Annals of Operations Research 49(1994)207–220

Fast simulation of buffer overflows in tandem networks of *GI/GI/1* queues*

Michael R. Frater
Department of Electrical Engineering, University College, Australian Defence Force Academy, Campbell, ACT 2600, Australia

Brian D.O. Anderson
Department of Systems Engineering and Cooperative Research Centre for Robust and Adaptive Systems, Australian National University, G.P.O. Box 4, Canberra, ACT 2601, Australia

Simply because of their rarity, the estimation of the statistics of buffer overflows in well-dimensioned queueing networks via direct simulation is extremely costly. One technique that can be used to reduce this cost is importance sampling, and it has been shown previously that large deviations theory can be used in conjunction with importance sampling to minimize the required simulation time. In this paper, we obtain results on the fast simulation of tandem networks of queues, and derive an analytic solution to the problem of finding an optimal simulation system for a class of tandem networks of *GI/GI/1* queues.

Keywords: Queueing networks, large deviations, importance sampling, fast simulation.

1. Introduction

In designing telecommunications networks, we are often interested in determining the rate at which data is lost due to buffer overflows. In general, it is not possible to eliminate these overflows in dimensioning the network, but because of the high cost of their occurrence, we would like to dimension the network such that the rate of data loss is minimized. In order to do this, we need a cheap, simple method for determining this rate.

There are many examples in the literature where importance sampling [1] is applied to rare event problems in telecommunications (e.g. [2–6].) Of particular

*Work supported by Australian Telecommunications and Electronics Research Board (ATERB). The authors wish to acknowledge the funding of the activities of the Cooperative Research Centre for Robust and Adaptive Systems by the Australian Commonwealth Government under the Cooperative Research Centres Program.

interest is [2], where large deviations theory is used in conjunction with importance sampling to minimize the simulation time required. These ideas are extended to queueing networks in [3]. However, no analytic expression for the parameters of the simulation system is given. This latter problem has been addressed previously for a number of special cases such as Jackson networks [7]. In applying queueing networks to the modeling of practical systems, it is often necessary to use more complicated models than the $M/M/1$ queue examined in the above examples. The specific case of queues with deterministic service times, which is of great practical interest for modeling modern fast packet-switching networks, is dealt with in [8]. In this paper, we look at tandem networks of $GI/GI/1$ queues, i.e. queues where the distributions of inter arrival and virtual service times are essentially arbitrary, except that successive inter arrival times are independent, as are successive service times. An analytic expression is derived for the simulation system for a particular subclass of such networks.

Section 2 introduces the idea of importance sampling, and the concept of maximizing speedup. The major results of [3] as they apply to this paper are summarized in section 3. Section 4 presents a number of lemmas necessary for the construction of an optimal simulation system for a tandem network of $GI/GI/1$ queues, which is carried out in section 5.

2. Importance sampling

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

(1) the events $\bar{A}$ in $\bar{S}$ are more frequent than the events A in S, and

(2) the connection between S and $\bar{S}$ allows one to infer $\mathrm{P}(A)$ if one knows $\bar{\mathrm{P}}(\bar{A})$. ($\bar{\mathrm{P}}(\bar{A})$ is the probability of the event $\bar{A}$ in $\bar{S}$.)

The problem of finding the best system to use in importance sampling can be posed as an optimization problem as follows. Let A be a rare event for a system S, with $\alpha = \mathrm{P}(A) \ll 1$. For a direct Monte Carlo simulation involving n independent experiments, we could estimate α via:

$$\hat{\alpha}_n = \frac{1}{n}\sum_{i=1}^{n} 1_A(\omega_i), \tag{1}$$

where the ω_i are the i.i.d. outcomes of the experiments. The variance of $\hat{\alpha}_n$ is easily

computed as

$$\mathrm{E}[\alpha - \hat{\alpha}_n]^2 = \frac{1}{n}(\alpha - \alpha^2). \tag{2}$$

