A Quick Simulation Method for Excessive Backlogs in Networks of Queues

SHYAM PAREKH, MEMBER, IEEE, AND JEAN WALRAND, MEMBER, IEEE

Abstract—We consider stable open Jackson networks and study the rare events of excessive backlogs. Although these events occur rarely, they can be critical, since they can impair the functioning of the network. We attempt to estimate the probabilities of these events by simulations. Since a direct simulation of a rare event takes a very long time, this procedure is very costly. Instead, we devise a method for changing the network to speed up simulations of rare events. We try to pursue this idea with the help of large deviation theory. This approach, under certain assumptions, results in a system of differential equations which may be difficult to solve. To circumvent this, we develop a heuristic method which gives the rule for changing the network for the purpose of simulations. We illustrate, by examples, that our method of simulation can be several orders of magnitude faster than direct simulations.

I. INTRODUCTION

A. Problem Description

We consider arbitrary open Jackson networks. A Jackson network is an interconnection of M/M/1 queues in which customers visit various nodes according to state and time independent (Markovian) routing probabilities. The heuristic that will be developed can be applied to networks of GI/GI/1 queues with Markovian routing. However, most of the discussion will be limited to the case of Jackson networks, since in this case it is easier to check our simulation results by numerical methods. A network is called open, if every arrival customer leaves the system with probability 1. Let us define T as the first time that the total population in the network reaches N. We are interested in estimating E[T], where E[T] denotes the expected value of T given that the system starts empty. Notice that we are interested in the transient behavior of the system.

Since very little is known about the transient behavior of networks, we will attempt to estimate E[T] by efficient simulations. Our method of simulation, besides saving simulation time, also sheds some light on the fundamentals of the dynamics of the system.

B. Principle (Importance Sampling)

For a stable system, the events of reaching a large total backlog are very infrequent. Hence, direct simulations are very slow and take up a lot of computer time. Besides, there is also the difficulty of implementing a pseudorandom generator that can function effectively during very long simulations. The central idea is to make the rare events under investigation more frequent by changing appropriately the probability measures governing the system and performing simulations on the changed system. We then obtain our answers by translating them back to the original system. This is done by using likelihood ratios.

C. Optimal Change of Measure (Largest Speed-Up)

Large deviation theory deals with certain Markov processes and determines the asymptotic (e.g., as the backlog size N grows for an M/M/1 queue, see Section III) exponential rate of diminishing probabilities as a solution of a variational problem. The solution of this variational problem also gives the optimal exponential change of measure (see Section III) for simulations. Unfortunately, the theory does not apply to general Jackson networks. Here a smoothness condition regarding the jump distributions (see Section III) is violated. To our knowledge, there are no known results of large deviation theory for excursions of Markov processes with discontinuous kernels which can be directly applied to the backlog process of a Jackson network. (For some partial results in that direction, see Weiss [15].) To circumvent this problem, we are going to rely on a heuristic of Borovkov, Rudin, et al. (e.g., see [9]) which gives certain tail probabilities for a GI/GI/1 queue (see Section IV). We utilize this heuristic for obtaining a change of measure that leads to substantial speed-up for simulations. We also generalize this heuristic to networks.

D. Outline of the Remaining Sections

In Section II, we motivate the idea of change of measure for simulations of certain rare events of an M/M/1 queue. In Section III, we present a few results of large deviation theory which are useful for simulations of rare events. We also point out the difficulties in applying this theory to general Jackson networks. In Section IV, we present a heuristic method for obtaining an optimal change of measure for simulations of rare events for Jackson networks. Next we extend this heuristic to networks of GI/GI/1 queues. Our hope is that the heuristic explanation and observations presented here will motivate more research in this area. Finally, we will summarize the results of this paper in Section V.

E. Some Relevant Contributions

Cottrell et al. [3] have recently illustrated the use of large deviation theory for simulations of rare events. In particular, they consider rare events for the Aloha protocol. The key large deviation theorems for this kind of applications are due to Azenkott et al. [1] and Ventsel [12]. More applications of large deviation theory can be found in works of Dupuis et al. [4], Weiss [14], etc. A good reference for the fundamental results of large deviation theory is the succinct monograph of Varadhan [11]. Some recent large deviation results for the empirical distributions of Markov chains are due to Ellis [5] and [6] and Natarajan [7].

II. M/M/1 EXAMPLE

A. Model and Problem

Consider an M/M/1 queue with arrival rate \( \lambda \) and service rate \( \mu \) such that \( \lambda < \mu \). Consider the embedded discrete-time Markov
chain \( \{X_m, m = 0, 1, 2, \cdots \} \) of the queue length at the epochs of arrivals and departures of the queue. We assume, without any loss of generality, \( \lambda + \mu = 1 \) (otherwise, we can rescale time). Fig. 1 depicts such a queue.

As described in Section I-A, we are interested in estimating, for large \( N \), \( E_0 \{T\} \) where \( T \) denotes the first time \( \{X_m\} \) reaches \( N \). Note that the number of times \( \{X_m\} \) returns to 0 before hitting \( N \) is geometrically distributed with parameter \( 1 - \alpha \), where \( \alpha \) is the probability that \( \{X_m\} \) reaches \( N \) before returning to 0 given that it starts from 0. For large \( N \), we can argue that
\[
E_0 \{T\} = \frac{1 - \alpha}{\alpha} \cdot E_0 \{T_0\} = \frac{1}{\alpha} \cdot E_0 \{T_0\}
\]
where \( T_0 \) denotes the time to hit 0 for the first time. Since, for stable systems, \( E_0 \{T_0\} \) can be easily estimated by direct simulations, the difficult part in estimating \( E_0 \{T\} \) is the estimation of \( \alpha \). So, from now on, our primary concern will be the estimation of \( \alpha \).

We define a cycle as the duration starting with an empty system and ending at the instant the system, for the first time, either becomes empty again or reaches \( N \). Let us define
\[
V_i := \mathbbm{1}\{X_m \text{ reaches } N \text{ in cycle } k\}
\]
where \( \mathbbm{1}\{B\} \) (or sometimes written \( \mathbbm{1}_B \)) denotes the indicator of an event \( B \). See Fig. 2. Notice that \( V_i \)'s are i.i.d. Also notice that, as shown in Fig. 2, we have modified \( \{X_m\} \) so that we restart \( \{X_m\} \) at 0, if it exceeds \( N \). Clearly, \( \alpha = P\{V_1 = 1\} \). Here we can find \( \alpha \) by the first step method. For this let, for \( 0 \leq i \leq N, P_i \) denote the probability that \( \{X_m\} \) hits \( N \) before 0 given that it starts from \( i \). Clearly, \( P_0 = 0, P_N = 1, \) and \( P_1 = \alpha \). The first step equations give
\[
P_i = \mu \cdot P_{i-1} + \lambda \cdot P_{i+1}, \quad 1 \leq i \leq N - 1.
\]
The solution of these linear equations can be seen to give
\[
\alpha = P_1 = \frac{\frac{\mu}{\lambda} - 1}{\left(\frac{\mu}{\lambda}\right)^N - 1}.
\]

For future calculations, let us derive the formula for \( E\{J_k\} \), where \( J_k \) denotes the number of random jumps in cycle \( k \). Notice that \( J_k \)'s are i.i.d. and that a cycle begins with a deterministic transition to 1. Let \( Z_1 \) denote a jump which takes values +1 and \( -1 \) w.p. \( \lambda \) and \( \mu \), respectively. Note that cycle \( k \) ends at \( N \) with probability \( \alpha \) and in this case \( Z_1 + Z_2 + \cdots + Z_k = N - 1 \). Similarly, cycle \( k \) ends at \( 0 \) with probability \( 1 - \alpha \) and in this case \( Z_1 + Z_2 + \cdots + Z_k = 1 - \alpha \).

Using Wald's identity we identify the left-hand side with
\[
E\{J_k\} \cdot E\{Z_k\} = E\{J_k\} \cdot (\lambda - \mu).
\]
This gives
\[
E\{J_k\} = \frac{1 - N \cdot \alpha}{\mu - \lambda}.
\]

In the following subsections we present the idea of change of measure for estimating \( \alpha \) by simulation.

