The Annals of Applied Probability
2003, Vol. 13, No. 2, 446–474

INVARIANT RATE FUNCTIONS FOR DISCRETE-TIME QUEUES

BY AYALVADI GANESH, NEIL O'CONNELL AND BALAJI PRABHAKAR[1]

Microsoft Research, BRIMS and Stanford University

We consider a discrete-time queue with general service distribution and characterize a class of arrival processes that possess a large deviation rate function that 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 quasireversibility for this class of arrival processes. Finally, we prove the existence of stationary point processes that have a probability law that is preserved by the queueing operator and conjecture that they have large deviation rate functions which belong to the class of invariant rate functions described above.

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 been shown [16] that a similar result holds for single-server queues with a general service time distribution: nontrivial 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 that have sample path large deviations behavior that is preserved by the queue and we establish a large deviations analogue of quasireversibility for this class of arrival processes. We 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? We also show that the invariant rate function corresponding to each mean arrival rate is unique in the class of rate functions satisfying the large deviations analogue of quasireversibility.

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

Received May 2001; revised April 2002.
[1]Supported in part by a Terman Fellowship and an Alfred P. Sloan Fellowship.
AMS 2000 subject classifications. 60K25, 60F10.
Key words and phrases. Large deviations, queueing theory.

results of Mairesse and Prabhakar [16] and Mountford and Prabhakar [17] 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 nth 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 nth 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 present the relevant large deviation results from [19] and [10], and identify a class of rate functions that are preserved by the queueing operator. We establish a large deviations (LD) analogue of quasireversibility for arrival proceses having rate functions in this class, and show that any invariant rate function that satisfies LD quasireversibility must belong to this class. 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 nonnegative 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 $t \ge 0$ 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 let $\mathcal{AC}(\mathbb{R}_+)$ denote 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

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

$$\Lambda_S(\theta) = \lim_{n\to\infty} \frac{1}{n} \log E \exp\theta(S_1 + \cdots + S_n),$$

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

$$\mathcal{I}_A(\phi) = \begin{cases} \displaystyle\int_0^\infty I_A\big(\dot{\phi}(t)\big)\, dt, & \text{if } \phi \in \mathcal{AC}(\mathbb{R}_+), \\ +\infty, & \text{otherwise,} \end{cases}$$

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, e.g., [4, 7, 8, 12]) 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 $w > 0$,

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

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 alternative characterization

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

Let $\tilde{D}_n(t)$ denote the scaled departure process, defined analogously 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 ([19], Theorem 3.3). *Under Assumptions* 1–3, *the sample mean* $\tilde{D}_n(1) = (D_1 + \cdots + D_n)/n$ *of the equilibrium departure process satisfies an LDP in* $\mathbb{R}$ *with rate function* I_D *given by*

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

subject to the constraints that

$$\begin{aligned} q, z_1, z_2, \beta_1, \beta_2, c_2, \tau \ge 0, \qquad & \beta_1 + \beta_2 + \tau \le 1, \\ \beta_2 c_2 \ge z_2, \qquad & z - z_1 - z_2 \ge 0. \end{aligned} \tag{7}$$

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 the time 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 a LDP in $\mathbb{R}^2$ with rate function $I_{D,W}$ given by

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

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

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

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) and (9) corresponds to determining the most likely path within these scenarios.

It was shown in [10] 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 analogously to $\tilde{A}_n$ and $\tilde{S}_n$), 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 $\tilde{D}_n(1) = d$, $P(\sup_{t \in [0,1]} |\tilde{D}_n(t) - dt| > \varepsilon) \to 0$ as $n \to \infty$ for every $\varepsilon > 0$. 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 quasireversibility 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

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

In Theorem 2 below, we consider a service process of unit rate; for each $\alpha\in(0,1)$ which is in the interior of the effective domain of the rate function for the service process, we construct an arrival process of rate α whose large deviation rate function is preserved by the queueing operator.

Suppose $\alpha\in(0,1)$ is not in the effective domain of the rate function for the service process. Then

$$\alpha<\alpha^*:=\inf(\text{dom } I_S)$$

and the service capacity in every time slot is bounded below by α^*. In this case, any arrival process of rate α for which the arrivals in a time slot are bounded deterministically by α^* will be a fixed point of the queue; in particular, its large deviation rate function will not change in passing through the queue.

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* $E[S_1]=\Lambda_S'(0)=1$. *Assume that* I_S *is strictly convex and let* $\alpha\in(0,1)$ *be in the interior of the effective domain of* I_S. *Define*

$$\lambda_\alpha=\inf\{\text{subdiff } I_S(\alpha)\}. \tag{10}$$

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

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

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

$$\mathcal{I}_{D,W}(\phi, w) = \begin{cases} \displaystyle\int_0^t I_A(\dot{\phi}(s))\, ds \;+\; \delta w, & \text{if } \phi \in \mathcal{AC}([0,t]), \\ +\infty, & \text{otherwise.} \end{cases} \tag{12}$$

Note that λ_α exists and is finite by the convexity of I_S and the assumption that α is in the interior of $\operatorname{dom} I_S$. 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 [6]). Consequently, by the convexity and nonnegativity of I_S, I_S is decreasing on $(-\infty, 1)$ and increasing on $(1, \infty)$. Since $\alpha \in (0,1)$, it follows that $\lambda_\alpha < 0$. Finally, it is not hard to verify from the definition that the subdifferential is a closed set. So, by (10), $\lambda_\alpha \in \operatorname{subdiff} I_S(\alpha)$.

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 $\lambda_\alpha \in \operatorname{subdiff} I_S(\alpha)$, that

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

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 strict convexity from I_S, and that $I_A(\alpha) = 0$. Therefore, I_A is a strictly convex rate function.

