TELETRAFFIC AND DATATRAFFIC
in a Period of Change, ITC-13
A. Jensen and V.B. Iversen (Editors)
Elsevier Science Publishers B.V. (North-Holland)

Fast Estimation of the Statistics of Rare Events in Jackson Networks†

Michael R. Frater‡ Tava M. Lennon‡ Brian D. O. Anderson‡

† Work supported by Australian Telecommunications and Electronics Research Board (ATERB)
‡ Dept Systems Engineering, Australian National University, G.P.O. Box 4 Canberra ACT 2601.

ABSTRACT: Because of their rarity, the estimation of the statistics of buffer overflows in networks of queues by direct simulation is very costly. An asymptotically optimal (as the overflow recurrence time becomes large) scheme has been proposed by others, using importance sampling. In the existing approach, a numerical minimization is required to generate the simulation network. This paper describes an equivalent analytic minimization. A simple procedure for constructing the optimal simulation network is included.

1 Introduction

The efficient estimation of the statistics of rare events has been of interest for a number of years. Large deviations theory has been applied to perform asymptotically optimal (in the sense of variance) simulations of rare events [1]. More recently, specific application of this theory has been made to buffer overflows in queueing networks [2, 3, 4, 5], in which large deviations theory and numerical techniques have been used to find optimally fast simulation systems for networks of queues.

This paper addresses the large deviations approach to finding optimal simulation systems for queueing networks, finding of the optimal simulation system, in which a numerical minimization has been required previously. A simple direct analytic solution to this problem is presented here. The solution to this problem has been presented previously for tandem networks of M/M/1 queues [4] and for isolated GI/GI/1 queues [6]. This paper demonstrates that a unique solution to the minimization problem exists for more general queueing networks. The construction of such networks is also discussed.

The standard theory used in this paper is outlined in Section 2. The analytic solution of the minimization problem is set out in Section 3, (the minimality of this solution is proved in Appendix A.)

2 Background Theory

2.1 Importance Sampling

In a queueing network with finite buffers, a certain proportion of packets are lost due to buffer overflows. While the mean time between overflows can be calculated analytically for a single M/M/1 queue (e.g. [3]), the first step equations for a network of queues cannot be solved analytically because the order of the characteristic equation becomes large. Therefore, simulation is often used to find the recurrence time of buffer overflows. Because the mean time between overflows is large in a properly dimensioned network, direct simulation may not be feasible, simply because of its large cost in computer time. However, using the idea of importance sampling, the mean time between overflows can be found by simulation without incurring the large cost involved in direct simulation.

The idea in importance sampling is as follows. Suppose we are interested in certain (rare) events occurring in a system S that we can simulate on a computer. Then instead of simulating S we simulate a second system $\overline{S}$, which has the property that events in S and $\overline{S}$ correspond in some way. In particular, to the rare events A in S correspond events $\overline{A}$ in $\overline{S}$. The correspondence is such that

1. the events $\overline{A}$ in $\overline{S}$ are more frequent than the events A in S, and
2. the connection between S and $\overline{S}$ allows one to infer $P(A)$ if one knows $\overline{P}(\overline{A})$. ($\overline{P}(\overline{A})$ is the probability of event $\overline{A}$ in system $\overline{S}$.)

Now let α be the probability that, starting at zero, the number of customers in a queueing network hits some (large) value N before hitting zero again. Let T denote the first time the number of customers reaches N and let J denote the time to hit either 0 or N for the first time after leaving 0. It is not hard to check (see also [3]) that

$$\mathrm{E}[T] = \frac{1}{\alpha}\mathrm{E}[J] \tag{1}$$

Our interest is to understand how E[T] can be obtained by (efficient) simulation. Now, E[J] can be easily obtained by direct simulation on S, while α, which will be small, is obtainable from $\overline{S}$, as follows. Let us call a cycle of the system S or $\overline{S}$ a movement from 0 to the first time either zero is reached again, or an overflow occurs. Define $V_k = \mathbf{1}_{\{X_m \text{reaches } N \text{ in cycle } k\}}$. For S, we have

$$\mathrm{E}[V_k] = \alpha \tag{2}$$

Let L_k denote the likelihood ratio $\frac{\mathrm{d}P}{\mathrm{d}\overline{P}}$ during cycle k. Notice that the L_k are i.i.d. and