### B. Direct Simulation

For direct Monte Carlo simulation, consider an unbiased and convergent estimator
\[
\alpha_n := \frac{V_1 + V_2 + \cdots + V_n}{n}.
\]
Observe that \( E\{V_i\} = \alpha \) and \( \text{Var}\{V_i\} = \alpha(1 - \alpha) \). Suppose we want to ensure that the relative error does not exceed \( \epsilon \) with probability more than \( \beta \). We will call such an estimator an \((\epsilon, \beta)\)-confidence estimator. The normal approximation then gives
\[
P\{||\alpha_n - \alpha| > \epsilon \cdot \alpha\} = \beta = n_\epsilon \frac{\epsilon^2}{\alpha^2} \cdot \text{Var}\{V_i\}
\]
where \( c = \Phi^{-1}(\beta/2) \), where \( \Phi \) denotes the distribution function of a Gaussian r.v. with the mean equal to 0 and variance equal to 1. Hence, \( n_\epsilon = \gamma(1 - \alpha)/\epsilon^2 \), where \( \gamma = c^2/\epsilon^2 \), cycles are necessary to achieve the \((\epsilon, \beta)\)-confidence estimator by a direct simulation. Let \( T_d \) denote the units of simulation time required for achieving the \((\epsilon, \beta)\)-confidence estimator by a direct simulation. Then,
\[
T_d = E\{J_k\} \cdot n_\epsilon.
\]
Since \( \lambda < \mu \), for large \( N \), \( E\{J_k\} = 1/\mu - \lambda \) [see (2)]. Hence,
\[
T_d = \gamma \cdot \frac{1}{\alpha} \cdot \frac{1}{\mu - \lambda}.
\]

### C. Change of Measure

For estimating \( \alpha \), we propose to consider the \( M/M/1 \) queue with arrival rate \( \mu \) and service rate \( \lambda \), i.e., the \( M/M/1 \) queue obtained by interchanging arrival rate and service rate of the original queue. Let \( P \) and \( P^* \) denote the measures induced by the corresponding Markov chains. Fig. 3 shows these queues.

In simulations under the changed measure, we observe \( V_i \)'s under \( P^* \). Let \( L_k \) denote the likelihood ratio \( dP/dP^* \) during cycle \( k \). Notice that \( L_k \)'s are i.i.d. and that \( E^*\{L_k \cdot V_k\} = E\{V_k\} = \alpha \), where \( E^*\{\cdot\} \) denotes the expectation under the measure \( P^* \).

Hence,
\[
\alpha^*_n := \frac{L_1 \cdot V_1 + L_2 \cdot V_2 + \cdots + L_n \cdot V_n}{n}
\]
is also an unbiased and convergent estimator of \( \alpha \). As before, to achieve \((\epsilon, \beta)\)-confidence estimator, now the minimum number of cycles required will be
\[
n_\epsilon = \frac{\gamma}{\alpha^2} \cdot \text{Var}\{L_k \cdot V_k\}
\]
where \( \text{Var}\{ \cdot \} \) denotes the variance under the measure \( P^\ast \).

Observe that by interchanging \( \lambda \) and \( \mu \) in (2), we have \( E^*\{ J_k \} = N/\mu \). Let \( T_c \) denote the units of simulation time required for achieving the \((\epsilon, \beta)\)-confidence estimator under the changed measure. Then

\[
T_c = E^*\{ J_k \} \cdot \eta = \frac{\sigma^2}{\alpha^2} \cdot \frac{N}{\lambda} \cdot \frac{\lambda}{\mu}
\]

where \( \sigma^2 = \text{Var}\{ L_k \cdot V_k \} \).

We should point out that in reality the simulation time will be somewhat larger like \((1 + \delta) \cdot T_c\), where \( \delta > 0 \) accounts for the time required to calculate likelihood ratios \( L_k \).

\section{Comparison of \( T_d \) and \( T_c \)}

Let us define the speed-up factor \( S := T_d/T_c \). From (3) and (4) we get

\[
S = \frac{1}{\alpha} \cdot \frac{\lambda}{\sigma^2} \cdot \frac{1}{1 - \lambda/\mu}
\]

Suppose that \( \omega \) is a realization such that \( V_k = 1 \) and there are \( l \) departures and \( N + l - 1 \) arrivals (not counting the first arrival) during cycle \( k \). So, \( J_k(\omega) = N + 2l - 1 \). Let \( \omega_k \) denote the section of \( \omega \) that pertains to cycle \( k \). Then, \( P\{ \omega_k \} = \lambda^{N+2l-1} \cdot \mu^l \) and \( P^\ast\{ \omega_k \} = \mu^{N+l-1} \cdot \lambda^l \). Therefore,

\[
L_k(\omega_k) = \left( \frac{\lambda}{\mu} \right)^{N-1}.
\]

This implies that, on the set \{ \( V_k = 1 \) \}

\[
L_k = \left( \frac{\lambda}{\mu} \right)^{N-1}.
\]

Hence,

\[
\sigma^2 = E^*\{(L_k \cdot V_k)^2\} - \alpha^2
= \left( \frac{\lambda}{\mu} \right)^{N-1} \cdot E^*\{L_k \cdot V_k\} - \alpha^2
= \left( \frac{\lambda}{\mu} \right)^{N-1} \cdot \alpha - \alpha^2
\]

where the second equality follows from (6). Now using (1), we get

\[
\frac{\sigma^2}{\alpha} = \left( \frac{\lambda}{\mu} \right)^N.
\]

Substituting (7) in (5), we get

\[
S = \left[ N \cdot \left( \frac{\lambda}{\mu} \right)^N \cdot \left( 1 - \frac{\lambda}{\mu} \right) \right]^{-1}.
\]

\section{Example}

Consider the \( M/M/1 \) queue with \( \lambda = 0.33 \) and \( \mu = 0.67 \). We want to estimate \( \alpha \) for \( N = 21 \). Equation (1) gives \( \alpha = 3.583 \times 10^{-1} \). For the \((\epsilon, \beta)\)-confidence estimator, (3) gives \( T_c = 1.32 \times 10^{10} \) units (4.42 \times 10^9 cycles), while (4) gives \( T_c = 4.96 \times 10^9 \) units (1.58 \times 10^9 cycles). Table I gives some simulation results for this example.

Table II gives results of a few more simulation experiments. It also shows the time required for a simulation and the corresponding number of calls to the random number generator (RNG). Table III gives the empirical standard deviations, means, and coefficients of variation of the results obtained by the change of measure for the same examples as in Table II. All the simulations were done on a VAX-750 machine. Notice that the convergence under the changed measure seems to be more rapid than predicted by (4). This is due to the uncertainty factor introduced in the derivation of (4) because of the use of the normal approximation.

\section{III. Large Deviation Theory and Optimal Change of Measure}

\subsection{A. Fundamental Theorem}

\textbf{Theorem 1 (Cramér's Theorem) [11]:} Let \( \xi_1, \xi_2, \cdots \) be i.i.d. r.v.'s taking values in \( R^d \). Let \( F \) denote the distribution function (d.f.) of \( \xi_1 \) and \( m \) its mean. Let \( P_n \) denote the d.f. of \( (\xi_1 + \xi_2 + \cdots + \xi_n)/n \). We assume that the Laplace transform of \( F \)

\[
M(s) := \int \exp \{s \cdot z\} dF(z), \quad z \in R^d
\]

is finite in a neighborhood of 0. Then, \( P_n \) satisfies the following:

\( i \) for each closed subset \( C \) of \( R^d \)

\[
\limsup_{n \to \infty} \frac{1}{n} \cdot \log P_n(C) \leq -\inf_{x \in C} h(x)
\]
TABLE III

EMPIRICAL STANDARD DEVIATION FOR AN M/M/1 QUEUE

<table>
<thead>
<tr>
<th>Example</th>
<th>( \lambda = 0.20 )</th>
<th>( \mu = 0.80 )</th>
<th>( N = 15 )</th>
<th>( \alpha = 2.794 \times 10^{-4} )</th>
<th># of Experiments = 20</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \lambda = 0.20 )</td>
<td>( \mu = 0.80 )</td>
<td>( N = 15 )</td>
<td>( \alpha = 2.794 \times 10^{-4} )</td>
<td># of Experiments = 20</td>
<td></td>
</tr>
<tr>
<td># of Cycles (a)</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td># of Cycles (b)</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

\[ h(u) = \frac{1 + u}{2} \cdot \log \left( \frac{1 + u}{2 - \mu} \right) \cdot \frac{1 - u}{2} \cdot \log \left( \frac{1 - u}{2 - \mu} \right), \quad -1 \leq u \leq 1, \quad \text{if } \mu < 1, \]
\[ \text{otherwise.} \]

(10)

Then

\[ h(u) = \nu \cdot u - 1 - \log (\nu \cdot u), \quad u > 0, \]
\[ \text{if } \nu > 0, \quad \text{otherwise.} \]

(11)

**B. Slow Markov Walk**

In this section we present a large deviation theorem due to Ventsel [12], regarding certain Markov chains. Cottrell et al. [3] have a more detailed discussion of this result.

