Invariant Rate Functions for Discrete-Time Queues

Ayalvadi Ganesh[1], Neil O'Connell, Balaji Prabhakar[2]
Basic Research Institute in the Mathematical Sciences
HP Laboratories Bristol

HPL-BRIMS-2000-29
December 18th , 2000*

large deviations,
fixed points,
tandem queues

We consider a discrete time queue with general service distribution and characterize a class of arrival processes whose large deviation rate function remains unchanged in passing through the queue. This invariant rate function corresponds to a kind of *exponential tilting* of the service distribution. We establish a large deviations analogue of quasi-reversibility for this class of arrival processes. Finally, we prove the existence of stationary point processes whose probability law is preserved by the queueing operator, and conjecture that these have large deviation rate functions which belong to the class of invariant rate functions described above.

* Internal Accession Date Only Approved for External Publication

[1] Microsoft Research, 1 Guildhall Street, Cambridge, CB2 3NH, UK

[2] Dept of Electrical Eng. Stanford University, Stanford, CA 94305, USA

1 Introduction

Burke's theorem says that if the arrival process to a $\cdot/M/1$ queue is Poisson with rate less than the service rate, then the departure process in equilibrium is also Poisson with the same rate. In other words, a Poisson process of rate α is a fixed point of the $\cdot/M/1$ queue with service rate 1, for every $\alpha < 1$. It has recently been shown [14] that a similar result holds for single-server queues with a general service time distribution: non-trivial stationary ergodic fixed points do exist. However, little is known about the properties of fixed points.

In this paper we consider the fixed point question at the large deviations scaling. Assuming the service process satisfies a sample path large deviation principle, we identify a class of arrival processes whose sample path large deviations behaviour is preserved by the queue, and conjecture that fixed points belong to this class. The invariant rate function corresponding to a given arrival rate is given by a kind of exponential tilting of the service distribution. This suggests that, in some sense, the fixed point is as similar in relative entropy to the service process as it can be, subject to its rate constraint. To make sense of this interpretation, however, raises more questions and seems to be an interesting topic for future research. For example, is the fixed point a Gibbs measure?

For completeness, we also present some results on the existence and attractiveness of fixed points for discrete-time queues. We show that the continuous-time results of [14] and [15] can be reproduced in discrete time with minor modifications.

The results in this paper are derived in the context of a discrete time queueing model which we now describe. The queue has arrival process $\{A_n, n \in \mathbb{Z}\}$, where A_n denotes the amount of work arriving in the n^{th} time slot. The service process is denoted by $\{S_n\}$, where S_n denotes the maximum amount of work that can be completed in the n^{th} time slot. The arrival and service processes are assumed to be stationary and ergodic sequences of positive real random variables. The workload process, $\{W_n\}$, is described by Lindley's recursion: $W_{n+1} = \max\{W_n + A_n - S_n, 0\}$. The amount of work departing in time slot n is given by

$$D_n = A_n + W_n - W_{n+1} = \min\{W_n + A_n, S_n\}. \tag{1}$$

If A_n and S_n are integer-valued for all n, then W_n can be thought of as the number of customers in the queue at time n.

In the next section we will present the relevant large deviation results from [17] and [8], and identify a class of rate functions which are preserved by the queueing operator. In sections 3 and 4, we present some results on the existence and attractiveness of fixed points.

2 Invariant rate functions for the single server queue

Let $\mathcal{X}$ be a Hausdorff topological space with Borel σ-algebra $\mathcal{B}$, and let X_n be a sequence of random variables taking values in $\mathcal{X}$. A *rate function* is a non-negative lower semicontinuous function on $\mathcal{X}$. We say that the sequence X_n satisfies the *large deviation principle* (LDP) with rate function I, if for all $B \in \mathcal{B}$,

$$-\inf_{x \in B^\circ} I(x) \le \liminf_n \frac{1}{n} \log P(X_n \in B) \le \limsup_n \frac{1}{n} \log P(X_n \in B) \le -\inf_{x \in \bar{B}} I(x).$$

Here B° and $\bar{B}$ denote the interior and closure of B, respectively. A large deviation rate function is *good* if it has compact level sets.

Let $\tilde{S}_n$ denote the polygonal approximation to the scaled service process, defined for $t \geq 0$ by:

$$\tilde{S}_n(t) = \hat{S}_n(t) + (nt - \lfloor nt \rfloor) \left(\hat{S}_n \left(\frac{\lfloor nt \rfloor + 1}{n} \right) - \hat{S}_n \left(\frac{\lfloor nt \rfloor}{n} \right) \right),$$

where

$$\hat{S}_n(t) = \frac{1}{n} \sum_{k=1}^{\lfloor nt \rfloor} S_k.$$

Given an arrival process A_n, we define $\tilde{A}_n(t)$ analogously. Let $\mathcal{C}(\mathbb{R}_+)$ denote the space of continuous functions on the positive real line and $\mathcal{AC}(\mathbb{R}_+)$ the subset of absolutely continuous functions. We now record some hypotheses.

Assumptions:

1. The sequences $\{A_n\}$ and $\{S_n\}$ are stationary and ergodic, and independent of each other. The limiting cumulant generating functions,
$$\Lambda_A(\theta) = \lim_{n\to\infty} \frac{1}{n} \log E \exp \theta(A_1 + \ldots + A_n),$$
$$\Lambda_S(\theta) = \lim_{n\to\infty} \frac{1}{n} \log E \exp \theta(S_1 + \ldots + S_n),$$
exist as extended real numbers for all $\theta \in \mathbb{R}$, are differentiable at the origin and lower semicontinuous.

2. The sequences $\tilde{A}_n$ and $\tilde{S}_n$ both satisfy the LDP in $\mathcal{C}(\mathbb{R}_+)$ equipped with the topology of uniform convergence on compacts, with respective rate functions $\mathcal{I}_A$ and $\mathcal{I}_S$ given by:
$$\mathcal{I}_A(\phi) = \begin{cases} \int_0^\infty I_A(\dot{\phi}(t))dt, & \text{if } \phi \in \mathcal{AC}(\mathbb{R}_+), \\ +\infty, & \text{otherwise,} \end{cases}$$
where
$$I_A(x) = \sup_{\theta\in\mathbb{R}} \{\theta x - \Lambda_A(\theta)\}$$
is the convex dual of Λ_A; $\mathcal{I}_S$ and I_S are described similarly in terms of Λ_S.

3. The stability condition $\Lambda'_A(0) < \Lambda'_S(0)$ holds.

4. $I_A(x) \le I_S(x)$ for all $x \le \Lambda'_A(0)$.

It has been shown by a number of authors under different levels of generality (see, for example, [2, 5, 6, 10]) that the tail of the workload distribution in equilibrium satisfies
$$\lim_{b\to\infty} \frac{1}{b} \log P(W > b) = -\delta, \tag{2}$$

where

$$\delta = \inf_{T>0} T I_W(1/T) \tag{3}$$

and, for $w > 0$,

$$I_W(w) = \inf_{a \geq w} \big[I_A(a) + I_S(a - w) \big]. \tag{4}$$

In order for the workload to build up at rate w over a long period of time, arrivals over this period must occur at some rate a exceeding the service rate by w; the most likely way for this to happen is found by minimizing the expression in (4) over all possible choices of a. Large workloads occur by the queue building up at rate $1/T$ over a period of (scaled) length T, chosen optimally according to (3). The decay rate δ has the following alternative characterization:

$$\delta = \sup\{\theta : \Lambda_A(\theta) + \Lambda_S(-\theta) \leq 0\}. \tag{5}$$

Let $\tilde{D}_n(t)$ denote the scaled departure process, defined analogous to $\tilde{A}_n(t)$ and $\tilde{S}_n(t)$, and let $\tilde{W}_n(t) = W(\lfloor nt \rfloor)/n$ denote the scaled workload at time $\lfloor nt \rfloor$.

