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 arriving 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 $E_0\{T\}$, where $E_0\{T\}$ 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 $E_0\{T\}$ 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.

Manuscript received September 19, 1986; revised February 15, 1988. Paper recommended by Past Associate Editor, A. Ephremides. This work was supported in part by the National Science Foundation under Grant ECS8421128, by Pacific Bell, and by a MICRO Grant from the State of California.

S. Parekh is with AT&T Bell Laboratories, Holmdel, NJ 07733.

J. Walrand is with the Department of Electrical Engineering and Computer Sciences and the Electronics Research Laboratory, University of California, Berkeley, CA 94720.

IEEE Log Number 8824269.

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, Ruget, etc. (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 Azencott *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 λ and service rate μ such that $\lambda < \mu$. Consider the embedded discrete-time Markov

0018-9286/89/0100-0054$01.00

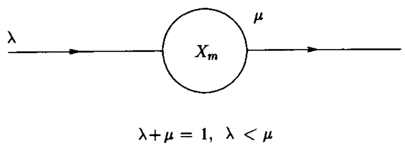


Fig. 1. $M/M/1$ queue.

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 α 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

$$E_0\{T\} \approx \frac{1-\alpha}{\alpha} \cdot E_0\{T_0\} \approx \frac{1}{\alpha} \cdot E_0\{T_0\}$$

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 α. So, from now on, our primary concern will be the estimation of α.

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

$$V_k := 1\{X_m \text{ reaches } N \text{ in cycle } k\}$$

where $1\{B\}$ (or sometimes written 1_B) denotes the indicator of an event B. See Fig. 2. Notice that V_k's are i.i.d. Also notice that, as shown in Fig. 2, we have modified $\{X_m\}$ in that we restart $\{X_m\}$ at 0, if it exceeds N. Clearly, $\alpha = P\{V_k = 1\}$. Here we can find α by the first step method. For this let, for $0 \le i \le N$, 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

$$P_i = \mu \cdot P_{i-1} + \lambda \cdot P_{i+1}, \qquad 1 \le i \le N-1.$$

The solution of these linear equations can be seen to give

$$\alpha = P_1 = \frac{\frac{\mu}{\lambda} - 1}{\left(\frac{\mu}{\lambda}\right)^N - 1}. \tag{1}$$

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_i denote a jump which takes values $+1$ and -1 w.p. λ and μ, respectively. Note that cycle k ends at N with probability α and in this case $Z_1 + Z_2 + \cdots + Z_{J_k} = N - 1$. Similarly, cycle k ends at 0 with probability $1 - \alpha$ and in this case $Z_1 + Z_2 + \cdots + Z_{J_k} = -1$. Then, for cycle k,

$$E\{Z_1 + Z_2 + \cdots + Z_{J_k}\} = \alpha \cdot (N-1) + (1-\alpha) \cdot (-1).$$

Using Wald's identity we identify the left-hand side with

$$E\{J_k\} \cdot E\{Z_i\} = E\{J_k\} \cdot (\lambda - \mu).$$

This gives

$$E\{J_k\} = \frac{1 - N \cdot \alpha}{\mu - \lambda}. \tag{2}$$

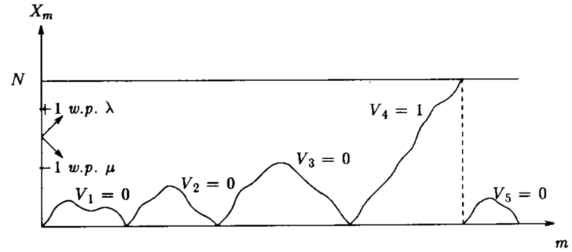


Fig. 2. Realization of $\{X_m\}$.

In the following subsections we present the idea of change of measure for estimating α 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 $E\{V_k\} = \alpha$ and $\text{Var}\{V_k\} = \alpha \cdot (1 - \alpha)$. Suppose we want to ensure that the relative error does not exceed $\epsilon\%$ with probability more than β. We will call such an estimator an (ϵ, β)-confidence estimator. The normal approximation then gives

$$P\{|\alpha_{n_d} - \alpha| > \epsilon \cdot \alpha\} \approx \beta \Leftrightarrow n_d \approx \frac{c^2}{\epsilon^2} \cdot \frac{\text{Var}\{V_k\}}{\alpha^2}$$

where $c = \Phi^{-1}(\beta/2)$, where Φ denotes the distribution function of a Gaussian r.v. with the mean equal to 0 and variance equal to 1. Hence, $n_d \approx \gamma \cdot (1 - \alpha)/\alpha$, where $\gamma = c^2/\epsilon^2$, cycles are necessary to achieve the (ϵ, β)-confidence estimator by a direct simulation. Let T_d denote the units of simulation time required for achieving the (ϵ, β)-confidence estimator by a direct simulation. Then,

$$T_d = E\{J_k\} \cdot n_d.$$

Since $\lambda < \mu$, for large N, $E\{J_k\} \approx 1/\mu - \lambda$ [see (2)]. Hence,

$$T_d \approx \gamma \cdot \frac{1}{\alpha} \cdot \frac{1}{\mu - \lambda}. \tag{3}$$

C. Change of Measure

For estimating α, we propose to consider the $M/M/1$ queue with arrival rate μ and service rate λ, 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_k'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 $E^*\{\ \}$ 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 α. As before, to achieve (ϵ, β)-confidence estimator, now the minimum number of cycles required will be

$$n_c \approx \gamma \cdot \frac{\text{Var}^*\{L_k \cdot V_k\}}{\alpha^2}$$

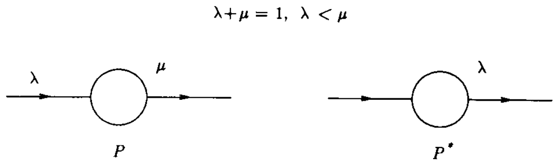


Fig. 3. Change of measure for an $M/M/1$ queue.

where Var*{ } denotes the variance under the measure P^*. Observe that by interchanging λ and μ in (2), we have $E^*\{J_k\} \approx N/\mu$. Let T_c denote the units of simulation time required for achieving the (ϵ, β)-confidence estimator under the changed measure. Then

$$T_c = E^*\{J_k\} \cdot n_c \approx \gamma \cdot \frac{\sigma^2}{\alpha^2} \cdot \frac{N}{\mu} \tag{4}$$

where $\sigma^2 := \mathrm{Var}^*\{L_k \cdot V_k\}$.

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's.

D. 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

$$S \approx \frac{1}{N} \cdot \frac{\alpha}{\sigma^2} \cdot \frac{1}{1 - \frac{\lambda}{\mu}}. \tag{5}$$

Suppose that ω 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, $J_k(\omega) = N + 2 \cdot l - 1$. Let ω_k denote the section of ω that pertains to cycle k. Then, $P\{\omega_k\} = \lambda^{N+l-1} \cdot \mu^l$ and $P^*\{\omega_k\} = \mu^{N+l-1} \cdot \lambda^l$. Therefore,

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

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

$$L_k \equiv \left(\frac{\lambda}{\mu}\right)^{N-1}. \tag{6}$$

Hence,

$$\begin{aligned}
\sigma^2 &= E^*\{(L_k \cdot V_k)^2\} - \alpha^2 \\
&= \left(\frac{\lambda}{\mu}\right)^{N-1} \cdot E^*\{L_k \cdot V_k\} - \alpha^2 \\
&= \left(\frac{\lambda}{\mu}\right)^{N-1} \cdot \alpha - \alpha^2
\end{aligned}$$

