RC 16280 (#71858) 10/11/90
Computer Science 19 pages

91A000644

Research Report

Analysis of Rare Events in Continuous Time Markov Chains via Time Reversal and Fluid Approximation

Venkat Anantharam

School of Electrical Engineering
Cornell University
Ithaca, NY 14853

Philip Heidelberger and Pantelis Tsoucas

IBM Research Division
T. J. Watson Research Center
Yorktown Heights, NY 10598

ANALYSIS OF RARE EVENTS IN CONTINUOUS TIME MARKOV CHAINS VIA TIME REVERSAL AND FLUID APPROXIMATION

Venkat Anantharam

School of Electrical Engineering
Cornell University
Ithaca, NY 14853

Philip Heidelberger and Pantelis Tsoucas

IBM Research Division
T.J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

ABSTRACT

This paper is concerned with the occurence of rare events in stationary, continuous time Markov chains (CTMC). The following observation is exploited: given that a stationary CTMC is in a set B at some time t_0, the path by which it got there has the distribution of the reverse time process given that it started in the set B. For rare events, the path of the process in forward time, up to time t_0, asymptotically approaches the fluid approximation limit of the reversed time process. Thus one derives direct analogs of results that are obtained using the theory of large deviations, while the method is substantially easier to apply whenever the time-reversed process is known. As an application, the build-up of large queue lengths in Jackson networks is analyzed. The utility of these results is that the paths obtained often suggest effective importance sampling schemes that can be used in simulation. This is demonstrated in a two queue tandem network.

1. Introduction

This paper is concerned with identifying how rare events occur in stationary continuous time Markov chains (CTMC's). Previous approaches to this problem mainly involve the Wentzel-Freidlin theory of large deviations and the solution of the so-called exit problem. (See [Va].) This is a variational problem the solution of which asymptotically determines the logarithm of the probability of occurence of a rare event during a regenerative cycle of the process. From the solution of the exit problem one also obtains an asymptotic characterization of the path of the process leading to the rare event. This path is useful in importance sampling methods for the estimation of probabilities of rare events by simulation. It provides parameters for the optimal, in a certain sense, simulation of the Markov chain. This connection was suggested in [CFM] and is further explored in [Bu].

A significant difficulty with this approach is that the classical theory of large deviations does not apply directly to certain Markov chains that arise frequently in applications. Roughly speaking, the difficulty is due to discontinuities of the transition rates at the boundary of the state space. An important such example is the problem of evaluating the probability that the total number of customers in an ergodic Jackson network of queues exceeds some level $N \to \infty$ during a busy cycle of the network. A substantial amount of work has been devoted to this problem. In [PW] a heuristic exit problem was proposed and the solution to the ensuing optimization problem was obtained in [$FA1$], [$FA2$] and [FLA]. In [DIM] a viscosity solution was obtained for an associated variational problem and a partial characterization of the action functional for large deviations of a related class of Markov chains was obtained in [DEW]. An action functional for networks of queues in series was provided in [Ts]. The solution of an associated exit problem for two queues in series was also obtained.

The point of departure for this paper is the observation in [FLA] that the exit path obtained by solving the heuristic variational problem in [PW], for the particular event described above, corresponds to time reversal of the Jackson network with respect to its stationary probability distribution. Our main result is the formulation, in Section 2, of this observation into a limit theorem which holds for general CTMC's and general events. Roughly speaking, the result can be stated as follows:

> given that a stationary CTMC is in a rare set B^N at some time (say 0), then as $N \to \infty$, the path by which it got there is the same path by which the reverse time process evolves, given that the reverse time process starts in the set B^N.

The above requires that an appropriately scaled version of the reverse time process has a deterministic fluid limit, which is a consequence of the law of large numbers. In addition, the limiting "hitting distribution" on B^N, given that the process ends up in B^N, is equal to the limiting conditional stationary measure on B^N. If this limiting conditional measure is concentrated on a single point, then the above result states that, with very high probability and asymptotically as $N \to \infty$, there is only one way in which the rare event occurs. If the conditional stationary measure on B^N has a nondegenerate limit, then the occurrence of the rare event can be described as follows. The hitting location is first chosen (according to the limiting conditional measure). Then, given the hitting location, the CTMC follows the fluid limit path on which the reversed process evolves from that location. These results are direct analogs of what one obtains from the theory of large deviations when it is applicable.

In Section 3 we specialize the results of Section 2 to Jackson networks. A derivation of the well known fluid limit is presented and in Section 3.5 we obtain limits of the conditional stationary distribution on a class of open convex sets. The latter belongs to a class of results on limiting conditional product form distributions. (See [$Ke2$] and references therein.) It may be of independent interest since it does not seem to follow from previous results.

Section 4 contains an application to the build-up of large queue lengths in a two queue tandem network.

In Section 5, simulation of rare events based on importance sampling is carried out for the two queue example of Section 4. For the CTMC's considered here, an interesting feature of our results is that they readily provide parameters on which effective importance sampling schemes could be based. The possibility is illustrated in Section 5 by a simulation of the example of Section 4. Finally, the results are summarized and related future research problems are described in Section 6.

2. The general scheme

Consider a family of CTMC's denoted by $(X^N(t))$. For each $N \geq 1$, X^N belongs to a subset S^N of

$$\mathbf{Z}/N \stackrel{\text{def}}{=} \{k/N : k \in \mathbf{Z}\}$$

which can be written as $\mathbf{Z}/N \cap S$ for a closed set $S \subset \mathbf{R}^K$. Attention is restricted to sample paths in $D([-T,T], \mathbf{Z}/N)$. Assume that X^N is positive recurrent on S^N and let π^N be the stationary probability measure of X^N. This implies that π^N is strictly positive for all open sets in S for large enough N. This fact will be used in defining conditional probabilities in the sequel.

For any stationary CTMC $(Y(t))_{t\in[-T,T]}$ with stationary probability distribution π and generator matrix Q, the time reversed process defined as $\tilde{Y}(t) \stackrel{\text{def}}{=} Y(-t)$ is also a stationary CTMC with stationary distribution π and generator matrix $\tilde{Q}$ where

$$\tilde{Q}_{i,j} = \pi(j)Q_{j,i}/\pi(i) \tag{2.1}$$

(see, e.g., [Kel].) Consider the time reversal of process X^N with respect to π^N denoted as $\tilde{X}^N(t) \stackrel{\text{def}}{=} X^N(-t)$ where $X^N(0)$ has distribution π^N. Consequently, $X^N(t)$ has distribution π^N for all $t \in [-T,T]$. The following limit is assumed for the sequence of processes $\tilde{X}^N$ which in our examples is a consequence of the law of large numbers. It is described in terms of a set of transformations on S parametrized by $t \in [-T,T]$ and denoted by $\left(\tilde{\phi}_t(\cdot)\right)$. In this paper such a set will be called a flow.

Assumption 1. There exists a flow $\tilde{\phi}$ such that for any $\epsilon_0 > 0$,

$$\lim_{N\to\infty} P\left\{ \sup_{0\leq s\leq T} \|\tilde{X}^N(s) - \tilde{\phi}_s(x)\| \geq \epsilon \;\middle|\; \|\tilde{X}^N(0) - x\| < \epsilon_0 \right\} = 0,$$

for all $\epsilon > \epsilon_0$.