Theorem 1 *[17, Theorem 3.3] Under assumptions 1-3, the sample mean $\tilde{D}_n(1) = D_n/n$ of the equilibrium departure process satisfies an LDP in $\mathbb{R}$ with rate function I_D given by*

$$\begin{aligned} I_D(z) \quad = \quad \inf\Bigg\{ & \delta q + \beta_1 \Big[I_A\Big(\frac{z_1 - q}{\beta_1}\Big) + I_S\Big(\frac{z_1}{\beta_1}\Big)\Big] + \beta_2 \Big[I_A\Big(\frac{z_2}{\beta_2}\Big) + I_S(c_2)\Big] \\ & + \tau I_A\Big(\frac{z - z_1 - z_2}{\tau}\Big) + (1 - \beta_1 - \beta_2) I_S\Big(\frac{z - z_1 - z_2}{1 - \beta_1 - \beta_2}\Big)\Bigg\}, \end{aligned} \tag{6}$$

subject to the constraints that

$$q, z_1, z_2, \beta_1, \beta_2, c_2, \tau \geq 0,\ \beta_1 + \beta_2 + \tau \leq 1,\ \beta_2 c_2 \geq z_2,\ z - z_1 - z_2 \geq 0. \tag{7}$$

The interpretation is as follows. Let $q, z_1, z_2, \beta_1, \beta_2, c_2, \tau$ achieve the infimum above subject to the constraints. The most likely path resulting in departures at rate z in equilibrium is the following. The system starts with an initial queue size q at time 0. Then, in the first phase of length β_1, arrivals occur at rate $(z_1 - q)/\beta_1$ and services at rate z_1/β_1, so that at the end of this period the queue is empty and z_1 customers have departed. In the next phase, of length β_2, customers arrive at rate z_2/β_2, which is no more than the available service rate c_2 during this period; hence the queue remains empty and an additional z_2 customers depart. The available service rate during the final phase of length $1-\beta_1-\beta_2$ is $(z-z_1-z_2)/(1-\beta_1-\beta_2)$. The arrival rate is $(z - z_1 - z_2)/\tau$ during the initial τ units of this phase, and is the mean arrival rate for the remainder, of length $1 - \beta_1 - \beta_2 - \tau$. Clearly, only $z - z_1 - z_2$ customers can depart during this final phase, bringing the total departures to z. The reason that the optimal path can have at most three phases has to do with the convexity of I_A and I_S. This implies that the arrival and service rates must be constant from when the queue is first empty until the time that it is last empty during the scaled time interval $[0, 1]$. Likewise the arrival and service rates must be constant from the start until the time the queue is first empty, and from the time the queue is last empty until the end of the time period. This interpretation helps us to write down the joint rate function for the sample mean of the scaled departure process during $[0, 1]$ and the scaled workload in queue at time 1. We also note that arrival and service rates must be constant through the first two phases; if not, 'straightening' by replacing the paths of the arrival and service processes over the first two phases with straight lines at the respective mean rates leaves the total departures unchanged but reduces the objective function in (6). Likewise, the arrival rate must be constant throughout the final phase. Thus, we can modify Theorem 1 as follows.

Under assumptions 1-3, the sample mean $\tilde{D}_n(1)$ of the equilibrium departure process over the period $(0, n)$ and the scaled workload $\tilde{W}_n(1)$ at time n

jointly satisfy an LDP in $\mathbb{R}^2$ with rate function $I_{D,W}$ given by

$$I_{D,W}(z,w) = \min\{\inf_{q\in C_1} f_1(q),\ \inf_{x\in C_2} f_2(x)\}, \tag{8}$$

where $x = (q, z_1, z_2, \beta)$,

$$\begin{aligned} f_1(q) &= \delta q + I_A(z+w-q) + I_S(z), \quad C_1 = \{q : 0 \le q \le z+w\}, \\ f_2(x) &= \delta q + \beta\Big[I_A\big(\frac{z_1-q}{\beta}\big) + I_S\big(\frac{z_2}{\beta}\big)\Big] + \\ &\qquad (1-\beta)\Big[I_A\big(\frac{z+w-z_1}{1-\beta}\big) + I_S\big(\frac{z-z_1}{1-\beta}\big)\Big], \\ C_2 &= \{x : 0 \le q \le z_1 \le z, z_2 \ge z_1, \beta \in [0,1]\}. \end{aligned} \tag{9}$$

We omit a detailed derivation of this result for brevity. The intuition behind it is that the most likely path leading to $\tilde{D}_n(1) = z$ and $\tilde{W}_n(1) = w$ can only be of one of the following two types. In the first case, we have an initial workload nq at time 0, arrivals at constant rate $z + w - q$ and constant service capacity z over the entire period $[0, n]$. The queue never empties on $[0, n]$ and no service capacity is wasted. In the second scenario, the optimal path has two distinct phases. The first phase begins at time 0 with workload nq. The arrival rate is $(z_1 - q)/\beta$ and the service capacity is z_2/β during this phase, which runs until time βn. Moreover, $z_1 \le z_2$, and so the queue is empty at the end of the first phase. During the second phase, which runs over $[\beta n, n]$, the arrival rate is $(z + w - z_1)/(1 - \beta)$, the service rate is $(z - z_1)/(1 - \beta)$ and the queue is never empty. The optimization problem in (8), (9) corresponds to determining the most likely path within these scenarios.

It was shown in [8] that, if assumption 4 is violated, then the rate function governing the sample path LDP for the scaled departure process in equilibrium, $\tilde{D}_n$ (defined analogous to $\tilde{A}_n, \tilde{S}_n$), is not convex; in particular, there is no convex function $I(\cdot)$ such that $\mathcal{I}_D(\phi) = \int I(\dot{\phi})(t)dt$ for $\phi \in \mathcal{AC}(\mathbb{R}_+)$. Assumption 4 guarantees that I_D is convex and that, conditional on $\tilde{D}_n(1) = d$, $P(\sup_{t\in[0,1]} |\tilde{D}_n(t) - dt| > \epsilon) \to 0$ as $n \to \infty$ for every $\epsilon > 0$. This 'linear geodesic property' is still not sufficient to guarantee that the rate function

$\mathcal{I}_D(\phi) = \int I(\dot{\phi})(t)dt$ for all $\phi \in \mathcal{AC}(\mathbb{R}_+)$. The main result of this section is that, given a service process which satisfies assumptions 1 and 2, we can find an arrival process such that assumptions 1-4 are satisfied, and such that the departure process satisfies the sample path LDP with rate function $\mathcal{I}_D = \mathcal{I}_A$. For this arrival process, we also show that a large deviations version of quasi-reversibility holds: the joint rate function for $\tilde{D}_n([0,1]), \tilde{W}_n(1)$ is the sum of the individual rate functions for $\tilde{D}_n([0,1])$ and $\tilde{W}_n(1)$ respectively. We state the result following some definitions.

Let $f : \mathbb{R} \to \mathbb{R} \cup \infty$ be a convex function. The effective domain of f, which we denote by dom f, is the set $\{x \in \mathbb{R} : f(x) < \infty\}$. For $x \in$ dom f, the subdifferential of f at x, denoted subdiff $f(x)$, is the set

$$\{\beta \in \mathbb{R} : \ f(y) \geq f(x) + \beta(y - x) \ \ \forall\, y \in \mathbb{R}\}\,.$$

It is convenient to work in a topology which is finer than the topology of uniform convergence on compacts. Set

$$\mathcal{Y} = \left\{ \phi \in \mathcal{C}(\mathbb{R}_+) : \lim_{t\to\infty} \frac{\phi(t)}{1+t} \text{ exists,} \right\}$$

and equip $\mathcal{Y}$ with the norm

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

Theorem 2 *Suppose the service process $\{S_n, n \in \mathbb{Z}\}$ satisfies assumptions 1 and 2, and assume without loss of generality that the mean service rate $E[S_1] = \Lambda_S'(0) = 1$. Let $\alpha \in (0,1)$ be in the interior of the effective domain of I_S. Define*

$$\lambda = \inf\{\textit{subdiff } I_S(\alpha)\}. \tag{10}$$

If the arrival process $\{A_n, n \in \mathbb{Z}\}$ satisfies assumptions 1 and 2 and

$$I_A(x) = I_S(x) - I_S(\alpha) - \lambda(x - \alpha), \tag{11}$$

then assumptions 3 and 4 hold as well, and the departure process $\tilde{D}_n$ satisfies the LDP in $\mathcal{Y}$ with good convex rate function $\mathcal{I}_D \equiv \mathcal{I}_A$. In addition, for any $t > 0$, $(\tilde{D}_n([0,t]), \tilde{W}_n(t))$ jointly satisfy the LDP in $\mathcal{C}([0,t]) \times \mathbb{R}$ with good convex rate function

$$\mathcal{I}_{D,W}(\phi, w) = \begin{cases} \int_0^t I_A(\dot{\phi}(s))ds \ + \ \delta w, & \text{if } \phi \in \mathcal{AC}([0,t]), \\ \quad +\infty, & \text{otherwise,} \end{cases}$$