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

$$\frac{\sigma^2}{\alpha} \approx \left(\frac{\lambda}{\mu}\right)^N. \tag{7}$$

Substituting (7) in (5), we get

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

TABLE I
EXAMPLE OF CHANGE OF MEASURE

# of cycles (n)	1000	2000	10000
α_n	0.0	0.0	0.0
α_n^*	3.440×10^{-7}	3.520×10^{-7}	3.708×10^{-7}

TABLE II
SIMULATIONS FOR AN $M/M/1$ QUEUE

Method	Direct Simulation			Quick Simulation		
Example-I: $\lambda = 0.20$ $\mu = 0.80$ $N = 15$; $\alpha = 2.794\times10^{-9}$; $\lambda^* = 0.80$ $\mu^* = 0.20$						
# of Cycles (n)	5000	10000	20000	50	100	200
α_n (α_n^*)	0.0	0.0	0.0	2.831×10^{-9}	2.682×10^{-9}	2.663×10^{-9}
CPU Time	2.5*Sec.*	5.4*Sec.*	10.6*Sec.*	0.3*Sec.*	0.6*Sec.*	1.2*Sec.*
Calls to RNG	8550	16712	33624	800	1656	3255
Example-II: $\lambda = 0.30$ $\mu = 0.70$ $N = 20$; $\alpha = 5.826\times10^{-8}$; $\lambda^* = 0.70$ $\mu^* = 0.30$						
# of Cycles (n)	5000	10000	20000	200	300	500
α_n (α_n^*)	0.0	0.0	0.0	6.322×10^{-8}	5.268×10^{-8}	5.955×10^{-8}
CPU Time	3.9*Sec.*	7.6*Sec.*	16.1*Sec.*	2.0*Sec.*	2.4*Sec.*	4.6*Sec.*
Calls to RNG	12492	25426	51052	5598	7084	13684
Example-III: $\lambda = 0.40$ $\mu = 0.60$ $N = 30$; $\alpha = 2.608\times10^{-6}$; $\lambda^* = 0.60$ $\mu^* = 0.40$						
# of Cycles (n)	20000	30000	40000	1000	2000	3000
α_n (α_n^*)	0.0	0.0	0.0	2.910×10^{-6}	2.401×10^{-6}	2.649×10^{-6}
CPU Time	30.2*Sec.*	44.4*Sec.*	56.2*Sec.*	16.4*Sec.*	30.4*Sec.*	43.2*Sec.*
Calls to RNG	105738	151322	195760	47466	82956	127832

E. Example

Consider the $M/M/1$ queue with $\lambda = 0.33$ and $\mu = 0.67$. We want to estimate α for $N = 21$. Equation (1) gives $\alpha = 3.583 \times 10^{-7}$. For the ($\epsilon = 0.05$, $\beta = 0.05$)-confidence estimator, (3) gives $T_d = 1.32 \times 10^{10}$ units (4.42×10^9 cycles), while (4) gives $T_c = 4.96 \times 10^4$ units (1.58×10^3 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 estimates 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.

III. Large Deviation Theory and Optimal Change of Measure

A. A Fundamental Theorem

Theorem 1 (Cramér's Theorem) [11]: Let $\xi_1, \xi_2, \cdots$ be i.i.d. r.v.'s taking values in $\boldsymbol{R}^d$. Let F denote the distribution function (d.f.) of ξ_k 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

$$M(s) := \int_{\boldsymbol{R}^d} \exp \langle s, z \rangle \, dF(z), \qquad s \in \boldsymbol{R}^d$$

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

i) for each closed subset C of $\boldsymbol{R}^d$

$$\limsup_{n \to \infty} \frac{1}{n} \cdot \log P_n\{C\} \leq -\inf_{x \in C} h(x) \text{ and}$$

TABLE III
EMPIRICAL STANDARD DEVIATION FOR AN $M/M/1$ QUEUE

Example-I $\lambda = 0.20$ $\mu = 0.80$ $N = 15$ $\alpha = 2.794\times10^{-9}$ # of Experiments = 20 $\lambda^* = 0.80$ $\mu^* = 0.20$		
# of Cycles (n)	100	200
Empirical Mean ($\hat{m}$)	2.744×10^{-9}	2.794×10^{-9}
Empirical Std. Dev. ($\hat{\sigma}$)	1.150×10^{-10}	1.019×10^{-10}
($\hat{\sigma}/\hat{m}$)$\times$100 %	4.1910 %	3.645 %
Example-II $\lambda = 0.70$ $\mu = 0.30$ $N = 20$ $\alpha = 5.826\times10^{-8}$ # of Experiments = 20 $\lambda^* = 0.70$ $\mu^* = 0.30$		
# of Cycles (n)	300	500
Empirical Mean ($\hat{m}$)	5.856×10^{-8}	5.906×10^{-8}
Empirical Std. Dev. ($\hat{\sigma}$)	2.803×10^{-9}	2.474×10^{-9}
($\hat{\sigma}/\hat{m}$)$\times$100 %	4.786 %	4.190 %
Example-III $\lambda = 0.40$ $\mu = 0.60$ $N = 30$ $\alpha = 2.608\times10^{-6}$ # of Experiments = 20 $\lambda^* = 0.60$ $\mu^* = 0.40$		
# of Cycles (n)	2000	3000
Empirical Mean ($\hat{m}$)	2.743×10^{-6}	2.680×10^{-6}
Empirical Std. Dev. ($\hat{\sigma}$)	2.652×10^{-7}	2.409×10^{-7}
($\hat{\sigma}/\hat{m}$)$\times$100 %	9.669 %	8.989 %

ii) for each open subset G of $\boldsymbol{R}^d$

$$\liminf_{n\to\infty} \frac{1}{n} \cdot \log P_n\{G\} \geq -\inf_{x\in G} h(x)$$

where the function h, called Cramér or Legendre transform, is defined as

$$h(y) = \sup_{s\in \boldsymbol{R}^d} [\langle s, y\rangle - \log M(s)], \qquad y \in \boldsymbol{R}^d. \qquad \blacksquare \quad (8)$$

Interested readers can find a simple proof of this theorem in the monograph by Varadhan [11]. This theorem gives the rate of convergence for the weak law of large numbers (WLLN). This is quite easily seen from an equivalent statement of this theorem in $\boldsymbol{R}^1$. For this, let $a > m$, then

$$\lim_{n\to\infty} \frac{1}{n} \cdot \log P\left\{\frac{\xi_1+\xi_2+\cdots+\xi_n}{n} > \alpha\right\} = -h(a). \qquad (9)$$

Intuitively, (9) states that $P\{S_n/n \approx a\} = \exp(-n \cdot h(a) + o(n))$, where $S_n = \xi_1 + \xi_2 + \cdots + \xi_n$.

Next, we list some properties of the Cramér transform defined in (8). We define

$$l(s) := \log M(s), \qquad s \in \boldsymbol{R}^d.$$

$P1$: h is convex and nonnegative lower semicontinuous.
$P2$: For each $b < \infty$, the set $\{u/h(u) \leq b\}$ is compact in $\boldsymbol{R}^d$.
$P3$: $h(y)$ has its minimum value 0 at $y = m$, i.e., $h(m) = 0$.
$P4$: l and h are convex dual of each other,

$$l(s) = \sup_u [\langle s, u\rangle - h(u)].$$

$P5$: Let V denote the interior of the set $\{s \in \boldsymbol{R}^d/M(s) < \infty\}$ and U denote the set $\{u \in \boldsymbol{R}^d/h(u) < \infty\}$. The derivatives h' and l' are reciprocals of each other, i.e.,

$$h'(l'(s)) = s, \qquad s \in V$$

and

$$l'(h'(u)) = u, \qquad u \in U.$$

Finally, we give a few examples of the Cramér transform that will be useful to us in the subsequent sections.

$E1$: ξ_k's take values $+1$ and -1 w.p. p_1 and p_2, respectively. Then

$$\begin{aligned} h(u) &= \frac{1+u}{2} \cdot \log\left(\frac{1+u}{2 \cdot p_1}\right) + \frac{1-u}{2} \\ &\quad \cdot \log\left(\frac{1-u}{2 \cdot p_2}\right), \qquad -1 \leq u \leq 1, \\ &= \infty, \qquad \text{otherwise.} \end{aligned} \qquad (10)$$

$E2$: ξ_k's are exponentially distributed with the parameter $\nu > 0$. Then,

$$\begin{aligned} h(u) &= \nu \cdot u - 1 - \log(\nu \cdot u), \qquad u > 0, \\ &= \infty, \qquad \text{otherwise.} \end{aligned} \qquad (11)$$

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 $\{X_n^\epsilon\} \in \boldsymbol{R}^d$ given by

$$\begin{aligned} X_0^\epsilon &= x_0, \\ X_{n+1}^\epsilon &= X_n^\epsilon + \epsilon \cdot V(X_n^\epsilon, \xi_n), \qquad n \geq 0 \end{aligned} \qquad (12)$$

where $\epsilon > 0$ is the parameter defining the Markov chain $\{X_n^\epsilon\}$, x_0 is the initial value, $V(\cdot, \cdot)$ is a function from $\boldsymbol{R}^d \times \boldsymbol{R}^1 \to \boldsymbol{R}^d$ and ξ_n's are i.i.d. r.v.'s. We are interested in analyzing $\{X_n^\epsilon\}$ when $\epsilon \to 0$. Let F_x denote the d.f. of $V(x, \xi_n)$. Let

$$m(x) = \int_{\boldsymbol{R}^d} z\,dF_x(z)$$

be the mean of F_x,

$$M_x(s) := \int_{\boldsymbol{R}^d} \exp\langle s, z\rangle\, dF_x(z)$$

be its Laplace transform, $l_x(s) := \log M_x(s)$ and

$$h_x(u) = \sup_{s\in\boldsymbol{R}^d} [\langle s, u\rangle - l_x(s)]$$

be its Cramér transform. Assume the following:

$A1$: $M_x(s) < \infty$ in a neighborhood of 0 for each $x \in \boldsymbol{R}^d$.
$A2$: $d(F_{x_1}, F_{x_2}) \leq c \cdot \|x_1 - x_2\|$, where d is the Prohorov distance (see [2]) and $c > 0$ is a constant, i.e., F_x is Lipschitz smooth in x.