$$\overline{\mathrm{E}}[L_k V_k] = \mathrm{E}[V_k] = \alpha \tag{3}$$

There are frequent occurrences of the set $\{V_k = 1\}$ for the system $\overline{S}$. We examine p cycles for $\overline{S}$ and estimate α

by

$$\widehat{\alpha} = \frac{L_1V_1 + L_2V_2 + \ldots + L_pV_p}{p} \quad (4)$$

In the above analysis we have shown how knowing and simulating $\overline{S}$, it can be used to compute α and thus important statistics concerning S. But we have not described how one might choose the system $\overline{S}$, given S. *A major issue in the use of importance sampling is how one should construct $\overline{S}$ from S.* To an extent, the problems of obtaining the probability of a rare event or the mean time between occurrences of a rare event (which are both problems of excessive computer time) are being replaced by another difficult problem ('How should we obtain $\overline{S}$ from S?') in importance sampling.

2.2 Large Deviations Theory

Large deviations theory has been used to obtain a number of asymptotic results [3] that apply not only to Jackson networks, but also to more general queueing networks. Some of these, that are relevant to the results presented below, are summarized here.

Let $\xi_1 \ldots \xi_n$ be i.i.d random variables in $\mathbb{R}^d$. Let F be the distribution function of the $\{\xi_k\}$ and m its mean. Assume that the Laplace transform of F

$$M(s) = \int_{\mathbb{R}^d} \exp < s, z > \mathrm{d}F(z) \quad (5)$$

is finite in a neighbourhood of 0. Then the Cramér or Legendre transform is defined as [1]:

$$h(y) = \sup_{s \in \mathbb{R}^d} [< s, y > - \log M(s)] \quad (6)$$

For example, the Cramér transform of an exponential distribution with parameter λ is

$$h_\lambda(u) = \begin{cases} \lambda u - \log(\lambda u) - 1 & u > 0 \\ \infty & \text{otherwise} \end{cases} \quad (7)$$

The following properties of the Cramér transform are used in this paper:

1. $h(\cdot)$ is convex;
2. $h(\cdot)$ is non-negative;
3. $h(y) = 0$, if and only if $y = m$, where m is the mean of the distribution function F;
4. $h'(m) = 0$

For a network of queues, call a *cycle* a piece of a trajectory starting at the zero state and terminating on the first occasion when either the total number of customers in the network exceeds some value (say N), or the state equals zero again. Call a cycle that terminates with the system in the empty state a cycle of the first kind, and one that terminates with the number of customers in the network greater than N a cycle of the second kind. Let d be the number of queues in the network, λ_i be the rate of external arrivals at queue i, γ_i be the total arrival rate at queue i, μ_i be the virtual service rate at queue i, p_{ij} be the routing probability from queue i to queue j and p_{i0} be the probability that a customer leaving queue i leaves the network. For current purposes, we will assume that all external arrival processes are poisson, that all the service rates are exponentially distributed, and that all queues are asymptotically stable (i.e. $\gamma_i < \mu_i \forall i$). All these parameters of the system S (i.e. γ_i, λ_i, μ_i and p_{ij}) are assumed constant. These parameters of the system satisfy the traffic equations:

$$\gamma_i = \sum_{j=1}^{d} p_{ji}\gamma_j + \lambda_i, \quad (8)$$

and the routing probabilities satisfy:

$$\sum_{j=0}^{d} p_{ij} = 1 \quad (9)$$

Suppose α is the probability that a cycle ends in a buffer overflow, (i.e. that it is of the second kind,). There is a relation between α and a system $\overline{S}(\lambda_i', \mu_i', \gamma_i', p_{ij}')$, which is obtained from S by varying its parameters, and which is used for estimating α by simulation. This relation is derived by heuristic argument in [3]. The parameters for a system $\overline{S}$ $(\lambda_i', \mu_i', \gamma_i', p_{ij}')$ can be found as the arguments achieving minimization in the following large-deviations approximation for α:

$$\alpha \simeq \exp -N \inf_{\lambda_i', \mu_i', \gamma_i', p_{ij}'} R \left[\sum_{i=1}^{d} \lambda_i' h_{\lambda_i} \left(\frac{1}{\lambda_i'} \right) + \sum_{i=1}^{d} \mu_i' h_{\mu_i} \left(\frac{1}{\mu_i'} \right) + \sum_{i=1}^{d} \min(\gamma_i', \mu_i') K_i \right] \quad (10)$$

where

$$K_i = \sum_{j=0}^{d} p_{ij}' \log \frac{p_{ij}'}{p_{ij}} \quad (11)$$

$$R = \frac{1}{\sum_i (\gamma_i' - \mu_i') \mathbf{1}_{\gamma_i' > \mu_i'}} \quad (12)$$

The infimum is subject to the following constraints:

$$\lambda_i', \mu_i', \gamma_i' \geq 0 \quad (13a)$$

$$0 \leq p_{ij}' \leq 1 \quad (13b)$$

$$\gamma_i' > \mu_i' \quad \text{for at least one } i \quad (13c)$$

$$\sum_{i=1}^{d} (\lambda_i' + \mu_i') = 1 \quad (13d)$$

$$\sum_{j=0}^{d} p_{ij}' = 1 \quad (13e)$$

$$\gamma_i' = \sum_{j=1}^{d} p_{ji}' \min(\gamma_j', \mu_j') + \lambda_i' \quad (13f)$$

It has been argued [1, 3] that if the system $\overline{S}$ defined by the parameters γ_i', λ_i', μ_i' and p_{ij}' is used to perform simulation, then this simulation is asymptotically optimal as the mean time between overflows grows large, i.e. as N becomes large. We will perform the minimization using the method of Lagrange multipliers to satisfy the equality constraints. The solution obtained will then be shown to satisfy the inequality constraints.

3 The Optimal Simulation System

In this section, a direct analytic solution is given to the minimization problem described above for Jackson networks, (i.e. we assume that all the external arrival streams are poisson, and that the service times are exponentially distributed.) After the mathematical details of the solution, some comments on interpretation are made. A proof of the optimality of the solution is contained in Appendix A.

3.1 Evaluation of $\overline{S}$

In order to find the optimal simulation system, the arguments achieving the infimum in the exponent of (10) must be found, subject to the constraints listed. In order to do this, we define a Lagrangian as follows, with Lagrange multipliers g, b_i and c_i:

$$\begin{aligned}\mathcal{L} &= R\left[\sum_{i=1}^{d}\lambda_i' h_{\lambda_i}\left(\frac{1}{\lambda_i'}\right)+\sum_{i=1}^{d}\mu_i' h_{\mu_i}\left(\frac{1}{\mu_i'}\right)+\right.\\ &\left.\sum_{i=1}^{d}\min(\gamma_i',\mu_i')K_i\right]+g\left(\sum_{i=1}^{d}(\lambda_i'+\mu_i')-1\right)+\\ &\sum_{i=1}^{d}b_i\left(\lambda_i'+\sum_{j=1}^{d}\min(\gamma_j',\mu_j')p_{j,}'-\gamma_i'\right)\\ &+\sum_{i=1}^{d}c_i\left(\sum_{j=0}^{d}p_{ij}'-1\right)\end{aligned}\tag{14}$$

Each of the equality constraints (13d) through (13f) is associated with a Lagrange multiplier. We will assume without real loss of generality that queue 1 has the largest load, (i.e. $\rho_1 > \rho_i$ for all $i \neq 1$, where $\rho_j = \frac{\gamma_j}{\mu_j}$.)

Define r_i as the expected number of times that a customer arriving at queue i will pass subsequently through queue 1 before leaving the network.[1] Because the routing is Markovian, r_i does not depend in any way on the previous history of a customer, e.g. whether the customer enters that network at queue i, or comes to queue i from within the network. Thus r_i is also the expected number of visits to queue 1 of a customer entering the network at queue i. Then:

$$\sum_{i=1}^{d} r_i\lambda_i = \gamma_1 \tag{15}$$

We note that $r_i = 0$ implies that customers arriving at queue i can never be routed through queue 1 before leaving the network. Also, since all customers arriving at queue 1 must pass through this queue before leaving the network, we must have $r_1 \geq 1$.