Note that λ exists and is finite by the convexity of I_S and the assumption that α is in the interior of dom I_S. Since Λ_S is differentiable at the origin, with $\Lambda_S'(0) = E[S_1] = 1$ by assumption, we have $I_S(1) = 0$ and $I_S(x) > 0$ for all $x \neq 1$ (see [4]). Consequently, by the convexity and non-negativity of I_S, I_S is decreasing on $(-\infty, 1)$ and increasing on $(1, \infty)$. Since $\alpha \in (0,1)$, it follows that $\lambda < 0$. Finally, it is not hard to verify from the definition that the subdifferential is a closed set. So, by(10), $\lambda \in \text{subdiff}\, I_S(\alpha)$.

We now verify that I_A defined by (11) is a rate function and that it is convex. We have, by definition of the subdifferential and the fact that $\lambda \in \text{subdiff}\, I_S(\alpha)$, that

$$I_S(x) \geq I_S(\alpha) + \lambda(x - \alpha) \ \ \forall\, x \in \mathbb{R},$$

Hence, by (11), $I_A(x) \geq 0$ for all $x \in \mathbb{R}$. It is also clear from (11) that I_A inherits lower semicontinuity and convexity from I_S, and that $I_A(\alpha) = 0$. Therefore, I_A is a convex rate function.

A continuous time queue is called quasi-reversible if, in stationarity, the state of the queue at any time t, the departure process before time t and the arrival process after time t are mutually independent (the state is the same as the queue length if service times are exponential but is more complex in general). It then follows that the arrival and departure processes are Poisson. The joint distribution in a network of quasi-reversible queues is product-form, which makes them analytically tractable and has contributed to the popularity of quasi-reversible queueing models in performance analysis. A more detailed discussion of quasi-reversibility can be found in [12, 16, 19].

In Theorem 2, we show that a large deviations analogue of this property, which we shall refer to as LD quasi-reversibility, holds for a general discrete time queue whose input has the invariant rate function given by (11). Specifically, the past of the departure process is independent of the current workload on a large deviations scale, in the sense that the joint rate function for the past departures and the current workload is the sum of their individual rate functions. We have from the definition of $\mathcal{I}_A, \mathcal{I}_S$ that the joint rate function for $(\tilde{A}_n((-\infty,t]), \tilde{S}_n((-\infty,t]), \tilde{A}_n(t,\infty)$ decomposes into a sum of their individual rate functions. Since the workload at t and the departures up to time t depend only on the arrivals and services up to time t, we see that in fact the past departures, the current workload, and the future arrivals are mutually independent on the large deviation scale, in the sense described above.

The proof of Theorem 2 proceeds through a sequence of lemmas.

Lemma 1 *Let $\{A_n\}$, $\{S_n\}$ satisfy the assumptions of Theorem 2, with I_A given by (10) and (11). Then, for δ defined by (5), $\delta = -\lambda$.*

Proof: Since I_A and Λ_A are convex duals, as are I_S and Λ_S, we obtain using (11) that

$$\begin{aligned} \Lambda_A(\theta) &= \sup_{x\in\mathbb{R}} [\theta x - I_A(x)] \ = \ \sup_{x\in\mathbb{R}} [(\theta+\lambda)x - I_S(x) + I_S(\alpha) - \lambda\alpha] \\ &= \Lambda_S(\theta+\lambda) + I_S(\alpha) - \lambda\alpha, \end{aligned} \tag{12}$$

and so

$$\Lambda_A(-\lambda) + \Lambda_S(\lambda) = \Lambda_S(0) + I_S(\alpha) - \lambda\alpha + \Lambda_S(\lambda). \tag{13}$$

We have from (10) and the definition of subdifferentials that $I_S(x) \geq I_S(\alpha) + \lambda(x-\alpha)$ for all $x \in \mathbb{R}$. Hence,

$$\Lambda_S(\lambda) = \sup_{x\in\mathbb{R}} [\lambda x - I_S(x)] = \lambda\alpha - I_S(\alpha). \tag{14}$$

Combining this with the fact that $\Lambda_S(0) = 0$, we get from (13) that $\Lambda_A(-\lambda) + \Lambda_S(\lambda) = 0$, so that, by (5), $\delta \geq -\lambda$.

We have shown that $f(\theta) := \Lambda_A(\theta) + \Lambda_S(-\theta) = 0$ at $\theta = -\lambda > 0$. Now f is convex and $f(0) = 0$ since $\Lambda_A(0) = \Lambda_S(0) = 0$. Moreover, $f'(0) = \Lambda'_A(0) - \Lambda'_S(0) < 0$ by assumption 3, so f isn't identically zero on $[0, -\lambda]$. Hence, 0 and $-\lambda$ are the only zeros of f and $f(\eta) > 0$ for all $\eta > -\lambda$. It follows from (5) that $\delta \leq -\lambda$. Combining this with the reverse inequality obtained earlier completes the proof of the lemma. ■

Lemma 2 *Suppose $\{A_n\}$, $\{S_n\}$ satisfy the assumptions of Theorem 2, with I_A given by (10) and (11). Let $z, w \geq 0$ be given. Then,*

$$f_1(q) \geq I_A(z) + \delta w \quad \textit{and} \quad f_2(x) \geq I_A(z) + \delta w$$

for any $q \in C_1$ and any $x \in C_2$.

Proof: For any $q \in [0, z + w]$, we have by (11) and Lemma 1 that

$$\begin{aligned} f_1(q) &= \delta q + I_A(z + w - q) + I_S(z) \\ &= \delta w + I_S(z + w - q) + I_A(z) \;\geq\; \delta w + I_A(z), \end{aligned} \tag{15}$$

where the inequality is seen to follow from the non-negativity of $I_S(\cdot)$.

Next, let $x = (q, z_1, z_2, \beta)$ achieve the infimum of f_2 over C_2. The infimum is attained at some $x \in C_2$ because f_2 is convex and lower semicontinuous with compact level sets (it inherits these properties from the rate functions I_A, I_S), and C_2 is closed. We shall show that $f_2(x) \geq I(z) + \delta w$.

We see from the definition of C_2 that $z_2 \geq z_1$. If $z_2 = z_1$, we obtain from the definition of f_2 in (9) and the convexity of I_A and I_S that

$$f_2(x) \geq \delta q + I_A(z + w - q) + I_S(z),$$

and so, by (15), $f_2(x) \geq I_A(z) + \delta w$.