This limit characterizes the past behavior of X^N given the current state. This can be seen from the equivalent form

$$\lim_{N\to\infty} P\left\{ \sup_{-T\leq s\leq 0} \|X^N(s) - \tilde{\phi}_{-s}(x)\| \geq \epsilon \;\middle|\; \|X^N(0) - x\| < \epsilon_0 \right\} = 0.$$

Next, consider (E,C) a pair of subsets of S with $C \subset E$ such that a limit exists for the conditional measure on E evaluated at C, i.e.,

$$\pi^N(C|E) \stackrel{\text{def}}{=} \frac{\pi^N(C)}{\pi^N(E)} \longrightarrow_{N\to\infty} \mu_E(C). \tag{2.2}$$

Denote the collection of such pairs by $\mathcal{D}(S)$. In order to formulate the main result of this section the following notation is needed. Define the open ϵ-neighborhood of a set A as

$$A^\epsilon \stackrel{\text{def}}{=} \left\{ x \in \mathbf{R}^K \;:\; \inf_{y\in A} \|x - y\| < \epsilon \right\}.$$

Also define the event

$$\mathcal{E}^N \stackrel{\text{def}}{=} \left\{ \inf_{x \in A} \sup_{-T \le s \le 0} \| X^N(s) - \tilde{\phi}_{-s}(x) \| \ge \epsilon \right\}. \tag{2.3}$$

Lemma 1. For A, B such that $A \subset B$, and ϵ, δ such that $\epsilon > \delta > 0$ and $A^{\epsilon+\delta} \subset B$, if $(B, A^{\epsilon+\delta})$ and $(B, A^{\epsilon-\delta})$ belong to $\mathcal{D}(S)$, then

$$1 - \mu_B\left(A^{\epsilon-\delta}\right) \ge \limsup_{N\to\infty} P\left(\mathcal{E}^N \mid X^N(0) \in B\right) \ge \liminf_{N\to\infty} P\left(\mathcal{E}^N \mid X^N(0) \in B\right) \ge 1 - \mu_B\left(A^{\epsilon+\delta}\right). \tag{2.4}$$

Proof. By conditioning further on $X^N(0)$ one has

$$\begin{aligned} P\left(\mathcal{E}^N \mid X^N(0) \in B\right) = & P\left(\mathcal{E}^N \mid X^N(0) \in A^{\epsilon-\delta}\right) \pi^N\left(A^{\epsilon-\delta} \mid B\right) \\ & + P\left(\mathcal{E}^N \mid X^N(0) \in B \backslash A^{\epsilon+\delta}\right)\left(1 - \pi^N\left(A^{\epsilon+\delta} \mid B\right)\right) \\ & + P\left(\mathcal{E}^N \mid X^N(0) \in A^{\epsilon+\delta} \backslash A^{\epsilon-\delta}\right) \pi^N\left(A^{\epsilon+\delta} \backslash A^{\epsilon-\delta} \mid B\right). \end{aligned} \tag{2.5}$$

Inequalities (2.4) follow from Assumption 2 and (2.5) if one establishes the limits

$$P\left(\mathcal{E}^N \mid X^N(0) \in B \backslash A^{\epsilon+\delta}\right) = 1$$

$$P\left(\mathcal{E}^N \mid X^N(0) \in A^{\epsilon-\delta}\right) = 0.$$

The first limit follows immediately by recalling the definition of $\mathcal{E}^N$ in (2.3). For the second limit pick $\delta > \eta > 0$ and for each $x \in A$ consider the ball

$$B_x \stackrel{\text{def}}{=} \left\{ y \in \mathbf{R}^K : \| y - x \| < \epsilon - \eta \right\}.$$

The sets $(B_x)_{x \in A}$ form an open cover of the closure of the open set $A^{\epsilon-\delta}$ in $\mathbf{R}^K$. Let $(B_{x_i})_{i=1}^{\infty}$ be a countable subcover of that set. By virtue of the inequality

$$P\left(\mathcal{E}^N | X^N(0) \in A^{\epsilon-\delta}\right) \le \sum_{i=1}^{\infty} P\left(\mathcal{E}^N | X^N(0) \in B_{x_i}\right) \pi^N\left(B_{x_i} | A^{\epsilon-\delta}\right),$$

and the monotone convergence theorem, it suffices to show that $\lim_{N\to\infty} P\left\{\mathcal{E}^N | X^N(0) \in B_{x_i}\right\} = 0$. But this is a consequence of Assumption 1. □

Remark. If μ_B were continuous in δ at A^{ϵ} then

$$\lim_{N\to\infty} P\left(\mathcal{E}^N | X^N(0) \in B\right) = 1 - \mu_B(A^{\epsilon}).$$

This is often the case for subsets of $\mathcal{D}(S)$ in specific applications.

3. Jackson networks

The scheme of the previous section is now employed in the study of paths of overflow in ergodic Jackson networks.

3.1. The evolution equations. Consider a network consisting of K nodes of $M/M/1$ queues. Denote the set of nodes by $\mathcal{K} = \{1, 2, \ldots, K\}$. By convention denote the outside world by node 0. Let the vectors of arrival and service processes be denoted by λ and μ in $\mathbf{R}_+^K$ respectively, and let the routing matrix be denoted by P. For simplicity assume that $p_{ii} = 0$ for $i = 1, \ldots, K$. Also assume that $(I - P)^{-1}$ exists and that the network is ergodic, i.e., $\lambda(I - P)^{-1} < \mu$.

With $Y(x)$ denoting a Poisson random variable with parameter $x \geq 0$, the evolution of the queueing processes $(Z_i(t))_{i=1}^K$ can be described as follows (see [*Ku*] for the notation used here.)

$$Z_i(t) = Z_i(0) + Y_i^a(\lambda_i t) + \sum_{j=1}^K Y_{ji}^d\left(\mu_j p_{ji} \int_0^t 1\{Z_j(s) > 0\}\, ds\right) - \sum_{j=0}^K Y_{ij}^d\left(\mu_i p_{ij} \int_0^t 1\{Z_i(s) > 0\}\, ds.\right) \quad (3.1)$$

The superscripts a and d in the above equations indicate arrival and departure processes respectively.

We are interested in determining asymptotically the past evolution of Z given that the process has the stationary distribution and is conditioned to be in a given set at time 0. The sets considered here are written in the form

$$B^N \stackrel{\text{def}}{=} \{k \in \mathbf{Z}_+^K \,:\, k/N \in B\}, \quad (3.2)$$

where $B \subset \mathbf{R}_+^K$ is open. Such a set is rare if its closure does not contain the origin. In order to study asymptotics as $N \to \infty$ it is convenient to define the scaled process

$$Z^N(\cdot) \stackrel{\text{def}}{=} Z(N\cdot)/N. \quad (3.3)$$

We now turn our attention to Assumption 1.

3.2. Convergence. The limiting flow of Assumption 1 is known as a fluid approximation limit. It has been previously obtained for networks of $GI/GI/1$ queues with Bernoulli routing in [*CM*]. Its derivation is a straightforward application of results in [*HR*] and will be described briefly for the special case of Jackson networks. Our presentation also borrows from [*GM*].

Adding and subtracting the means of the Poisson random variables in (3.1) gives

$$Z_i(t) = Z_i(0) + \lambda_i t + \sum_{j=1}^K \mu_j p_{ji} \int_0^t 1\{Z_j(s) > 0\}\, ds - \mu_i \int_0^t 1\{Z_i(s) > 0\}\, ds + M_i(t), \quad (3.4)$$

where $M_i(t)$ is a martingale with quadratic variation

$$< M_i >_t = \lambda_i t + \sum_{j=1}^K \mu_j p_{ji} \int_0^t 1\{Z_j(s) > 0\}\, ds - \mu_i \int_0^t 1\{Z_i(s) > 0\}\, ds. \quad (3.5)$$

Denoting by $L_i(t)$ the expected number of virtual departures in node i when the node is empty, i.e.,

$$L_i(t) \stackrel{\text{def}}{=} \mu_i \int_0^t 1\{Z_i(s) = 0\}\, ds,$$

equation (3.4) can be written in vector form as

$$\begin{aligned} Z(t) &= Z(0) + (\lambda - \mu(I-P))t + M(t) + L(t)(I-P) \\ &\stackrel{\text{def}}{=} X(t) + L(t)(I-P), \end{aligned}$$

where we have set $X(t) \stackrel{\text{def}}{=} Z(0) + (\lambda - \mu(I-P))t + M(t)$.