Next, we verify that Assumptions 1 and 2 imply Assumptions 3 and 4 for I_A defined by (11). Since $I_A(\alpha) = 0$ and I_A is strictly convex and nonnegative, I_A has a unique zero at α. By the strong law of large numbers, $\tilde{A}_n(t) \to tE[A_1]$ as $n \to \infty$, and it follows from Assumption 2 that $E[A_1] = \alpha$. Since $E[S_1] = 1$ and $\alpha < 1$ by assumption, the stability condition in Assumption 3 holds. Recalling that $\lambda_\alpha < 0$, we have $\lambda_\alpha(x - \alpha) > 0$ for all $x < \alpha = E[A_1] = \Lambda_A'(0)$. Since $I_S(\alpha) \geq 0$, it follows from (11) that $I_A(x) < I_S(x)$ for all $x < \Lambda_A'(0)$, that is, Assumption 4 is satisfied.

Suppose S_n is an i.i.d. sequence and S_1 has probability law μ. We say that a probability law ν is an exponential tilting of μ if ν is absolutely continuous with respect to μ and there is a $\lambda \in \mathbb{R}$ such that $\Lambda_S(\lambda)$ is finite, and

$$\frac{d\nu}{d\mu}(x) = \exp\bigl(\lambda x - \Lambda_S(\lambda)\bigr)$$

for all $x \in \mathbb{R}$. If X_1 is a random variable with law ν, then its cumulant generating function is given by

$$\begin{aligned} \Lambda_X(\theta) &= \log \int_{\mathbb{R}} e^{\theta x}\, d\nu(x) \\ &= \log \exp\bigl(-\Lambda_S(\lambda)\bigr) \int_{\mathbb{R}} e^{(\theta+\lambda)x}\, d\mu(x) = \Lambda_S(\theta + \lambda) - \Lambda_S(\lambda). \end{aligned}$$

Taking convex duals of the above, we get

$$I_X(x) = \sup_\theta [\theta x - \Lambda_X(\theta)]$$
$$= \sup_\theta [\theta x - \Lambda_S(\theta + \lambda)] + \Lambda_S(\lambda) = I_S(x) - \lambda x + \Lambda_S(\lambda).$$

Now if λ is in the interior of the domain of Λ_S, define β to be $\Lambda_S'(\lambda)$, which exists and is finite. It is easy to verify that $I_S(\beta) = \lambda\beta - \Lambda_S(\lambda)$. Thus, we have from above that

$$I_X(x) = I_S(x) - I_S(\beta) - \lambda(x - \beta).$$

Comparing this with (11), we see that the invariant rate function corresponds to an exponential tilting of the service distribution when the service process is an i.i.d. sequence. Note that even if the service process is i.i.d., the fixed point of the queue, whose existence is established in Section 3, may not be an i.i.d. sequence. Nevertheless, our conjecture is that its large deviation rate function will correspond to an exponential tilting of the service distribution, as above.

A continuous-time queue is called quasireversible 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 quasireversible queues is product-form, which makes them analytically tractable and has contributed to the popularity of quasireversible queueing models in performance analysis. A more detailed discussion of quasireversibility can be found in [14, 18, 23].

In Theorem 2, we show that a large deviations analogue of this property, which we shall refer to as LD quasireversibility, holds for a general discrete time queue, the input of which 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 $\mathcal{I}_A$ and $\mathcal{I}_S$ 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* $\{A_n\}$ *and* $\{S_n\}$ *satisfy the assumptions of Theorem* 2, *with* I_A *given by* (10) *and* (11). *Then, for* δ *defined by* (5), $\delta = -\lambda_\alpha$.

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

$$\begin{aligned}\Lambda_A(\theta) &= \sup_{x\in\mathbb{R}}[\theta x - I_A(x)] = \sup_{x\in\mathbb{R}}[(\theta+\lambda_\alpha)x - I_S(x) + I_S(\alpha) - \lambda_\alpha\alpha] \\ &= \Lambda_S(\theta+\lambda_\alpha) + I_S(\alpha) - \lambda_\alpha\alpha,\end{aligned} \tag{13}$$

and so

$$\Lambda_A(-\lambda_\alpha) + \Lambda_S(\lambda_\alpha) = \Lambda_S(0) + I_S(\alpha) - \lambda_\alpha\alpha + \Lambda_S(\lambda_\alpha). \tag{14}$$

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

$$\Lambda_S(\lambda_\alpha) = \sup_{x\in\mathbb{R}}[\lambda_\alpha x - I_S(x)] = \lambda_\alpha\alpha - I_S(\alpha). \tag{15}$$

Combining this with the fact that $\Lambda_S(0) = 0$, we get from (14) that $\Lambda_A(-\lambda_\alpha) + \Lambda_S(\lambda_\alpha) = 0$, so that, by (5), $\delta \geq -\lambda_\alpha$.

We have shown that $f(\theta) := \Lambda_A(\theta) + \Lambda_S(-\theta) = 0$ at $\theta = -\lambda_\alpha > 0$. 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 is not identically 0 on $[0, -\lambda_\alpha]$. Hence, 0 and $-\lambda_\alpha$ are the only 0's of f and $f(\eta) > 0$ for all $\eta > -\lambda_\alpha$. It follows from (5) that $\delta \leq -\lambda_\alpha$. Combining this with the reverse inequality obtained earlier completes the proof of the lemma. □

LEMMA 2. *Suppose* $\{A_n\}$ *and* $\{S_n\}$ *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

$$\begin{aligned}f_1(q) &= \delta q + I_A(z+w-q) + I_S(z) \\ &= \delta w + I_S(z+w-q) + I_A(z) \;\geq\; \delta w + I_A(z),\end{aligned} \tag{16}$$

where the inequality is seen to follow from the nonnegativity 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 and 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 (16), $f_2(x) \geq I_A(z) + \delta w$.