On the other hand, if $z_2 > z_1$, then the constraint on z_2 in the definition of C_2 is slack, so f_2 must attain an unconstrained minimum with respect to z_2, i.e., z_2/β is a local minimizer of $I_S(\cdot)$. Since $I_S(x)$ is convex and achieves its minimum value of zero uniquely at $x = 1$, we have $z_2/\beta = 1$. We also note that I_S is non-increasing on $(-\infty, 1]$ and so

$$I_S\big(\frac{z_1 - q}{\beta}\big) \geq I_S\big(\frac{z_1}{\beta}\big),$$

since $z_1 < z_2$ and $q \geq 0$. Hence, by (11) and Lemma 1,

$$I_A\big(\frac{z_1 - q}{\beta}\big) - I_A\big(\frac{z_1}{\beta}\big) = I_S\big(\frac{z_1 - q}{\beta}\big) - I_S\big(\frac{z_1}{\beta}\big) - \delta\frac{q}{\beta} \geq -\delta\frac{q}{\beta}. \tag{16}$$

We obtain from (9,16) and the equality $I_S(z_2/\beta) = I_S(1) = 0$, that

$$f_2(x) \geq \beta I_A\big(\frac{z_1}{\beta}\big) + (1-\beta)\Big[I_A\big(\frac{z + w - z_1}{1-\beta}\big) + I_S\big(\frac{z - z_1}{1-\beta}\big)\Big]. \tag{17}$$

Using (11) and Lemma 1 again, we see that

$$I_A\big(\frac{z + w - z_1}{1-\beta}\big) + I_S\big(\frac{z - z_1}{1-\beta}\big) = I_S\big(\frac{z + w - z_1}{1-\beta}\big) + I_A\big(\frac{z - z_1}{1-\beta}\big) + \delta w.$$

Substituting this in (17) and noting that

$$\beta I_A(z_1/\beta) + (1-\beta) I_A((z - z_1)/(1-\beta)) \geq I_A(z)$$

by the convexity of I_A, we get

$$f_2(x) \geq I_A(z) + \delta w.$$

Since x minimizes f_2 over C_2 by assumption, the above inequality also holds for any $y \in C_2$. This completes the proof of the lemma. ■

Lemma 3 *Let $w, z \geq 0$ be given. If $z + w \geq 1$, then the infimum in (8) is achieved by f_1 at $q^* = z + w - 1$, whereas, if $z + w \leq 1$, then the infimum in (8) is achieved by f_2 at*

$$x^* = (q, z_1, z_2, \beta) = \big(0, \frac{z(1-w-z)}{1-z}, \frac{1-w-z}{1-z}, \frac{1-w-z}{1-z}\big).$$

In either case, the minimum value, $I_{D,W}(z, w)$, is $I_A(z) + \delta w$.

Proof: We have from (15) that

$$f_1(q^*) = \delta w + I_S(z + w - q^*) + I_A(z) = \delta w + I_A(z),$$

since $I_S(z + w - q^*) = I_S(1) = 0$.

Using the definition of f_2 in (9), we obtain after some simplification that

$$f_2(x^*) = \frac{1 - w - z}{1 - z}\left[I_A(z) + I_S(1)\right] + \frac{w}{1 - z}\left[I_A(1) + I_S(z)\right]. \tag{18}$$

Now $I_S(1) = 0$ and we obtain from (11) and Lemma 1 that

$$I_A(1) + I_S(z) = I_A(z) + I_S(1) + \delta(1 - z) = I_A(z) + \delta(1 - z).$$

Thus, we have from (18) that $f_2(x^*) = \delta w + I_A(z)$.

It can readily be verified that $q^* \in C_1$ and $x^* \in C_2$. The optimality of q^* and x^* is now immediate from the lower bounds on f_1 and f_2 obtained in Lemma 2 above. This establishes the claim of the lemma. ■

Lemma 3 establishes an LD quasi-reversibility property: the joint rate function for the mean departure rate on $(0, n)$ and the workload at time n is the sum of the corresponding individual rate functions. In other words, the queue is approximately in equilibrium at time n (the rate function for the workload is the same as the equilibrium rate function) irrespective of the mean rate of departures on $(0, n)$. This property turns out to be crucial to the proof of Lemma 4 below and thereby to the proof of Theorem 2.

Lemma 4 *For any $k \in \mathbb{N}$ and $0 = t_0 < t_1 < \ldots < t_k$, the random vector $(\tilde{D}_n(t_1), \ldots, \tilde{D}_n(t_k), \tilde{W}_n(t_k))$ satisfies the LDP in $\mathbb{R}^{k+1}$ with rate function*

$$I^k_{D,W}(z_1, \ldots, z_k, w) = \sum_{i=1}^{k} (t_i - t_{i-1}) I_A\Big(\frac{z_i - z_{i-1}}{t_i - t_{i-1}}\Big) + \delta w.$$

Proof: The proof is by induction on k. The basis $k = 1$ was established in Lemma 3 for $t_1 = 1$, but can easily be extended to arbitrary $t_1 > 0$ by simply rescaling the most likely path leading to the event $\tilde{D}_n(1) = z$, $\tilde{W}_n(1) = w$, which was identified in Lemma 3.

Assume the claim of the lemma holds for $k-1$. Fix $\epsilon > 0$ and let $E_k(w)$ denote the event

$$E_k(w) = \{|\tilde{D}_n(t_i) - z_i| < \epsilon,\ i = 1, \ldots, k,\ |\tilde{W}_n(t_k) - w| < \epsilon\},$$

where the dependence of $E_k(w)$ on n, ϵ and (t_i, z_i), $i = 1, \ldots, k$ is suppressed in the notation. For notational simplicity, we shall write $a \approx b$ for $|a-b| < \epsilon$. We have

$$\mathbf{P}(E_k(w)) \geq \mathbf{P}(E_{k-1}(q)) \times \mathbf{P}\left(\tilde{D}_n(t_k) \approx z_k, \tilde{W}_n(t_k) \approx w \mid E_{k-1}(q)\right) \quad (19)$$

for all $q \geq 0$. By the induction hypothesis,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}(E_{k-1}(q)) = -\sum_{i=1}^{k-1} (t_i - t_{i-1}) I_A\Big(\frac{z_i - z_{i-1}}{t_i - t_{i-1}}\Big) + \delta q + O(\epsilon). \quad (20)$$

Now, conditional on $E_{k-1}(q)$, $\tilde{D}_n(t_k)$ and $\tilde{W}_n(t_k)$ depend only on the arrival and service processes on $[t_{k-1}, t_k]$ and on q, t_{k-1} and z_{k-1}. Consequently, it is clear from the form of the rate functions $\mathcal{I}_A$ and $\mathcal{I}_D$ in assumption 2 that the joint rate function of $(\tilde{D}_n(t_k), \tilde{W}_n(t_k))$ conditional on $E_{k-1}(q)$ depends on the past up to t_{k-1} only through q, t_{k-1} and z_{k-1}. Therefore, we have from (19) and (20) that

$$\begin{aligned}\liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}(E_k(w)) \quad &\geq \quad -\sum_{i=1}^{k-1} (t_i - t_{i-1}) I_A\Big(\frac{z_i - z_{i-1}}{t_i - t_{i-1}}\Big) + O(\epsilon) - \\ &\qquad \inf_{q\geq 0}\Big[\delta q - \liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}(F_k | \tilde{W}_n(t_{k-1}) = q)\Big],\end{aligned}$$

where F_k denotes the event $\tilde{D}_n(t_k) - \tilde{D}_n(t_{k-1}) \approx z_k - z_{k-1}, \tilde{W}_n(t_k) \approx w$. We recognize the infimum over $q \geq 0$ above as the limit of the scaled logarithm of the probability that $\tilde{D}_n(t_k) - \tilde{D}_n(t_{k-1}) \approx z_k - z_{k-1}$ and that $\tilde{W}_n(t_k) \approx w$ given that the queue is in equilibrium at time t_{k-1}. Thus, by the induction hypothesis, the infimum is simply $(t_k - t_{k-1}) I_{D,W}((z_k - z_{k-1})/(t_k - t_{k-1}), w)$, for $I_{D,W}$ given by (8), and we obtain using Lemma 3 that

$$\begin{aligned} \lim_{\epsilon \to 0} \liminf_{n \to \infty} \frac{1}{n} \log \mathbf{P}(E_k(w)) &\geq -\sum_{i=1}^{k} (t_i - t_{i-1}) I_A\big(\frac{z_i - z_{i-1}}{t_i - t_{i-1}}\big) - \delta w \\ &= -I^k_{D,W}(z_1, \ldots, z_k, w). \end{aligned} \tag{21}$$

The corresponding upper bound can be obtained using the principle of the largest term. We note that $\mathbf{P}(E_k(w))$ is bounded above by

$$\sum_{i=1}^{n} \mathbf{P}(E_{k-1}(i\epsilon)) \mathbf{P}(\tilde{D}_n(t_k) \approx z_k, \tilde{W}_n(t_k) \approx w | E_{k-1}(i\epsilon)) + \mathbf{P}(\tilde{W}_n(t_{k-1}) \geq n\epsilon).$$