Lemma 2. (a) The process L satisfies the fixed point equation

$$L(t) = \sup_{0 \leq s \leq t} [L(s)P - X(s)]^+, \quad (3.6)$$

where sup and $[\cdot]^+$ are taken component-wise. (For $x \in \mathbf{R}$, $[x]^+ \stackrel{\text{def}}{=} \max(x, 0)$.)

(b) For each $X \in D([-T,T],\mathbf{R}_+^K)$, equation (3.6) has a unique solution in $D([-T,T],\mathbf{R}_+^K)$.

(c) The process L is a continuous causal functional of process X which we denote by

$$L(t) \stackrel{\text{def}}{=} \Psi \circ [X(s)]_0^t . \tag{3.7}$$

Proof. It is included in the proof of Theorem 1 of [*HR*] with the modification that the underlying space is $D([-T,T],\mathbf{R}_+^K)$. Note that this space is complete in the uniform metric (see [*Bi*1], p.150.) □

We are now ready to prove the existence of a limiting flow in forward time. Recall the definition of Z^N from (3.3) and let M^N denote the scaled version of process M in (3.4).

Lemma 3. For all $\epsilon > 0$,

$$\lim_{N\to\infty} P\left\{ \sup_{0\le s\le T} \|M^N(s)\| \ge \epsilon \right\} = 0.$$

Proof. It follows from martingale maximal inequalities. Apply for instance Lenglart's inequality (see [*JS*]) and note that

$$< M^N >_t = \frac{1}{N} < M >_t .$$

□

For $Z_0^N \in \mathbf{Z}_+^K/N$, the process Z^N is given by

$$Z^N(t) = Z_0^N + (\lambda - \mu(I-P))t + M^N(t) + \Psi \circ \left[Z_0^N + (\lambda - \mu(I-P))s + M^N(s)\right]_0^t (I-P).$$

Lemma 3 then implies that a limiting flow ϕ exists for Z^N. For $x \in \mathbf{R}_+^K$ it is given by

$$\phi_t(x) \stackrel{\text{def}}{=} x + (\lambda - \mu(I-P))t + \Psi \circ [x + (\lambda - \mu(I-P))s]_0^t (I-P). \tag{3.8}$$

Corollary 1. For any $\epsilon_0 > 0$ and for all $\epsilon > \epsilon_0$

$$\lim_{N\to\infty} P\left\{ \sup_{0\le s\le T} \|X^N(s) - \phi_s(x)\| \ge \epsilon \;\middle|\; \|X^N(0) - x\| < \epsilon_0 \right\} = 0.$$

3.3. The limit. We now give an expression for the flow in (3.8). To this end consider the so-called generalized throughput equation

$$\xi = \lambda + (\xi \wedge \mu)P, \tag{3.9}$$

for λ, μ, P not necessarily satisfying the stability condition $\lambda(I-P)^{-1} < \mu$ of Section 3.1. The next lemma and its proof appear in [*GM*]. We include the existence proof as the notation will prove useful in Lemma 5.

Lemma 4. Equation (3.9) has a unique non-negative solution.

Proof. Set $\psi^{(0)} \stackrel{\text{def}}{=} \lambda + \mu P$, $U_0 \stackrel{\text{def}}{=} \{j \in \mathcal{K} : \psi^{(0)}{}_j < \mu_j\}$, $V_0 \stackrel{\text{def}}{=} \mathcal{K}\backslash U_0$. If $U_0 = \emptyset$ then $\psi^{(0)}$ solves (3.9). Otherwise define inductively for $n = 0, 1, \ldots,$

$$\begin{aligned} \psi_{U_n}^{(n+1)} &= (\lambda_{U_n} + \mu_{V_n} P_{V_n U_n})(I - P_{U_n U_n})^{-1}, \\ \psi_{V_n}^{(n+1)} &= \lambda_{V_n} + \mu_{V_n} P_{V_n V_n} + \psi_{U_n}^{(n+1)} P_{U_n V_n}, \\ U_{n+1} &= \left\{ j \in \mathcal{K} : \psi_j^{(n+1)} < \mu_j \right\}. \qquad V_{n+1} = \mathcal{K}\backslash U_{n+1}. \end{aligned} \tag{3.10}$$

By verifying that $\psi^{(n+1)} \le \cdots \le \psi^{(0)}$ one has that $U_0 \subseteq \cdots \subseteq U_{n+1}$ and if $U_n = U_{n+1}$ then $\psi^{(n+1)}$ solves (3.9). □

Note that in the Jackson network under consideration $\lambda(I-P)^{-1} < \mu$ and hence

$$\rho \stackrel{\text{def}}{=} \lambda(I-P)^{-1} \tag{3.11}$$

is the unique solution of (3.9).

We will need the following notation. For $z \in \mathbf{R}_+^K$ set

$$E(z) \stackrel{\text{def}}{=} \{j \in \mathcal{K} \,:\, z_j = 0\}, \qquad F(z) \stackrel{\text{def}}{=} \mathcal{K} \backslash E(z),$$

for the set of empty and non-empty nodes respectively, corresponding to a limiting queue length vector z. For a vector $z \in \mathbf{R}^K$, subsets S, S' of $\mathcal{K}$ and a matrix A in $\mathbf{R}^{K \times K}$, let $z_S \in \mathbf{R}^{|S|}$ be the restriction of the coordinates of z on S and let $A_{SS'}$ be the restriction of A consisting of entries from rows in S and columns in S'. In this notation, for a vector $z \in \mathbf{R}_+^K$ consider the solution $\xi \geq 0$ of the following equations

$$\xi_{E(z)} = \left(\lambda_{E(z)} + \mu_{F(z)} P_{F(z)E(z)}\right) + \left(\xi_{E(z)} \wedge \mu_{E(z)}\right) P_{E(z)E(z)}, \tag{3.12}$$

$$\xi_{F(z)} = \left(\lambda_{F(z)} + \mu_{F(z)} P_{F(z)F(z)}\right) + \left(\xi_{E(z)} \wedge \mu_{E(z)}\right) P_{E(z)F(z)}. \tag{3.13}$$

Note that existence and uniqueness of $\xi_{E(z)}$ follows from Lemma 4(b) and $\xi_{F(z)}$ is directly computed from (3.13).

Intuitively, for a given $z \in \mathbf{R}_+^K$ nodes in $E(z)$ receive jobs at rate $\mu_{F(z)} P_{F(z)E(z)}$ which is added to the external arrival rate $\lambda_{E(z)}$, hence equation (3.12). It is therefore plausible that, for $z_0 \in \mathbf{R}_+^K$, the image of the function

$$x(t) = z_0 + (\lambda - \mu(I-P))t, \; t \in [-T, T]$$

under flow ϕ in (3.8) is given by the solution to the integral equation

$$z(t) = z_0 + \int_0^t \left(\left[\xi_{E(z(s))} - \mu_{E(z(s))}\right]^+, \xi_{F(z(s))} - \mu_{F(z(s))} \right) ds. \tag{3.14}$$

Again, $[\cdot]^+$ is taken component-wise. Observe that the solution of (3.14) has piecewise constant time derivative. Changes can only occur at instants when the sets $E(z(t))$ and $F(z(t))$ change. To verify that the solution of (3.14) satisfies $z(t) = \phi_t(z_0)$ consider the function l, taking values in $\mathbf{R}_+^K$, given by

$$l(t) \stackrel{\text{def}}{=} \int_0^t \left(\mu_{E(z(s))} - \xi_{E(z(s))} \wedge \mu_{E(z(s))}, 0\right) ds$$

Noting that $[a-b]^+ = a - a \wedge b$ for $a, b \in \mathbf{R}$, it is seen that the pair of functions $(l(t), x(t))$ satisfies equation (3.6) and one has

$$z(t) = x(t) + l(t)(I-P).$$

We conclude this subsection with a result whose proof is obtained from the method of Lemma 4.