Next, construct continuous-time paths from the realizations of $\{X_n^\epsilon\}$. To do this, at the epochs

$$t = n \cdot \epsilon \text{ define } X^\epsilon(t) := X_n^\epsilon \qquad (13)$$

and interpolate piecewise linearly. Let $\boldsymbol{C}_T$ denote the set of the continuously piecewise differentiable functions ϕ:[0, T] $\to \boldsymbol{R}^d$ such that $\phi(0) = x_0$ is fixed. Let P^ϵ denote the measure induced by the Markov chain $\{X_n^\epsilon\}$ on the Borel σ-field Σ of $\boldsymbol{C}_T$ endowed with the Skorohod topology [2]. Define the action integral

$$I(\phi) := \int_0^T h_{\phi(t)}(\phi'(t))\, dt.$$

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

Theorem 2 (Ventsel) [12]: Let ϕ be a path in $\boldsymbol{C}_T$. Define a tube of diameter d around ϕ, $T_d(\phi)$, as the set of trajectories $\eta(t)$'s such that

$$|\eta(t) - \phi(t)| < d, \qquad \text{for all } t \in [0, T].$$

Then, there exists δ_0 such that, for $0 < \delta < \delta_0$,

$$\lim_{\epsilon\to 0} (-\epsilon \cdot \log P^\epsilon\{T_\delta(\phi)\}) = I(\phi) + e(\delta)$$

with $\lim_{\delta\to 0} e(\delta) = 0$. ■

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

Corollary 1 (Ventsel) [12]: Let $S \in \Sigma$ be such that

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

then

$$\lim_{\epsilon\to 0} (-\epsilon \cdot \log P^\epsilon\{S\}) = \inf_{\phi\in S} I(\phi). \qquad \square \quad (15)$$

Corollary 1 suggests that

$$P^\epsilon\{S\} \approx \sum_k P^\epsilon\{T_\delta(\phi_k)\} \approx \sum_k \exp\left(-\frac{1}{\epsilon} \cdot I(\phi_k)\right)$$
$$\approx \exp\left(-\frac{1}{\epsilon} \cdot \inf_{\phi\in S} I(\phi)\right) \text{ (UTLE)}$$

where the second approximation follows from Theorem 1, the last one follows from Corollary 1, and UTLE is the acronym for up to logarithmic 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 $\text{argmin}_{\phi\in S} I(\phi)$ by ϕ_{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

$$C(x) := \inf\left\{\int_0^{T(\phi)} H(\phi(t), \phi'(t))\, dt/\phi(0) = x, \; \phi \in C,\, T(\phi) < \infty\right\} \quad (16)$$

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

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

$$\frac{\partial^2 C}{\partial x_{(i)}\partial x_{(j)}} = \frac{\partial^2 C}{\partial x_{(j)}\partial x_{(i)}}, \qquad 1 \le i \le d,\ 1 \le j \le d.$$

Let us define

$$\theta_{(i)}(x) := -\frac{\partial C}{\partial x_{(i)}}(x), \qquad 1 \le i \le d. \quad (17)$$

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

$$L(x, \theta(x)) = 0 \quad (18)$$

and ϕ_{opt} is a solution of the following system of differential equations:

$$\frac{d\theta_{(i)}}{dt} = -\frac{\partial L}{\partial x_{(i)}}(x, \theta), \qquad 1 \le i \le d, \quad (19)$$

$$\frac{dx_{(i)}}{dt} = \frac{\partial L}{\partial \theta_{(i)}}(x, \theta), \qquad 1 \le i \le d. \quad (20)$$

Proof: First, we expand $C(x)$ as

$$C(x) = \inf_v \{H(x, v) \cdot \Delta t + C(x + v \cdot \Delta t) + o(\Delta t)\}$$
$$= \inf_v \{H(x, v) \cdot \Delta t + C(x) - \sum_{i=1}^d v_{(i)} \cdot \theta_{(i)}(x) \cdot \Delta t + o(\Delta t)\}$$

where we have used the definition of θ in (17). Canceling $C(x)$ from both the sides, dividing by Δt, and letting $\Delta t \to 0$, we get

$$\inf_v \{H(x, v) - \sum_{i=1}^d v_{(i)} \cdot \theta_{(i)}(x)\} = 0,$$

i.e.,

$$\sup_v \{\langle\theta, v\rangle - H(x, v)\} = 0. \quad (21)$$

Using (21) and the convex duality property P4 of Cramér transform, Section III-A, we get

$$L(x, \theta(x)) = 0.$$

Suppose that the supremum in (21) is achieved at $\bar{v}$, then by differentiating, we get

$$\theta_{(i)} = \frac{\partial H}{\partial v_{(i)}}(x, \bar{v}).$$

Now using the reciprocity property of l' and h', property P5 of the Cramér transform, Section III-A, along ϕ_{opt}, we get

$$\bar{v}_{(i)} = \frac{\partial L}{\partial \theta_{(i)}}(x, \theta), \quad 1 \le i \le d. \quad (22)$$

Observe from (18) that

$$0 = \frac{dL}{dx_{(i)}}(x, \theta(x))$$
$$= \frac{\partial L}{\partial x_{(i)}}(x, \theta) + \sum_{k=1}^d \frac{\partial L}{\partial \theta_{(k)}}(x, \theta) \cdot \frac{\partial \theta_{(k)}}{\partial x_{(i)}}(x). \quad (23)$$

But, along ϕ_{opt},

$$\frac{d\theta_{(i)}}{dt} = \sum_{k=1}^d \frac{\partial \theta_{(i)}}{\partial x_{(k)}}(x) \cdot \frac{dx_{(k)}}{dt}$$
$$= \sum_{k=1}^d \frac{\partial \theta_{(i)}}{\partial x_{(k)}}(x) \cdot \frac{\partial L}{\partial \theta_{(k)}}(x, \theta) \quad (24)$$

by using (22). Now by the assumption regarding smoothness of $C(x)$ and the definition of $\theta(x)$ in (17), we get

$$\frac{\partial \theta_{(i)}}{\partial x_{(k)}}(x) = \frac{\partial \theta_{(k)}}{\partial x_{(i)}}(x).$$

Using this in (24), we have

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

Now using (23), along ϕ_{opt}, we get

$$\frac{d\theta_{(i)}}{dt} = -\frac{\partial L}{\partial x_{(i)}}(x, \theta), \qquad 1 \le i \le d.$$

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 $x(T) \in \partial A$ have ϕ_{opt} as a solution. To solve for ϕ_{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 θ. For this, define a new probability measure F_x^* from F_x as

$$dF_x^*(z) := \frac{e^{\langle \theta_x, z\rangle}\, dF_x(z)}{M_x(\theta_x)} \tag{25}$$

where the parameter $\theta_x \in \boldsymbol{R}^d$. This is called the *exponential change of measure* with the parameter θ_x.

Suppose we want to select θ_x along ϕ_{opt} in such a way that

$$\phi'_{\text{opt}}(t) = m^*(\phi_{\text{opt}}(t)) \tag{26}$$

where $m^*(x)$ denotes the mean of F_x^*. Then,

$$m^*(x) = \int_{\boldsymbol{R}^d} z\,dF_x^*(z) = \frac{\int_{\boldsymbol{R}^d} z \cdot e^{\langle \theta_x, z\rangle}}{M_x(\theta_x)} = \frac{M'_x(\theta_x)}{M_x(\theta_x)} = l'_x(\theta_x). \tag{27}$$

Equations (26) and (27) indicate that the parameter of the exponential change of measure that makes the trajectory ϕ'_{opt} most likely satisfies

$$\phi'_{\text{opt}}(t) = l'_{\phi_{\text{opt}}(t)}(\theta_{\phi_{\text{opt}}(t)}). \tag{28}$$

Recalling our notation that $L(x, \theta) = l_x(\theta)$ and comparing (19) and (28), it is clear that the variable θ 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. $\{X_n, n = 0, 1, 2, \cdots\}$ and let (Ω, Σ, 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 (Ω, Σ) such that P is absolutely continuous with respect to P'. Denote the Radon–Nikodym derivative (likelihood ratio) by $L := dP/dP'$. We consider α_n and α'_n as two convergent and unbiased estimators of α, where

$$\alpha_n := \frac{1}{n} \cdot \sum_{i=1}^{n} 1_S(\omega_i) \tag{29}$$

and

$$\alpha'_n = \frac{1}{n} \cdot \sum_{i=1}^{n} 1_S(\omega_i) \cdot L(\omega_i). \tag{30}$$