Alternatively, consider a probability measure $\bar{P}$ associated with a system $\bar{S}$, with P absolutely continuous with respect to $\bar{P}$, such that the same event spaces apply for S and $\bar{S}$. Using $\bar{S}$ we can obtain a second estimate

$$\hat{\bar{\alpha}}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_A(\bar{\omega}_i)L(\bar{\omega}_i), \tag{3}$$

where $L = \mathrm{d}P/\mathrm{d}\bar{P}$ and the $\bar{\omega}_i$ are the i.i.d. outcomes of n experiments using $\bar{S}$. The quantity $\mathrm{d}P/\mathrm{d}\bar{P}$ is known as the likelihood ratio, or Radon–Nikodym derivative. The variance of $\hat{\bar{\alpha}}$ is different to (2), and is obtainable as

$$\frac{1}{n}\left(\int_A L^2(\omega)\,\mathrm{d}\bar{P}(\omega) - \alpha^2\right). \tag{4}$$

We want the estimate of $\hat{\bar{\alpha}}$ to be as accurate as possible. Therefore, we want to replace the probabilities of all events in S to new ones in $\bar{S}$ (which has the same event space) so that the variance

$$(\sigma^*)^2 \triangleq \int_A L^2(\omega)\,\mathrm{d}\bar{P}(\omega) \tag{5}$$

is minimized [2].

Let $V_k = \mathbf{1}_{\{\text{the event } A \text{ occurs in trial } k\}}$. Then in our original system S we have:

$$\mathrm{E}[V_k] = \alpha. \tag{6}$$

Let L_k denote the likelihood ratio $\mathrm{d}P/\mathrm{d}\bar{P}$ during trial k, i.e. the ratio of the probabilities of the trajectories under the measures P and $\bar{P}$ in S and $\bar{S}$. We observe that the L_k are i.i.d. and

$$\bar{\mathrm{E}}[L_k V_k] = \mathrm{E}[V_k] = \alpha. \tag{7}$$

Hence, if we simulate the system $\bar{S}$ for p trials, and if $\bar{S}$ is chosen so that L_k is known for each event, or at least those for which V_k is non-zero, then from (7), it is clear that we can estimate the probability of the event A, α via:

$$\hat{\alpha} = \frac{L_1V_1 + L_2V_2 + \ldots + L_pV_p}{p}. \tag{8}$$

Given a system $\bar{S}$ minimizing $(\sigma^*)^2$, we can use (8) to find the estimate of α for the original system S from (much faster) simulation performed on $\bar{S}$.

Now we have not yet suggested how the system $\bar{S}$ might be chosen in order to ensure that a good speedup is obtained, or better still, to maximize this speedup, i.e. minimizing $(\sigma^*)^2$ in (5). In many ways, we have replaced one difficult problem (finding the probability of overflow) with another (finding $\bar{S}$ given S).

3. Fast simulation of networks of *GI/GI/1* queues

We consider a general open network of *GI/GI/1* queues. 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. We will assume that all queues are stable in the sense that $\gamma_i < \mu_i$, $\forall i$. All the 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*:

$$\gamma_i = \sum_{j=1}^{d} p_{ji}\gamma_j + \lambda_i, \quad i = 1, 2, \ldots, d, \tag{9}$$

and the routing probabilities satisfy:

$$\sum_{j=0}^{d} p_{ij} = 1, \quad i = 1, 2, \ldots, d. \tag{10}$$

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 an optimal simulation system $\bar{S}(\lambda_i', \mu_i', \gamma_i', p_{ij}')$, which is structurally the same as S, and is obtained from S by varying its parameters from λ_i, μ_i, γ_i, p_{ij} to λ_i', μ_i', γ_i', p_{ij}', and which is used for estimating α by simulation. Using a similar heuristic justification to that presented in the previous section, it is argued in [3] that the parameters for an optimal simulation system $\bar{S}$ $(\gamma_i', \mu_i', \gamma_i', p_{pj}')$ can be found as the arguments achieving