When the derivatives of the Lagrangian with respect to the parameters of the system $\overline{S}$ are evaluated (see Appendix A), it can be shown that the following values of the parameters of the system $\overline{S}$ correspond to a turning point of the Lagrangian, and are in fact the required infimum.

$$\gamma_i' = \gamma_i\left[1+\frac{r_i}{r_1}\frac{\mu_1-\gamma_1}{\gamma_1}\right] \tag{16a}$$

$$\lambda_i' = \lambda_i\frac{\gamma_i'}{\gamma_i} \tag{16b}$$

$$\mu_i' = \begin{cases}\gamma_1+\frac{(r_1-1)}{r_1}(\mu_1-\gamma_1) & \text{for } i=1\\ \mu_i & \text{for } i>1\end{cases} \tag{16c}$$

$$p_{i0}' = p_{i0}\frac{\gamma_i}{\min(\gamma_i',\mu_i')} \tag{16d}$$

$$p_{ij}' = p_{ij}\frac{\gamma_i}{\min(\gamma_i',\mu_i')}\frac{\gamma_j'}{\gamma_j} \qquad \text{for } j>0 \tag{16e}$$

The Lagrange multipliers take values

$$g = 0 \tag{17a}$$

$$b_i = -R\log\frac{\gamma_i'}{\gamma_i} \tag{17b}$$

$$c_i = R\min(\gamma_i',\mu_i')\left[\log\frac{\min(\gamma_i',\mu_i')}{\gamma_i}-1\right] \tag{17c}$$

In the simulation system defined here, we can show that queue 1 becomes unstable in $\overline{S}$, and that all other queues remain stable in $\overline{S}$. From (16a) and (16c), we can see that $\gamma_1' > \mu_1'$ if and only if:

$$\mu_1 > \frac{\gamma_1+(r_1-1)\mu_1}{r_1}, \tag{18}$$

which is true if and only if $\gamma_1 < \mu_1$, which is given. Hence, queue 1 is unstable in $\overline{S}$, in the sense that the number of customers resident in this queue will, on average, increase with time.

Next, we show that no other queues are unstable in $\overline{S}$. It has been assumed that $\rho_1 > \rho_j \forall j \neq 1$. Also, we must have $r_1 \geq r_j$, because all customers arriving at queue 1 are counted in r_1, and it is not possible to have more of these customers counted in r_j. Therefore:

$$\rho_1^{-1} - 1 < \rho_j^{-1} - 1 \tag{19}$$

Hence,

$$\frac{\rho_1^{-1}-1}{r_1} < \frac{\rho_j^{-1}-1}{r_j} \tag{20}$$

[1] r_i is easily calculated as the value of γ_1 when the values $\lambda_i = 1$ and $\lambda_j = 0$ for $j \neq i$ are substituted into the traffic equations (8). This is possible because the r_i depend only on the routing in the network, and therefore do not change when the external arrival rates are changed.

and simple manipulation yields:

$$\gamma_j \left[1 + \frac{r_j}{r_1} \frac{\mu_1 - \gamma_1}{\gamma_1}\right] < \mu_j \qquad \forall j \neq 1 \tag{21}$$

i.e., substituting from (16a) and (16c), we must have

$$\gamma'_j < \mu'_j \qquad \text{for } j \neq 1 \tag{22}$$

In other words, all queues are stable in the optimal simulation system except for queue 1.

This instability of queue 1 satisfies the last of the inequality constraints in (24) (that $\gamma'_i > \mu'_i$ for some i). It is shown in the appendix that the other inequality constraints are satisfied.

3.2 Interpretation

While no physical understanding is necessary to generate an asymptotically optimal simulation system using the above equations, it is nonetheless useful to see the meaning of the above transformation.