Here ω_i's are the i.i.d. outcomes of experiments on (Ω, Σ, P). As discussed in Section II α_n is more efficient than α'_n if and only if $\text{Var}'\{\alpha'_1\} < \text{Var}\,\{\alpha_1\}$, which will be the case if and only if

$$\int_S L^2(\omega)\, dP'(\omega) < \alpha. \tag{31}$$

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

In the previous section we discussed the Markov chain $\{X_n^\epsilon\} \in \boldsymbol{R}^d$, 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 $P^{\epsilon *}$, obtained by an exponential change of measure, from P^ϵ. Their theorem is presented in [3] for the case of $\boldsymbol{R}^1$. However, it can be generalized to the case of $\boldsymbol{R}^d$. It is assumed that the mean drift function $\psi(x) := E\{V(X_n^\epsilon, \xi_n)/X_n^\epsilon = x\}$ is such that the O.D.E., $x'(t) = \psi(x(t))$, with $x(0)$ specified, has 0 as a stable equilibrium point. See [3] for details.

Suppose that we want to estimate, for small $\epsilon > 0$, $P_0^\epsilon(S)$, probability of the event

$$S := \{\omega/\{X_n^\epsilon\} \text{ exceeds 1 before hitting } 0\}$$

given that $X_0^\epsilon = 0$. Let us define a probability measure $P^{\epsilon *}$ as the resultant measure when F_x^* is taken as defined by (25), with θ_x being the solution of

$$M_x(\theta_x) = 1, \qquad \theta_x > 0. \tag{32}$$

The probability measure $P^{\epsilon *}$ 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 $X_n^\epsilon \in \boldsymbol{R}^1$, defined in (12), assumptions A1 and A2 hold. Then among all the exponential changes of measure, the transformation $P^\epsilon \to P^{\epsilon *}$ is asymptotically optimal in the sense of the variance, i.e., for $P^{\epsilon *}$

$$\lim_{\epsilon \to 0} \int_S L^2(\omega)\, dP^{\epsilon *}(\omega)$$

where $L = dP^\epsilon/dP^{\epsilon *}$ is minimum. ∎

D. Applications and Difficulties

Consider an open Jackson network of $d > 0$ nodes with infinite buffers. Let $\{X_n, n = 0, 1, 2, \cdots\} \in \boldsymbol{R}^d$ 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 $X_n = (X_{n(1)}, X_{n(2)}, \cdots, X_{n(d)}) \in \boldsymbol{R}^d$ (actually, $X_n \in N^d$). 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, $x_{(1)} + x_{(2)} + \cdots + x_{(d)} \ge N$, before hitting 0, $x_{(1)} = x_{(2)} = \cdots = x_{(d)} = 0$. We want to estimate the probability $\alpha := P_0\{S\}$, the probability of S given that $X_0 = 0$.

We can represent $\{X_n\}$ as

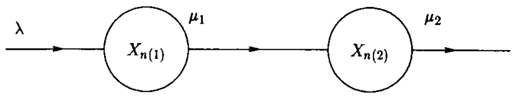
$$X_0 = x_0,$$
$$X_{n+1} = X_n + V(X_n, \xi_n), \qquad n \ge 0 \tag{33}$$

where $V(x, \xi_n)$ 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 $\lambda < \mu_2$. We also assume, without any loss of generality, that $\lambda + \mu_1 + \mu_2 = 1$. For simplicity, we will refer to such a system by a (λ, μ_1, μ_2)-network. Now $\{X_n\}$ is a Markov chain in $\boldsymbol{R}^2$ defined by (33), where the distributions of $V(\cdot, \xi_n)$ 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 $X_n^N := X_n/N$. Then,

$$\begin{aligned} X_{n+1}^N &= X_n^N + \frac{1}{N} \cdot V(X_n, \xi_n) = X_n^N + \frac{1}{N} \cdot V(N \cdot X_n^N, \xi_n) \\ &= X_n^N + \frac{1}{N} \cdot V(X_n^N, \xi_n). \end{aligned} \tag{34}$$

The last equality follows from the fact that in Jackson networks the distributions of $V(x, \xi_n)$ and $V(c \cdot x, \xi_n)$ 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 $\{X_n^N\}$ we are interested in estimating $\alpha = P_0\{S^N\}$, where S^N is the set of the realizations of $\{X_n^N\}$ that reach the region of the state space where the sum of its coordinates exceeds 1.



$\lambda+\mu_1+\mu_2 = 1,\ \lambda < \mu_1,\ \lambda < \mu_2$

Fig. 4. $M/M/1$ queues in tandem.

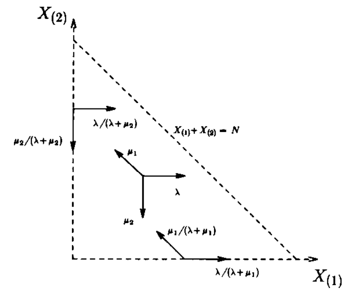


Fig. 5. Jump distributions of $M/M/1$ queues in tandem.

M/M/1 Queue: Let X_n denote the backlog of a stable $M/M/1$ queue with rates λ and μ. For $\{X_n\}$, note that

$$M_x(s)=\lambda\cdot e^{s}+\mu\cdot e^{-s}, \qquad x>0.$$

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

$$\theta_x=\log\left(\frac{\mu}{\lambda}\right), \qquad x>0. \tag{35}$$

Equation (20) gives

$$\phi'_{\text{opt}}=\mu-\lambda. \tag{36}$$

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

$$h_{\phi(t)}(\phi'(t))=(\mu-\lambda)\cdot\log\left(\frac{\mu}{\lambda}\right), \qquad t>0.$$

Now we can use Corollary 1, Section III-B, to evaluate $P_0\{S\}$. Noting that, for ϕ_{opt} 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 $P_0\{S\} \approx (\lambda/\mu)^N$ (UTLE). Observe that this matches well with the exact expression for $P_0\{S\}$ given by (1). Also observe that θ_x given by (40) gives the exponential change of measure [see (25)] that corresponds to the $M/M/1$ queue with arrival rate μ and service rate λ (see Fig. 3).

M/M/1 Queues in Tandem: We consider a (λ, μ_1, μ_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 $x_{(1)}$-axis (second queue empty) and $x_{(2)}$-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 ($x_{(1)}$-axis and $x_{(2)}$-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 $\{X_n^N\}$, this construction is illustrated in Fig. 6. If such a construction were indeed valid, we could use (18), (19), and (20) to find ϕ_{opt} 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.

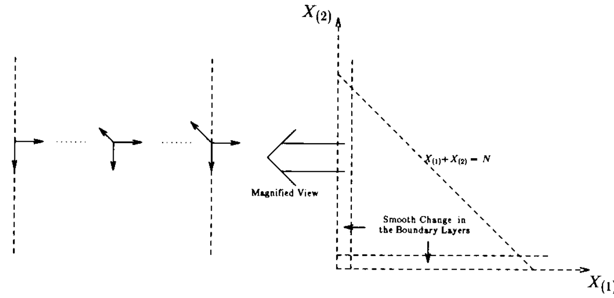


Fig. 6. Boundary layer construction.

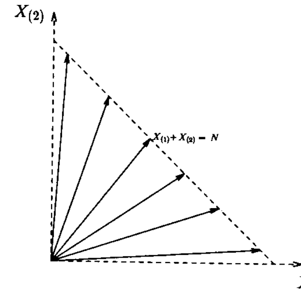


Fig. 7. ϕ_{opt}, neglecting the boundaries.

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 $\phi'(t)$ is constant. Hence, ϕ_{opt} will be one of the rays through R (see Fig. 7). Solving $l_x(\theta) = 0$ [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 θ can be seen to give the (μ_2, μ_1, λ)-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 step equations numerically to be 3.759×10^{-4}. If we simulate the (0.50, 0.30, 0.20)-network, as suggested by the above discussion, we get $\tilde{\alpha}_{1000} = 8.388 \times 10^{-5}$, while simulating the (0.30, 0.20, 0.50)-network we get $\alpha^*_{1000} = 3.595 \times 10^{-4}$. 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 heuristic explanations presented

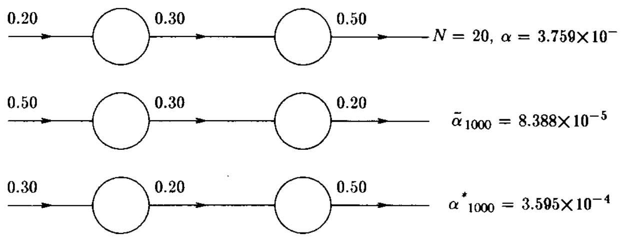


Fig. 8. Comparison of changes of measure.