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, that is, 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 nonincreasing on $(-\infty, 1]$ and so

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

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

$$I_A\left(\frac{z_1 - q}{\beta}\right) - I_A\left(\frac{z_1}{\beta}\right) = I_S\left(\frac{z_1 - q}{\beta}\right) - I_S\left(\frac{z_1}{\beta}\right) - \delta\frac{q}{\beta} \geq -\delta\frac{q}{\beta}. \tag{17}$$

We obtain from (9), (17) and the equality $I_S(z_2/\beta) = I_S(1) = 0$ that

$$f_2(x) \geq \beta I_A\left(\frac{z_1}{\beta}\right) + (1-\beta)\left[I_A\left(\frac{z + w - z_1}{1-\beta}\right) + I_S\left(\frac{z - z_1}{1-\beta}\right)\right]. \tag{18}$$

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

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

Substituting this in (18) and noting that

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

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*

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

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

PROOF. We have from (16) 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

$$f_2(x^*) = \frac{1-w-z}{1-z}[I_A(z) + I_S(1)] + \frac{w}{1-z}[I_A(1) + I_S(z)]. \tag{19}$$

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 (19) 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. This establishes the claim of the lemma. □

Lemma 3 establishes a LD quasireversibility 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 $0 = t_0 < t_1 < \cdots < t_k$, 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*

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

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 $\varepsilon > 0$ and let $E_k(w)$ denote the event

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

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 $|a - b| < \varepsilon$. We have

$$\mathbf{P}\big(E_k(w)\big) \geq \mathbf{P}\big(E_{k-1}(q)\big) \times \mathbf{P}\big(\tilde{D}_n(t_k) \approx z_k, \tilde{W}_n(t_k) \approx w | E_{k-1}(q)\big) \tag{20}$$

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

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}\big(E_{k-1}(q)\big) = -\sum_{i=1}^{k-1} (t_i - t_{i-1}) I_A\left(\frac{z_i - z_{i-1}}{t_i - t_{i-1}}\right) + \delta q + O(\varepsilon). \tag{21}$$

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 (20) and (21) that

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}\big(E_k(w)\big) \geq -\sum_{i=1}^{k-1} (t_i - t_{i-1}) I_A\left(\frac{z_i - z_{i-1}}{t_i - t_{i-1}}\right) + O(\varepsilon)$$
$$- \inf_{q\geq 0}\left[\delta q - \liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}\big(F_k | \tilde{W}_n(t_{k-1}) = q\big)\right],$$

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

$$\lim_{\varepsilon\to 0} \liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}\big(E_k(w)\big) \geq -\sum_{i=1}^{k} (t_i - t_{i-1}) I_A\left(\frac{z_i - z_{i-1}}{t_i - t_{i-1}}\right) - \delta w$$
$$= -I^k_{D,W}(z_1, \ldots, z_k, w). \tag{22}$$

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

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

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

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

The second term is negligible in comparison to the first for large n. The first term is simply $n\varepsilon$ times the supremum over q of the right-hand side of (20), which was used to obtain the lower bound in (22). Thus, we get

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

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 [6], Theorem 4.1.11 and 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 $0 < t_1 < \cdots < t_k = t$. 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 5.1.2 in [6], 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 Theorem 1 in [11] and is omitted. □

Recall that the rate function for the sample path of the departure process is given by the solution of a variational problem, which is not tractable in general. In the discussion above, we exploited the property of LD quasireversibility to compute this rate function explicitly for a special class of arrival processes. We now show, under mild additional assumptions, that these are the only invariant rate functions for which LD quasireversibility holds. In other words, the invariant rate function is unique (at each arrival rate α) within the class of arrival processes satisfying LD quasireversibility.

THEOREM 3. *Suppose the arrival process* $\{A_n\}$ *and the service process* $\{S_n\}$ *satisfy Assumptions* 1–4, *and assume without loss of generality that the mean service rate* $E[S_1] = \Lambda_S'(0) = 1$. *Denote the mean arrival rate* $E[A_1]$ *by* α *and assume that* $\alpha \in (0,1)$ *is in the interior of the effective domain of* I_S. *Assume that* I_S *is strictly convex and differentiable in the interior of its domain. Suppose the sequence* $\{\tilde{D}_n(1), \tilde{W}_n(1)\}$ *satisfies an LDP in* $\mathbb{R}^2$ *with rate function* $I_{D,W}$ *given by*

$$I_{D,W}(z,w) = I_A(z) + \delta w, \tag{23}$$

where $\delta > 0$ *solves* $\Lambda_A(\delta) + \Lambda_S(-\delta) = 0$. *Then* $\delta = -I_S'(\alpha)$ *and* I_A *is given by* (11), *with* $\lambda_\alpha = I_S'(\alpha)$.

A corollary is that if the sequence $\{\tilde{D}_n([0,t]), \tilde{W}_n(t)\}$, satisfies an LDP with rate function of the form (12), then I_A is given by (11). To see this, note that the LDP for $\{\tilde{D}_n(1), \tilde{W}_n(1)\}$ assumed in the theorem follows from that for $\{\tilde{D}_n([0,t]), \tilde{W}_n(t)\}$ by the contraction principle, and the rate function is given by (23). In other words, (11) specifies the unique arrival rate function (for an arrival process of rate α) which is left invariant by the queue and for which the past of the departure process is independent of the current workload in the LD scaling.

The proof of the theorem proceeds through a sequence of lemmas.

LEMMA 5. *Under the assumptions of Theorem* 3, $\delta \in \text{subdiff}\, I_A(1)$ *and* $\Lambda_S(-\delta) = I_A(1) - \delta$.