Consider the Markov chain \( \{X_n\} \in \mathbb{R}^d \) given by

\[ X_{n+1} = X_n + \epsilon \cdot V(X_n, \xi_n), \quad n \geq 0 \]

where \( \epsilon > 0 \) is the parameter defining the Markov chain \( \{X_n\} \), \( X_0 \) is the initial value, \( V(\cdot, \cdot) \) is a function from \( \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d \) and \( \xi_n \)'s are i.i.d. r.v.'s. We are interested in analyzing \( \{X_n\} \) when \( \epsilon \to 0 \). Let \( F_\epsilon \) denote the d.f. of \( V(X_n, \xi_n) \). Let

\[ m(x) = \int_{\mathbb{R}^d} \exp (s, z) dF_\epsilon(z) \]

be the mean of \( F_\epsilon \),

\[ M_\epsilon(s) := \int_{\mathbb{R}^d} \exp (s, z) dF_\epsilon(z) \]

be its Laplace transform, \( L_\epsilon(s) := \log M_\epsilon(s) \) and

\[ h_\epsilon(u) := \sup_{s \in \mathbb{R}^d} \{ (s, u) - L_\epsilon(s) \} \]

be its Cramér transform. Assume the following:

A1: \( M_\epsilon(s) < \infty \) in a neighborhood of 0 for each \( x \in \mathbb{R}^d \).

A2: \( d(F_{n+1}, F_n) \leq c \| x_{n+1} - x_n \| \), where \( d \) is the Prohorov distance (see [2]) and \( c > 0 \) is a constant, i.e., \( F_n \) is Lipschitz smooth in \( x \).

Next, construct continuous-time paths from the realizations of \( \{X_n\} \). To do this, at each epoch

\[ t = n \cdot \epsilon \]

define \( Z(t) := X_n \quad \text{for } \epsilon \to 0 \), and interpolate piecewise linearly. Let \( C_T \) denote the set of the continuously piecewise differentiable functions \( \phi : [0, T] \to \mathbb{R}^d \) such that \( \phi(0) = x_0 \) is fixed. Let \( P_T^\epsilon \) denote the measure induced by the Markov chain \( \{X_n\} \) on the Borel \( \sigma \)-field \( \Sigma \) of \( C_T \) endowed with the Skorohod topology [2]. Define the action integral

\[ I(\phi) := \int_0^T h_{\epsilon(t)}(\phi'(t)) \ dt. \]

Under some additional assumptions, with A1 and A2 being the most crucial ones, we have the following result.

**Theorem 2** (Ventsel) [12]: Let \( \phi \) be a path in \( C_T \). Define a tube of diameter \( d \) around \( \phi \), \( T_d(\phi) \), as the set of trajectories \( \eta(t) \)'s such that

\[ |\eta(t) - \phi(t)| < d, \quad \text{for all } t \in [0, T]. \]
Then, there exists \( \delta_0 \) such that, for \( 0 < \delta < \delta_0 \),

\[
\lim_{c \to 0} (-c \cdot \log P^c\{T_S(\phi)\}) = I(\phi) + \epsilon(\delta)
\]

with \( \lim_{\delta \to 0} \epsilon(\delta) = 0 \).

Next, we present a consequence of Theorem 2 that will enable us to estimate \( P^c\{S\} \) for \( S \subseteq \Sigma \) whose boundary satisfies certain smoothness conditions.

Corollary 1 (Ventzel) [12]: Let \( S \subseteq \Sigma \) be such that

\[
\inf \{ I(\phi) / \phi \in \text{int}(S) \} = \inf \{ I(\phi) / \phi \in \text{cl}(S) \},
\]

then

\[
\lim_{c \to 0} (-c \cdot \log P^c\{S\}) = \inf_{\phi \in S} I(\phi).
\]

(15)

Corollary 1 suggests that

\[
P^c\{S\} = \sum_k P^c\{T_S(\theta_k)\} = \sum_k \exp \left( -\frac{1}{c} \cdot I(\theta_k) \right)
\]

\[
= \exp \left( -\frac{1}{c} \cdot \inf_{\phi \in S} I(\phi) \right) \quad \text{(UTLE)}
\]

where the second approximation follows from Theorem 1, the last one follows from Corollary 1, and UTLE is the acronym for upper turnpike law equivalence. Using lower semicontinuity of \( I(\phi) \) and the condition in (15), it is not difficult to show that \( \inf_{\phi \in S} I(\phi) \) is achievable. Let us denote \( \arg\inf_{\phi \in S} I(\phi) \) by \( \phi_{\text{opt}} \).

Suppose we are interested in the probability of the set \( S \) of trajectories which hit a rare set \( A \) before hitting 0 given that we start from 0. Let us assume that the conditions leading to Corollary 1 are satisfied. It follows from the above discussion that asymptotically it is sufficient to find \( \inf_{\phi \in S} I(\phi) \). For this, define

\[
C(x) := \inf \left\{ \int_0^{T(x)} H(\phi(t), \phi'(t)) dt / \phi(0) = x \right\}
\]

\[
\phi \in C, T(\phi) < \infty \}
\]

(16)

where \( x = (x_{(1)}, \cdots, x_{(d)}) \) and \( \nu = (\nu_{(1)}, \cdots, \nu_{(d)}) \) are vectors in \( R^d \), \( C \) denotes the set of the continuously piecewise differentiable functions \( \phi:[0, \infty) \to R^d \) and \( H(\phi(t), \phi'(t)) = h_{\nu}(\phi'(t)) \). We denote \( L(\theta) \) by \( L(x, \theta) \). Notice that \( \phi_{\text{opt}} \) is a trajectory that achieves the infimum for \( C(0) \). The following result gives a recipe for finding \( \phi_{\text{opt}} \).

Theorem 3: Assume that \( C(x) \) is smooth enough to satisfy

\[
\frac{\partial^2 C}{\partial x_{(i)} \partial x_{(j)}} = -\frac{\partial^2 C}{\partial x_{(j)} \partial x_{(i)}}, \quad 1 \leq i \leq d, 1 \leq j \leq d.
\]

Let us define

\[
\theta_{(i)}(x) := -\frac{\partial C}{\partial x_{(i)}} (x), \quad 1 \leq i \leq d.
\]

(17)

Then, for each \( x \) that is on some \( \phi \in S \),

\[
L(x, \theta(x)) = 0
\]

(18)

and \( \phi_{\text{opt}} \) is a solution of the following system of differential equations:

\[
\frac{d\theta_{(i)}}{dt} = \frac{\partial L}{\partial \theta_{(i)}} (x, \theta), \quad 1 \leq i \leq d.
\]

(19)

\[
\frac{dx_{(i)}}{dt} = \frac{\partial L}{\partial x_{(i)}} (x, \theta), \quad 1 \leq i \leq d.
\]

(20)

Using this in (24), we have

\[
\frac{d\theta_{(i)}}{dt} = \sum_{k=1}^d \frac{\partial L}{\partial \theta_{(i)}} (x, \theta) \cdot \frac{\partial \theta_{(k)}}{\partial x_{(i)}} (x).
\]

(23)
Now using (23), along \( \phi_{\text{opt}} \), we get
\[
\frac{db_{(i)}(t)}{dt} = -\frac{\partial L}{\partial x(t)} (x, \theta), \quad 1 \leq i \leq d.
\]

Note that the assertion in (20) is equivalent to (22).

Notice that (19) and (20), the initial condition \( x(0) = x_0 \), and the terminal condition \( x(T) \) \( \in \Delta \) have \( \phi_{\text{opt}} \) as a solution. To solve for \( \phi_{\text{opt}} \), sometimes it is convenient also to use (18). This will be illustrated in an example in Section III-D.

Next, we explain the role played by the variable \( \theta \). For this, define a new probability measure \( F^*_\theta \) from \( F^* \), as
\[
dF^*_\theta(z) := \frac{e^{\theta \cdot z} dF^*(z)}{M_\theta(\theta)}
\]
(25)
where the parameter \( \theta \in \mathbb{R}^d \). This is called the exponential change of measure with the parameter \( \theta \).