Lemma 5. For each $Z_0 \in \mathbf{R}_+^K$, $\tau(Z_0) \stackrel{\text{def}}{=} \inf\{t > 0 \,:\, Z(t) = 0\}$ is finite and $Z(t) = 0$ for $t \geq \tau(Z_0)$.

Proof. For $Z \in \mathbf{R}_+^K$ denote by ξ^* the solution of (3.12) and (3.13). Also, consider the sequence $(\psi^{(n)}, U_n)$ from (3.10). Suppose it terminates at $k^* \leq K$ and observe that the stability condition of Section 3.1 implies that $U_{k^*} = \mathcal{K}$. Next, observe that $\xi^*_{U_0} \leq \psi^{(0)}_{U_0} < \mu_{U_0}$. This implies that all nodes in U_0 will empty in finite time and will remain empty. Similarly, observe that if $U_l \subseteq E(Z) \subseteq U_{l+1}$, $l = 1, \ldots, k^* - 1$, then $\xi^*_{U_{l+1}} \leq \psi^{(l+1)}_{U_{l+1}} < \mu_{U_{l+1}}$, and the proof is complete. □

3.4. Time reversal. We have thus far proved the existence of and have described a limiting flow in forward time for the scaled process of queue lengths Z^N. It is well known that the time reversal of process Z^N with

respect to its invariant probability measure, denoted here by π^N, yields a process $\hat{Z}^N$ of queue lengths in a K node Jackson network with parameters $\tilde{\lambda}$, $\tilde{\mu}$, $\tilde{P}$, where the obvious correspondence is implied. This can be obtained from (2.1) which also yields the parameters of the reversed time network. The precise statement is that if $Z^N(0)$ has distribution π^N then the process $\tilde{Z}^N(-t) \stackrel{\text{def}}{=} Z^N(t) \in \mathbf{Z}_+^K/N$ is given by

$$\tilde{Z}_i^N(t) = Z_i^N(0) + \frac{1}{N} Y_i^a\left(N\tilde{\lambda}_i t\right) + \frac{1}{N}\sum_{j=1}^K Y_{ji}^d\left(N\tilde{\mu}_j\tilde{p}_{ji}\int_0^t 1\left\{\tilde{Z}_j(s)>0\right\}ds\right)$$
$$- \frac{1}{N}\sum_{j=0}^K Y_{ij}^d\left(N\tilde{\mu}_i\tilde{p}_{ij}\int_0^t 1\left\{\tilde{Z}_i(s)>0\right\}ds.\right)$$

It follows that the results of the previous subsections apply to the process $\tilde{Z}^N$ with parameters $\tilde{\lambda}$, $\tilde{\mu}$ and $\tilde{P}$ replacing λ, μ and P respectively. Applying Corollary 1 we have thus verified Assumption 1 of Section 2.

3.5. Asymptotics. In this section we demonstrate that Lemma 1 is of interest for Jackson networks by showing that the class $\mathcal{D}(\mathbf{R}_+^K)$ defined in Section 2 contains pairs of sets that frequently arise in applications. The main result of this section is condition (C) below, a sufficient condition for pairs of convex, open sets to belong to $\mathcal{D}(\mathbf{R}_+^K)$. A calculation of the corresponding limit (2.2) is given in Lemmas 7 and 9.

The stationary probability distribution of a Jackson network has the form

$$\pi(n) = \prod_{i=1}^K (1-\rho_i)\rho_i^{n_i}, \quad n \in \mathbf{Z}_+^K,$$

where ρ is given by (3.11). Our aim is to give a sufficient condition on open convex subsets A, B of $\mathbf{R}_+^K$ such that $(B,A) \in \mathcal{D}(\mathbf{R}_+^K)$ and determine the limit (2.2) rewritten here as

$$\lim_{N\to\infty} \frac{F_A(N)}{F_B(N)}, \tag{3.15}$$

where

$$F_Q(N) \stackrel{\text{def}}{=} \sum_{n\in Q^N} e^{N\left(\sum_{i=1}^K n_i/N \ln\rho_i\right)}, \qquad Q = A, B.$$

The existence of this limit and its value depends crucially on the behavior of the function

$$h(x) \stackrel{\text{def}}{=} \sum_{i=1}^K x_i \ln\rho_i$$

on the sets A and B. The idea is to express the functions $F_Q(N)$, $Q = A, B$, as one parameter sums to which Laplace's method can be applied. This motivates the following developments. Define $y_Q \stackrel{\text{def}}{=} \max\left\{h(x) : x \in \bar{Q}\right\}$, $Q = A, B$. At this point we discriminate between two possibilities and first consider

Case I: $y_A = y_B \stackrel{\text{def}}{=} y$. Define the hyperplanes $H_z \stackrel{\text{def}}{=} \left\{x \in \mathbf{R}_+^K : h(x) = z\right\}$ and for $\eta > 0$ set

$$C_Q^N \stackrel{\text{def}}{=} \left\{h(n) : h(n) \geq y - \eta, n/N \in Q^N\right\}, \quad Q = A, B. \tag{3.16}$$

List the elements of C_Q^N in decreasing order as $\left\{y_1^{Q,N}, \ldots, y_{M_N^Q}^{Q,N}\right\}$ and observe that

$$F_Q(N) = \sum_{m=1}^{M_N^Q} e^{Ny_m^{Q,N}}\left|H_{y_m^{Q,N}} \cap Q^N\right| + O\left(e^{N(y-\eta)}\right), \quad Q = A, B. \tag{3.17}$$

To determine the limit in (3.15) it is therefore sufficient to only consider the first term. We next turn our attention to the asymptotics of the cardinality of the sets $H_{y,\underline{m}}^{Q,\kappa} \cap Q^N$, $m = 1, \ldots, M_N^Q$. This will be determined by considering the set

$$V \stackrel{\text{def}}{=} \left\{ n \in \mathbf{Z}^K : \sum_{i=1}^{K} n_i \ln \rho_i = 0 \right\}.$$

The following lemma relates a representation of V with the numbers $(\ln \rho_i)_{i=1}^K$ viewed as elements of the vector space $\mathbf{R}$ over the field of rationals $\mathbf{Q}$ (see [*Bi2*], p.170.)

Lemma 6. For $l = 1, \ldots, K$, if the rank κ of $\{\ln \rho_1, \ldots, \ln \rho_K\}$ in the vector space $\mathbf{R}$ over the field $\mathbf{Q}$ is l, then there are linearly independent vectors $\{n^{(1)}, \ldots, n^{(K-l)}\} \subset \mathbf{Z}^K$ such that each $m \in V$ has a unique representation

$$m = \sum_{i=1}^{K-l} k_i n^{(i)}; \qquad k_i \in \mathbf{Z}, \quad i = 1, \ldots, K-l.$$

Proof. If $l = K$ then clearly $V = \{0\}$. For economy of notation we only prove the case $l = 1$. The general case is similar. For $i = 1, \ldots, K-1$, since a_1, a_{i+1} are linearly independent over $\mathbf{Q}$, one can find relatively prime, non-zero integers $n_1^{(i)}, n_{i+1}^{(i)}$, such that

$$a_1 n_1^{(i)} + a_{i+1} n_{i+1}^{(i)} = 0. \tag{3.18}$$

Next, form the vectors

$$n^{(i)} = \left(n_1^{(i)}, 0, \ldots, n_{i+1}^{(i)}, \ldots, 0 \right), \qquad i = 1, \ldots, K-1,$$

and consider any $m = (m_1, \ldots, m_K) \in V$ with $m_1 \neq 0$. From (3.18) one obtains that $a_{i+1}/a_1 = -n_1^{(i)}/n_{i+1}^{(i)}$, $i = 1, \ldots, K$, and since $m \in V$,

$$m_1 = \sum_{i=1}^{K-1} m_{i+1} \frac{n_1^{(i)}}{n_{i+1}^{(i)}}.$$

This in turn implies that

$$m_1 \prod_{i=1}^{K-1} n_{i+1}^{(i)} = \sum_{i=1}^{K-1} m_{i+1} n_1^{(i)} \prod_{i \neq j} n_{i+1}^{(i)}.$$

Therefore, $n_{i+1}^{(i)}$ divides m_{i+1}, $i = 1, \ldots, K-1$. Uniqueness of the representation follows since $\left(n^{(i)}\right)_{i=1}^{k_1}$ are linearly independent. □

For $z \in [y - \eta, y]$ and $Q = A, B$ define

$$l_Q(z) \stackrel{\text{def}}{=} \mathcal{L}^{K-1}(H_z \cap Q),$$

where $\mathcal{L}^{k-1}$ is the Lebesgue measure on $\mathbf{R}^{K-1}$. We can now state a condition on l_Q, $Q = A, B$, sufficient for the existence of limit (3.15).