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_d and h_d 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 α, 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., $\alpha = P_0\{S\}$.

Let X_i^d denote the ith i.i.d. copy of a random variable distributed with the d.f. D. Then, X_i^a denotes the ith interarrival time and X_i^b denotes the ith 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,

$$\begin{aligned} &P\{X_1^a + \cdots + X^a_{\lambda' \cdot T} \approx T\} \\ &= P\left\{\frac{X_1^a + \cdots + X^a_{\lambda'\cdot T}}{\lambda' \cdot T} \approx \frac{1}{\lambda'}\right\} \\ &\approx \exp\left(-\lambda' \cdot T \cdot h_a\left(\frac{1}{\lambda'}\right)\right) \text{ (UTLE)} \end{aligned}$$

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

$$\begin{aligned} &P\{X_1^b + \cdots + X^b_{\mu' \cdot T} \approx T\} \\ &\approx \exp\left(-\mu' \cdot T \cdot h_b\left(\frac{1}{\mu'}\right)\right) \text{ (UTLE)}. \end{aligned}$$

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

$$\begin{aligned} \alpha &\approx \sum_T \sum_{\substack{\lambda' > \mu' \geq 0 \\ N = T\cdot(\lambda' - \mu')}} \exp\left\{-T \cdot \left(\lambda' \cdot h_a\left(\frac{1}{\lambda'}\right) + \mu' \cdot h_b\left(\frac{1}{\mu'}\right)\right)\right\} \text{ (UTLE)} \\ &= \sum_{\lambda' > \mu' \geq 0} \exp\left\{-\frac{N}{\lambda' - \mu'} \cdot \left(\lambda' \cdot h_a\left(\frac{1}{\lambda'}\right) + \mu' \cdot h_b\left(\frac{1}{\mu'}\right)\right)\right\}. \end{aligned}$$

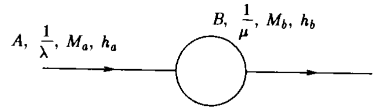


Fig. 9. $GI/GI/1$ queue.

Hence, for large N,

$$\alpha \approx \exp\left\{-N \cdot \inf_{\lambda' > \mu' \geq 0} \left[\frac{1}{\lambda' - \mu'} \cdot \left(\lambda' \cdot h_a\left(\frac{1}{\lambda'}\right) + \mu' \cdot h_b\left(\frac{1}{\mu'}\right)\right)\right]\right\} \text{ (UTLE)}. \tag{37}$$

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 λ' and μ' and equate the results to 0. This gives

$$\begin{aligned} h_a\left(\frac{1}{\lambda'}\right) + h_b\left(\frac{1}{\mu'}\right) &= \left(\frac{1}{\lambda'} - \frac{1}{\mu'}\right) \cdot h_a'\left(\frac{1}{\lambda'}\right) \\ &= \left(\frac{1}{\mu'} - \frac{1}{\lambda'}\right) \cdot h_b'\left(\frac{1}{\mu'}\right). \end{aligned} \tag{38}$$

Suppose that λ^* and μ^* achieve the infimum. Then, from (38),

$$-h_a'\left(\frac{1}{\lambda^*}\right) = h_b'\left(\frac{1}{\mu^*}\right) = \theta^* \text{ (say)}. \tag{39}$$

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

$$\theta^* \cdot \frac{1}{\lambda^*} + h_a\left(\frac{1}{\lambda^*}\right) = \theta^* \cdot \frac{1}{\mu^*} - h_b\left(\frac{1}{\mu^*}\right). \tag{40}$$

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

$$\log M_b(\theta^*) = \theta^* \cdot \frac{1}{\mu^*} - h_b\left(\frac{1}{\mu^*}\right). \tag{41}$$

Therefore,

$$\log M_a(-\theta^*) = -\log M_b(\theta^*), \tag{42}$$

i.e., the conditions for determining θ^* are

$$\theta^* > 0 \text{ and } M_b(\theta^*) \cdot M_a(-\theta^*) = 1. \tag{43}$$

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

$$\alpha \approx \exp(-N \cdot \log M_b(\theta^*)) \text{ (UTLE)}. \tag{44}$$

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 θ_a^* satisfies

$$dA^*(z) = \frac{e^{\theta_a^* \cdot z}\, dA(z)}{M_a(\theta_a^*)}$$

and

$$\frac{1}{\lambda^*}=\int \frac{z \cdot e^{\theta_a^* \cdot z} dA(z)}{M_a(\theta_a^*)}=\frac{d}{d\theta} \log M_a(\theta_a^*).$$

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

$$\theta_a^* = -\theta^*.$$

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 θ_b^* can be seen to satisfy

$$\theta_b^* = \theta^*.$$

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 λ^*, μ^*, and P^* suggest that, for large N,

$$\frac{dP}{dP^*} \ll 1$$

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

M/M/1 Example: Let λ and μ ($0 < \lambda < \mu$) denote arrival and service rates. If D is the exponential d.f. with the mean $1/v$, then we denote by M_v and h_v its Laplace and Cramér transforms, respectively. Recall that

$$M_\nu(s)=\frac{\nu}{\nu-s}, \qquad s<\nu,$$
$$=\infty, \qquad \text{otherwise}.$$

Equation (43) gives

$$\theta^*>0 \text{ and } \frac{\lambda}{\lambda+\theta^*} \cdot \frac{\mu}{\mu-\theta^*}=1. \tag{45}$$

It is easily checked that the solution of (45) is

$$\theta^*=\theta-\lambda.$$

Then, (45) gives

$$\alpha \approx \left(\frac{\lambda}{\mu}\right)^N \text{ (UTLE)}.$$

Observe that this matches well with the exact expression for α 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 α by simulations is the one that corresponds to the interchange of λ and μ.

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 $\{X_n, n = 0, 1, 2, \cdots\} \in \mathbf{R}^d$ 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 $\alpha \equiv P_0\{S\}$, 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 $\{X_n\} \in \mathbf{R}^2$, 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 λ', μ_1', and μ_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

$$\alpha \approx \sum_{\substack{\lambda'>0,\mu_1'\ge 0,\mu_2'\ge 0\\ \lambda'+\mu_1'+\mu_2'=1\\ \lambda'>\mu_1' \text{ or } \lambda'>\mu_2'}} \exp\left\{-T(\lambda',\mu_1',\mu_2') \cdot \left(\lambda' \cdot h_\lambda\left(\frac{1}{\lambda'}\right)+\mu_1' \cdot h_{\mu_1}\left(\frac{1}{\mu_1'}\right) + \mu_2' \cdot h_{\mu_2}\left(\frac{1}{\mu_2'}\right)\right)\right\} \text{ (UTLE)}$$

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 λ', μ_1', and μ_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 $\lambda' > \mu_2'$, $T(\lambda', \mu_1', \mu_2') \approx N \cdot R(\lambda', \mu_1', \mu_2')$, where

$$R=\begin{cases} 1/(\lambda'-\mu_1'), & \text{if } \lambda'>\mu_1' \text{ and } \mu_1' \le \mu_2', \\ 1/(\lambda'-\mu_2'), & \text{otherwise.}\end{cases}$$

Therefore, for large N,

$$\alpha \approx \exp\left\{-N \cdot \inf_{\substack{\lambda'>0,\mu_1'\ge 0,\mu_2'\ge 0\\ \lambda'+\mu_1'+\mu_2'=1\\ \lambda'>\mu_1' \text{ or } \lambda'>\mu_2'}} \left[R(\lambda',\mu_1',\mu_2') \cdot \left(\lambda' \cdot h_\lambda\left(\frac{1}{\lambda'}\right)+\mu_1' \cdot h_{\mu_1}\left(\frac{1}{\mu_1'}\right) + \mu_2' \cdot h_{\mu_2}\left(\frac{1}{\mu_2'}\right)\right)\right]\right\} \text{ (UTLE)}. \tag{46}$$

Numerical minimization gives λ^*, μ_1^*, and μ_2^* that correspond to the interchange of λ with the smallest of μ_1 and μ_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 α, it will be faster to simulate the embedded Markov chain of the (λ^*, μ_1^*, μ_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 λ_i and μ_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 (λ_1, $\mu_1 | \lambda_2$, μ_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 λ_1^*, μ_1^*, λ_2^*, and μ_2^* that correspond to the