Suppose we want to select \( \theta \), along \( \phi_{\text{opt}} \) in such a way that
\[
\phi_{\text{opt}}(t) = m^*(\phi_{\text{opt}}(t))
\]
(26)
where \( m^*(x) \) denotes the mean of \( F^*_\theta \). Then,
\[
m^*(x) = \int_{\mathbb{R}^d} zdF^*_\theta(z) = \int_{\mathbb{R}^d} z \cdot \frac{e^{\theta \cdot z}}{M_\theta(\theta)} \frac{M^*_\theta(\theta)}{M_\theta(\theta)} dL(z) = l^*_\theta(\theta).
\]
(27)

Equations (26) and (27) indicate that the parameter of the exponential change of measure that makes the trajectory \( \phi_{\text{opt}} \) most likely satisfies
\[
\phi_{\text{opt}}(t) = l^*_\theta(\theta)(\phi_{\text{opt}}(t)).
\]
(28)
Recalling our notation that \( L(x, \theta) = L(\theta) \) and comparing (19) and (28), it is clear that the variable \( \theta \) in the system of differential equations (19) and (20) represent the parameter for the exponential change of measure required to achieve the condition in (26).

C. Quick Simulation Method (Optimal Exponential Change of Measure)

Consider a discrete-time M.C. \( \{X_n, n = 0, 1, 2, \ldots \} \) and let \((\Omega, \Sigma, P)\) be the corresponding probability space. Let \( S \in \Sigma \) be a rare event, i.e., \( \alpha := P[\{S\}] \ll 1 \). Let \( P' \) be another probability measure on \((\Omega, \Sigma)\) such that \( P \) is absolutely continuous with respect to \( P' \). Denote the Radon–Nikodym derivative (likelihood ratio) by \( L := dP/dP' \). We consider \( \alpha_n \) and \( \alpha_n^* \) as two convergent and unbiased estimators of \( \alpha \), where
\[
\alpha_n := \frac{1}{n} \sum_{i=1}^{n} l_j(\omega_i)
\]
(29)
and
\[
\alpha_n^* = \frac{1}{n} \sum_{i=1}^{n} l_j(\omega_i) \cdot L(\omega_i).
\]
(30)

Here \( \omega_i \)'s are the i.i.d. outcomes of experiments on \((\Omega, \Sigma, P)\). As discussed in Section II, \( \alpha_n \) is more efficient than \( \alpha_n^* \) if and only if \( \text{Var} \{\alpha_n^*\} < \text{Var} \{\alpha_n\} \), which will be the case if and only if
\[
\int_S L^*(\omega) dP^*(\omega) < \alpha.
\]
(31)

Obviously, if \( L(\omega) < 1 \) whenever \( \omega \in S \), then this condition is satisfied.

In the previous section we discussed the Markov chain \( \{X^*_n\} \in \mathbb{R}^d \) defined in (12). We now present a theorem due to Cottrell et al. [3] that gives, for the simulation purpose, the optimality of a measure \( P^*_\alpha \), obtained by an exponential change of measure, from
\( P^* \). Their theorem is presented in [3] for the case of \( R^1 \). However, it can be generalized to the case of \( R^d \). It is assumed that the mean drift function \( \psi(x) := E[V(X^*_n, \xi_n) | X^*_n = x] \) is such that the O.D.E., \( x'(t) = \psi(x(t)) \), with \( x(0) \) specified, has \( \theta \) as a stable equilibrium point. See [3] for details.

Suppose that we want to estimate, for small \( \epsilon > 0 \), \( P_\alpha(S) \), probability of the event
\[
S := \{\omega | X^*_\epsilon > 0 \}
\]
given that \( X_0 = 0 \). Let us define a probability measure \( P^*_\alpha \) as the resultant measure when \( F^*_\theta \) is taken as defined by (25), with \( \theta \) being the solution of
\[
M_\theta(\theta) = 1, \quad \theta > 0.
\]
(32)

The probability measure \( P^*_\alpha \) is optimal in the sense made precise by the following theorem due to Cottrell et al. [3].

Theorem 4 (Cottrell et al. [3]): Suppose that for the Markov chain \( X^*_n \in \mathbb{R}^d \), defined in (12), assumptions A1 and A2 hold. Then among all the exponential changes of measure, the transformation \( P \rightarrow P^*_\alpha \) is asymptotically optimal in the sense of the variance, i.e., for \( P^*_\alpha \)
\[
\lim_{\epsilon \to 0} \int_S L^*(\omega) dP^*_\alpha(\omega)
\]
where \( L = dP/dP^*_\alpha \) is minimum.

D. Applications and Difficulties

Consider an open Jackson network of \( d > 0 \) nodes with infinite buffers. Let \( \{X_n, n = 0, 1, 2, \ldots \} \in \mathbb{R}^d \) denote the embedded discrete-time Markov chain representing queue-lengths of the nodes at the epochs of the jumps in the network (arrivals, departures, and transfers), where \( X_n = (X_{n(1)}, X_{n(2)}, \ldots, X_{n(d)}) \in \mathbb{R}^d \) (actually, \( X_n \in \mathbb{N}^d \)). Let \( S \) denote the set of the realizations of \( \{X_n\} \) that reach the region of the state space where the total backlog exceeds \( n \cdot x_1 + x_2 + \cdots + x_d \gtrsim N \), before hitting 0, \( X_{n(1)} = \cdots = x_{n(d)} = 0 \). We want to estimate the probability \( \alpha := P_\alpha(S) \), the probability of \( S \) given that \( X_0 = 0 \).

We can represent \( \{X_n\} \), as
\[
X_0 = x_0, \quad X_{n+1} = X_n + V(X_n, \xi_n), \quad n \geq 0
\]
(33)
where \( V(x, \xi) \) denotes the r.v. representing the jump from \( X_n = x \). For example, consider \( M/M/1 \) queues in tandem (see Fig. 4). We assume, for stability, \( \lambda < \mu_1 \) and \( \lambda < \mu_d \). We also assume, without any loss of generality, that \( \lambda + \mu_1 + \mu_d = 1 \). For simplicity, we will refer to such a system by a \((\lambda, \mu_1, \mu_d)\)-network. Now \( \{X_n\} \) is a Markov chain in \( \mathbb{R}^d \) defined by (33), where the distributions of \( V(\cdot, \xi) \) are as depicted in Fig. 5.

Let us return to the discussion of general Jackson networks. It is possible to represent the embedded Markov chain \( \{X_n\} \) in the form of (12). For this define \( X^*_n = X_n/N \). Then,
\[
X^*_{n+1} = X^*_{n} + \frac{1}{N} \cdot V(X^*_n, \xi_n) = X^*_n + \frac{1}{N} \cdot V(N \cdot X^*_n, \xi_n)
\]
\[
= X^*_n + \frac{1}{N} \cdot V(X^*_n, \xi_n).
\]
(34)

The last equality follows from the fact that in Jackson networks the distributions of \( V(x, \xi) \) and \( V(c \cdot x, \xi) \) are the same for all \( x \) and all \( c > 0 \). Because of (34), we have an equivalent representation of \( \{X_n\} \) which is in the same form as (12) with \( \epsilon = 1/N \). For the process \( \{X^*_n\} \) we are interested in estimating \( \alpha = P_\alpha(S^*) \), where \( S^* \) is the set of the realizations of \( \{X^*_n\} \) that reach the region of the state space where the sum of its coordinates exceeds 1.
M/M/1 Queue: Let $X_\lambda$ denote the backlog of a stable M/M/1 queue with rates $\lambda$ and $\mu$. For $\{X_\lambda\}$, note that

$$M_\lambda(s) = \lambda \cdot e^{-s} + \mu \cdot e^{-\lambda s}, \quad s > 0.$$  

Equation (32), along with the condition $\phi_{\mu}(T) = 1$, gives

$$\theta = \log \left( \frac{\mu}{\lambda} \right), \quad x > 0.$$  

Equation (20) gives

$$\phi_{\mu}(\tau) = \mu - \lambda.$$  

From example E1 of Section III-A (10) we have

$$h_{\mu}(\phi^*(t)) = (\mu - \lambda) \cdot \log \left( \frac{\mu}{\lambda} \right), \quad t > 0.$$  

Now we can use Corollary 1, Section III-B, to evaluate $P_0$ of $S$. Noting that, for $\phi_{\mu}$ defined in (36), $T = 1/(\mu - \lambda)$, we get

$$\lim_{N \to \infty} \frac{1}{N} \cdot \log P_0(S) = \log \left( \frac{\mu}{\lambda} \right).$$  

This gives $P_0(S) = (\mu/\lambda)^N$ (UTLE). Observe that this matches well with the exact expression for $P_0$ of $S$ given by (1). Also observe that $\theta$ given by (40) gives the exponential change of measure [see (25)] that corresponds to the M/M/1 queue with arrival rate $\mu$ and service rate $\lambda$ (see Fig. 3).