minimization in the following large-deviations approximation[1] for α:

$$\alpha \sim \exp\left\{-N \inf_{\lambda_i', \mu_i', \gamma_i', p_{ij}'} R\left[\sum_{i=1}^{d} \lambda_i' h_{\lambda_i}\left(\frac{1}{\lambda_i'}\right) + \sum_{i=1}^{d} \mu_i' h_{\mu_i}\left(\frac{1}{\mu_i'}\right) + \sum_{i=1}^{d} \min(\gamma_i', \mu_i') K_i\right]\right\}, \tag{11}$$

where

$$K_i = \sum_{j=0}^{d} p_{ij}' \log \frac{p_{ij}'}{p_{ij}}, \tag{12}$$

$$R = \frac{1}{\sum_i (\gamma_i' - \mu_i') 1_{\gamma_i' > \mu_i'}}, \tag{13}$$

where $h_{\lambda_i}(\cdot)$ is the Cramér transform [9] of the distribution of the external inter arrival times at queue i, and $h_{\mu_i}(\cdot)$ is the Cramér transform of the distribution of the virtual service times at queue i.

The infimum is subject to the following constraints:

$$\lambda_i', \mu_i', \gamma_i' \geq 0, \tag{14a}$$

$$0 \leq p_{ij}' \leq 1, \tag{14b}$$

$$\gamma_i' > \mu_i' \quad \text{for at least one } i, \tag{14c}$$

$$\sum_{j=0}^{d} p_{ij}' = 1, \tag{14d}$$

$$\gamma_i' = \sum_{j=1}^{d} p_{ji}' \min(\gamma_j', \mu_j') + \lambda_i'. \tag{14e}$$

We will use the symbols γ_i^*, λ_i^*, μ_i^* and p_{ij}^* to denote the optimal values of the γ_i', λ_i', μ_i' and p_{ij}' respectively. It has been argued [2, 3] that if the system $\bar{S}$ 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 N becomes large. In the next section, we will perform the minimization for a tandem network of $GI/GI/1$ queues using the method of Lagrange multipliers to satisfy the equality constraints.

[1] This approximation is quite crude, and is not of practical use for estimating α. However, arguments associated with the minimization (i.e. λ_i', μ_i', γ_i', p_{ij}') will be useful in estimating the statistics via simulation, even though the system generated will only be approximately optimal.

4. Technical lemmas

In this section, we state and prove a number of lemmas that will be used later in this paper.

LEMMA 1

Consider the problem of the previous section, in which the network under consideration is a tandem network of queues. For convenience, we assume that the queues are arranged with increasing numbers from left to right, so that a customer leaving one queue will enter the queue immediately to its right. The formula (13) for R can be written

$$R = \frac{1}{\lambda_1' - \mu_M'}, \tag{15}$$

where M is the unstable queue with the highest index, for the particular values of λ_1', $\mu_1', \ldots, \mu_d'$ chosen[2].

Proof

Let k be the index of the leftmost unstable queue in the simulation system. Then it is clear that

$$\gamma_k' = \lambda_1', \tag{16}$$

since the average rate of customers leaving a queue is the minimum of its arrival and virtual service rates. Similarly, let i be a queue that is unstable, and let j be the index of the first queue to the right of queue i that is also unstable. Then we have

$$\gamma_j' = \mu_i'. \tag{17}$$

Because of this,

$$\sum_{i=1}^{d} (\gamma_i' - \mu_i') \mathbf{1}_{\gamma_i' > \mu_i'} = \gamma_k' - \mu_M', \tag{18}$$

where M is the index of the rightmost unstable queue. Then, combining the above statements, we must have

$$R = \frac{1}{\lambda_1' - \mu_M'}. \tag{19}$$

□

[2] This makes queue M the rightmost unstable queue.