Now $\mathbf{P}(\tilde{W}_n(t_{k-1}) \geq n\epsilon) = \mathbf{P}(W(\lfloor nt_{k-1} \rfloor) \geq n^2\epsilon) \leq \exp(-\delta n^2 \epsilon / 2t_{k-1})$ for large enough n. Hence, $\mathbf{P}(E_k(w))$ is bounded above by

$$n\epsilon \sup_{q \geq 0} \mathbf{P}(E_{k-1}(q)) \mathbf{P}(\tilde{D}_n(t_k) \approx z_k, \tilde{W}_n(t_k) \approx w | E_{k-1}(q)) + \exp\big(-\frac{\delta n^2 \epsilon}{2t_{k-1}}\big).$$

The second term is negligible in comparison to the first for large n. The first term is simply $n\epsilon$ times the supremum over q of the right hand side of (19), which was used to obtain the lower bound in (21). Thus, we get

$$\lim_{\epsilon \to 0} \liminf_{n \to \infty} \frac{1}{n} \log \mathbf{P}(E_k(w)) \leq -I^k_{D,W}(z_1, \ldots, z_k, w).$$

We have thus established the large deviation upper and lower bounds for a base of the topology on $\mathbb{R}^{k+1}$. Together with the exponential tightness of $(\tilde{D}_n(t_1), \ldots, \tilde{D}_n(t_k), \tilde{W}_n(t_k))$, this implies the full LDP on $\mathbb{R}^{k+1}$ with rate function $I^k_{D,W}$ (see [4, Theorem 4.1.11, Lemma 1.2.18]). ■

Proof of Theorem 2: For each $t > 0$, Lemma 4 establishes the LDP for every finite-dimensional distribution $(\tilde{D}_n(t_1), \ldots, \tilde{D}_n(t_k), \tilde{W}_n(t_k))$, where $0 < t_1 < \ldots < t_k = t$. These can be extended to an LDP for $(\tilde{D}_n([0,t]), \tilde{W}_n(t))$ on $\mathcal{C}([0,t]) \times \mathbb{R}$ by the method of projective limits. The argument is identical to the proof of Mogulskii's theorem in [4, Theorem 5.1.2] and is omitted. It is not hard to see that the rate function for this LDP is indeed $\mathcal{I}_{D,W}$. By contraction, we also obtain the LDP for $\tilde{D}_n([0,t])$ in $\mathcal{C}([0,t])$, for each $t \geq 0$. By taking projective limits, these imply the LDP for $\tilde{D}_n([0,\infty))$ in $\mathcal{C}(\mathbb{R}_+)$ equipped with the topology of uniform convergence on compacts, which is the projective limit topology. We can strengthen this result to an LDP in $\mathcal{Y}$ by showing that $\tilde{D}_n([0,\infty))$ is an exponentially tight sequence in $\mathcal{Y}$. The argument is the same as in the proof of [9, Theorem 1] and is omitted. ■

3 Existence of fixed points

In this section we present some results on the existence of fixed points in a discrete-time setting, mostly using arguments analogous to those presented in [14] for the continuous time setting.

Consider the space $\mathbb{R}^{\mathbb{Z}}$ equipped with the topology of coordinatewise convergence, which is metrizable using the metric

$$d(\mathbf{x},\mathbf{y}) = \sum_{i\in\mathbb{Z}} \frac{1}{2^{|i|}} \frac{|x_i - y_i|}{1 + |x_i - y_i|}.$$

We let M be the space of stationary probability measures on $\mathbb{R}^{\mathbb{Z}}$ which are stochastically dominated by the service process and equip it with the weak topology generated by the metric $d(\cdot,\cdot)$. More precisely, let ν_n denote the distribution of $S_1 + \ldots + S_n$, where $(S_n, n \in \mathbb{Z})$ is a realization of the service process, and define $f_n : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}$ by $f_n(\mathbf{x}) = x_1 + \ldots + x_n$. We say that a stationary probability measure λ on $\mathbb{R}^{\mathbb{Z}}$ is in M if, for each $n \in \mathbb{N}$, $\lambda \circ f_n^{-1}$ is stochastically dominated by ν_n. Weak convergence in M coincides with convergence in distribution of all finite-dimensional marginals and can be metrized using, for instance, the Prohorov metric. Thus, M is a closed subset of a Polish space, it is clearly convex and it can be shown to be compact. We denote by M_e the subset of M consisting of ergodic measures and by M^{α} (resp. M_e^{α}) the subset consisting of measures (resp. ergodic measures) whose one-dimensional marginals have mean $\alpha \in \mathbb{R}$.

Consider an infinite queueing tandem. Let A_n denote the amount of work entering the first queue of the tandem in time slot n and let S_n^k denote the amount of work that can be served by queue k in time slot n, $k \in \mathbb{N}$, $n \in \mathbb{Z}$. Let W_n^k denote the workload in queue k at the beginning of time slot n and D_n^k the amount of work departing queue k and entering queue $(k+1)$ during time slot n. We assume the following in the remainder of this section.

Assumptions S_n^k is an iid sequence for each fixed k and identically distributed for all k,

$$ES_1^1 = 1, \quad \Lambda_S(\theta) := \log\ E \exp \theta S_1^1 < \infty \quad \text{for all } \theta \text{ in a neighbourhood of } 0.$$

The service distribution is non-degenerate, i.e., $P(S_1^1 \neq 1) > 0$. The arrival process A_n and the service processes S_n^k at the different queues are mutually independent, A_n is stationary and ergodic with rate $\alpha < 1$, i.e.,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} A_i = E[A_1] = \alpha \quad a.s.,$$

and A_n is stochastically dominated by the service process (at any queue). In addition,

$$\Lambda_A(\theta) := \lim_{n\to\infty} \frac{1}{n} \log E \exp\ \theta(A_1 + \ldots + A_n)$$

exists as an extended real number for all $\theta \in \mathbb{R}$, Λ_A is differentiable in the interior of its domain (the set where Λ_A is finite) and steep, i.e., $|\Lambda'_A(\theta)| \to \infty$ as θ approaches the boundary of its domain.

It follows from the above assumptions that the departure process $(D_n^k, n \in \mathbb{Z})$ is stationary and ergodic with rate α, for each $k \in \mathbb{N}$. Recall that Λ_A and Λ_S are convex functions and that Λ_S has infinitely many derivatives in the interior of its domain (see [4], e.g., for proofs). Since the arrival process was assumed to be stochastically dominated by the service process, it follows that $\Lambda_A(\theta) \leq \Lambda_S(\theta)$ for $\theta > 0$ and so Λ_A is finite in some neighbourhood of zero (finiteness for $\theta < 0$ is not an issue since the A_n are non-negative). We have,

$$\Lambda_A(0) = \Lambda_S(0) = 0, \quad \Lambda'_A(0) = E[A_1] = \alpha < 1 = E[S_1^1] = \Lambda'_S(0),$$

and so

$$\exists\,\theta_0 > 0: \quad \Lambda_A(\theta) + \Lambda_S(-\theta) < 0 \quad \forall\,\theta \in (0,\theta_0). \tag{22}$$

Let $M_0 \subset M$ be the set of stationary probability measures on $\mathbb{R}^{\mathbb{Z}}$ whose ergodic decompositions do not contain an atom at the service distribution. We can define the queueing operator $\mathcal{Q}$ on M_0 by setting $\mathcal{Q}(\nu)$ to be the law of the departure process corresponding to an arrival process which is independent of the service process and has law $\nu \in M_0$. It follows from Loynes' construction [13] that $\mathcal{Q}$ is well-defined and maps M_0 into itself, that it preserves ergodicity, i.e., $\mathcal{Q}(M_0 \cap M_e) \subseteq M_0 \cap M_e$, and that it is mean-preserving in the sense that $\mathcal{Q}(M_e^\alpha) \subseteq (M_e^\alpha)$ for all $\alpha < 1$. Moreover, $\mathcal{Q}$ is linear, i.e.,

$$\mathcal{Q}(\beta\nu_1 + (1-\beta)\nu_2) = \beta\mathcal{Q}(\nu_1) + (1-\beta)\mathcal{Q}(\nu_2),$$

for all $\nu_1, \nu_2 \in M_0$ and $\beta \in [0,1]$. Finally, $\mathcal{Q}$ is continuous in the weak topology restricted to M_0. The proof of the last statement is virtually identical to that of [14, Theorem 4.3] and is omitted.