PROOF. Observe from (8) and (9) that $I_{D,W}(z, w) \le \inf_{q \in C_1} f_1(q)$. Hence, by the assumption of (23), we have

$$I_A(z) + \delta w \le \delta q + I_A(z + w - q) + I_S(z) \qquad \forall\, 0 \le q \le z + w. \tag{24}$$

Taking $z = 1$ and noting that $I_S(1) = 0$, we get $I_A(1 + x) \ge I_A(1) + \delta x$ for all $x \ge -1$. The inequality also holds for $x < -1$ since $I_A(z)$ is infinite for $z < 0$. Hence, $\delta \in \text{subdiff}\ I_A(1)$. Consequently, $\Lambda_A(\delta) = \delta - I_A(1)$, and since $\Lambda_A(\delta) + \Lambda_S(-\delta) = 0$, the lemma is proved. □

LEMMA 6. *Under the assumptions of Theorem* 3, $\delta = -I_S'(\alpha)$ *and* $I_S(\alpha) = (1 - \alpha)\delta - I_A(1)$.

PROOF. Setting $z + w - q = 1$ in (24), we observe that $I_A(z) \le \delta(z - 1) + I_A(1) + I_S(z)$. Taking duals now yields

$$\Lambda_A(\theta) \ge \sup_z [\theta z - \delta(z - 1) - I_A(1) - I_S(z)] = \delta - I_A(1) + \Lambda_S(\theta - \delta).$$

We observe from Lemma 5 that equality holds at $\theta = 0$ since $\Lambda_A(0) = 0$. Thus, the convex function $f(\theta) := \delta - I_A(1) + \Lambda_S(\theta - \delta)$ is dominated by the convex function $\Lambda_A(\theta)$, and they coincide at $\theta = 0$. Consequently, subdiff $f(0) \subseteq$ subdiff $\Lambda_A(0) = \{\alpha\}$, where the latter equality holds because Λ_A is assumed to be differentiable at the origin, and its derivative is $E[A_1] = \alpha$. Now, subdiff $f(0) =$ subdiff $\Lambda_S(-\delta)$ is nonempty because Λ_S is finite at $-\delta$ and does not take the value $-\infty$ anywhere. Hence, subdiff $\Lambda_S(-\delta) = \{\alpha\}$ and, by duality, $-\delta \in$ subdiff $I_S(\alpha)$. Now, by the assumption that I_S is differentiable in the interior of its domain, we get $I_S'(\alpha) = -\delta$.

An immediate consequence is that $I_S(\alpha) = -\alpha\delta - \Lambda_S(-\delta)$. Combining this with Lemma 5 yields $I_S(\alpha) = (1 - \alpha)\delta - I_A(1)$. □

LEMMA 7. *Under the assumptions of Theorem* 3,

$$I_A(z) = I_S(z) - I_S(\alpha) + \delta(z - \alpha) \qquad \textit{for all } z \ge 1.$$

PROOF. We have from (8) and the assumptions of Theorem 3 that

$$I_A(z) + \delta w = \min\left\{ \inf_{q \in C_1} f_1(q),\ \inf_{x \in C_2} f_2(x) \right\},$$

where f_1, f_2, C_1 and C_2 are defined in (9). Now, by the convexity of I_A and I_S, we have

$$f_2(x) \ge \delta q + I_A(z + w - q) + I_S(z - z_1 + z_2).$$

Since I_S is nondecreasing on $[1, \infty)$ and $z_2 \geq z_1$ for all $x \in C_2$, we obtain for all $z \geq 1$ that

$$f_2(x) \geq \delta q + I_A(z + w - q) + I_S(z) = f_1(q).$$

Observe that if $x \in C_2$, then $q \in C_1$. Hence,

$$I_A(z) + \delta w = \inf_{0 \leq q \leq z+w} \delta q + I_A(z + w - q) + I_S(z) \qquad \forall\, z \geq 1.$$

Since $\delta \in \operatorname{subdiff} I_A(1)$ by Lemma 5, the infimum above is achieved at $q = z + w - 1$ and we have

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

Substituting for $I_A(1)$ from Lemma 6 yields the claim of the lemma. □

REMARK. The claim of Lemma 7 holds even without the assumption of LD quasireversibility. In other words, if the arrival and service processes satisfy Assumptions 1–4 and $I_D(z) = I_A(z)$, then, for $z \geq 1 = E[S_1]$, $I_A(z)$ must have the form claimed in the lemma. To see this, we first observe by comparing (6) and (7) with (8) and (9) that $I_D(z) = I_{D,W}(z, 0)$. We then note that the proof of the lemma carries through unchanged if we set $w = 0$.

PROOF OF THEOREM 3. Let $z \in [0, 1)$. Choose $w > 1$ and observe that

$$\begin{aligned} \inf_{q \in C_1} f_1(q) &= \inf_{0 \leq q \leq z+w} \delta q + I_A(z + w - q) + I_S(z) \\ &= \delta(z + w - 1) + I_A(1) + I_S(z). \end{aligned}$$

Substituting for $I_A(1)$ from Lemma 6, we get

$$\inf_{q \in C_1} f_1(q) = \delta w + I_S(z) - I_S(\alpha) + \delta(z - \alpha). \tag{25}$$

Next, we observe that it is not possible to have $\beta = 1$ at the minimizer of $f_2(x)$ over C_2. Indeed, having $\beta = 1$ would require that $z_1 = z + w$, which is impossible since $z_1 \leq z$ for all $x \in C_2$ and we have chosen $w > 1$.