(*C*) The functions l_Q, $Q = A, B$, are twice continuously differentiable in the interval $[y - \eta, y]$ and the ratio $l_A'(y)/l_B'(y)$ is not indeterminate whenever $l_B(y) = 0$.

The method of determining the asymptotics of $|H_{y,\underline{m}}^{Q,\kappa} \cap Q^N|$ in (3.17) depends on the rank κ of $\{\ln \rho_i\}_{i=1}^K$ in $\mathbf{R}$ over $\mathbf{Q}$. We only present the extreme cases $\kappa = 1$ and $\kappa = K$. Intermediate cases can be handled by combining the two methods.

Case I(a): $\kappa = 1$. Note that in this case the number $\eta > 0$ can be picked small enough so that $\mathcal{C}_A^N = \mathcal{C}_B^N \stackrel{\text{def}}{=} \mathcal{C}^N$. By α denote the volume in $\mathbf{R}^{K-1}$ of the parallelopiped formed by vectors $\left(n^{(i)}\right)_{i=1}^{K-1}$ of Lemma 6. Then, one has

Lemma 7. (a) $\left|H_{y_m^N} \cap Q^N\right| = l_Q\left(y_m^N\right) N^{K-1}/\alpha + r_{N,m}^Q$, where $\lim_{N\to\infty} \sup_m \left|r_{N,m}^Q\right|/N^{K-1} = 0, \quad Q = A, B.$

(b) $\delta^N = c/N$, where $c > 0$ and $\delta^N \stackrel{\text{def}}{=} y_m^N - y_{m+1}^N$, $m = 1, \ldots, M_N - 1$.

Proof. Part (a) follows from condition (C) above and the fact that, by Lemma 6, the set $H_{y_m^N} \cap Q^N$ becomes a Riemann partition for the functions l_Q, $Q = A, B$. Part (b) is immediate. □

Consequently, it suffices to consider

$$\lim_{N\to\infty} \frac{\sum_{m=1}^{M_N} e^{Ny_m^N} l_A\left(y_m^N\right) \delta^N}{\sum_{m=1}^{M_N} e^{Ny_m^N} l_B\left(y_m^N\right) \delta^N}.$$

The sums in the above ratio resemble Riemann approximations of the Laplace integrals

$$\int_{y-\eta}^{y} e^{Nz} l_Q(z)\, dz, \quad Q = A, B.$$

By repeating the arguments of Laplace's method (see [*Co*], p.36,) one obtains the following result.

Lemma 8. Assume condition (C). Then, if $y_A = y_B = y$,

$$\lim_{N\to\infty} F_A(N)/F_B(N) = \begin{cases} l_A(y)/l_B(y), & \text{if } l_B(y) > 0; \\ l_A'(y)/l_B'(y), & \text{if } l_B(y) = l_A(y) = 0. \end{cases}$$

Note that since the sets A, B are convex, $l_A'(y) \le l_B'(y)$ whenever $l_A(y) = l_B(y) = 0$.

Case I(b): $\kappa = K$. As before, pick $\eta > 0$ and note that the limit in (3.15) can, as in (3.17), be determined by only considering the limit of the ratio of the terms

$$F_Q^\eta(N) \stackrel{\text{def}}{=} \sum_{\{n \in Q^N : h(n) \ge y-\eta\}} e^{Nh(n)}, \qquad Q = A, B.$$

Recall the definition of the sets $\mathcal{C}_Q^N$ from (3.16) and define for $\epsilon > 0$

$$\mathcal{I}_k^{Q,N} \stackrel{\text{def}}{=} \mathcal{C}_Q^N \cap \left(y - \frac{(k+1)\epsilon}{N}, y - \frac{k\epsilon}{N}\right], \qquad k = 0, \ldots, \left[\frac{\eta N}{\epsilon}\right].$$

Then, one obtains the bounds

$$e^{-\epsilon} \sum_k e^{N(y-k\epsilon/N)} \left|\mathcal{I}_k^{Q,N}\right| \le F_Q^\eta(N) \le \sum_k e^{N(y-k\epsilon/N)} \left|\mathcal{I}_k^{Q,N}\right|, Q = A, B.$$

Since ϵ is chosen arbitrarily, it suffices to consider the limit

$$\lim_{N\to\infty} \frac{\sum_k e^{N(y-k\epsilon/N)} \left|\mathcal{I}_k^{A,N}\right|}{\sum_k e^{N(y-k\epsilon/N)} \left|\mathcal{I}_k^{B,N}\right|}. \tag{3.19}$$

The counterpart of Lemma 7 now reads,

Lemma 9. $\left|\mathcal{I}_k^{Q,N}\right| = \epsilon\, l_Q(y - k\epsilon/N) N^{K-1} + \tilde{r}_{N,k}^Q$, where $\lim_{N\to\infty} \sup_k \left|\tilde{r}_{N,k}^Q\right|/N^{K-1} = 0.$

From this lemma and (3.19) the statement of Lemma 8 holds again. Observe that this method cannot be applied when $\kappa = 1$ since some of the sets $\mathcal{I}_k^{Q,N}$ are empty for ϵ small enough.

Case II: $y_A < y_B$. This can be handled by similar but easier means without requiring condition (C) above.

Lemma 9. If $y_A < y_B$ then $\lim_{N\to\infty} F_A(N)/F_B(N) = 0$.

4. A tandem queue example

In this section, we apply the results of Sections 2 and 3 to a simple Jackson network which serves to illustrate how the methodology can be used to determine the way in which large queue lengths build up. The example is a two queue tandem network with arrival rate λ, and service rates μ_1 and μ_2. We denote this as a (λ, μ_1, μ_2) system. The reverse time process is also a two queue tandem network with arrival rate λ, and service rates μ_1 and μ_2. The forward and reverse time networks are shown in Figure 1; note that arrivals enter queue 1 in the forward time network, but they enter queue 2 in the reverse time network. As in Section 3, let

$$Z^N(t) = (Z_1(Nt)/N, Z_2(Nt)/N),$$

where $Z_i(t)$ is the queue length at server i at time t. We consider several variations of rare events associated with this network. (In the discussion that follows, some of the technical details that are required to apply the results of Sections 2 and 3.5 are omitted in the interest of clarity.)

We first consider the event $Z_1^N(0) + Z_2^N(0) \geq 1$, where the total network population first exceeds N. Parekh and Walrand [*PW*] present a heuristic action functional and show that if $\lambda < \mu_1 \leq \mu_2$, then as $N \to \infty$ the minimizing path has queue 1 building up at rate $\mu_1 - \lambda$, with queue 2 remaining stable. To estimate the probability of this event, this result suggests using the importance sampling distribution corresponding to an unstable (μ_1, λ, μ_2) system. (Similarly, if $\lambda < \mu_2 \leq \mu_1$, queue 1 remains stable and queue 2 builds up at rate $\mu_2 - \lambda$, which corresponds to using the (μ_2, μ_1, λ) system for importance sampling.) This heuristic result has been extended to $n > 2$ queues in tandem in [*FA1*].