- Only the dominating queue, (i.e. the queue with the largest load,) becomes unstable in the simulation system, as was shown above.
- $r_i = 0$ implies that no customers passing through queue i reach queue 1, and, from (16a), that $\gamma'_i = \gamma_i$. Hence, (16b) implies that $\lambda'_i = \lambda_i$, i.e. arrival rates are changed only at external inputs from which customers may be routed to the dominating queue.
- The arrival rate at the dominating queue in the simulation system is always the same as the service rate of this same queue in the original system.
- The rate of customers leaving the network at external outputs remains unchanged in the simulation system (16d). This can be seen in (16d), where it is clear that $p'_{i0} \min(\gamma'_i, \mu'_i) = p_{i0}\gamma_i$.
- Only those parts of the network that can contribute to contribute to overflows in the original system contribute to overflows in the simulation system. That is, it is possible for customers to be routed from one queue to another in the simulation system if and only if it is possible in the original system. This can be seen from (16e), where it is clear that $p'_{ij} > 0$ if and only if $p_{ij} > 0$.
- The average behaviour of the simulation system described here is the same as that of the reverse-time model in the period immediately before an overflow occurs [5, 7].

It should also be noted (perhaps surprisingly) that the distributions of service times for queues other than that dominating the overflow statistics are not required to be exponential, and the external arrival processes at queues from which there is no direct path to the input of queue 1 (i.e. $r_j = 0$) need not be poisson. Both of these facts are demonstrated in Appendix A.

4 Conclusion

This paper has extended the previously known theory for generating optimal (in the sense of variance) importance sampling simulation systems, demonstrating a simple analytic method for the construction of such systems, removing the need to perform numerical minimizations. This derivation did not require the assumption of exponential service rates on queues other than that dominating the overflow statistics. In some cases, it is also possible to remove the assumption that arrival streams are poisson.

This work could be extended further by finding the optimal simulation system for networks where the arrival streams are not poisson, and to the case where the service times of the dominating queue are not exponentially distributed.

A Proof of Optimality

In this section, it is shown that the set of equations given in Section 3 defines a global minimum of the exponent of (10). First, it will be shown that these equations correspond to a turning point, and then that this point is in fact the required minimum. As in the previous text, we will use the symbol S to represent the original system, and $\overline{S}$ to represent the importance-sampling system generated by large deviations.

A.1 Evaluation of Derivatives of Lagrangian

Let

$$H = R\left[\sum_{i=1}^{d} \lambda'_i h_{\lambda_i}\left(\frac{1}{\lambda'_i}\right) + \sum_{i=1}^{d} \mu'_i h_{\mu_i}\left(\frac{1}{\mu'_i}\right) + \sum_{i=1}^{d} \min(\gamma'_i, \mu'_i) K_i\right] \tag{23}$$

where R and K_i are as defined in the main text. Then the extrema of H subject to the equality constraints (13d) through (13f) are found by setting the partial derivatives of the lagrangian to zero; the relevant equations are:

$$\frac{\partial \mathcal{L}}{\partial \lambda'_i} = R \log \frac{\lambda'_i}{\lambda_i} + y + b_i = 0 \tag{24}$$

$$\frac{\partial \mathcal{L}}{\partial \mu'_i} = \begin{cases} R\left[H + \log \frac{\mu'_i}{\mu_i} + \sum_{j=0}^{d} p'_{ij} \log \frac{p'_{ij}}{p_{ij}}\right] & \\ \quad + y + \sum_{j=1}^{d} b_j p'_{ij} & \gamma'_i > \mu'_i \\ R \log \frac{\mu'_i}{\mu_i} + y & \text{otherwise} \end{cases} \tag{25}$$

$$= 0$$

$$\frac{\partial \mathcal{L}}{\partial p'_{i0}} = R \min(\gamma'_i, \mu'_i)\left[1 + \log \frac{p'_{i0}}{p_{i0}}\right] + c_i = 0 \tag{26}$$

$$\frac{\partial \mathcal{L}}{\partial p'_{ij}} = R \min(\gamma'_i, \mu'_i)\left[1 + \log \frac{p'_{ij}}{p_{ij}}\right] + b_j \min(\gamma'_i, \mu'_i) + c_i = 0 \qquad \text{for } j \geq 1 \tag{27}$$

$$\frac{\partial \mathcal{L}}{\partial \gamma_i'} = \begin{cases} -RH - b_i & \gamma_i' > \mu_i' \\ R\sum_{j=0}^{d} p_{ij}' \log \frac{p_{ij}'}{p_{ij}} & \\ \quad + \sum_{j=1}^{d} b_j p_{ij}' - b_i & \text{otherwise} \end{cases} \tag{28}$$

$$= 0$$

$$\frac{\partial \mathcal{L}}{\partial g} = \sum_{i=1}^{d} \left(\lambda_i' + \mu_i'\right) - 1 = 0 \tag{29}$$

$$\frac{\partial \mathcal{L}}{\partial b_i} = \lambda_i' + \sum_{j=1}^{d} \min(\gamma_j', \mu_j') p_{ji}' - \gamma_i' = 0 \tag{30}$$

$$\frac{\partial \mathcal{L}}{\partial c_i} = \sum_{j=0}^{d} p_{ij}' - 1 = 0 \tag{31}$$