Thus, $(z + w - z_1)/(1 - \beta)$ is finite and greater than 1 for all $x \in C_2$, and it follows from Lemma 7 that

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

Now I_S is nonnegative, $I_S(1) = 0$ and I_S was assumed to be strictly convex. Hence, there is an $\varepsilon > 0$ and an $\eta > 0$ such that

$$I_S\left(\frac{z + w - z_1}{1 - \beta}\right) \geq \varepsilon\left(\frac{z + w - z_1}{1 - \beta} - 1 - \eta\right).$$

Combining the two equations above yields

$$I_A\left(\frac{z+w-z_1}{1-\beta}\right) \geq (\delta+\varepsilon)\left(\frac{z+w-z_1}{1-\beta}\right) - \delta\alpha - \varepsilon(1+\eta) - I_S(\alpha).$$

We now obtain from (9) and the nonnegativity of I_A and I_S that

$$f_2(x) \geq \delta q + (\delta+\varepsilon)(z+w-z_1) - (1-\beta)\big(\delta\alpha + \varepsilon(1+\eta) + I_S(\alpha)\big).$$

Since $z_1 \leq z$, $q \geq 0$ and $\beta \in [0,1]$, for all $x \in C_2$, we have

$$\inf_{x\in C_2} f_2(x) \geq (\delta+\varepsilon)w - \delta\alpha - 1 - \eta - I_S(\alpha). \tag{26}$$

Fixing z and comparing (25) and (26), we see that for large enough w, the infimum of f_1 over C_1 is smaller than the infimum of f_2 over C_2. Thus, it follows from (8), (23) and (25) that

$$I_A(z) + \delta w = \delta w + I_S(z) - I_S(\alpha) + \delta(z-\alpha)$$

for all $z \in [0,1)$ and w sufficiently large. In particular, $I_A(z) = I_S(z) - I_S(\alpha) + \delta(z-\alpha)$ for all $z \in [0,1)$. The same equality was established for $z \geq 1$ in Lemma 7, and is trivial for $z < 0$ since I_A and I_S are both infinite on the negative half-line. This completes the proof of the theorem. □

EXAMPLES. We now describe some examples of arrival and service processes where the law of the departure process is known explicitly and is the same as that of the arrival process.

Suppose the service process $\{S_n, n \in \mathbb{Z}\}$ is an i.i.d. sequence of geometric random variables of unit mean and suppose the arrival process $\{A_n, n \in \mathbb{Z}\}$ is an i.i.d. sequence of geometric random variables of mean $\alpha < 1$. In other words,

$$P(S_1 = k) = \left(\frac{1}{2}\right)^{k+1}, \qquad P(A_1 = k) = \frac{1}{1+\alpha}\left(\frac{\alpha}{1+\alpha}\right)^k, \qquad k \geq 0.$$

It was shown by Bedekar and Azizoglu [2] that, in equilibrium, the departures are i.i.d. geometric with mean α and the queue length is independent of the past departures, that is, the queue is quasireversible.

It is easy to verify for this example that the assumptions of Theorem 2 hold, with

$$I_S(x) = x\log x - (1+x)\log\frac{1+x}{2}.$$

Hence, by Theorem 2, $\delta = -I_S'(\alpha) = -\log(2\alpha/(1+\alpha))$ and the invariant rate function for an arrival process of rate α is given by

$$\begin{aligned} I_A(x) &= I_S(x) - I_S(\alpha) - I_S'(\alpha)(x-\alpha) \\ &= x\log\frac{2x}{1+x} - \log\frac{1+x}{2} - \alpha\log\frac{2\alpha}{1+\alpha} + \log\frac{1+\alpha}{2} - (x-\alpha)\log\frac{2\alpha}{1+\alpha} \\ &= x\log\frac{x}{\alpha} - (1+x)\log\frac{1+x}{1+\alpha}. \end{aligned}$$

The last expression is the rate function for the arrival process $\{A_n\}$ described above.

Recall that the arrival and service processes to the queue are not constrained to be integer-valued. A continuum version of the geometric model described above has i.i.d. exponentially distributed $\{S_n\}$ and $\{A_n\}$ with $E[S_1] = 1$ and $E[A_1] = \alpha < 1$. It was shown by O'Connell [20] that the departure process in this model has the same law as the arrival process and the workload is independent of the past departures. For this model, we have

$$I_S(x) = x - 1 - \log x$$

and the assumptions of Theorem 2 are satisfied. Thus, the invariant rate function corresponding to an arrival rate of α is given by

$$\begin{aligned} I_A(x) &= I_S(x) - I_S(\alpha) - I_S'(\alpha)(x - \alpha) \\ &= x - \alpha - \log \frac{x}{\alpha} - \left(1 - \frac{1}{\alpha}\right)(x - \alpha) \\ &= \frac{x}{\alpha} - 1 - \log \frac{x}{\alpha}. \end{aligned}$$

The last is the rate function for the fixed point described above, namely, an i.i.d. $\exp(1/\alpha)$ sequence of arrivals.

The last example we consider involves i.i.d. Bernoulli arrivals and services. There are either one or no arrivals in each time slot, an arrival occurring with probability p. In each time slot, there is a probability q that one customer can be served and a probability $1 - q$ that no service occurs (the mean service rate is q rather than 1, but Theorem 2 is still applicable). We obtain that

$$I_S(x) = \begin{cases} x \log \dfrac{x}{q} + (1 - x) \log \dfrac{1 - x}{1 - q}, & x \in [0, 1], \\ \infty, & \text{otherwise.} \end{cases}$$

Using Theorem 2, we compute the invariant rate function for an arrival process of rate p. We obtain that $I_A(x) = I_S(x) - I_S(p) - I_S'(p)(x - p)$ is infinite for $x \notin [0, 1]$, whereas for $x \in [0, 1]$,

$$\begin{aligned} I_A(x) &= x \log \frac{x}{q} + (1 - x) \log \frac{1 - x}{1 - q} \\ &\quad - p \log \frac{p}{q} - (1 - p) \log \frac{1 - p}{1 - q} \\ &\quad - (x - p)\left(\log \frac{p}{q} - \log \frac{1 - p}{1 - q}\right) \\ &= x \log \frac{x}{p} + (1 - x) \log \frac{1 - x}{1 - p}. \end{aligned}$$

This coincides with the rate function for the Bernoulli(p) arrival process, which is known to be a fixed point for the Bernoulli(q) service process [2].