Let μ_0 denote the law of $(A_n, n \in \mathbb{Z})$, μ_k the law of $(D_n^k, n \in \mathbb{Z})$ and μ^S the law of the service process, $(S_n, n \in \mathbb{Z})$. We have assumed that $\mu_0 \in M_e^\alpha$ for some $\alpha < 1$, whereas $\mu^S \in M_e^1$, so μ_0 is not the service distribution. Since μ_0 consists of a single ergodic component, it follows that $\mu_0 \in M_0$. Hence, so is $\mu_k = \mathcal{Q}^k(\mu_0)$ for any $k \in \mathbb{N}$, where $\mathcal{Q}^k$ denotes the k^{th} iterate of $\mathcal{Q}$. Since M_0 is clearly convex,

$$\lambda_k := \frac{1}{k}\sum_{i=0}^{k-1} \mu_k \in M_0 \quad \text{for all } k \in \mathbb{N}. \tag{23}$$

Since M is compact, there is a subsequence $k(j)$ of $\mathbb{N}$ such that $\lambda_{k(j)} \to \lambda$ for some $\lambda \in M$. We shall show that λ is a fixed point of the queueing operator.

Theorem 3 *Let $\lambda \in M$ be defined as above as a subsequential limit of the λ_k's, where λ_k is the Cesaro average of the distributions of the departures* *from the first k queues in the tandem. Then $\lambda \in M_0$ and $\mathcal{Q}(\lambda) = \lambda$, i.e., λ is a fixed point of the queueing operator.*

Proof: Since $\lambda_k \in M_0$ and $\mathcal{Q} : M_0 \to M_0$, we have $\mathcal{Q}(\lambda_k) \in M_0$ for all k. But,

$$\mathcal{Q}(\lambda_k) = \mathcal{Q}\left(\frac{1}{k}\sum_{i=0}^{k-1}\mu_i\right) = \frac{1}{k}\sum_{i=1}^{k}\mu_i = \lambda_k + \frac{1}{k}(\mu_k - \mu_0). \tag{24}$$

To obtain the second equality, we have used the fact that $\mathcal{Q}$ is linear and that $\mathcal{Q}(\mu_i) = \mu_{i+1}$ by definition of the μ_i. It is clear from (24) that,

$$\lim_{j\to\infty} \mathcal{Q}(\lambda_{k(j)}) = \lim_{j\to\infty} \lambda_{k(j)} = \lambda. \tag{25}$$

We show in Lemma 7 below that $\lambda \in M_0$. Since $\mathcal{Q} : M_0 \to M_0$ is continuous in the weak topology and $\lambda_{k(j)} \to \lambda$ in this topology, it follows that

$$\lim_{j\to\infty} \mathcal{Q}(\lambda_{k(j)}) = \mathcal{Q}(\lambda). \tag{26}$$

By (25) and (26), $\mathcal{Q}(\lambda) = \lambda$. ■

Lemma 5 *Consider a sequence of stationary arrival distributions $\nu_k \in M$, converging weakly to a stationary arrival distribution $\nu \in M$. Let $W_0(k)$ (resp. W_0) denote a random variable with the distribution of the workload at the beginning of time slot zero, when the arrival process has distribution ν_k (resp. ν) and is independent of the service process. Then, we have*

$$\liminf_{k\to\infty} E[W_0(k)] \geq E[W_0].$$

The result holds even if $E[W_0] = +\infty$.

The proof proceeds along the lines of the proof of [14, Lemma 4.4] and is omitted.

Lemma 6 *Let $W_0(k)$ denote a random variable with the distribution of the workload at the beginning of time slot zero, when the arrival process has distribution λ_k and the service process has distribution μ^S. Then, we have*

$$\limsup_{k\to\infty} E[W_0(k)] < +\infty.$$

Proof: Recall that W_0^k is the waiting time at queue k at the beginning of time slot zero, when the arrival process into this queue has distribution μ_k and is independent of the service process at this queue, which has distribution μ^S. It is now immediate from the definition of λ_k that

$$W_0(k) \overset{d}{=} \frac{1}{k}\sum_{i=1}^{k} W_0^i \quad \text{and so} \quad E[W_0(k)] = \frac{1}{k}\sum_{i=1}^{k} E[W_0^i], \tag{27}$$

where $\overset{d}{=}$ denotes equality in distribution. But, by Loynes' construction,

$$W_0^1 = \sup_{n\geq 0} \sum_{i=-n}^{-1} A_i - S_i^1, \tag{28}$$

where, as usual, we take the empty sum to be zero. We also have that

$$D_n^k = D_n^{k-1} + W_{n-1}^k - W_n^k, \quad n \in \mathbb{Z}, k = 1, 2, 3, \ldots \tag{29}$$

where D_n^0 is identified with A_n. Using (28) and (29), it can be shown inductively (see, for example, [7] or [1, Proposition 5.4]) that

$$\sum_{i=1}^{k} W_0^i = \sup_{n_k\geq\ldots\geq n_1\geq 0} \sum_{i=-n_k}^{-1} A_i - \sum_{j=1}^{k} \sum_{i=-n_j}^{-n_{j-1}-1} S_i^j, \tag{30}$$

where n_0 is defined to be zero. Hence, by the mutual independence of the arrival process and the service processes at the different queues, we have for all x and any $\theta > 0$, that

$$\begin{aligned}
&\mathbf{P}\left(\sum_{i=1}^{k} W_0^i > kx\right) \\
\leq\; & e^{-\theta kx} E\left[\sup_{n_k \geq \ldots \geq n_1 \geq 0} \exp\ \theta(\sum_{i=-n_k}^{-1} A_i - \sum_{j=1}^{k} \sum_{i=-n_j}^{-n_{j-1}-1} S_i^j)\right] \\
\leq\; & e^{-\theta kx} \sum_{n_k=0}^{\infty} E \exp(\theta \sum_{i=-n_k}^{-1} A_i)\ E\left[\sup_{n_k \geq \ldots \geq n_1 \geq 0} \exp(-\theta \sum_{j=1}^{k} \sum_{i=-n_j}^{-n_{j-1}-1} S_i^j)\right]
\end{aligned}$$

To obtain the last equality above, we have used the fact that the expectation of the supremum of a collection of non-negative random variables is no more than the sum of their expectations. Now, the number of terms over which the supremum in the last line above is taken is the number of ways of partitioning n_k into k non-negative integers, which is $\binom{n+k}{k}$. Moreover, since the S_i^j for different i, j are iid, the random variables over which the supremum is taken are identically distributed, with the distribution of $\exp -\theta \sum_{i=-n_k}^{-1} S_i^1$. Thus, we obtain that

$$\mathbf{P}\left(\sum_{i=1}^{k} W_0^i > kx\right) \leq e^{-\theta kx} \sum_{n=0}^{\infty} \binom{n+k}{k} E \exp\left[\theta \sum_{i=-n}^{-1} (A_i - S_i^1)\right]. \qquad (31)$$

Since Λ_A is convex, it is continuous on the interior of its domain, and on this set it is the pointwise limit of continuous functions,

$$\Lambda_n(\theta) := \frac{1}{n} \log E \exp\ \theta(A_1 + \ldots + A_n).$$

Hence Λ_A is uniformly continuous on compact subsets of its domain and the convergence of Λ_n to Λ_A is uniform on these subsets. Let $\theta_0 > 0$ be in the interior of the domain of Λ. Then, for any $\epsilon > 0$, there is an $N < \infty$ such that

$$|\Lambda_n(\theta) - \Lambda_A(\theta)| < \epsilon \quad \forall\ n \geq N,\ \theta \in [0, \theta_0]. \qquad (32)$$