It can be shown by direct substitution of (16a) through (16e) and (17a) through (17c) that these equations are satisfied by the solution outlined in Section 3. Therefore, it is clear that these equations define at least a turning point of H, under the equality constraints in (13d) through (13f). It will be shown later that this solution also satisfies the inequality constraints (13a) through (13c). In the following sections it will be shown that this solution is in fact a global minimum.

A.2 Evaluation of Optimal Simulation System

Given a system $\overline{S}(\lambda_i', \mu_i', \gamma_i', p_{ij}')$ whose parameters define a turning point of H, we assume without loss of generality that I queues are unstable in $\overline{S}$, and that (after renumbering if necessary) the queues are numbered such that queues 1 through I are unstable and all others are stable. For the moment, we do not suppose that under this numbering, queue 1 is the most heavily loaded.

Let r_{ij} be the expected number of times that a customer arriving at queue i will pass through queue j before leaving the network, and let $\mathbf{r} = (r_{ij})$ and $\mathbf{p} = (p_{ij})$ for $i, j \in [1, d]$, (i.e. $\mathbf{p}$ and $\mathbf{r}$ are square matrices of dimension d.)

Just as in the main text, where r_i was the value of γ_1 when $\lambda_i = 1$ and $\lambda_j = 0$ for $j \neq i$, here we have that r_{ij} is the value of γ_j when $\lambda_i = 1$ and $\lambda_k = 0$ for $k \neq i$. Hence,

$$\gamma_j = \sum_{i=1}^{d} r_{ij} \lambda_i \tag{32}$$

and

$$\mathbf{r} = \mathbf{I} + \mathbf{pr} \tag{33}$$

Therefore,

$$\left(\mathbf{I} - \mathbf{p}\right) \mathbf{r} = \mathbf{I} \tag{34}$$

As has been argued previously for tandem networks of queues [4], physical constraints require that $g = 0$. The reason for this is that g is the rate of change of $\mathcal{L}$ with the sum of the arrival and service rates, i.e. the scaling of time, and we do not expect the probability that a cycle exits rather than returns to zero to depend on the scaling of time. Therefore, we require $g = 0$. Then (25) yields $\mu_i' = \mu_i$ for $i > I$.

Using (24) to substitute for b_i in (27), and using also (26), we have for $i, j \in [1, d]$:

$$\frac{p_{ij}'}{p_{ij}} = \frac{\lambda_j'}{\lambda_j} \frac{p_{i0}'}{p_{i0}} \tag{35}$$

Now consider the first equation of (25). Recognize from (28) that $RH = -b_i = R \log \frac{\lambda_i'}{\lambda_i}$ for $i \in [1, I]$. Substitute also for $\frac{p_{ij}'}{p_{ij}}$ using (35).s There results the first equation of (36) below, after simplifying. In a similar way, using the second equation of (28). and substituting for b_i and $\frac{p_{ij}'}{p_{ij}}$, the second equation of (36) results: (25) and (28), we obtain:

$$\frac{p_{i0}'}{p_{i0}} = \begin{cases} \frac{\lambda_i}{\lambda_i'} \frac{\mu_i}{\mu_i'} & i \leq I \\ \frac{\lambda_i}{\lambda_i'} & i > I \end{cases} \tag{36}$$

Hence from (35), for $j \geq 1$:

$$\frac{p_{ij}'}{p_{ij}} = \begin{cases} \frac{\lambda_j'}{\lambda_j} \frac{\lambda_i}{\lambda_i'} \frac{\mu_i}{\mu_i'} & i \leq I \\ \frac{\lambda_j'}{\lambda_j} \frac{\lambda_i}{\lambda_i'} & i > I \end{cases} \tag{37}$$