The examples described above are, to the best of our knowledge, the only instances in which the law of the fixed point is known explicitly. In all these cases, the large deviation rate function of the fixed point is as specified in Theorem 2.

In the next section, we show that if the service process is an i.i.d. sequence, then there is a fixed point at each rate $\alpha < 1$, that is, an arrival process that has its probability law preserved by the queueing operator. It was shown by Chang [5] that there is at most one such probability law for each $\alpha < 1$. We conjecture that each of these fixed points satisfies a sample path LDP with rate function specified by Theorem 2. If the fixed points satisfy a sample path LDP and LD quasireversibility, then it follows from Theorem 3 that the rate function should be of the form specified by Theorem 2.

We now present some heuristic arguments that motivate the conjecture. Recall that if the arrival process is an i.i.d. sequence, then Λ_A is analytic on the interior of its domain, so that specifying Λ_A on an interval specifies it everywhere on its domain. In fact, this is true even if the arrival process is not i.i.d., provided it does not have long-range interactions (see, e.g., Theorems 5.6.2 and 5.6.5 in Ruelle [22]). Since the service process was assumed to have an exponentially decaying tail and correlations, and stochastically dominates the departure process, it is plausible to expect that the fixed point does not have long-range correlations. In particular, we expect that the fixed point satisfies a sample path LDP with rate function of the form in Assumption 2 and that Λ_A is real analytic on the interior of its domain. Now, by the remark following the proof of Lemma 7, I_A is uniquely specified on $[1, \infty)$, and so Λ_A is uniquely specified on $[I'_A(1), \infty) \cap \mathrm{dom}(\Lambda_A)$. Hence, Λ_A should have a unique analytic extension on its domain. This suggests that even without the assumption of LD quasireversibility, there is a unique invariant rate function and, therefore, that this is the rate function of the fixed point.

In Section 4, we consider an infinite tandem of queues with independent and identically distributed service processes. We show that for an arbitrary arrival process of rate α entering the first queue in the tandem, the departure process from the nth queue in the tandem converges in law to the fixed point at rate α as $n \to \infty$. Consider an arrival process A^1 into the first queue of the tandem and let A^k denote the arrival process into the kth queue. If A^1 satisfies a sample path LDP with rate function specified by Theorem 2, then so does A^k for every k. The fact that the A^k's converge in law to the fixed point provides additional motivation for our conjecture, but we have not been able to show convergence in a strong enough topology to establish the conjecture.

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 [16] 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 $S_1 + \cdots + S_n$, 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 $f_n(\mathbf{x}) = x_1 + \cdots + x_n$. 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 [3]. 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 let D_n^k denote 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 i.i.d. sequence for each fixed k and is identically distributed for all k,

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

The service distribution is nondegenerate, that is, $P(S_1^1 \neq 1) > 0$. 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 $\alpha < 1$, that is,

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

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

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

exists as an extended real number for all $\theta \in \mathbb{R}$, and Λ_A is differentiable in the interior of its domain (the set where Λ_A is finite) and steep, that is, $|\Lambda_A'(\theta)| \to \infty$ 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 [6], e.g., for proofs). Since the arrival process was assumed to be stochastically dominated by the service process, it follows that $\Lambda_A(\theta) \le \Lambda_S(\theta)$ for $\theta > 0$ and so Λ_A is finite in some neighborhood of 0 (finiteness for $\theta < 0$ is not an issue since the A_n are nonnegative). We have

$$\Lambda_A(0) = \Lambda_S(0) = 0, \qquad \Lambda_A'(0) = E[A_1] = \alpha < 1 = E[S_1^1] = \Lambda_S'(0)$$

and so

$$\exists \theta_0 > 0: \qquad \Lambda_A(\theta) + \Lambda_S(-\theta) < 0 \qquad \forall \theta \in (0, \theta_0). \tag{27}$$

Let $M_0 \subset M$ be the set of stationary probability measures on $\mathbb{R}^{\mathbb{Z}}$ that have ergodic decompositions that 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's construction [15] that $\mathcal{Q}$ is well defined and maps M_0 into itself, that it preserves ergodicity, that is, $\mathcal{Q}(M_0 \cap M_e) \subseteq M_0 \cap M_e$, 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, that is,

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

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 Theorem 4.3 in [16] and is omitted.

Let μ_0 denote the law of $(A_n, n \in \mathbb{Z})$, μ_k denote the law of $(D_n^k, n \in \mathbb{Z})$ and μ^S denote 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 kth iterate of $\mathcal{Q}$. Since M_0 is clearly convex,

$$\lambda_k := \frac{1}{k}\sum_{i=0}^{k-1} \mu_k \in M_0 \qquad \text{for all } k \in \mathbb{N}. \tag{28}$$

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 4. *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 $\mathcal{Q}(\lambda) = \lambda$, that is, 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

$$\mathcal{Q}(\lambda_k) = \mathcal{Q}\left(\frac{1}{k}\sum_{i=0}^{k-1}\mu_i\right) = \frac{1}{k}\sum_{i=1}^{k}\mu_i = \lambda_k + \frac{1}{k}(\mu_k - \mu_0). \tag{29}$$

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 (29) that

$$\lim_{j\to\infty} \mathcal{Q}(\lambda_{k(j)}) = \lim_{j\to\infty} \lambda_{k(j)} = \lambda. \tag{30}$$

We show in Lemma 10 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

$$\lim_{j\to\infty} \mathcal{Q}(\lambda_{k(j)}) = \mathcal{Q}(\lambda). \tag{31}$$

By (30) and (31), $\mathcal{Q}(\lambda) = \lambda$. □

LEMMA 8. *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 Lemma 4.4 in [16] and is omitted.