M/M/1 Queues in Tandem: We consider a $(\lambda, \mu_1, \mu_2)$-network defined in the beginning of this section (see Figs. 4 and 5). This simple Jackson network illustrates the difficulties in applying the results of the previous two sections to Jackson networks.

Observe from Fig. 5 that the jump distributions change abruptly near the $x_{\lambda_2}$-axis (second queue empty) and $x_{\lambda_2}$-axis (first queue empty) if we move from these axes to $R$, the interior region (both the queues nonempty). This violates the smoothness assumption A2 of Section III-B. Hence, the results of the previous two sections are not applicable here.

As a remedy to this difficulty, we may consider a process which has the jump distributions modified near the boundaries ($x_{\lambda_1}$-axis and $x_{\lambda_2}$-axis) such that over a thin layer they make smooth transitions. We call such a construction a boundary layer construction. One might argue that $P_0(S)$ does not change much by such a modification. For the scaled process $\{X_\mu^a\}$, this construction is illustrated in Fig. 6. If such a construction were indeed valid, we could use (18), (19), and (20) to find $\phi_{\mu_2}$ and $P_0(S)$ by the quick simulation method. However, we find this numerical approach rather formidable because of the need to solve a system of differential equations with mixed initial and terminal conditions. Since our purpose is to suggest a simple alternative method, we will not pursue further this approach here.

One can avoid the complications by neglecting the boundaries. However, this results in a poor approximation. Suppose we assume that the jump distributions are identical everywhere to that of the interior region $R$. Then, from (19) and (20), we see that $\phi_{\mu}(\tau)$ is constant. Hence, $\phi_{\mu}$ will be one of the rays through $R$ (see Fig. 7). Solving $I(\theta) = 0$ [see (18)], with the constraint that $\theta_{(1)} = \theta_{(2)} (\neq 0)$ and a boundary condition (see Parekh [8] for details), we get

$$\theta_{(1)} = \theta_{(2)} = \log \left( \frac{\mu_2}{\lambda} \right).$$  

The exponential change of measure with the parameter $\theta$ can be seen to give the $(\mu_2, \mu_1, \lambda)$-network.

It is easy to convince oneself by simulations that the above is a poor change of measure. For example, for the $(\lambda = 0.20, \mu_1 = 0.30, \mu_2 = 0.50)$-network and $N = 20$, $\alpha = P_0(S)$ is found by solving the first set of equations numerically to be $3.759 \times 10^{-6}$. If we simulate the (0.50, 0.30, 0.20)-network, as suggested by the above discussion, we get $\alpha_{(000)} = 3.288 \times 10^{-6}$, while simulating the (0.30, 0.20, 0.50)-network we get $\alpha_{(000)} = 3.659 \times 10^{-6}$. This example is illustrated in Fig. 8. Note that the (0.30, 0.20, 0.50)-network is also obtained from the original network by an exponential change of measure. In the next section we will present a heuristic that will justify the optimality of this change of measure.

IV. SIMULATION OF EVENTS OF EXCESSIVE BACKLOG—A HEURISTIC APPROACH

The purpose of this section is to report some very interesting observations. Our hope is that the intuitive explanations presented
here will motivate more research in the area. Some limiting cases for our heuristic are reported in Section IV-D.

A. Heuristic of Borovkov, Ruget [9], Etc., for a GI/GI/1 Queue and Its Application to Simulations

Consider a GI/GI/1 queue. Let A and B denote the interarrival and service time d.f.'s, respectively. Generically, let $M_A$ and $h_A$ denote the Laplace and Cramér transforms of a d.f. $D$. Let $1/\lambda$ and $1/\mu$ denote the means of $A$ and $B$, respectively. Such a queue is shown in Fig. 9. For stability, we assume $1/\lambda > 1/\mu$. Let $P$ denote the measure induced by the stochastic process describing the queue. We want to calculate $\alpha$, the probability of the backlog exceeding $N$ in a cycle, i.e., the probability of hitting $N$ before returning to 0 given that the system starts empty. Let $S$ denote the corresponding event, i.e., $\alpha = P_S\{S\}$.

Let $X_t^d$ denote the i.i.d. copy of a random variable distributed with the d.f. $D$. Then, $X_t^d$ denotes the $i$th interarrival time and $X_t^q$ denotes the $i$th virtual service time. Consider the subset of $S$ where the system reaches $N$ at time $T$ and the average interarrival and the virtual service times are $1/\lambda'$ and $1/\mu'$, respectively, with $1/\lambda' < 1/\mu'$. Now, by Cramér's theorem, Theorem 1,

$$P\{X_t^d + \cdots + X_{\lambda'-T}^d = N\} = \exp \left( -\lambda' \cdot T \cdot h_A \left( \frac{1}{\lambda'} \right) \right) \text{ (UTLE)}$$

where UTLE is the acrony for up to logarithmic equivalence. Similarly,

$$P\{X_t^q + \cdots + X_{\lambda'-T}^q = N\} = \exp \left( -\mu' \cdot T \cdot h_B \left( \frac{1}{\mu'} \right) \right) \text{ (UTLE)}.$$

Since $1/\lambda' < 1/\mu'$, for large $T$, we assume that most of the virtual services were the actual services. Then, $T = N/(\lambda' - \mu')$. Since, the interarrival times and the virtual service times are independent,

$$\alpha = \sum_{\lambda' > \mu'} \frac{1}{N - T}(\lambda' - \mu') \cdot \exp \left\{ -T \cdot \left( \lambda' \cdot h_A \left( \frac{1}{\lambda'} \right) \right) + \mu' \cdot h_B \left( \frac{1}{\mu'} \right) \right\} \text{ (UTLE)}$$

$$= \sum_{\lambda' > \mu'} \frac{N}{\lambda' - \mu'} \cdot \exp \left\{ -T \cdot \left( \lambda' \cdot h_A \left( \frac{1}{\lambda'} \right) \right) + \mu' \cdot h_B \left( \frac{1}{\mu'} \right) \right\}.$$

Hence, for large $N$,

$$\alpha = \exp \left\{ -N \cdot \inf_{\lambda' > \mu'} \left( \lambda' \cdot h_A \left( \frac{1}{\lambda'} \right) \right) + \mu' \cdot h_B \left( \frac{1}{\mu'} \right) \right\} \text{ (UTLE)} \quad (37)$$

and

$$\alpha = \exp \left\{ -\lambda' \cdot T \cdot h_A \left( \frac{1}{\lambda'} \right) \right\} \text{ (UTLE)} \quad (38)$$

To obtain the exponent, we differentiate

$$\frac{1}{\lambda' - \mu'} \cdot \left( \lambda' \cdot h_A \left( \frac{1}{\lambda'} \right) + \mu' \cdot h_B \left( \frac{1}{\mu'} \right) \right)$$

with respect to $\lambda'$ and $\mu'$ and to the results to 0. This gives

$$h_A \left( \frac{1}{\lambda'} \right) + h_B \left( \frac{1}{\mu'} \right) = \left( \frac{1}{\lambda'} \right) \cdot \left( \frac{1}{\mu'} \right)$$

$$= \left( \frac{1}{\lambda'} - \frac{1}{\mu'} \right) \cdot h_A \left( \frac{1}{\lambda'} \right) + h_B \left( \frac{1}{\mu'} \right). \quad (39)$$

Suppose that $\lambda^*$ and $\mu^*$ achieve the infimum. Then, from (38),

$$h_A \left( \frac{1}{\lambda^*} \right) = h_B \left( \frac{1}{\mu^*} \right) = \theta^* \quad (40)$$

We can argue from the convexity of $h_A$ and $h_B$ that $\theta^* > 0$. Also, from (38), we have

$$\theta^* \cdot \frac{1}{\lambda^*} + h_A \left( \frac{1}{\lambda^*} \right) = \theta^* - h_B \left( \frac{1}{\mu^*} \right). \quad (41)$$