Recall that $\Lambda_n(\theta) \leq \Lambda_S(\theta)$ for all $\theta > 0$ and $n \in \mathbb{N}$ since the service process was assumed to stochastically dominate the arrival process. Hence, we have from (31) and (32) that, for all $\theta \in (0, \theta_0)$,

$$\begin{aligned} \mathbf{P}\left(\sum_{i=1}^{k} W_0^i > kx \right) &\leq e^{-\theta kx} \left[\sum_{n=0}^{N-1} \binom{n+k}{k} e^{n(\Lambda_S(\theta)+\Lambda_S(-\theta))} \right. \\ &\qquad \left. + \sum_{n=N}^{\infty} \binom{n+k}{k} e^{n(\Lambda_A(\theta)+\epsilon+\Lambda_S(-\theta))} \right]. \end{aligned}$$

Observe from (22) that we can find $\theta \in (0, \theta_0)$ and $\epsilon > 0$ sufficiently small that $\Lambda_A(\theta) + \epsilon + \Lambda_S(-\theta) < -\epsilon$. For such a θ and ϵ, we get

$$\begin{aligned} &\mathbf{P}\left(\frac{1}{k} \sum_{i=1}^{k} W_0^i > x \right) \\ \leq\ & e^{-\theta kx} \left[\sum_{n=0}^{N-1} \binom{n+k}{k} e^{n(\Lambda_S(\theta)+\Lambda_S(-\theta))} + \sum_{n=N}^{\infty} \binom{n+k}{k} e^{-n\epsilon} \right] \\ \leq\ & cN^k e^{-\theta kx} \ = \ ce^{-k(\theta x - \ln N)}, \end{aligned}$$

where c is a constant that may depend on θ, ϵ and N but does not depend on k. Thus, we obtain using (27) that

$$\begin{aligned} E[W_0(k)] &= \int_0^\infty \mathbf{P}(W_0(k) \geq x) dx \\ &\leq \int_0^{2\ln N/\theta} dx + \int_{2\ln N/\theta}^\infty ce^{-k\theta x/2} dx \ \leq \ \frac{2\ln N}{\theta} + \frac{2c}{k\theta}. \end{aligned}$$

The above quantity is bounded uniformly in k, which establishes the claim of the lemma. ■

Lemma 7 *The distribution λ, which was defined in the statement of the theorem as a subsequential limit of the λ_k's (mixtures of departure distributions from successive queues in the tandem), does not contain an atom at the service distribution. In other words, $\lambda \in M_0$.*

Proof: Since the service process was assumed to be non-deterministic, it follows from Loynes' construction that if the arrival process is independent of the service process but has the same distribution, then the expected workload at time zero is infinite. By the linearity of the queueing operator, the same is true if the ergodic decomposition of the arrival distribution contains an atom at the service distribution. In other words, if W_0 denotes the workload at time zero when the arrival process has distribution λ and is independent of the service process, then

$$\lambda \in M \setminus M_0 \quad \Rightarrow \quad E[W_0] = +\infty.$$

Now $\lambda_{k(j)} \to \lambda \in M$ by definition, so it follows from Lemmas 5 and 6 that $E[W_0] < +\infty$. Hence $\lambda \in M_0$. ■

Now λ is a stationary process belonging to M_0 and hence could consist of stationary components at different rates. Define M_{sp}^{ζ}, the set of stationary measures of "pathwise rate ζ", as those measures in M^{ζ} whose ergodic components belong *only to* M_e^{ζ}. Thus if a process $X = \{X_n, n \in \mathbb{Z}\}$ is distributed according to some $\nu \in M_{sp}^{\zeta}$, then a.s.

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X(i) = \zeta.$$

The fixed point λ obtained above can be decomposed into its components in $\cup_{\zeta\in[0,1)} M_{sp}^{\zeta}$ as

$$\lambda = \int_0^1 \lambda_\zeta \Phi(d\zeta),$$

where Φ is some measure on [0,1). By linearity of $\mathcal{Q}$, $\mathcal{Q}(\lambda) = \int_0^1 \mathcal{Q}(\lambda_\zeta)\Phi(d\zeta)$. But the queueing operator also preserves rates: λ_ζ and $\mathcal{Q}(\lambda_\zeta)$ must have the same rate for all ζ in the support of Φ. Thus $\mathcal{Q}(\lambda) = \lambda$ implies $\mathcal{Q}(\lambda_\zeta) = \lambda_\zeta$, Φ a.s. Therefore there exists a fixed point for $\mathcal{Q}$ in M_{sp}^{ζ}, for ζ belonging to the support of Φ.

However, the question remains as to whether $\mathcal{Q}$ has an *ergodic* fixed point of rate ζ. We shall settle this question in Theorem 5 below as a corollary of Theorem 4.

4 Attractiveness of fixed points

In this section we present some results on the attractiveness of fixed points, which are discrete analogues of those obtained by Mountford and Prabhakar [15] in the continuous-time setting.

Consider an infinite tandem of queues indexed by the non-negative integers. Let $S = \{S_n, n \in \mathbb{Z}\}$ be an iid family of non-negative integer valued random variables, where S_n denotes the maximum amount of service effort available at queue 0 in the n^{th} time slot. For $n \in \mathbb{Z}$, $k \geq 1$, let S_n^k be the maximum amount of service effort in the n^{th} time slot at queue number k. The processes $S^k = \{S_n^k, n \in \mathbb{Z}\}$ are iid, independent of S, and $S_n^k \overset{d}{=} S_1$ for all n and k. Consider a stationary and ergodic arrival process $A = \{A_n, n \in \mathbb{Z}\}$, where A_n takes values in the non-negative integers, $E(A_1) = \alpha < 1$. We shall assume that A is independent of the service processes S and S^k, $k \geq 1$.

Suppose that A is input to queue 0 and let $A^k = \{A_n^k, n \in \mathbb{Z}\}$ be the arrival process to queue k. The result of Loynes [13] asserts that each A^k is stationary and ergodic, and $E(A_1^k) = \alpha$. In what is to come, it is convenient to use the notation $A^1 = \mathcal{Q}(A, S)$ to denote that A^1 is the departure process from a queue with arrival process A and service process S. Similarly, write $A^{k+1} = \mathcal{Q}(A^k, S^k)$.

We proceed as follows. First, by assuming the existence of an ergodic fixed point F at mean α, we show that A^k converges to F in the $\bar{\rho}$ metric (defined below).

Definition 1 *The $\bar{\rho}$ (Rho Bar) distance between two stationary and ergodic sequences $X = \{X_n, n \in \mathbb{Z}\}$ and $Y = \{Y_n, n \in \mathbb{Z}\}$ of mean α is given by*

$$\bar{\rho}(X, Y) = \inf_{\gamma} E_{\gamma}|\hat{X}_1 - \hat{Y}_1|,$$

where γ is a distribution on $M_e^\alpha \times M_e^\alpha$ – the space of jointly stationary and ergodic sequences $(\hat{X}, \hat{Y})$, with marginals $\hat{X}_1$ and $\hat{Y}_1$ distributed as X_1 and

Y_1. *(See, e.g. Gray[11] or Chang[3], Definition 2.3, for further details of the $\bar{\rho}$ metric.)*

Theorem 4 *Consider the infinite queueing tandem described above. Suppose queue 0, and hence queue $k, k \geq 1$, admits a mean α stationary and ergodic fixed point. Suppose also that $P(S_n = 0) > 0$. Then $\bar{\rho}(A^k, F) \rightarrow 0$ as k goes to infinity.*

Proof: Our method of proof will closely follow that of Mountford and Prabhakar [15]; we shall merely set up the language and notation needed to import the argument in [15].

We use the coupling in [15]. Let F be distributed as the fixed point, independent of A and of all service variables. The coupling is achieved by allowing the service process S to serve both the processes A and F. Thus $F^1 = \mathcal{Q}(F, S)$ is the arrival process to queue 1, and for each $k \geq 1$ $F^{k+1} = \mathcal{Q}(F^k, S^k)$ is the arrival process to queue $k+1$. Note that the processes F^k are all ergodic, of mean α, and distributed as F. It is helpful to imagine that there are two separate buffers at each queue k, one for the A-customers and one for the F-customers. This makes explicit the notion that customers of one process do not influence the waiting of the customers of the other process. The coupling between the two processes at each queue merely consists of using the same service process for both the A- and the F-customers.