LEMMA 9. *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

$$W_0(k) \stackrel{d}{=} \frac{1}{k}\sum_{i=1}^{k} W_0^i \quad \text{and so} \quad E[W_0(k)] = \frac{1}{k}\sum_{i=1}^{k} E[W_0^i], \tag{32}$$

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

$$W_0^1 = \sup_{n\geq 0} \sum_{i=-n}^{-1} A_i - S_i^1, \tag{33}$$

where, as usual, we take the empty sum to be 0. We also have that

$$D_n^k = D_n^{k-1} + W_{n-1}^k - W_n^k, \qquad n \in \mathbb{Z},\ k = 1, 2, 3, \ldots, \tag{34}$$

where D_n^0 is identified with A_n. Using (33) and (34), it can be shown inductively (see, e.g., [9] or [1], Proposition 5.4) that

$$\sum_{i=1}^{k} W_0^i = \sup_{n_k \geq \cdots \geq n_1 \geq 0} \sum_{i=-n_k}^{-1} A_i - \sum_{j=1}^{k} \sum_{i=-n_j}^{-n_{j-1}-1} S_i^j, \tag{35}$$

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

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

To obtain the last equality above, we have used the fact that the expectation of the supremum of a collection of nonnegative 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 to partition n_k into k nonnegative integers, which is $\binom{n_k+k}{k}$. Moreover, since the S_i^j for different i, j are i.i.d., 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

$$\begin{aligned}
&\mathbf{P}\left(\sum_{i=1}^{k} W_0^i > kx \right) \\
&\qquad \leq e^{-\theta kx} \sum_{n=0}^{\infty} \binom{n+k}{k} E \exp\left[\theta \sum_{i=-n}^{-1} (A_i - S_i^1) \right].
\end{aligned} \tag{36}$$

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,

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

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 $\varepsilon > 0$, there is an $N < \infty$ such that

$$|\Lambda_n(\theta) - \Lambda_A(\theta)| < \varepsilon \qquad \forall n \geq N,\ \theta \in [0, \theta_0]. \tag{37}$$

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 (36) and (37) that, for all $\theta \in (0, \theta_0)$,

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

Observe from (27) that we can find $\theta \in (0, \theta_0)$ and $\varepsilon > 0$ sufficiently small that $\Lambda_A(\theta) + \varepsilon + \Lambda_S(-\theta) < -\varepsilon$. For such θ and ε, we get

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

where c is a constant that may depend on θ, ε and N, but does not depend on k. Thus, we obtain using (32) that

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

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

LEMMA 10. *The distribution* λ, *which was defined in the statement of Theorem* 4 *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 nondeterministic, it follows from Loynes's construction that if the arrival process is independent of the service process but has the same distribution, then the expected workload at time 0 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 0 when the arrival process has distribution λ and is independent of the service process, then

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

Now $\lambda_{k(j)} \to \lambda \in M$ by definition, so it follows from Lemmas 8 and 9 that $E[W_0] < +\infty$. Hence $\lambda \in M_0$. $\square$

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} that have ergodic components that belong *only to* M_{e}^{ζ}. Thus if a process $X = \{X_n,\ n \in \mathbb{Z}\}$ is distributed according to some $\nu \in M_{\mathrm{sp}}^{\zeta}$, 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 $\bigcup_{\zeta\in[0,1)} M_{\mathrm{sp}}^{\zeta}$ as

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

where Φ is some measure on $[0, 1)$. By linearity of $\mathcal{Q}$, $\mathcal{Q}(\lambda) = \int_0^1 \mathcal{Q}(\lambda_\zeta)\Phi(d\zeta)$. However, 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 6 below as a corollary to Theorem 5.

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 [17] in the continuous-time setting.

Consider an infinite tandem of queues indexed by the nonnegative integers. Let $S = \{S_n, n \in \mathbb{Z}\}$ be an i.i.d. family of nonnegative integer-valued random variables, where S_n denotes the maximum amount of service effort available at queue 0 in the nth time slot. For $n \in \mathbb{Z}$, $k \geq 1$, let S_n^k be the maximum amount of service effort in the nth time slot at queue number k. The processes $S^k = \{S_n^k, n \in \mathbb{Z}\}$ are i.i.d. and independent of S, and $S_n^k \stackrel{d}{=} S_1$ 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 nonnegative integers, $E(A_1) = \alpha < E(S_1)$. 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 [15] 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 $\bar{\rho}$ 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 [13] or Chang [5], Definition 2.3, for further details of the $\bar{\rho}$ metric.)

THEOREM 5. *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* F. *Suppose also that* $P(S_n = 0) > 0$. *Then* $\bar{\rho}(A^k, F) \to 0$ *as* k *goes to infinity*.

PROOF. Our method of proof will closely follow that of [17]; we shall merely set up the language and notation needed to import the argument in [17].

We use the coupling in [17]. 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 [17], 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 (nonrandom) 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 [17] 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 kth queue. From the service policy and coloring scheme, we readily obtain:

LEMMA 11. *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 colored yellow after $k+1$ queues.*

The importance of Lemma 11 is that among customers who 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 [17], this order preservation property can be used to obtain the following lemma (identical to Lemma 3.1 of [17]).