From the convex duality property P4 of the Cramér transform (see Section III-A) and (39) and (40), we have

$$\log M_A(-\theta^*) = -\theta^* \cdot \frac{1}{\lambda^*} - h_A \left( \frac{1}{\lambda^*} \right)$$

and

$$\log M_B(\theta^*) = \theta^* \cdot \frac{1}{\mu^*} - h_B \left( \frac{1}{\mu^*} \right). \quad (42)$$

Therefore,

$$\log M_A(-\theta^*) = \log M_B(\theta^*), \quad (42)$$

i.e., the conditions for determining $\theta^*$ are

$$\theta^* > 0 \text{ and } M_A(\theta^*) = M_B(-\theta^*). \quad (43)$$

From (37) and (41), for large $N$, we also have

$$\alpha = \exp \left\{ -N \cdot \log M_B(\theta^*) \right\} \text{ (UTLE)} \quad (44)$$

Let $A^*$ denote the measure obtained by an exponential change of measure from $A$ such that its mean is $1/\lambda^*$, i.e., the parameter for the exponential change of measure $\theta^*$ satisfies

$$dA^*(z) = \frac{e^{z \cdot \theta^*}}{M_B(\theta^*)} dA(z) \quad (45)$$
and
\[
\frac{1}{\lambda^*} \int z \cdot e^{\theta^* z} c \cdot dA(z) = d \log M_\theta(\theta^*).
\]

Using (39) and the property of reciprocity of the derivatives of the Cramér and the log-Laplace transforms (property P5 of the Cramér transform, Section III-A), we get
\[
\theta^* = -\theta^*.
\]

Similarly, let \( B^* \) denote the measure obtained by an exponential change of measure from \( B \) such that its mean is \( 1/\mu^* \). Then, the required parameter for the exponential change of measure \( \theta^* \) can be seen to satisfy
\[
\theta^* = \theta^*.
\]

Now define a transformed GI/GI/1 queue with \( A^* \) and \( B^* \) as its interarrival time and service time d.f.'s, respectively. Let \( P^* \) denote the measure induced by the transformed stochastic process. The definitions of \( \lambda^*, \mu^*, \) and \( P^* \) suggest that, for large \( N \),
\[
\frac{dP}{dP^*} \to 1
\]

almost everywhere (under measure \( P \)) on the event \( S \). Then, (31) indicates that it will be faster to estimate \( \alpha \) under the measure \( P^* \) than under \( P \).

**M/M/1 Example:** Let \( \lambda < \mu \) denote interarrival and service rates. If \( D \) is the exponential d.f. with mean \( 1/\mu \), then we denote by \( M_\lambda \) and \( h_\lambda \) its Laplace and Cramér transforms, respectively. Recall that
\[
M_\lambda(s) = \frac{s}{s - \lambda}, \quad s < \lambda,
\]
\[
= \infty, \quad \text{otherwise}.
\]

Equation (43) gives
\[
\theta^* > 0 \quad \text{and} \quad \frac{\alpha}{\lambda + \theta^*} - \frac{\mu}{\mu - \theta^*} = 1. \quad (45)
\]

It is easily checked that the solution of (45) is
\[
\theta^* = \theta - \lambda.
\]

Then, (45) gives
\[
\alpha = \left( \frac{\lambda}{\mu} \right)^N \quad \text{(UTLE).}
\]

Observe that this matches well with the exact expression for \( \alpha \) given by (1). Also, calculations of \( G^* \) and \( F^* \), as defined above, show that the transformed M/M/1 queue for the purpose of estimating \( \alpha \) by simulations is the one that corresponds to the interchange of \( \lambda \) and \( \mu \).

The above argument is presented for the continuous time variables. We can emulate the same argument for the embedded M.C. of a Jackson network.

**B. Extension to Simple Jackson Networks (M/M/1 Queues in Tandem and in Parallel)**

As in Section III-D for an open Jackson network of \( d > 0 \) nodes with infinite buffers, let \( \{X_n, n = 0, 1, 2, \cdots \} \in R^d \) denote the embedded discrete-time M.C. representing queue-lengths of the nodes at the epochs of the jumps in the network (arrivals, departures, and transfers). We want to estimate \( \alpha = P_0[S] \), where \( S \) is the set of the realizations of \( \{X_n\} \) that reach the region of the state-space where the total backlog exceeds \( N \), before hitting 0.

**M/M/1 Queues in Tandem:** For the embedded Markov chain \( \{X_n\} \in R^2 \), Fig. 5 gives the jump distributions. Recall that we have uniformized the M.C., i.e., \( \lambda + \mu_1 + \mu_2 = 1 \).

Consider the paths of \( S \) which require \( T \) transitions and have \( \lambda^*, \mu_1^*, \) and \( \mu_2^* \) proportions for the arrivals, virtual departures from the first queue and that from the second queue, respectively. Continuing the same line of heuristic as in Section IV-A, we can write
\[
\alpha = \sum_{\lambda^*, \mu_1^*, \mu_2^*} \frac{1}{T(\lambda^*, \mu_1^*, \mu_2^*)} \left( \begin{array}{c}
\left( \lambda^* \cdot h_0 \left( \frac{1}{\lambda^*} \right) + \mu_1^* \cdot h_1 \left( \frac{1}{\mu_1^*} \right) \\
+ \mu_2^* \cdot h_2 \left( \frac{1}{\mu_2^*} \right) \end{array} \right) \right) \quad \text{(UTLE)}
\]

where \( T(\lambda^*, \mu_1^*, \mu_2^*) \) is the total number of transitions (which equals the number of time units due to the uniformization) required for the realizations belonging to \( S \) with \( \lambda^*, \mu_1^*, \) and \( \mu_2^* \) proportions of arrivals and virtual services from the queues, respectively.

It can be heuristically argued that, for large \( N \) and when \( \lambda^* > \mu_1^* \) or \( \lambda^* > \mu_2^* \), \( T(\lambda^*, \mu_1^*, \mu_2^*) = N \cdot R(\lambda', \mu_1^*, \mu_2^*) \), where
\[
R = \begin{cases} 
1/(\lambda' - \mu_1^*), & \text{if } \lambda' > \mu_1^* \text{ and } \mu_2^* \leq \mu_2^*, \\
1/(\lambda' - \mu_2^*), & \text{otherwise.}
\end{cases}
\]

Therefore, for large \( N \),
\[
\alpha = \exp \left[ -N \inf_{\lambda^*, \mu_1^*, \mu_2^*} \left( R(\lambda', \mu_1^*, \mu_2^*) \right) \right]
\]

Numerical minimization gives \( \lambda^*, \mu_1^*, \) and \( \mu_2^* \) that correspond to the interchange of \( \lambda \) with the smallest of \( \mu_1 \) and \( \mu_2 \). (For the limiting case where \( \mu_1 = \mu_2 \), see Section IV-D.) As explained for the case of an M/M/1 queue in Section IV-A to estimate \( \alpha \), it will be faster to simulate the embedded Markov chain of the \((\lambda^*, \mu_1^*, \mu_2^*)\)-network.

Tables IV and V show the results of some experiments with M/M/1 queues in tandem. All the simulations were done on a VAX-750 machine and the first step equations were solved using the IMSL routine LEQT2F.

**M/M/1 Queues in Parallel:** Consider two M/M/1 queues in parallel with \( \lambda_i \) and \( \mu_i, i = 1, 2, \) as their arrival and service rates, respectively. We assume that \( \lambda_i < \mu_i \), \( i = 1, 2, \) and \( \lambda_1 + \mu_1 + \lambda_2 + \mu_2 = 1 \). We denote such a system by the \((\lambda_1, \mu_1, \lambda_2, \mu_2)\)-network.

As for M/M/1 queues in tandem, we can approximate the probability of interest by an exponential term. Minimization of the exponent gives \( \lambda_1^*, \mu_1^*, \lambda_2^*, \) and \( \mu_2^* \) that correspond to the
interchange of $\lambda$ and $\mu$, with the larger traffic intensity $\lambda/\mu$. (For the limiting case where $\lambda_1/\mu_1 = \lambda_2/\mu_2$, see Section IV-D) Tables VI and VII show the results of some experiments with M/M/1 queues in parallel.

C. Extension to Networks with Routing

In Section IV-B we extended our heuristic to M/M/1 queues in tandem and in parallel. In this subsection we will extend it further to networks where probabilistic routing may be present. By doing so, we will have extended the heuristic to arbitrary open Jackson networks. For this purpose, we need the following theorem due to Sanyo [10].