TABLE IV
SIMULATIONS FOR $M/M/1$ QUEUES IN TANDEM

Method	Direct Simulation			Quick Simulation		
Example-I $\lambda = 0.05$ $\mu_1 = 0.10$ $\mu_2 = 0.85$ $N = 15$; $\alpha = 3.459\times 10^{-5}$ *CPU Time* = 61.1*Sec.*; $\lambda^* = 0.10$ $\mu_1^* = 0.05$ $\mu_2^* = 0.85$						
# of Cycles (n)	10000	20000	40000	200	500	1000
α_n (α_n^*)	0.0	0.0	0.0	3.338×10^{-5}	3.577×10^{-5}	3.448×10^{-5}
CPU Time	17.0*Sec.*	33.3*Sec.*	69.9*Sec.*	2.5*Sec.*	5.6*Sec.*	10.4*Sec.*
Calls to RNG	52769	109573	216395	5512	13595	26303
Example-II $\lambda = 0.10$ $\mu_1 = 0.50$ $\mu_2 = 0.40$ $N = 13$; $\alpha = 2.104\times 10^{-7}$ *CPU Time* = 29.6*Sec.*; $\lambda^* = 0.40$ $\mu_1^* = 0.50$ $\mu_2^* = 0.10$						
# of Cycles (n)	20000	30000	50000	700	1000	1500
α_n (α_n^*)	0.0	0.0	0.0	1.979×10^{-7}	2.159×10^{-7}	1.594×10^{-7}
CPU Time	25.4*Sec.*	38.1*Sec.*	67.3*Sec.*	7.5*Sec.*	11.7*Sec.*	16.9*Sec.*
Calls to RNG	79816	120270	200917	18920	27529	40763
Example-III $\lambda = 0.20$ $\mu_1 = 0.30$ $\mu_2 = 0.50$ $N = 20$; $\alpha = 3.759\times 10^{-4}$ *CPU Time* = 310.3*Sec.*; $\lambda^* = 0.30$ $\mu_1^* = 0.20$ $\mu_2^* = 0.50$						
# of Cycles (n)	5000	10000	20000	300	500	1000
α_n (α_n^*)	2.000×10^{-4}	1.000×10^{-4}	7.500×10^{-4}	3.848×10^{-4}	3.734×10^{-4}	3.595×10^{-4}
CPU Time	24.6*Sec.*	48.1*Sec.*	92.3*Sec.*	7.5*Sec.*	12.2*Sec.*	23.0*Sec.*
Calls to RNG	72234	144913	286539	18006	29854	56489

TABLE VI
SIMULATIONS FOR $M/M/1$ QUEUES IN PARALLEL

Method	Direct Simulation			Quick Simulation		
Example-I $\lambda_1 = 0.10$ $\mu_1 = 0.20$ $\lambda_2 = 0.30$ $\mu_2 = 0.40$ $N = 23$; $\alpha = 1.213\times 10^{-8}$ *CPU Time* = 624.4*Sec.*; $\lambda_1^* = 0.10$ $\mu_1^* = 0.20$ $\lambda_2^* = 0.40$ $\mu_2^* = 0.30$						
# of Cycles (n)	10000	20000	30000	2000	2500	3000
α_n (α_n^*)	1.000×10^{-8}	1.350×10^{-8}	1.267×10^{-8}	1.228×10^{-8}	1.188×10^{-8}	1.120×10^{-8}
CPU Time	57.0*Sec.*	120.4*Sec.*	173.9*Sec.*	41.6*Sec.*	52.5*Sec.*	60.4*Sec.*
Calls to RNG	160278	321031	482256	99576	127937	145472
Example-II $\lambda_1 = 0.10$ $\mu_1 = 0.40$ $\lambda_2 = 0.15$ $\mu_2 = 0.35$ $N = 18$; $\alpha = 6.935\times 10^{-7}$ *CPU Time* = 158.1*Sec.*; $\lambda_1^* = 0.10$ $\mu_1^* = 0.40$ $\lambda_2^* = 0.35$ $\mu_2^* = 0.15$						
# of Cycles (n)	10000	20000	50000	3000	5000	6000
α_n (α_n^*)	0.0	0.0	0.0	7.016×10^{-7}	6.264×10^{-7}	6.483×10^{-7}
CPU Time	16.7*Sec.*	36.0*Sec.*	84.1*Sec.*	42.3*Sec.*	68.0*Sec.*	80.7*Sec.*
Calls to RNG	46758	94602	236618	93984	150190	179600
Example-III $\lambda_1 = 0.08$ $\mu_1 = 0.12$ $\lambda_2 = 0.20$ $\mu_2 = 0.60$ $N = 23$; $\alpha = 4.635\times 10^{-5}$ *CPU Time* = 635.0*Sec.*; $\lambda_1^* = 0.12$ $\mu_1^* = 0.08$ $\lambda_2^* = 0.20$ $\mu_2^* = 0.60$						
# of Cycles (n)	10000	20000	50000	3000	4000	5000
α_n (α_n^*)	0.0	5.000×10^{-5}	0.0	4.725×10^{-5}	4.435×10^{-5}	4.725×10^{-5}
CPU Time	31.2*Sec.*	66.9*Sec.*	154.3*Sec.*	68.1*Sec.*	85.0*Sec.*	110.9*Sec.*
Calls to RNG	87180	186083	417832	153426	202164	262938

TABLE V
EMPIRICAL STANDARD DEVIATION FOR $M/M/1$ QUEUES IN TANDEM

Example-I $\lambda = 0.05$ $\mu_1 = 0.10$ $\mu_2 = 0.85$ $N = 15$; $\alpha = 3.459\times 10^{-5}$ *# of Experiments* = 20; $\lambda^* = 0.10$ $\mu_1^* = 0.05$ $\mu_2^* = 0.85$		
# of Cycles (n)	500	1000
Empirical Mean ($\hat{m}$)	3.493×10^{-5}	3.385×10^{-5}
Empirical Std. Dev. ($\hat{\sigma}$)	8.971×10^{-7}	7.985×10^{-7}
$(\hat{\sigma}/\hat{m})\times 100$ %	2.568 %	2.359 %
Example-II $\lambda = 0.10$ $\mu_1 = 0.50$ $\mu_2 = 0.40$ $N = 13$; $\alpha = 2.104\times 10^{-7}$ *# of Experiments* = 20; $\lambda^* = 0.40$ $\mu_1^* = 0.50$ $\mu_2^* = 0.10$		
# of Cycles (n)	700	1500
Empirical Mean ($\hat{m}$)	2.223×10^{-7}	2.116×10^{-7}
Empirical Std. Dev. ($\hat{\sigma}$)	2.320×10^{-8}	1.610×10^{-8}
$(\hat{\sigma}/\hat{m})\times 100$ %	10.437 %	7.608 %
Example-III $\lambda = 0.20$ $\mu_1 = 0.30$ $\mu_2 = 0.50$ $N = 20$; $\alpha = 3.759\times 10^{-4}$ *# of Experiments* = 20; $\lambda^* = 0.30$ $\mu_1^* = 0.20$ $\mu_2^* = 0.50$		
# of Cycles (n)	500	1000
Empirical Mean ($\hat{m}$)	3.765×10^{-4}	3.805×10^{-4}
Empirical Std. Dev. ($\hat{\sigma}$)	2.481×10^{-5}	2.095×10^{-5}
$(\hat{\sigma}/\hat{m})\times 100$ %	6.588 %	5.500 %

TABLE VII
EMPIRICAL STANDARD DEVIATION FOR $M/M/1$ QUEUES IN PARALLEL

Example-I $\lambda_1 = 0.10$ $\mu_1 = 0.20$ $\lambda_2 = 0.30$ $\mu_2 = 0.40$ $N = 23$; $\alpha = 1.213\times 10^{-3}$ *# of Experiments* = 20; $\lambda_1^* = 0.10$ $\mu_1^* = 0.20$ $\lambda_2^* = 0.40$ $\mu_2^* = 0.30$		
# of Cycles (n)	2500	3000
Empirical Mean ($\hat{m}$)	1.193×10^{-3}	1.250×10^{-3}
Empirical Std. Dev. ($\hat{\sigma}$)	9.634×10^{-5}	7.571×10^{-5}
$(\hat{\sigma}/\hat{m})\times 100$ %	8.072 %	7.571 %
Example-II $\lambda_1 = 0.10$ $\mu_1 = 0.40$ $\lambda_2 = 0.15$ $\mu_2 = 0.35$ $N = 18$; $\alpha = 6.935\times 10^{-7}$ *# of Experiments* = 20; $\lambda_1^* = 0.10$ $\mu_1^* = 0.40$ $\lambda_2^* = 0.35$ $\mu_2^* = 0.15$		
# of Cycles (n)	5000	6000
Empirical Mean ($\hat{m}$)	6.905×10^{-7}	6.992×10^{-7}
Empirical Std. Dev. ($\hat{\sigma}$)	6.793×10^{-8}	5.028×10^{-8}
$(\hat{\sigma}/\hat{m})\times 100$ %	9.839 %	7.191 %
Example-III $\lambda_1 = 0.08$ $\mu_1 = 0.12$ $\lambda_2 = 0.20$ $\mu_2 = 0.60$ $N = 23$; $\alpha = 4.635\times 10^{-5}$ *# of Experiments* = 20; $\lambda_1^* = 0.12$ $\mu_1^* = 0.08$ $\lambda_2^* = 0.20$ $\mu_2^* = 0.60$		
# of Cycles (n)	3000	5000
Empirical Mean ($\hat{m}$)	4.471×10^{-5}	4.623×10^{-5}
Empirical Std. Dev. ($\hat{\sigma}$)	2.998×10^{-6}	2.004×10^{-6}
$(\hat{\sigma}/\hat{m})\times 100$ %	6.705 %	4.334 %

interchange of λ_i and μ_i with the larger traffic intensity λ_i/μ_i. (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 Sanov [10].