LEMMA 12. *If the density of ever-blues is strictly positive, then there exists an ε, not depending on k, such that the (nonrandom) density in F^k of red customers C satisfying "there exist blue customers of A^k in $(C(k), C(k) + 2/\varepsilon]$" must be at least $\varepsilon/2$.*

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 $\varepsilon/4$.

Given this and the conclusion of Lemma 12, we obtain the next lemma (similar to Lemma 3.2 of [17]).

LEMMA 13. *Under the assumptions of Lemma 12, 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 $(C(k), C(k) + 2/\varepsilon]$ and*
(b) *$P(C$ arrives at an empty queue $|F^k) > \delta$*

have density at least $\varepsilon/4$.

Consider a red customer R who satisfies properties (a) and (b) of Lemma 13. Because of property (b) the chance that R finds queue k empty upon arrival is at least δ. Since the process S^k is i.i.d., independent of F^k and $s = P(S_1 = 0) > 0$, the chance that R waits at least $2/\varepsilon$ units of time at queue k before departing is at least $s^{\lceil 2/\varepsilon \rceil - 1}$. 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 12, $\mathcal{D}(Y^{k+1}) - \mathcal{D}(Y^k) \geq \delta s^{\lceil 2/\varepsilon \rceil - 1}\varepsilon/4$ 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, as in Corollary 2 of [21], we obtain

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

This concludes the proof of Theorem 5. □

THEOREM 6. *If* $\lambda \in \mathcal{M}_{\mathrm{sp}}^{\alpha}$ *is a fixed point for the queue, then it is necessarily ergodic; that is,* $\lambda \in M_{\mathrm{e}}^{\alpha}$.

PROOF. Given Theorem 5, the proof is identical to the proof of Theorem 5.2 in [16] and is omitted. □

REFERENCES

[1] BACCELLI, F., BOROVKOV, A. and MAIRESSE, J. (2000). Asymptotic results on infinite tandem queueing networks. *Probab. Theory Related Fields* **118** 365–405.

[2] BEDEKAR, A. S. and AZIZOGLU, M. (1998). The information–theoretic capacity of discrete-time queues. *IEEE Trans. Inform. Theory* **44** 446–461.

[3] BILLINGSLEY, P. (1995). *Probability and Measure*. Wiley, New York.

[4] CHANG, C. S. (1994). Stability, queue length and delay of deterministic and stochastic queueing networks. *IEEE Trans. Automat. Control* **39** 913–931.

[5] CHANG, C. S. (1994). On the input–output map of a $G/G/1$ queue. *J. Appl. Probab.* **31** 1128–1133.

[6] DEMBO, A. and ZEITOUNI, O. (1993). *Large Deviations Techniques and Applications*. Jones and Bartlett, Boston.

[7] DE VECIANA, G. and WALRAND, J. (1995). Effective bandwidths: Call admission, traffic policing and filtering for ATM networks. *Queueing Systems Theory Appl.* **20** 37–59.

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

[9] GANESH, A. J. (1998). Large deviations of the sojourn time for queues in series. *Ann. Oper. Res.* **79** 3–26.

[10] GANESH, A. J. and O'CONNELL, N. (1998). The linear geodesic property is not generally preserved by a FIFO queue. *Ann. Appl. Probab.* **8** 98–111.

[11] GANESH, A. J. and O'CONNELL, N. (2002). A large deviation principle with queueing applications. *Stochastics Stochastics Rep.* **73** 23–35.

[12] GLYNN, P. and WHITT, W. (1994). Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *J. Appl. Probab.* **31A** 131–156.

[13] GRAY, R. M. (1988). *Probability, Random Processes and Ergodic Properties*. Springer, New York.

[14] KELLY, F. P. (1979). *Reversibility and Stochastic Networks*. Wiley, New York.

[15] LOYNES, R. (1962). The stability of a queue with non-independent interarrival and service times. *Proceedings of the Cambridge Philosophical Society* **58** 497–520.

[16] MAIRESSE, J. and PRABHAKAR, B. (1999). On the existence of fixed points for the $\cdot/GI/1/\infty$ queue. *Ann. Probab.* To appear.

[17] MOUNTFORD, T. and PRABHAKAR, B. (1995). On the weak convergence of departures from an infinite sequence of $\cdot/M/1$ queues. *Ann. Appl. Probab.* **5** 121–127.

[18] MUNTZ, R. R. (1972). Poisson departure processes and queueing networks. Research Report RC 4145, IBM.

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

[20] O'CONNELL, N. (1999). Directed percolation and tandem queues. Technical Report STP-99-12, DIAS.
[21] PRABHAKAR, B. (2000). The attractiveness of the fixed points of a $\cdot/GI/1$ queueing operator. *Ann. Probab.* To appear.
[22] RUELLE, D. (1969). *Statistical Mechanics*. Benjamin, Elmsford, NY.
[23] WALRAND, J. (1988). *An Introduction to Queueing Networks*. Prentice–Hall, Englewood Cliffs, NJ.

A. GANESH
MICROSOFT RESEARCH
7 J J THOMSON AVENUE
CAMBRIDGE CB3 0FB
UNITED KINGDOM
E-MAIL: ajg@microsoft.com

N. O'CONNELL
BRIMS
HEWLETT–PACKARD LABORATORIES
FILTON ROAD
BRISTOL BS34 6QZ
UNITED KINGDOM
E-MAIL: noc@hplb.hpl.hp.com

B. PRABHAKAR
DEPARTMEN OF ELECTRICAL ENGINEERING
AND COMPUTER SCIENCE
PACKARD BUILDING
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305
E-MAIL: balaji@stanford.edu