LEMMA 2

Using the same notation as above, for given $h_\mu(\cdot)$, there is a unique value of μ' that solves

$$\frac{1}{\mu'} h'_\mu\left(\frac{1}{\mu'}\right) - h_\mu\left(\frac{1}{\mu'}\right) = 0 \tag{20}$$

and this value is $\mu' = \mu$.

Proof

Consider the function

$$g(x) = h_\mu(x) - x h'_\mu(x), \tag{21}$$

for which

$$g'(x) = -x h''_\mu(x), \tag{22}$$

which is less than zero for all $x > 0$. It follows that $g(x)$ can have at most one zero in $x > 0$. By the properties of the Cramér transform, $x = \mu^{-1}$ is one such value, i.e. the $\mu' = \mu$ is the only solution. □

LEMMA 3

Let $F(\cdot)$ and $G(\cdot)$ be two probability distributions related by

$$G(z) = F(az + b). \tag{23}$$

Then their Cramér transforms $h_F(\cdot)$ and $h_G(\cdot)$ are related by

$$h_G(y) = h_F(ay + b). \tag{24}$$

Proof

The Laplace transform of $F(\cdot)$ is given by:

$$M_F(s) = \int e^{sz} \, dF(z) \tag{25}$$

and its Cramér transform by

$$h_F(y) = \sup_{s \in \mathbb{R}} [sy - \log M_F(s)]. \tag{26}$$

The Laplace transform of $G(\cdot)$ is given by

$$M_G(s) = \int e^{sz} dG(z) \tag{27}$$

$$= \int e^{(s(z-b))/a} dF(z)$$

$$= e^{-sb/a} M_F\left(\frac{s}{a}\right), \tag{28}$$

and its Cramér transform by

$$h_G(y) = \sup_{s \in \mathbb{R}} [sy - \log M_G(s)] \tag{29}$$

$$= \sup_{s \in \mathbb{R}} \left[sy + \frac{sb}{a} - \log M_F\left(\frac{s}{a}\right) \right]$$

$$= \sup_{s \in \mathbb{R}} [s(ay + b) - \log M_F(s)]$$

$$= h_F(ay + b). \tag{30}$$

□

LEMMA 4

Consider a single $GI/GI/1$ queue with average arrival rate λ and average virtual service rate μ, and let the average arrival and virtual service rates in the fast simulation system to be λ^* and μ^* respectively. Then

$$\lambda^* \geq \lambda, \tag{31a}$$

$$\mu^* \leq \mu. \tag{31b}$$

Proof

Let λ^*, μ^* be the values of λ' and μ' achieving the infimum in the minimization problem, obtained when the setup of section 3 is specialized to a single queue. It can be shown that the minimum satisfies [3]:

$$h_A\left(\frac{1}{\lambda^*}\right) + h_B\left(\frac{1}{\mu^*}\right) = \left(\frac{1}{\lambda^*} - \frac{1}{\mu^*}\right) h_A'\left(\frac{1}{\lambda^*}\right) \tag{32}$$

$$= \left(\frac{1}{\mu^*} - \frac{1}{\lambda^*}\right) h_B'\left(\frac{1}{\mu^*}\right). \tag{33}$$

(This can also be established from (11).) The values of λ^* and μ^* satisfying these

equations are the average arrival and service rates for the optimal simulation system.