Theorem 5 (Sanov) [10]: Let $Z_k$, $k \geq 1$, be random variables whose possible values are $a_1, \ldots, a_n$ with $p_1, \ldots, p_n$ as respective probabilities. For $N > 1$, define $m(N) := \#$ of $Z_k$'s,

| TABLE IV | SIMULATIONS FOR M/M/1 QUEUES IN TANDEM |
| Method | \(\lambda = 0.05 \mu = 0.03 \lambda = 0.05 \mu = 0.05 N = 15\) | \(\lambda = 0.05 \mu = 0.05 \lambda = 0.05 \mu = 0.05 N = 15\) |
| \# of Cycles (s) | 10000 | 20000 | 40000 |
| CPU Time | 1.09e-09 | 3.06e-09 | 5.18e-09 |
| Calls to RNG | 52790 | 105730 | 21086 |
| Example-I | \(\lambda = 0.05 \mu = 0.03 \lambda = 0.03 \mu = 0.03 N = 15\) | \(\lambda = 0.05 \mu = 0.03 \lambda = 0.03 \mu = 0.03 N = 15\) |
| \# of Cycles (s) | 10000 | 20000 | 40000 |
| CPU Time | 1.09e-09 | 3.06e-09 | 5.18e-09 |
| Calls to RNG | 52790 | 105730 | 21086 |
| Example-II | \(\lambda = 0.05 \mu = 0.05 \lambda = 0.05 \mu = 0.05 N = 15\) | \(\lambda = 0.05 \mu = 0.05 \lambda = 0.05 \mu = 0.05 N = 15\) |
| \# of Cycles (s) | 10000 | 20000 | 40000 |
| CPU Time | 1.09e-09 | 3.06e-09 | 5.18e-09 |
| Calls to RNG | 52790 | 105730 | 21086 |

| TABLE V | EMPIRICAL STANDARD DEVIATION FOR M/M/1 QUEUES IN TANDEM |
| Example-I | \(\lambda = 0.05 \mu = 0.03 \lambda = 0.05 \mu = 0.05 N = 15\) | \(\lambda = 0.05 \mu = 0.05 \lambda = 0.05 \mu = 0.05 N = 15\) |
| \# of Cycles (s) | 500 | 1000 |
| Empirical Mean (s) | 3.490s | 3.385s |
| Empirical Std. Dev. (s) | 2.958e-09 | 7.958e-09 |
| \(\mu/\mu_1\times 100\% | 2.588 | 2.589 |
| Example-II | \(\lambda = 0.10 \mu = 0.03 \lambda = 0.03 \mu = 0.03 N = 15\) | \(\lambda = 0.10 \mu = 0.03 \lambda = 0.03 \mu = 0.03 N = 15\) |
| \# of Cycles (s) | 500 | 1000 |
| Empirical Mean (s) | 2.212s | 2.118s |
| Empirical Std. Dev. (s) | 2.320e-09 | 1.850e-09 |
| \(\mu/\mu_1\times 100\% | 10.437 | 7.808 |
| Example-III | \(\lambda = 0.20 \mu = 0.03 \lambda = 0.03 \mu = 0.03 N = 15\) | \(\lambda = 0.20 \mu = 0.03 \lambda = 0.03 \mu = 0.03 N = 15\) |
| \# of Cycles (s) | 500 | 1000 |
| Empirical Mean (s) | 7.816s | 5.295s |
| Empirical Std. Dev. (s) | 7.915e-09 | 7.125e-09 |
| \(\mu/\mu_1\times 100\% | 5.886 | 5.886 |

1 \leq k \leq N, that are equal to $a_i$. Define the relative frequency

\[ p_i = \frac{m(N)}{N}, \quad 1 \leq i \leq n. \]

Let $q_1, \ldots, q_n$ be real numbers satisfying $q_i \geq 0, 1 \leq i \leq n$, and $q_1 + \cdots + q_n = 1$. Then,

\[ \lim_{N \to \infty} \frac{1}{N} \log P\{\lambda_i(N) - q_i \leq \epsilon, \ldots, \lambda_i(N) - q_i \leq \epsilon\} = -K(q, p) + e(\epsilon) \]

where

\[ K(q, p) = \sum_{i=1}^{n} q_i \cdot \log \left( \frac{q_i}{p_i} \right) \]
and the term $e(e)$ is $O(e \log (1/e))$. (If $q_i > 0$, $1 \leq i \leq n$, then $O(e \log (1/e)$ can be replaced by $O(e)$.)

The above theorem suggests that

$$P\{m_1(N) = q_1 \cdot N, \ldots, m_n(N) = q_n \cdot N\} = \exp(-N \cdot K(q, p)) \text{(UTLE).}$$

Now consider the network shown in Fig. 10. For stability, we assume that $\lambda < \mu_1$ and $\lambda(1-p) < \mu_2$. We also assume, without any loss of generality, that $\lambda + \mu_1 + \mu_2 = 1$. We consider the embedded Markov chain $\{X_t\}$.

As in the cases of $M/M/1$ queues in tandem and in parallel, consider the paths of $S$ which require $T$ transitions, have $\lambda^*, \mu_1^*, \mu_2^*$ proportions for the arrivals, virtual departures from the first queue and that from the second queue, respectively, and have $p^*$ and $(1-p)^*$ proportions of customers routed out of the network and to the second queue, respectively, from the outset of the first queue. Then, as in the last subsection, we can argue heuristically that, for large $N$,

$$\alpha = \exp \left\{-N \cdot \inf_{\lambda^*, \mu_1^*, \mu_2^*, p^*} \left( R(\lambda^*, \mu_1^*, \mu_2^*, p^*)^\lambda + R(\lambda^*, \mu_1^*, \mu_2^*, p^*)^\lambda \right) \right\} \text{(UTLE)}$$

where

$$R = \left\{ \begin{array}{ll}
(1/(\lambda^* - \mu_1^*) + (1/(\lambda^* - \mu_2^*)) - (1-p)^* / \mu_2^*) & \text{if } \lambda^* > \mu_1^* \text{ and } (1-p)^* > \mu_2^*; \\
(1/\lambda^* + (1/(\lambda^* - \mu_1^*) - (1-p)^*) / \mu_1^*) & \text{if } \lambda^* > \mu_1^* \text{ and } (1-p)^* > \mu_2^*; \\
1/\lambda^* & \text{otherwise.} \\
\end{array} \right.$$

Numerical minimization gives us $\lambda^*, \mu_1^*, \mu_2^*$, and $p^*$ as the parameters of the network obtained by an optimal exponential change of measure. Examples show that the node with higher traffic intensity blows up while the other one remains stable. The limiting case occurs when the traffic intensities are equal (see Section IV-D).

Tables VIII and IX list some illustrations of simulation speed-ups when simulated under the transformed system.

D. Some Observations

1) On $M/M/1$ Queues in Tandem:

a) If the set of arguments for the minimization in (46) is not unique, i.e., if there is more than one set of parameters $(\lambda^*, \mu_1^*, \mu_2^*)$ then, even for large $N$, it is not possible to have a single most dominant tube of paths in $S$. This case occurs when $\mu_1 = \mu_2$. For example, for the $(\lambda = 0.20, \mu_1 = 0.40, \mu_2 = 0.40)$-network, we get $(0.40, 0.40, 0.20)$ and $(0.40, 0.20, 0.40)$ as two sets of optimal parameters. In this limiting case the speed-up due to the change of measure is less than that for the examples shown in Table IV (at least for the small $N'$s that were feasible for us to consider), e.g., for $N = 20$, $\alpha = 1.812 \times 10^{-5}$. After simulating the $(0.40, 0.20, 0.40)$-network for 20000 cycles we obtained $\alpha^* = 1.764 \times 10^{-3}$ as an estimate. Our estimates had intolerable errors for less number of cycles. In summary, if $\mu_1 = \mu_2$ then we have observed speed-ups as compared to the direct
simulations but they are less than that for the examples of Table IV. Furthermore, if \( \mu_1 \) and \( \mu_2 \) are not much apart, then we need larger \( N \)'s to isolate the dominant tube of \( S \). In this case, we may not get very reliable estimates for small \( N \)'s without the simulations for relatively more (as compared to the numbers in Table IV) number of cycles under the changed measure.

b) It follows from the result of Weber [7] that the \((\lambda, \mu_1, \mu_2)\)-network and \((\lambda, \mu_1, \mu_2)\)-network with \( \mu_1 \geq \mu_2 \). Then the corresponding \((\lambda^*, \mu_1^*, \mu_2^*)\)-network as given by our heuristic will be the interchange of \( \lambda \) with \( \mu_2 \). For example, for the \((\lambda = 0.10, \mu_1 = 0.50, \mu_2 = 0.40)\)-network \( \alpha = 1.327 \times 10^{-14} \) for \( N = 25 \). A simulation of the \((0.40, 0.50, 0.10)\)-network gives \( 1.265 \times 10^{-14} \) after 20 000 cycles while a simulation of the \((0.40, 0.10, 0.50)\)-network gives \( 1.114 \times 10^{-14} \) after 40 000 cycles.

