to estimating asymptotics of the queue lengths in a network does not appear feasible.

We now consider the same queueing system in stationarity, rather than started empty. It is shown in O'Connell [4] that in stationarity, the scaled queue lengths $\mathbf{Q}_0/n$ satisfy an LDP in $\mathbb{R}^2$ with a rate function L that can be computed explicitly. Here $\mathbf{Q}_0 = (Q_0^1, Q_0^2)$ denotes the number of customers of each of the two types in the queue at time zero. It suffices for our purposes to note that $L(\mathbf{q}) \geq 0$ for all $\mathbf{q} \geq 0$, with equality if $\mathbf{q} = \mathbf{0}$.

Consider the system starting at time zero with scaled queue length $\mathbf{Q}_0/n = \mathbf{q}$. Suppose the arrival and service processes are given by $\phi = (\phi^1, \phi^2, \phi^3)$, and that $\psi = (\psi^1, \psi^2)$ is the corresponding departure process. Let a and c be defined as above to be the deterministic rate of the first arrival process and the service process respectively. Let $\epsilon \in (0, a)$ and $b > 0$ be such that $a + b - \epsilon < c$. We shall show that if $\psi(t) = \mathbf{z}t$ for all $t \in [0, 1]$, where $\mathbf{z} = (a - \epsilon, b)$, then $L(\mathbf{q}) + I(\phi) = \infty$.

Analogous to (33), the scaled process of aggregate departures up to time t is given by

$$D(\mathbf{q},\phi)(t) = \phi^3(t) \wedge \inf_{0\le\nu\le 1}[q + A(\phi)(\nu t) - \phi^3(\nu t) + \phi^3(t)], \tag{47}$$

where $q \triangleq q^1+q^2$ is the total number in queue at time zero, and $A(\phi) \triangleq \phi^1+\phi^2$ is the aggregate arrival process. Note that setting $q=0$ above recovers (10) since $\phi(0)=\mathbf{0}$. Now, if ψ is to be the departure process, then we must have $D(\mathbf{q},\phi) = \psi^1+\psi^2$. Since $\psi(t) = \mathbf{z}t$, with $\mathbf{z} = (a-\epsilon, b)$, the above implies that $D(\mathbf{q},\phi) = (a+b-\epsilon)t$ for all $t \in [0,1]$. Recall that if $I(\phi)$ is to be finite, then we must have $\phi^1(t) = at$ and $\phi^3(t) = ct$ for all $t \in [0,1]$, since the first arrival process and the service process are deterministic with rates a, c respectively. Therefore, for all such ϕ, (47) implies that

$$(a+b-\epsilon)t = ct \wedge \inf_{0\le s\le t}[q + A(\phi)(s) - cs] + ct.$$

Since $a+b-\epsilon < c$, the above implies that $\inf_{0\le s\le t}[q+A(\phi)(s)-cs]$ is strictly negative for all $t>0$. Now $A(\phi)(s) = \phi^1(s)+\phi^2(s)$, $\phi(0)=\mathbf{0}$ and ϕ is continuous if $I(\phi)$ is finite. Therefore, it follows from the above that $\mathbf{q}=0$, i.e., the queue must start empty. Then, by the argument above for departures from an empty queue, there is no process ϕ of arrivals and services such that the departure process is ψ and $I(\phi)$ is finite. We have also shown that this conclusion does not change if we allow any positive initial queue size, $\mathbf{q}$. This completes the proof that $I_d(\psi) = \infty$ even in equilibrium, where I_d denotes the rate function of the departure process.

We argued above that $\Lambda_d^*(\mathbf{z}) < \infty$ for departures from an empty queue. Since $L(0)=0$, see [4], this argument applies to departures in equilibrium as well. Thus, the most likely path leading to a mean departure rate $\mathbf{z}$ is not linear. This also implies that the rate function $I_d(\psi)$ for equilibrium departures cannot be of the form $\int_0^1 \Lambda^*(\dot\psi)(s)ds$, for any convex function Λ. Therefore, the departure process does not satisfy Assumption A2, needed to apply the results of [4] inductively to feed-forward multi-class queueing networks.

References

[1] D. Bertsimas, I. Paschalidis and J. Tsitsiklis, On the large deviations behaviour of acyclic networks of G/G/1 queues, *Preprint*, 1994.

[2] C.S. Chang, Sample path large deviations and intree networks, *Queueing Systems*, 20(1-2): 7-36, 1995.

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

[4] N. O'Connell, Large deviations in queueing networks, DIAS Technical Report DIAS-APG-9413.