In our framework, an analogous result is simply obtained as follows. Let $\rho_i = \lambda/\mu_i$, $i = 1, 2$, and consider the optimization problem

$$\max\left\{(1-\rho_1)\rho_1^{Nz_1}(1-\rho_2)\rho_2^{Nz_2} \ : \ z_1 + z_2 \geq 1\right\},$$

where $z_i = k_i/N$ for some integer k_i. This problem corresponds to maximizing the steady-state distribution over the set $B = \{(z_1, z_2) : z_1 + z_2 \geq 1\}$. In the limit, as $N \to \infty$, for $\rho_1 < \rho_2$ the optimum occurs at $z^* = (0, 1)$. In addition, if $(z_1(N), z_2(N)) = (k_1(N)/N, k_2(N)/N) \to (z_1, z_2) \neq (0, 1)$, then $\pi^N(z_1(N), z_2(N))/\pi^N(0, 1) \to 0$. Thus, the limiting hitting distribution is $\mu_B(A) = 1\{(0, 1) \in A\}$. From Figure 1, it is clear that starting in the state $z^N(0) = (0, 1)$, the reverse time process has queue 2 departure rate μ_2 and arrival rate λ, and thus its rate of change is $\lambda - \mu_2$ (until queue 2 empties out). In reverse time, queue 1 has arrival rate μ_2 and service rate μ_1. (The departure rate of queue 1 is μ_2 since $\mu_2 < \mu_1$.) Since arrivals and departures exchange meanings when one switches from reverse to forward time, the build-up path of the forward time fluid limit of Lemma 1 corresponds to the (μ_2, μ_1, λ) system. If $\rho_2 < \rho_1$ we would similarly obtain the (μ_1, λ, μ_2) system.

If $\rho_1 = \rho_2$, the results of [*PW*] and [*FLA*] fail to predict the existence of a continuum of exit paths. In this case, $\pi^N(z_1(N), z_2(N))$ is constant over the set

$$B^N \stackrel{\text{def}}{=} \{(z_1(N), z_2(N)) = (k_1/N, k_2/N) \ : \ k_1 + k_2 = N\}$$

and so the limiting hitting distribution is uniform on $B = \{(z_1, z_2) : z_1 + z_2 = 1\}$ in which B^N becomes dense as $N \to \infty$. Now given that $Z^N(0) = (z_1, z_2)$ where both z_1 and z_2 are positive, the reverse time

process has queue 2 arrival rate λ, queue 2 service rate μ ($\mu = \mu_1 = \mu_2$), and queue 1 service rate μ. Thus the queue 2 rate of change is $\lambda - \mu$ and the queue 1 rate of change is 0 until queue 2 empties out. At this time, queue 1 begins emptying out at rate $\lambda - \mu$. Thus in forward time, the build-up path corresponds to a (μ, λ, μ) system until queue 1 reaches a level NU where U is uniformly distributed on (0,1). The build-up path then corresponds to a (μ, μ, λ) system until queue 2 reaches the $N(1-U)$ level. A typical (limiting) sample path is illustrated in Figure 2.

We next consider the case when both queue lengths are large but the utilizations are unequal. We first consider the event B^N where $B = \{w \in \mathbf{R}_+^2 : w_1 \geq z_1,\ w_2 \geq z_2\}$. where both $z_1 > 0$, $z_2 > 0$. From the results of the previous section π^N asymptotically concentrates on the point (z_1, z_2). We can therefore assume that $\|Z^N(0) - (z_1, z_2)\| < \epsilon$ for some small ϵ.) If queue 2 is the bottleneck ($\rho_1 < \rho_2$), then, in reverse time, queue 2 initially empties at rate $\lambda - \mu_2$ and queue 1 empties at rate $\mu_2 - \mu_1$. This occurs until one of the queues goes empty. Note that queue 1 will empty first if $z_1/|\mu_2 - \mu_1| < z_2/|\lambda - \mu_2|$. If queue 1 empties first (case A), then queue 2 continues to empty at rate $\lambda - \mu_2$. If queue 2 empties first (case B), then queue 1 subsequently empties at rate $\lambda - \mu_1$. Figure 3 shows the build-up paths for the forward time process. In case A, the build-up path corresponds initially to a (μ_2, μ_1, λ) system, in which queue 2 is unstable, while in case B, the initial build-up path corresponds to a (μ_1, λ, μ_2) system, where queue 1 is unstable. In the initial phase, only one of the queues is unstable. In both cases, the build-up path then changes to a (μ_1, μ_2, λ) system in which both queues are unstable.

If, instead, queue 1 is the bottleneck ($\lambda < \mu_1 < \mu_2$), then we get quite different build-up paths as indicated in Figure 4. In this case, in reverse time, queue 2 empties at rate $\lambda - \mu_2$, but queue 1 increases at rate $\mu_2 - \mu_1$ (until queue 2 is empty). Subsequently, queue 1 empties at rate $\lambda - \mu_1$. Thus, in forward time, the build-up path has a phase corresponding to a (μ_1, λ, μ_2) system in which queue 1 is unstable, followed by a phase corresponding to a (μ_1, μ_2, λ) system in which queue 1 is stable but queue 2 is unstable. Notice that in this case, during phase 1 queue 1 actually overshoots its final target value Nz_1.

5. Simulation Results

In this section, we apply the results of the previous sections to simulations of rare events in the two queue tandem Jackson network. Specifically, we use importance sampling to follow the fluid limit path corresponding to the build-up of large queue lengths.

For any set A, let τ_A be time of first entering the set A, and let 0 denote the state when both queue lengths are equal to 0. Consider the rare event $B^N = \{(Z_1, Z_2) : Z_1 \geq N_1, Z_2 \geq N_2\}$. In estimating the mean time until B^N occurs (given that the process starts in 0), it is of interest to estimate $\alpha = P\{\tau_{B^N} < \tau_0 | Z(0) = 0\}$ (see [*Kei*] or [*PW*]). For an arbitrary time t, if we know that $Z_1(t) \geq N_1$ and $Z_2(t) \geq N_2$, then by the previous results, we know the asymptotic build-up path. This suggests that to estimate α, one should use the appropriate importance sampling change of measure in an attempt to follow this build-up path. Referring to Figure 3, if queue 2 is the bottleneck and the parameters are such that Case A applies, then this suggests using (i.e., simulating with) the importance sampling distribution corresponding to a (μ_2, μ_1, λ) system until queue 2 reaches the appropriate level, and then switching the importance sampling distribution to correspond to a (μ_1, μ_2, λ) system until B^N is hit. Similarly, in Case B (or when queue 1 is the bottleneck - see Figure 4), the initial importance sampling distribution corresponds to a (μ_1, λ, μ_2) system until queue 1 reaches the appropriate level, followed by a phase corresponding to a (μ_1, μ_2, λ) system until B^N is hit.

We wrote a simple simulator suitable for testing the effectiveness of these fluid limit importance sampling distributions for estimating α. As in [*PW*]), we simulated the embedded discrete time Markov chain rather than the CTMC. This typically results in a variance reduction (see [*HLS*]). (Although we have observed this in practice, we have not proved it in the current setting.) We used the combined generator described in in [*Lec*] as a source of randomness. We also wrote a program to numerically compute α which allows comparison

of simulation results to their analytical (i.e., numerical) values. Computation of α involves solving a system of linear equations (see [*Chu*]). In order to keep the number of equations finite, we also truncated the state space at a maximum of 40 customers per queue. This truncation point was large enough relative to the other parameters chosen so as to make the truncation error negligible.

We ran experiments with five sets of parameters. The parameter settings for each experiment are listed in Table 1, along with the computed value of α. Queue 2 is the bottleneck in experiments I - IV, and queue 1 is the bottleneck in experiment set V. Experiments I and II correspond to Case B, with queue 1 building up first, while experiments III and IV correspond to Case A. With these parameter settings, α ranges from 10^{-7} to 10^{-12}.