(We observe that these equations are consistent with the solution of the minimization problem set out in the main text, where $I = 1$.) Replacing p_{ij}' in (31) with the expansion available in (36) and (37) gives:

$$\left(\mathbf{I} - \mathbf{p}\right) \begin{pmatrix} \frac{\lambda_1'}{\lambda_1} \\ \frac{\lambda_2'}{\lambda_2} \\ \vdots \\ \frac{\lambda_d'}{\lambda_d} \end{pmatrix} = \begin{pmatrix} p_{10} \\ p_{20} \\ \vdots \\ p_{d0} \end{pmatrix} + \begin{pmatrix} \frac{\lambda_1'}{\lambda_1}\left(1 - \frac{\mu_1'}{\mu_1}\right) \\ \vdots \\ \frac{\lambda_I'}{\lambda_I}\left(1 - \frac{\mu_I'}{\mu_I}\right) \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{38}$$

Also, from (9), we have:

$$\begin{pmatrix} p_{10} \\ p_{20} \\ \vdots \\ p_{d0} \end{pmatrix} = \left(\mathbf{I} - \mathbf{p}\right) \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \tag{39}$$

Eliminating the p_{i0} terms from (38), using (39), and substituting $\mathbf{r}$ for $(\mathbf{I} - \mathbf{p})^{-1}$, we obtain:

$$\frac{\lambda_i'}{\lambda_i} = 1 + \sum_{j=1}^{I} r_{ij} \frac{\lambda_j'}{\lambda_j} \left(1 - \frac{\mu_j'}{\mu_j}\right) \tag{40}$$

From (25), for queues with index $i \leq I$, we obtain:

$$0 = R\left[H + \log \frac{\mu_i'}{\mu_i} + \sum_{j=0}^{d} p_{ij}' \log \frac{p_{ij}'}{p_{ij}}\right] + \sum_{j=1}^{d} b_j p_{ij}' \tag{41}$$

$$= R\left[\sum_{j=1}^{d} \left(p_{ij}' \log \frac{p_{ij}'}{p_{ij}} \frac{\lambda_j}{\lambda_j'}\right)\right] \quad \text{from (36) (42)}$$

$$= R(1 - p_{i0}') \log \frac{p_{i0}'}{p_{i0}} \quad \text{from (31) and (35) (43)}$$

i.e. since R cannot be zero, for $i \in [1, I]$, $p'_{i0} = p_{i0}$, or $p'_{i0} = 1$. Clearly, it is the first solution in which we are interested, since $p'_{i0} = 1$ implies $p'_{ij} = 0$ for all $j > 0$, and hence from (35), that $\lambda_j = 0$ for all queues to which customers can be routed in one step from queue i. Therefore, from (36), we have:

$$\frac{\mu'_i}{\mu_i} = \frac{\lambda_i}{\lambda'_i} = \text{constant } \forall i \leq I \tag{44}$$

i.e. for all queues that are unstable in $\overline{S}$, the ratio $\frac{\mu'_i}{\mu_i}$ takes the same value.

Now, in order for (29) to hold, we must have:

$$\begin{aligned} 0 &= \sum_{i=1}^{d} (\lambda'_i + \mu'_i) - 1 \\ &= \sum_{i=1}^{I} \gamma_i \frac{\lambda'_i}{\lambda_i} \left(1 - \frac{\mu'_i}{\mu_i}\right) - \sum_{i=1}^{I} \mu_i \left(1 - \frac{\mu'_i}{\mu_i}\right) \end{aligned} \tag{45}$$

With a small amount of manipulation, in particular using the fact that $\frac{\lambda'_i}{\lambda_i} = \frac{\mu_i}{\mu'_i}$ is constant for $i \leq I$, it can be shown that:

$$\lambda'_i = \lambda_i \left(\frac{\mu_1 + \ldots + \mu_I}{\gamma_1 + \ldots + \gamma_I}\right) \qquad \text{for } i \leq I \tag{46}$$

Therefore, assuming that a solution exists for which there are I unstable queues in $\overline{S}$, as postulated above, and substituting for b_i from (24) in (28) for $i \leq I$, we have:

$$H = \log \left(\frac{\mu_1 + \ldots + \mu_I}{\gamma_1 + \ldots + \gamma_I}\right) \tag{47}$$