2) On \( M/M/1 \) Queues in Parallel: As in the previous observation, we have the limiting case when the set of arguments of the minimization problem is not unique. In this case, it is not clear which one is the optimal set of arguments. For example, the \((\lambda_1 = 0.20, \mu_1 = 0.30, \lambda_2 = 0.20, \mu_2 = 0.30)\)-network has three sets of minimizing arguments, namely, \((0.30, 0.20)(0.20, 0.30), (0.30, 0.20)(0.30, 0.20), (0.30, 0.20)(0.30, 0.20) \). For \( N = 25 \), \( \alpha = 4.156 \times 10^{-4} \). After 20 000 cycles, these networks gave different results. It seems to us that in this limiting case, it might be faster to simulate the network where both the queue service and arrival rates have been interchanged. This observation also suggests that if the traffic intensities of the two queues are not much apart, then we will require larger \( N \)'s to single out the dominant paths of \( S \).

3) On the Network Shown in Fig. 10: If the traffic intensities of the two queues are not much apart, we need larger \( N \)'s for our method of simulation to be effective.

E. Extension to Networks of \( G I/G I/1 \) Queues

In this subsection, we extend the heuristic of the previous four subsections to networks of \( G I/G I/1 \) queues. Observe that for estimating \( \alpha \), we no longer have an embedded Markov chain to work with. Now we have to simulate the network, i.e., by generating various random times (service times and interarrival times).

Consider a general open network of \( G I/G I/1 \) queues shown in Fig. 11. Suppose there are \( d > 0 \) nodes. Let \( 1/\lambda_i, 1 \leq i \leq d \), and \( 1/\mu_i, 1 \leq i \leq d \) denote, respectively, the means of \( A_i \), the interarrival time d.f. of the external input process to the node \( i \) and \( B_i \), the service time d.f. at the node \( i \). Let \( P_{ij} \) denote the probability of routing from the node \( i \) to the node \( j \). By \( p_{ij} \), we denote the probability of leaving the network after the service completion at the node \( i \).

Consider the paths of \( S \) which require \( T \) time units to have the backlog build up to \( N \), have \( 1/\lambda_i \) and \( 1/\mu_i \) average interarrival times and virtual service times, respectively, and have \( P^* = \{p_{ij}^*\} \) as the apparent routing probabilities. Let \( L^* \) and \( M^* \) denote the \( d \)-dimensional vectors \( \{\lambda_i^*\} \) and \( \{\mu_i^*\} \), respectively. Let \( G^* = \{\lambda_i^*\} \) denote the effective rate for these paths which we can find approximately (because \( \mu_i^* \)'s are the virtual service rates) by solving the flow balance equations

\[
\gamma_i^* = \lambda_i^* + \sum_{j=1}^{d} \min(\gamma_j^*, \mu_j^*) \cdot p_{ij}^* \quad 1 \leq i \leq d. \tag{48}
\]

As in the previous subsections, we can argue heuristically to get the following relationship between \( T \), \( G^* \), and \( M^* \).

\[
T = N \cdot R, \text{ where}
\]

\[
R = \frac{1}{\sum_{i=1}^{d} (\gamma_i^* - \mu_i^*) \cdot 1\{\gamma_i^* > \mu_i^*\}}.
\]

Finally, the same line of heuristic gives

\[
\alpha = \sum_{L^*, M^*, P^*} \exp \{-N \cdot H(L^*, M^*, P^*)\} \quad \text{(UTLE)}
\]

where

\[
H(L^*, M^*, P^*) = R \cdot \left( \sum_{i=1}^{d} \frac{1}{\lambda_i^*} \cdot h_{A_i} \left( \frac{1}{\lambda_i^*} \right) + \sum_{i=1}^{d} \frac{1}{\mu_i^*} \cdot h_{B_i} \left( \frac{1}{\mu_i^*} \right) + \sum_{i=1}^{d} \min(\gamma_i^*, \mu_i^*) \cdot K(p_{ij}^*, p_i) \right)
\]

and \( p_i \) and \( p_i \) are the \( i \)-th rows of the matrices \( P^* \) and \( P \), respectively. Hence, for large \( N \),

\[
\alpha = \exp \{-N \cdot H^*\} \quad \text{(UTLE)}
\]

where

\[
H^* = \inf_{L^*, M^*, P^*} H(L^*, M^*, P^*)
\]

with \( G^* \) given by (48). Let \( L^*, M^*, \) and \( P^* \) denote the arguments achieving this infimum. Define new service time distributions \( B_i^* \)'s by

\[
dB_i^* (z) = e^{\theta z} dB_i (z)
\]

where \( \theta \) is such that it satisfies \( \int z dB_i^* (z) = 1/\mu_i^* \). Similarly, define new interarrival time distributions \( A_i^* \)'s by

\[
daA_i^* (z) = e^{\theta z} dA_i (z)
\]

where \( \theta \) is such that it satisfies \( \int z dA_i^* (z) = 1/\lambda_i^* \). Then, for large \( N \) we employ the network of \( G I/G I/1 \) queues with the parameters \( L^*, M^*, \) and \( P^* \) for estimating \( \alpha \).

V. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we have used some techniques inspired by the large deviation theory for obtaining a simulation method for events of excessive backlogs in networks of queues that is much faster than the direct Monte Carlo simulation method. We have seen that the classical large deviation results of Ventzel [12] and Azencott et al. [1] are not directly applicable to networks of queues. The main difficulty arises from the fact that Markov processes describing these networks have discontinuous kernels. To circumvent this difficulty, a heuristic method based on the
work by Borovkov, Ruget, etc., for a GI/GI/1 queue has been
developed for simulation purposes and has also been extended to
open networks of GI/GI/1 queues.

Further work is needed to justify analytically our heuristic
method and also to connect the transient and steady-state
behaviors for rare events in networks of queues.

REFERENCES

grands écarts à la loi des grands nombres," Z. Wahrscheinlich-
[3] M. Cottrell, J.-C. Fort, and G. Malgouyres, "Large deviations and
analysis and simulation; A phase locked loop example," SIAM J.
[6] ---, "Large deviations for the empirical measure of a Markov chain
[7] S. Natarajan, "Large deviation hypotheses testing, and coding for
[10] I. Sanov, "On the probability of large deviation of random variables,"
[12] A. D. Ventcel, "Rough limit theorems on large deviation for Markov

Shyam Parekh (S’82–S’84–M’84–M’86) received the B.E. degree in electrical and electronics
engineering from Birla Institute of Technology and Science, Pilani, India, in 1980, the M.S. degree in
electrical engineering, in 1984, the M.A. degree in statistics in 1984, and the Ph.D. degree in electrical
engineering, in 1986, all from the University of California, Berkeley.

Since 1987, he has been with the Teletraffic Theory and System Performance Department,
AT&T Bell Laboratories, Holmdel, NJ. His current research interests are rare events in stochastic processes and queuing theory.

Dr. Parekh is a member of the IEEE Control Systems Society and the
Information Theory Society.

Jean Walrand (S’71–M’74–M’80) received the ingénieur civil degree in electrical engineering from
the Université de Liège, Liège, Belgium, in 1974 and the Ph.D. degree in electrical engineering from
the University of California, Berkeley, in 1979.

From 1979 to 1981 he taught at Cornell University, Ithaca, NY. Since 1981 he has been
with the Department of Electrical Engineering and Computer Sciences, University of California,
Berkeley. His research interests are in queuing networks, communication networks, stochastic
control, and stochastic processes.

Dr. Walrand is the author of An Introduction to Queuing Networks
(Englewood Cliffs, NJ: Prentice-Hall, 1988). He has served as an Associate
Editor for the IEEE TRANSACTIONS ON AUTOMATIC CONTROL, and is on
the Editorial Board of Systems & Control Letters; Queueing Systems, and
Probability in the Engineering and Informational Sciences.