For each parameter setting, we ran $M = 1,000,000$ replications, with each replication corresponding to a simulation starting in state 0 and continuing until either 0 or B^N is hit. For replication r, let L_r denote the likelihood ratio (i.e., the probability of the sample path with the original parameters divided by the probability of the sample path with the importance sampling parameters). Let $I_r = 1$ if the process hits B^N before 0, and 0 otherwise. Then $E(L_r I_r) = \alpha$, where the expectation is understood to be with respect to the importance sampling distribution. We estimate α by

$$\bar{\alpha} \stackrel{\text{def}}{=} \frac{1}{M} \sum_{r=1}^{M} L_r I_r. \tag{5.1}$$

We computed the relative error, $(\bar{\alpha} - \alpha)/\alpha$, along with an estimate of its standard deviation, $\hat{\sigma}_1$. Note that $E((\bar{\alpha} - \alpha)/\alpha) = 0$. The results of the simulations are listed in Table 2.

In order to compare the efficiency of importance sampling relative to standard simulation, we computed the variance improvement ratio, which is also listed in Table 2. The variance improvement ratio is defined as follows. Suppose one uses standard simulation to estimate α. After M' replications, the estimator would be

$$\hat{\alpha} \stackrel{\text{def}}{=} \frac{1}{M'} \sum_{r=1}^{M'} I_r. \tag{5.2}$$

The relative error using standard simulation is therefore $(\hat{\alpha} - \alpha)/\alpha$ which has variance $\sigma_2^2 = (1 - \alpha)/(M'\alpha)$. Now if M and M' are chosen so that the same expected number of transitions are simulated using both standard simulation and importance sampling, then the variance improvement ratio

$$V \stackrel{\text{def}}{=} \frac{\sigma_2^2}{\hat{\sigma}_1^2} \tag{5.3}$$

is an estimate of the variance ratio obtained using standard simulation to that obtained using importance sampling for (approximately) an equal amount of computer work. Our simulation program counted the total number of transitions, and our numerical program also computed the expected number of transitions per replication using standard simulation. Thus, for a given M, we were able to estimate M'.

The results listed in Table 2 show variance improvement ratios ranging from 57 to over 10^6. The largest variance improvements are for experiments I and III, in which α is the smallest. The smallest variance improvement is for experiment II. In this experiment, we have observed sample paths that wander rather far from the piece-wise linear exit path that one obtains as the center of a fluid limit tube by applying the asymptotics. This results in increased variability in the likelihood ratio.

We note that estimating this type of rare event can be more difficult than estimating α for sets B^N of the form $B^N = \{(Z_1, Z_2) : Z_1 + Z_2 = N\}$, which was one of the problems considered in [*PW*]. When we tested our simulator on problems of this type, we obtained comparable results to those listed in Table V of [*PW*].

6. Summary

In this paper we have described a simple relationship between time reversal and rare events in continuous time Markov chains. For CTMCs for which the time reversed process is known (e.g., product-form networks), this relationship presents a simple alternative to the large deviations approach of describing how rare events occur. Assuming that both methods can be applied, our approach gives less information than large deviations since large deviations gives not only the path, but also the exponential decay rate describing the probability of the rare event. Note, however, that the results in Section 8 of [*Kel*] also imply that the time until the rare event occurs converges to an exponential distribution.

The mean of this exponential distribution can be estimated by simulation using importance sampling. The results of this paper immediately suggest that the importance sampling distribution should be chosen so as to follow the fluid limit path. Note that in the example of Section 5, this involves changing the importance sampling distributions for the service and interarrival times at certain instances as the queue lengths increase. We are currently studying further the practical effectiveness of following these fluid limit paths. This, the optimality (in terms of simulation variance) and questions on the relation between our results and those cited in the Introduction remain topics for further research.

7. References

[Bi1] Billingsley, P. (1968) *Convergence of Probability Measures.* Wiley, New York.

[Bi2] Billingsley, P. (1979) *Probability and Measure.* Wiley, New York.

[Bu] Bucklew, J.A. (1990) *Large Deviation Techniques in Decision, Simulation, and Estimation.* Wiley, New York.

[CM] Chen, H. and Mandelbaum, A. (1988) Discrete flow networks: bottleneck analysis and fluid approximations. *Preprint.*

[Chu] Chung, K.L. (1967) *Markov Chains With Stationary Transition Probabilities*, Second Edition. Springer-Verlag, New York.

[Co] Copson, E.T. (1965) *Asymptotic Expansions.* Cambridge University Press.

[CFM] Cottrell, M., Fort, J.-C. and Malgouyres, G. (1983) Large deviations and rare events in the study of stochastic algorithms. *IEEE Trans. Aut. Cont.* 28, 9, 907-920.

[DEW] Dupuis, P., Ellis, R. and Weiss, A. (1989) Large deviations for Markov processes with discontinuous statistics, I: general upper bounds. *Preprint.* Center for Appl. Math. and Math. Comp., University of Massachusetts-Amherst.

[DIM] Dupuis, P., Ishii, H. and Meté, S. (1990) A viscosity solution approach to the asymptotic analysis of queueing systems. *Ann. of Prob.* 18, 226-255.

[FA1] Frater, M.A. and Anderson, B.D.O. (1989) Fast estimation of the statistics of excessive backlogs in tandem networks of queues. *Australian Telecommunications Research* 23, 1, 49-55.

[FA2] Frater, M.A. and Anderson, B.D.O. (1989) Queueing systems, optimal control, reverse-time modeling and large deviations. *Preprint.*

[FLA] Frater, M.A., Lennon, T.M. and Anderson, B.D.O. (1989) Optimally efficient estimation of the statistics of rare events in queueing networks. *Preprint.*

[GM] Goodman, J.B. and Massey, W.A. (1984) The non-ergodic Jackson network. *J. Appl. Prob.* 21, 860-869.

[HR] Harrison, J.M. and Reiman, M.I. (1981) Reflected brownian motion on an orthant. *Ann. of Prob.* 9, 302-308.

[HIS] Hordijk, A., Iglehart, D.L. and Schassberger, R. (1976) Discrete-time methods of simulating continuous time Markov chains. *Adv. Appl. Prob.* 8, 772-788.

[JS] Jacod, J. and Shiryaev, A.N. (1987) *Limit Theorems for Stochastic Processes.* Springer-Verlag, New York.

[Kei] Keilson, J. (1979) *Markov Chain Models - Rarity and Exponentiality.* Springer-Verlag, New York.

[Ke1] Kelly, F.P. (1979) *Reversibility and Stochastic Networks.* Wiley, Chichester.

[Ke2] Kelly, F.P. (1985) Blocking probabilities in large circuit-switched networks. *Adv. Appl. Prob.* 18, 473-505.

[Ku] Kurtz, T.G. (1986) *Approximation of Population Processes.* CBMS-NSF Regional Conf. Series in Appl. Math. 36. SIAM, Philadelphia.

[Lec] L'Ecuyer, P. (1988) Efficient and portable combined random number generators. *Communications of the ACM* 31, 6, 742-749, and 774.

[PW] Parekh, S. and Walrand, J. (1989) A quick simulation method for excessive backlogs in networks of queues. *IEEE Trans. Aut. Cont.* 34, 1, 54-66.

[Ts] Tsoucas, P. (1990) Rare events in series of queues. *J. of Appl. Prob.* To appear.

[Va] Varadhan, S.R.S. (1984) *Large Deviations and Applications.* CBMS-NSF Conf. Series on Appl. Math. 46. SIAM, Philadelphia.

Experiment	ρ_1	ρ_2	N_1	N_2	α
I	0.25	0.30	10	10	4.32×10^{-11}
II	0.375	0.45	10	10	9.86×10^{-8}
III	0.25	0.33	6	14	3.80×10^{-10}
IV	0.375	0.50	6	14	8.82×10^{-7}
V	0.45	0.375	10	10	9.86×10^{-8}

Table 1. Parameters for simulations in two queue tandem network.