Theorem 5 (Sanov) [10]: Let Z_i, $i \geq 1$, be random variables whose possible values are $a_1, \cdots, a_n$ with $p_1, \cdots, p_n$ as respective probabilities. For $N > 1$, define $m_i(N) := \#$ of Z_k's, $1 \leq k \leq N$, that are equal to a_i. Define the relative frequency

$$\nu_i := \frac{m_i(N)}{N}, \qquad 1 \leq i \leq n.$$

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

$$\lim_{N\to\infty} \frac{1}{N} \cdot \log P\{|\nu_1(N) - q_1| \leq \epsilon, \cdots, |\nu_n(N) - q_n| \leq \epsilon\}$$
$$= -K(q, p) + e(\epsilon)$$

where

$$K(q, p) = \sum_{i=1}^{n} q_i \cdot \log\left(\frac{q_i}{p_i}\right)$$

and the term $e(\epsilon)$ is $O(\epsilon \cdot \log(1/\epsilon)$. (If $q_i > 0$, $1 \le i \le n$, then $O(\epsilon \cdot \log(1/\epsilon)$ can be replaced by $O(\epsilon)$.) ■

The above theorem suggests that

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

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

As in the cases of $M/M/1$ queues in tandem and in parallel, consider the paths of S which require T transitions, have λ', μ_1', and μ_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 outout of the first queue. Then, as in the last subsection, we can argue heuristically that, for large N,

$$\alpha \approx \exp\Bigg\{-N \cdot \inf_{\substack{\lambda'>0, \mu_1' \ge 0, \mu_2' \ge 0, 0 \le p' \le 1 \\ \lambda' + \mu_1' + \mu_2' = 1 \\ \lambda' > \mu_1' \text{ or } \lambda' \cdot (1-p') > \mu_2'}} \Bigg[R(\lambda', \mu_1', \mu_2', p') \cdot \left(\lambda' \cdot h_\lambda\left(\frac{1}{\lambda'}\right) + \mu_1' \cdot h_{\mu_1}\left(\frac{1}{\mu_1'}\right) + \mu_2' \cdot h_{\mu_2}\left(\frac{1}{\mu_2'}\right) + \min(\lambda', \mu_1') \cdot K(p', p)\right)\Bigg]\Bigg\} \text{ (UTLE)} \tag{47}$$

where

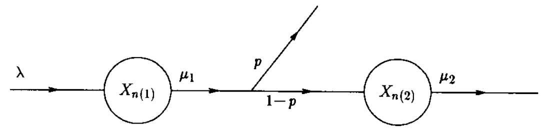
$$K(p', p) = p' \cdot \log\left(\frac{p'}{p}\right) + (1-p') \cdot \log\left(\frac{1-p'}{1-p}\right)$$

and (when $\lambda' > \mu_1'$ or $\lambda' \cdot (1-p') > \mu_2'$)

$$R = \begin{cases} 1/(\lambda' - \mu_1'), \text{ if } \lambda' > \mu_1' \text{ and } \mu_1' \cdot (1-p') \le \mu_2', \\ 1/((\lambda' - \mu_1') + (\mu_1' \cdot (1-p') - \mu_2')), \text{ if } \lambda' > \mu_1' \\ \quad \text{and } \mu_1' \cdot (1-p') > \mu_2', \\ 1/(\lambda' \cdot (1-p') - \mu_2'), \quad \text{otherwise.} \end{cases}$$

Numerical minimization gives us λ^*, μ_1^*, μ_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.



$\lambda + \mu_1 + \mu_2 = 1$, $\lambda < \mu_1$, $\lambda.(1-p) < \mu_2$

Fig. 10. Example of a network with routing.

TABLE VIII
EXAMPLE OF A NETWORK WITH ROUTING (SEE FIG. 10)

Method	Direct Simulation			Quick Simulation		
Example-I $\lambda = 0.20$ $\mu_1 = 0.30$ $\mu_2 = 0.50$ $p = 0.10$ $N = 20$; $\alpha = 3.269\times10^{-4}$ *CPU Time* = 288.8*Sec.*; $\lambda^* = 0.30$ $\mu_1^* = 0.20$ $\mu_2^* = 0.50$ $p^* = 0.10$						
# of Cycles (n)	5000	10000	15000	1000	1500	2000
α_n (α_n^*)	0.0	2.000×10^{-4}	2.667×10^{-4}	3.097×10^{-4}	3.479×10^{-4}	3.272×10^{-4}
CPU Time	28.6*Sec.*	60.1*Sec.*	79.4*Sec.*	25.8*Sec.*	39.3*Sec.*	53.0*Sec.*
Calls to RNG	84434	170263	245106	67941	106382	138763
Example-II $\lambda = 0.20$ $\mu_1 = 0.60$ $\mu_2 = 0.20$ $p = 0.50$ $N = 20$; $\alpha = 2.349\times10^{-6}$ *CPU Time* = 281.7*Sec.*; $\lambda^* = 0.30$ $\mu_1^* = 0.60$ $\mu_2^* = 0.10$ $p^* = 0.33$						
# of Cycles (n)	5000	10000	20000	500	1000	1500
α_n (α_n^*)	0.0	0.0	0.0	2.441×10^{-6}	2.447×10^{-6}	2.363×10^{-6}
CPU Time	15.7*Sec.*	31.3*Sec.*	63.4*Sec.*	14.4*Sec.*	31.8*Sec.*	45.0*Sec.*
Calls to RNG	45873	91234	193429	37768	80186	119749
Example-III $\lambda = 0.10$ $\mu_1 = 0.70$ $\mu_2 = 0.20$ $p = 0.20$ $N = 20$; $\alpha = 2.390\times10^{-8}$ *CPU Time* = 295.0*Sec.*; $\lambda^* = 0.22$ $\mu_1^* = 0.70$ $\mu_2^* = 0.08$ $p^* = 0.09$						
# of Cycles (n)	10000	30000	50000	1000	2000	5000
α_n (α_n^*)	0.0	0.0	0.0	2.364×10^{-8}	2.595×10^{-8}	2.425×10^{-8}
CPU Time	20.7*Sec.*	60.3*Sec.*	101.8*Sec.*	25.2*Sec.*	57.2*Sec.*	79.3*Sec.*
Calls to RNG	63703	193930	319010	68535	141488	210525

TABLE IX
EMPIRICAL STANDARD DEVIATION FOR THE EXAMPLES OF TABLE IX

Example-I $\lambda = 0.20$ $\mu_1 = 0.30$ $\mu_2 = 0.50$ $p = 0.10$ $N = 20$; $\alpha = 3.269\times10^{-4}$ # *of Experiments* = 20; $\lambda^* = 0.30$ $\mu_1^* = 0.20$ $\mu_2^* = 0.50$ $p^* = 0.10$		
# of Cycles (n)	1500	2000
Empirical Mean ($\hat{m}$)	3.194×10^{-4}	3.255×10^{-4}
Empirical Std. Dev. ($\hat{\sigma}$)	1.370×10^{-5}	1.011×10^{-5}
($\hat{\sigma}/\hat{m}$)× 100 %	4.288 %	3.107 %
Example-II $\lambda = 0.20$ $\mu_1 = 0.60$ $\mu_2 = 0.20$ $p = 0.50$ $N = 13$; $\alpha = 2.349\times10^{-6}$ # *of Experiments* = 20; $\lambda^* = 0.30$ $\mu_1^* = 0.60$ $\mu_2^* = 0.10$ $p^* = 0.33$		
# of Cycles (n)	1000	1500
Empirical Mean ($\hat{m}$)	2.366×10^{-6}	2.333×10^{-6}
Empirical Std. Dev. ($\hat{\sigma}$)	1.485×10^{-7}	1.294×10^{-7}
($\hat{\sigma}/\hat{m}$)× 100 %	6.278 %	5.546 %
Example-III $\lambda = 0.10$ $\mu_1 = 0.70$ $\mu_2 = 0.20$ $p = 0.20$ $N = 20$; $\alpha = 2.390\times10^{-8}$ # *of Experiments* = 20; $\lambda^* = 0.22$ $\mu_1^* = 0.70$ $\mu_2^* = 0.08$ $p^* = 0.09$		
# of Cycles (n)	2000	3000
Empirical Mean ($\hat{m}$)	2.405×10^{-8}	2.390×10^{-8}
Empirical Std. Dev. ($\hat{\sigma}$)	5.110×10^{-10}	4.357×10^{-10}
($\hat{\sigma}/\hat{m}$)× 100 %	2.125 %	1.823 %