In (32), $h_A(\cdot)$ is the Cramér transform of the distribution of inter arrival times. This distribution has mean $1/\lambda$. Now, the left hand side of (32) is non-negative, since the Cramér transform is non-negative for all values of its argument [9]. Therefore, the right hand side must also be non-negative. The constraint (14c) in conjunction with (11) means that $\lambda^* > \mu^*$. Hence, applying a known convexity property of the Cramér transform [9] gives $\lambda^* \geq \lambda$. Similarly, we can show $\mu^* \leq \mu$. □

5. Tandem network of *GI/GI/1* queues

For a tandem network of d *GI/GI/1* queues, the cost function H that we must minimize to find the optimal simulation system is:

$$H = R\left[\lambda_1' h_{\lambda_1}\left(\frac{1}{\lambda_1'}\right) + \sum_{i=1}^{d} \mu_i' h_{\mu_i}\left(\frac{1}{\mu_i'}\right)\right], \tag{34}$$

where

$$R = \frac{1}{\sum_i (\gamma_i' - \mu_i') 1_{\gamma_i' > \mu_i'}}. \tag{35}$$

The minimization is to be performed subject to the constraint

$$\gamma_i' > \mu_i' \quad \text{for at least one } i. \tag{36}$$

These equations are obtained from (11) by eliminating the routing probabilities p_{ij} and the external arrival streams at all queues other than the first (i.e. the leftmost) queue, and letting λ_1 be the external arrival rate at queue 1.

Using lemma 1, we can rewrite (34) as

$$H = \frac{1}{\lambda_1' - \mu_M'}\left[\lambda_1' h_{\lambda_1}\left(\frac{1}{\lambda_1'}\right) + \sum_{i=1}^{d} \mu_i' h_{\mu_i}\left(\frac{1}{\mu_i'}\right)\right], \tag{37}$$

where M is the index of the rightmost unstable queue in the simulation system.

The minimum will be achieved where the partial derivatives of H are zero:

$$\frac{\partial H}{\partial \lambda_1'} = -RH - \frac{1}{\lambda_1'} h'_{\lambda_1}\left(\frac{1}{\lambda_1'}\right) + h_{\lambda_1}\left(\frac{1}{\lambda_1'}\right) \tag{38a}$$

$$= 0,$$

$$\frac{\partial H}{\partial \mu'} = \begin{cases} RH - \dfrac{1}{\mu_i'} h'_{\mu_i}\left(\dfrac{1}{\mu_i'}\right) + h_{\mu_i}\left(\dfrac{1}{\mu_i'}\right) & i = M, \\ -\dfrac{1}{\mu_i'} h'_{\mu_i}\left(\dfrac{1}{\mu_i'}\right) + h_{\mu_i}\left(\dfrac{1}{\mu_i'}\right) & \text{otherwise,} \end{cases} \tag{38b}$$

$$= 0.$$

Therefore, by lemma 2, (38b) implies that for $i \neq M$, $\mu_i^* = \mu_i$. We are now left with two problems that must be solved to find the optimal simulation system:

(1) to find the value of M that is optimal;
(2) given M, to find the values of λ_1^* and μ_M^*.

Calculation of λ_1^* and μ_M^* given M

Given M, and the fact that the optimal value of μ_i' is $\mu_i^* = \mu_i$ for $i \neq M$, the optimization problem of (11) reduces to that which arises for the case of an isolated $GI/GI/1$ queue. This has been solved previously [3]. Hence λ_1^* and μ_M^* are the unique solution of

$$h_{\lambda_1}\left(\frac{1}{\lambda_1^*}\right) + h_{\mu_M}\left(\frac{1}{\mu_M^*}\right) = \left(\frac{1}{\lambda_1^*} - \frac{1}{\mu_M^*}\right) h'_{\lambda_1}\left(\frac{1}{\lambda_1^*}\right) \tag{39}$$

$$= \left(\frac{1}{\mu_M^*} - \frac{1}{\lambda_1^*}\right) h'_{\mu_M}\left(\frac{1}{\mu_M^*}\right), \tag{40}$$

with $\lambda_1^* > \lambda_1$ and $\mu_M^* > \mu_M$.

In other words, if M is known, the parameters of the optimal simulation system can be calculated analytically.

Calculation of M

In the general case, we cannot offer an analytic solution to the problem of finding M, and a numerical solution must be found. We note that, at worst, this involves a search over d possible values of M. However, in an important class of problems, a simple analytic solution does exist.