The customers of $A \cup F$ are colored yellow, blue or red according to these rules

- customers in $A \cap F$ are colored yellow
- customers in A but not in F are colored blue
- customers in F but not in A are colored red.

Let Y, B and R be the process of yellow, blue and red customers respectively. For each k, color the points of $A^k \cup F^k$ in a similar fashion and define Y^k, B^k and R^k to be the corresponding processes of yellow, blue and red

customers. As in [15], we adopt the following service policy to ensure that once a customer is yellow, it remains yellow forever. Thus at each queue:
a) Yellow customers observe a "first in, first out" rule.
b) Yellow customers take priority over any blue or red customers.
c) If a blue customer arrives at a queue at which there are red customers, then it immediately "couples" with the red customer who arrived first and has not yet coupled. Both the "coupled" customers will be colored yellow in future queues. A similar rule applies for red customers.

Given the joint ergodicity of the trio (A^k, F^k, S^k), it is not hard to see that the process (Y^k, B^k, R^k) is jointly ergodic. The problem is that a limit of the (Y^k, B^k, R^k) need not be ergodic. However, as a result of the above service policy, the (non-random) density of yellow customers increases with k. Using $\mathcal{D}$ to denote density, we wish to show that $\mathcal{D}(Y^k)$ increases to α.

Following [15] we argue by contradiction and hence suppose that there exist customers in the initial arrival processes A and F that never couple and therefore never become yellow. We call these customers "ever-blues" and "ever-reds" respectively. Given a customer V (in either A or F), write $V(k)$ for their departure time from the k^{th} queue. From the service policy and coloring scheme, we readily obtain

Lemma 8 *Let V and U be two customers (in A or F, not necessarily belonging to the same initial point process) such that $V(k) > U(k)$ for some k. If $U(k+1) > V(k+1)$, then customer V must be coloured yellow after $k+1$ queues.*

The importance of Lemma 8 is that among customers that never become yellow order is preserved: if an ever-blue in A arrives before an ever-red in F, then it will arrive before the ever-red after passing through any number of queues. In a manner entirely analogous to [15], this order preservation property can be used to obtain the following lemma (identical to Lemma 3.1 of [15]).

Lemma 9 *If the density of ever-blues is strictly positive, then there exists an ϵ, not depending on k, such that the (non-random) density in F^k of red customers C satisfying "there exist blue customers of A^k in $(C(k),\ C(k) + 2\ /\epsilon]$" must be at least $\epsilon/2$.*

Now by the stability of queue 0 under input F and the joint ergodicity of (F, S), the conditional probability, p, that an arrival of F sees an empty queue given past arrivals is a nonzero random variable. Because F is a fixed point, the pairs (F^k, A^k) are distributed as (F, A^k) and p is also the conditional probability that an arrival of any F^k sees an empty queue. Take $\delta > 0$ to be such that the density of customers in F^k for whom $p < \delta$ is less than $\epsilon/4$.

Given this and the conclusion of Lemma 9, we obtain the next lemma (similar to Lemma 3.2 of [15]).

Lemma 10 *Under the assumptions of Lemma 9, there exist strictly positive ϵ and δ such that for every k, red customers C in F^k with the properties*
(a) there exists a blue customer of A^k in $(C(k),\ C(k) + 2\ /\epsilon]$
(b) $P(C$ arrives at an empty queue $\mid F^k)) > \delta$
have density at least $\epsilon/4$.

Consider a red customer R that satisfies properties (a) and (b) of Lemma 10. Because of property (b) the chance that R finds queue k empty upon arrival is at least δ. Since the process S^k is iid, independent of F^k and $s = P(S_1 = 0) > 0$, the chance that R waits at least $2/\epsilon$ units of time at queue k before departing is at least $s^{\lceil 2/\epsilon \rceil - 1}$. Property (a) guarantees that a blue customer will arrive at queue k while R is waiting. This implies that R will be yellow in F^{k+1}. Therefore, under the assumptions of Lemma 9, $\mathcal{D}(Y^{k+1}) - \mathcal{D}(Y^k) \geq \delta s^{\lceil 2/\epsilon \rceil - 1} \epsilon/4$ for all k. This contradiction establishes that $\mathcal{D}(Y^k)$ increases to α.

Let ν and ν^k be the joint distributions of the processes (A, F) and (A^k, F^k), respectively. Since A and F are independent, ν equals the product measure $\mathcal{L}(A) \times \mathcal{L}(F)$ – clearly a member of $M_e^\alpha \times M_e^\alpha$. The translation invariant nature of the queueing operation preserves joint ergodicity. Therefore each ν^k is also a member of $M_e^\alpha \times M_e^\alpha$.

Now $\mathcal{D}(Y^k) = E_{\nu^k} \min(A_1^k, F_1^k)$. Therefore

$$\begin{aligned}
\bar{\rho}(A^k, F^k) &= \inf_\gamma E_\gamma |\hat{A}_1^k - \hat{F}_1^k| \\
&\leq E_{\nu^k} |A_1^k - F_1^k| \\
&= E_{\nu^k} \left(A_1^k + F_1^k - 2\min(A_1^k, F_1^k) \right) \\
&= 2\left(\alpha - \mathcal{D}(Y^k)\right) \\
&\overset{k\to\infty}{\longrightarrow} 0.
\end{aligned}$$

This concludes the proof of Theorem 4. ■

Theorem 5 *If $\lambda \in \mathcal{M}_{sp}^\alpha$ is a fixed point for the queue, then it is necessarily ergodic. That is, $\lambda \in M_e^\alpha$.*

Proof: Given Theorem 4, the proof is identical to the proof of Theorem 5.2 in [14] and is omitted. ■

References

[1] F. Bacelli, A. Borovkov and J. Mairesse. Asymptotic results on infinite tandem queueing networks. *Preprint*, 1999.

[2] C.S. Chang. Stability, queue length and delay of deterministic and stochastic queueing networks. *IEEE Trans. Autom. Control* 39 (1994) 913-931.

[3] C.S. Chang. On the input-output map of a G/G/1 queue. *Journal of Applied Probability,* **31**, 4, pp 1128-1133, 1994.

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

[5] G. de Veciana and J. Walrand. Effective bandwidths: call admission, traffic policing and filtering for ATM networks. *Queueing Systems*, 20 (1995) 37-59.

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

[7] A. J. Ganesh. Large deviations of the sojourn time for queues in series. *Ann. Oper. Res.* 79:3-26, 1998.

[8] A. J. Ganesh and Neil O'Connell. The linear geodesic property is not generally preserved by a FIFO queue. *Ann. Appl. Probab.*, 8 (1998) 98-111.

[9] A. J. Ganesh and Neil O'Connell. A large deviation principle with queueing applications. *Stochastics and Stochastic Reports*, to appear.

[10] Peter W. Glynn and Ward Whitt. Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *J. Appl. Prob.*, 31A (1994) 131-156.

[11] R.M. Gray, *Probability, Random Processes and Ergodic Properties,* Springer-Verlag, New York, 1988.

[12] F.P. Kelly, *Reversibility and Stochastic Networks*, John Wiley & Sons, New York, 1979.

[13] R. Loynes. The stability of a queue with non-independent interarrival and service times. *proc. Camb. Phil. Soc.* 58:497-520, 1962.

[14] Jean Mairesse and Balaji Prabhakar. On the existence of fixed points for the $\cdot/GI/1/\infty$ queue. Preprint, 1999.

[15] Tom Mountford and Balaji Prabhakar. On the weak convergence of departures from an infinite sequence of $\cdot$/M/1 queues. *Annals of Applied Probability*, **5**, 1, pp 121-127, 1995.

[16] R.R. Muntz. Poisson departure processes and queueing networks. *IBM Research Report RC 4145*, 1972.

[17] Neil O'Connell. Large deviations for departures from a shared buffer. *J. Appl. Prob.*, 34(3):753-766, 1997.

[18] R. T. Rockafellar, *Conjugate Duality and Optimization*, Society of Industrial and Applied Mathematics, 1974.

[19] J. Walrand, *An Introduction to Queueing Networks*, Prentice-Hall, New Jersey, 1988.