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 (λ^*, μ_1^*, μ_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 20 000 cycles we obtained $\alpha_n^* = 1.764 \times 10^{-5}$ 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

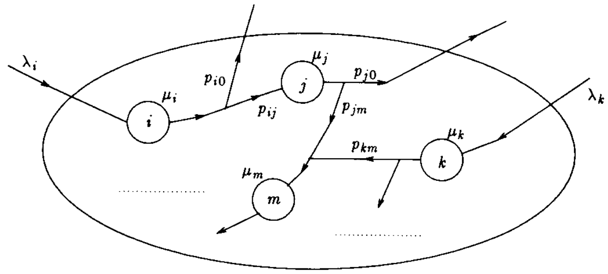


Fig. 11. Network of $GI/GI/1$ queues.

simulations but they are less than that for the examples of Table IV. Furthermore, if μ_1 and μ_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 running 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 (λ, μ_1, μ_2)-network and (λ, μ_2, μ_1)-network have identical α's for all N. For sufficiently large N, we have observed that it may be better to start with the (λ, μ_1, μ_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 λ with μ_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 1.265×10^{-14} after 20 000 cycles while a simulation of the (0.40, 0.10, 0.50)-network gives 1.114×10^{-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), and (0.20, 0.30 | 0.30, 0.20). For $N = 25, \alpha = 4.156 \times 10^{-4}$. After 20 000 cycles, these networks gave 3.337×10^{-4}, 3.855×10^{-4}, and 3.100×10^{-4}, respectively. It seems to us that in this limiting case, it might be faster to simulate the network where both the queues arrival and service 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 tube of 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 GI/GI/1 Queues

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

Consider a general open network of $GI/GI/1$ 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_{i0} 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 $\boldsymbol{P}' = \{p_{ij}'\}$ as the apparent routing probabilities. Let $\boldsymbol{L}'$ and $\boldsymbol{M}'$ denote the d-dimensional vectors $\{\lambda_i'\}$ and $\{\mu_i'\}$, respectively. Let $\boldsymbol{G}' = \{\gamma_i'\}$ denote the effective rate for these paths which we can find approximately (because μ_i''s are the virtual service rates) by solving the flow balance equations

$$\gamma_i' = \lambda_i' + \sum_{j=1}^{d} \min(\gamma_j', \mu_j') \cdot p_{ji}', \qquad 1 \leq i \leq d. \tag{48}$$

As in the previous subsections, we can argue heuristically to get the following relationship between T, $\boldsymbol{G}'$, and $\boldsymbol{M}'$.

$$T \approx N \cdot R, \text{ where}$$

$$R = \frac{1}{\sum_{i=1}^{d} (\gamma_i' - \mu_i') \cdot 1\{\gamma_i' > \mu_i'\}}.$$

Finally, the same line of heuristic gives

$$\alpha \approx \sum_{\boldsymbol{L}', \boldsymbol{M}', \boldsymbol{P}'} \exp\{-N \cdot H(\boldsymbol{L}', \boldsymbol{M}', \boldsymbol{P}')\} \text{ (UTLE)}$$

where

$$H(\boldsymbol{L}', \boldsymbol{M}', \boldsymbol{P}') = R \cdot \sum_{i=1}^{d} \lambda_i' \cdot h_{A_i}\left(\frac{1}{\lambda_i'}\right) + \sum_{i=1}^{d} \mu_i' \cdot h_{B_i}\left(\frac{1}{\mu_i'}\right) + \sum_{i=1}^{d} \min(\gamma_i', \mu_i') \cdot K(\boldsymbol{p}_i', \boldsymbol{p}_i)$$

and $\boldsymbol{p}_i'$ and $\boldsymbol{p}_i$ are the ith rows of the matrices $\boldsymbol{P}'$ and $\boldsymbol{P}$, respectively. Hence, for large N,

$$\alpha \approx \exp\{-N \cdot H^*\} \text{ (UTLE)}$$

where

$$H^* = \inf_{\boldsymbol{L}', \boldsymbol{M}', \boldsymbol{P}'} H(\boldsymbol{L}', \boldsymbol{M}', \boldsymbol{P}')$$

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

$$dB_i^*(z) = \frac{e^{\theta \cdot z}\, dB_i(z)}{\int e^{\theta \cdot z}\, dB_i(z)}$$

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

$$dA_i^*(z) = \frac{e^{\theta \cdot z}\, dA_i(z)}{\int e^{\theta \cdot z}\, dA_i(z)}$$

where θ is such that it satisfies $\int z dA_i^*(z) = 1/\lambda_i^*$. Then, for large N, we propose to use the network of $GI/GI/1$ queues with the parameters $\boldsymbol{L}^*$, $\boldsymbol{M}^*$, and $\boldsymbol{P}^*$ for estimating α.

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 Ventsel [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

[1] R. Azencott and G. Ruget, "Mélanges d'équations différentielles et grands écarts à la loi des grands nombres," *Z. Wahrscheinlichkeitstheorie verm. Gebiete*, vol. 38, pp. 1-54, 1977.

[2] P. Billingsley, *Convergence of Probability Measures*. New York: Wiley, 1968.

[3] M. Cottrell, J.-C. Fort, and G. Malgouyres, "Large deviations and rare events in the study of stochastic algorithms," *IEEE Trans. Automat. Contr.*, vol. AC-28, no. 9, pp. 907-920, 1983.

[4] P. Dupuis and J. Kushner, "Stochastic systems with small noise, analysis and simulation; A phase locked loop example," *SIAM J. Appl. Math.*, vol. 47, pp. 643-661, June 1987.

[5] R. Ellis, *Entropy, Large Deviations and Statistical Mechanics*. New York: Springer-Verlag, 1986.

[6] ——, "Large deviations for the empirical measure of a Markov chain with an application to the multivariate empirical measure," Dep. Math. Stat., Univ. Massachusetts, Amherst, MA, Int. Memo., 1987.

[7] S. Natarajan, "Large deviation, hypotheses testing, and coding for finite Markov chains," *IEEE Trans. Inform. Theory*, vol. IT-31, no. 3, pp. 360-365, 1985.

[8] S. Parekh, "Rare events in networks," Ph.D. dissertation, Dep. EECS, Univ. Calif., Berkeley, CA, 1986.

[9] G. Ruget, "Quelques Occurences Des Grands Ecarts Dans La Littérature Electronique," *Asterisque*, vol. 68, pp. 187-199, 1979.

[10] I. Sanov, "On the probability of large deviation of random variables," *Sel. Trans. Math. Statist. Prob.*, vol. I, pp. 213-244, 1957.

[11] S. R. S. Varadhan, "Large deviations and applications," *SIAM*, 1984.

[12] A. D. Ventsel, "Rough limit theorems on large deviation for Markov stochastic processes—II," *Theory Prob. Appl.* (USSR), vol. 21, pp. 499-512, 1976.

[13] R. Weber, "The interchangeability of *./M/*1 queues in series," *J. Appl. Prob.*, vol. 16, pp. 690-695, 1979.

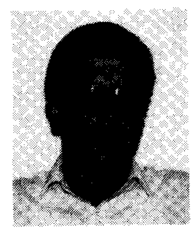
[14] A. Weiss, "A new technique for analyzing large traffic systems," *Adv. Appl. Prob.*, vol. 18, pp. 506-532, 1986.

[15] ——, "Large deviations for a Markov process with one boundary," presented at Conf. Stochastic Processes and Applications, Stanford, CA, Aug. 1987.

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 queueing 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 queueing networks, communication networks, stochastic control, and stochastic processes.

Dr. Walrand is the author of *An Introduction to Queueing 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*.