The postulate that there are a particular I unstable queues in $\overline{S}$ corresponding to a turning point of H has led to the conclusion that only one set of values $(\lambda'_i, \mu'_i, \gamma'_i, p'_{ij})$ apparently give a turning point; note however that for these values to actually give a turning point, there would have to be satisfied at the solution point the conditions $\gamma'_i > \mu'_i$ for $1 \leq i \leq I$, as well as the other inequality conditions in (24). If these conditions are not satisfied, the postulate that a particular I queues are unstable is inconsistent with there being an associated turning point of H.

Now consider all possible values of I, and all possible selections of I queues. With each such selection, there is a potential turning point for H, (which will actually be a turning point only if the postulated instability condition is actually fulfilled at the solution point.) The set of associated values of H has a minimum element, obtained by choosing only the queue with the highest load in S to be unstable in $\overline{S}$, since if $\frac{\gamma_1}{\mu_1} > \frac{\gamma_i}{\mu_i}$ for all $i \neq 1$,

$$\log \frac{\mu_1}{\gamma_1} < \log \left(\frac{\mu_{i_1} + \ldots + \mu_{i_I}}{\gamma_{i_1} + \ldots + \gamma_{i_I}}\right) \tag{48}$$

unless $i_1 = 1$ and $i_2 \ldots i_I$ are empty.

The solution point obtained above for the assumption that queue 1, and no other queue, is unstable in $\overline{S}$ turns out to always lead to queue 1, and no other queue, being unstable; this is established in the text. Further, the other inequality constraints in (24) hold; this is established in the next section. Hence, the minimum of the set of possible values of H is actually attained.

A.3 Satisfaction of Inequality Constraints

The use of the method of Lagrange multipliers ensures that the equality constraints of (24) are satisfied. We have already shown in the main text that the third of the inequality constraints, requiring that at least one queue in the simulation system becomes unstable, is satisfied by the solution presented here. (In fact, we showed that just one queue is unstable.) It remains to be shown that $\lambda'_i, \mu'_i, \gamma'_i \geq 0$ and $0 \leq p'_{ij} \leq 1$.

Firstly, it is clear from (13a) that $\gamma'_i > 0 \ \forall i$, since we know that all queues in S are stable (i.e. $\mu_i > \gamma_i$.) Hence, (16b) shows that $\lambda'_i > 0 \ \forall i$. From (16c), because $r_1 \geq 1$, it is clear that we must have $\mu'_i > 0 \ \forall i$.

Given the above, (16d) and (16e) imply $p'_{ij} \geq 0$. The requirement $p'_{ij} \leq 1$ is enforced by the equality constraint $\sum_j p'_{ij} = 1$. Hence, all the inequality constraints are satisfied.

References

[1] M. Cottrell, J. C. Fort, and G. Malgouyres, "Large deviations and rare events in the study of stochastic algorithms," *IEEE Trans. Automatic Control*, vol. AC-28, pp. 907 – 918, September 1983.

[2] J. Walrand, *An Introduction to Queueing Networks*. Englewood Cliffs, New Jersey: Prentice-Hall, 1988.

[3] S. Parekh and J. Walrand, "A quick simulation of excessive backlogs in networks of queues," *IEEE Trans. Automatic Control*, vol. 34, pp. 54 – 66, January 1989.

[4] M. R. Frater and B. D. O. Anderson, "Fast estimation of the statistics of excessive backlogs in tandem networks of queues," *Australian Telecommunication Research*, vol. 23, pp. 49 – 55, May 1989.

[5] M. R. Frater, T. M. Lennon, and B. D. O. Anderson, "Optimally efficient estimation of the statistics rare events in queueing networks," *IEEE Trans. Automatic Control*. Submitted for publication.

[6] M. R. Frater, J. Walrand, and B. D. O. Anderson, "Optimally efficient simulation of buffer overflows in queues with deterministic service times," *Australian Telecommunication Research*, vol. 24, no. 1, pp. 1 – 8, 1990.

[7] M. R. Frater, *Fast Estimation of the Statistics of Rare Events in Data Communications Systems*. PhD thesis, Australian National University, November 1990.