Experiment	$(\frac{\hat{\alpha}-\alpha}{\alpha})$	Std. Dev. $(\frac{\hat{\alpha}-\alpha}{\alpha})$	Variance Improvement
I	0.084	0.105	2.1×10^5
II	0.076	0.152	5.7×10^1
III	-0.007	0.009	3.5×10^6
IV	-0.004	0.013	1.1×10^3
V	0.067	0.073	2.3×10^2

Table 2. Simulation results for two queue tandem network based on 1,000,000 replications.

Figure 1

Two Queues in Tandem: Forward and Reverse Time Networks

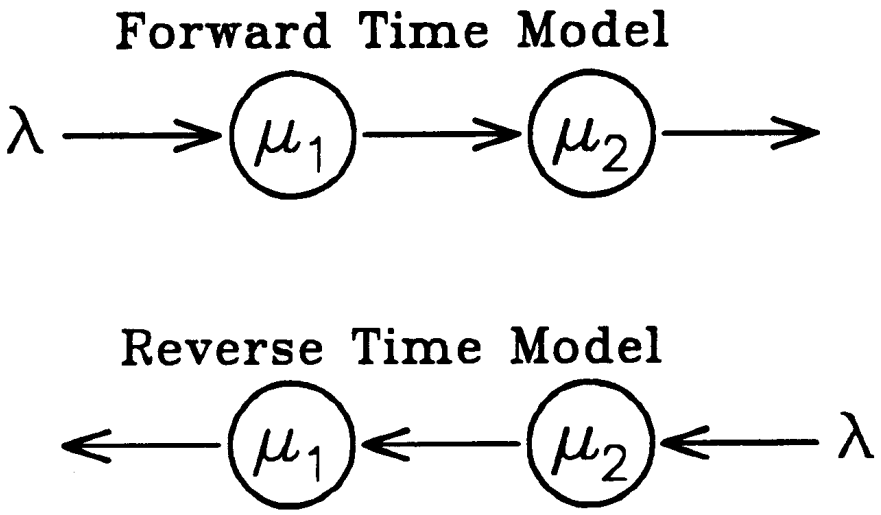

Figure 2

Typical System Build-up Path in the Two Queue Tandem Network with Equal Service Rates

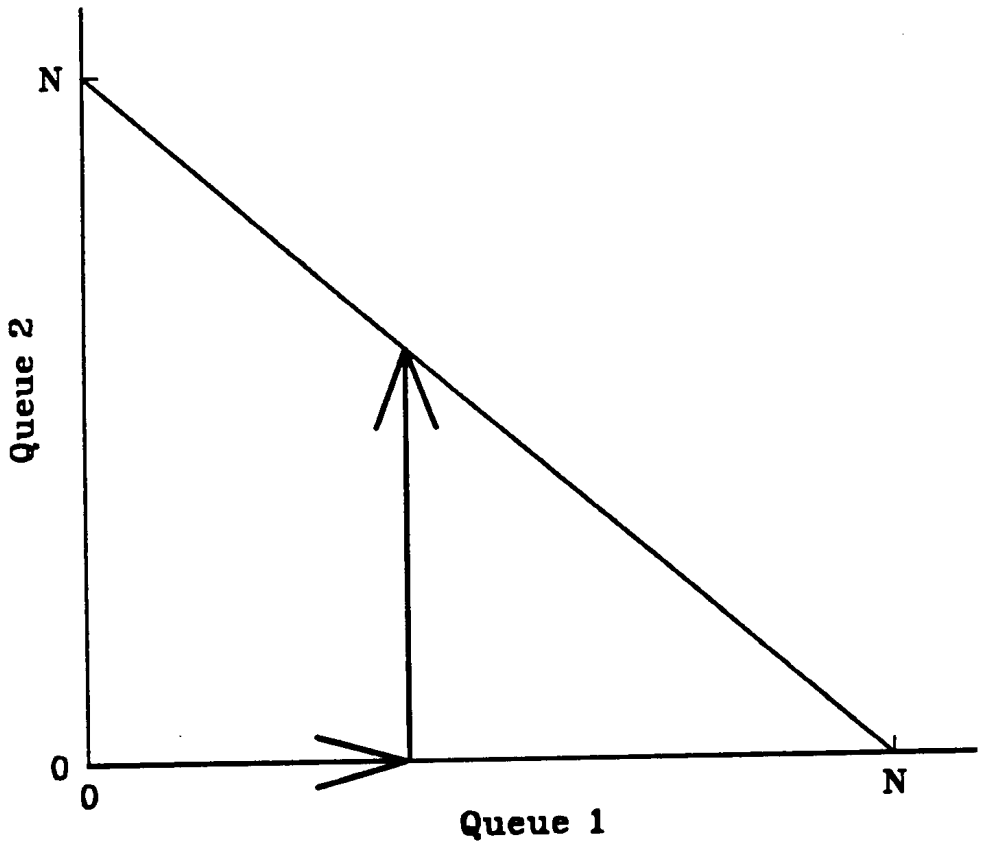

Figure 3

Build-up Paths in the (λ, μ_1, μ_2) Two Queue Tandem Network
in Which Queue 2 is the Bottleneck $(\lambda < \mu_2 < \mu_1)$

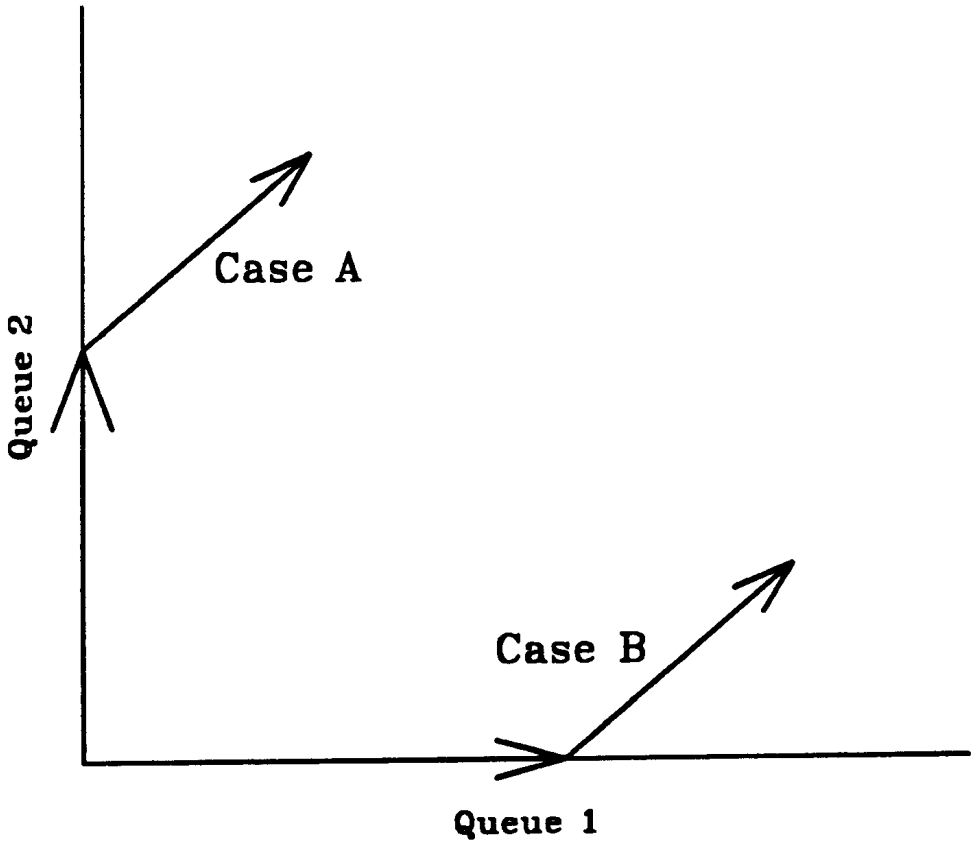

Figure 4

Build-up Paths in the (λ, μ_1, μ_2) Two Queue Tandem Network
in Which Queue 1 is the Bottleneck $(\lambda < \mu_1 < \mu_2)$