Assume there exists a function $h_C(\cdot,\cdot)$ with the properties that

(1) for specific values of the first argument, it interpolates the Cramér transform $h_{\mu_i}(\cdot)$ by

$$h_C(\mu_i,\cdot) = h_{\mu_i}(\cdot); \tag{41}$$

(2) it is differentiable with respect to its first argument, and

$$\frac{\partial h_C\left(\mu,\frac{1}{\mu'}\right)}{\partial \mu} > 0 \tag{42}$$

for all $\mu \in [\min_i(\mu_i), \max_i(\mu_i)]$ and $\mu' < \mu$.

Observe now, using (38b) that

$$H = \frac{1}{\lambda_1' - \mu_M'}\left[\lambda_1' h_{\lambda_1}\left(\frac{1}{\lambda_1'}\right) + \mu_M' h_{\mu_M}\left(\frac{1}{\mu_M'}\right)\right] \tag{43}$$

$$= \frac{1}{\lambda_1' - \mu_M'}\left[\lambda_1' h_{\lambda_1}\left(\frac{1}{\lambda_1'}\right) + \mu_M' h_C\left(\mu_M,\frac{1}{\mu_M'}\right)\right]. \tag{44}$$

We now wish to minimize H with respect to M, λ_1' and μ_M'. Clearly, since $h_C(\cdot,\cdot)$ is an increasing function of its first parameter, we must choose this to be as small as possible, i.e.

$$M = \underset{n\in[1\ldots d]}{\operatorname{argmin}}\, \mu_n. \tag{45}$$

In other words, M is the index of the queue with the smallest service rate, i.e. the most heavily loaded queue. The values of λ_1^* and μ_M^* can now be found using the procedure described above.

Further, we note that because the optimization problem solved for the tandem network is the same as that solved for the isolated $GI/GI/1$ queue, lemma 4 requires $\mu_M^* \leq \mu_{M'}$. Since μ_M is the smallest of the service rates, we have that $\gamma_{M+1} = \mu_M^* \leq \mu_M < \mu_{M+1} = \mu_{M+1}^*$. In other words, the arrival rate at queue $M+1$ is less than its virtual service rate in the simulation system. Noting that $\mu_i^* = \mu_i$ for $i \neq M$, and that this implies that $h_{\mu_i}(1/\mu_i^*) = 0$ for $i \neq M$ [7], and applying the above argument to each queue to the right of queue M in sequence, we are guaranteed that no queue to the right of queue M is unstable in the simulation system, i.e. $\gamma_i^* < \mu_i^*$ for all $i > M$.

Two examples that fit this prescription are:

(1) Where the distribution ($F_i(\cdot)$ at queue i) of the rate of virtual services can be expressed in the form

$$F_i(x) = F(a_i x), \tag{46}$$

for all $1 \leq i \leq d$, and some $F(\cdot)$ and positive constants a_i.

Let μ be the rate associated with $F(x)$ and μ_i the rate associated with $F_i(x)$. Also, let $h(\cdot)$ be the Cramér transform of $F(\cdot)$ and $h_{\mu_i}(\cdot)$ the Cramér transform of $F_i(\cdot)$. Then by lemma 3

$$h_{\mu_i}(y) = h(a_i y). \tag{47}$$

Let $a_i = \mu_i/\mu$. We can define $h_C(\cdot,\cdot)$ as

$$h_C\left(\mu_i, \frac{1}{\mu_i'}\right) = h_{\mu_i}\left(\frac{1}{\mu_i'}\right). \tag{48}$$

We observe also that

$$h_C\left(\mu_i, \frac{1}{\mu_i'}\right) = h_C\left(a_i\mu, \frac{1}{\mu_i'}\right) \tag{49}$$

$$= h\left(\frac{a_i}{\mu_i'}\right). \tag{50}$$

Then the derivative of $h_C(\cdot,\cdot)$ with respect to its first parameter can be found via:

$$\frac{\partial h_C\left(\mu_i, \frac{1}{\mu_i'}\right)}{\partial \mu_i} = \frac{1}{\mu}\frac{\partial h_C\left(a_i\mu, \frac{1}{\mu_i'}\right)}{\partial a_i} \tag{51}$$

$$= \frac{1}{\mu}\frac{\partial h\left(\frac{a_i}{\mu_i'}\right)}{\partial a_i}, \tag{52}$$

which is positive for $\mu_i' < \mu_i$, by the properties of the Cramér transform [9].

This form applies where the difference between the distributions of the virtual service times at the queues in the network is simply a time scaling. This may be relevant where a number of identical servers are connected in parallel to provide increased service capacity at particular queues.

(2) Where the distribution of the rate of virtual services can be expressed in the form

$$F_i(x) = F(x + b_i) \tag{53}$$

for all $1 \le i \le d$, and some $F(\cdot)$ and constants b_i. Applying an argument very similar to that in the previous example, it can be shown that this case fits the above assumptions concerning the existence of $h_C(\cdot,\cdot)$.

This form corresponds to the case where the service times at each queue in the network have the same distributions, except for a different constant delay at each node.

In both of these cases, as would be expected from the above result, the optimal solution involves changing the external arrival rate and the virtual service rate at the most heavily loaded queue, i.e. the queue with the smallest service rate.

An example of a network for which we can calculate analytically the optimal simulation system is a tandem network of $M/D/1$ queues. The parameters of the original and optimal simulation systems for a number of such networks are shown in table 1. In this particular example, all the virtual service rates remain unchanged in the simulation system, including that at node M; only the rate of external arrivals (λ_1) is takes a different value in the simulation system.

6. Conclusion

In this paper, we have shown how the results of [3] can be used to find an optimal importance sampling simulation system for a class of tandem networks of $GI/GI/1$ queues, without the need for a numerical minimization. In this process, we have derived some results that are applicable to the generation of a fast simulation system for an arbitrary tandem network of $GI/GI/1$ queues. In particular, we have shown that in the fast simulation system, the service distribution is different from that in the original system for only one queue in the network.

Further work is required to extend these results to arbitrary networks of $GI/GI/1$ queues. The principal added difficulty in the general case comes from the presence of the routing probabilities in the cost function to be minimized in finding the optimal simulation system, especially where there is feedback around the queue that dominates the overflow statistics.

Table 1
Optimal simulation systems for a number of tandem networks of $M/D/1$ queues.

λ_1	μ_1	μ_2	λ_1^*	μ_1^*	μ_2^*
0.5	1	2	1.76	1	2
0.5	2	1	1.76	2	1
0.7	1	2	1.38	1	2
0.7	2	1	1.38	2	1
0.9	1	2	1.11	1	2
0.9	2	1	1.11	2	1

References

[1] H. Kahn, Use of different Monte Carlo sampling techniques, in: *Proc. Symp. on Monte Carlo Methods*, University of Florida, ed. H.A. Meyer (Wiley, 1954).

[2] M. Cottrell, J.C. Fort and G. Malgouyres, Large deviations and rare events in the study of stochastic algorithms, IEEE Trans. Auto. Contr. AC-28 (1983) 907–918.

[3] S. Parekh and J. Walrand, A quick simulation of excessive backlogs in networks of queues, IEEE Trans. Auto. Contr. AC-34 (1989) 54–66.

[4] K.S. Shanmugam and P. Balaban, A modified Monte-Carlo simulation technique for the evaluation of error rate in digital communication systems, IEEE Trans. Commun. COM-28 (1980).

[5] M.C. Jeruchim, Techniques for estimating bit error rate in the simulation of digital communications systems, IEEE J. Sel. Areas Commun. SAC-2 (1984).

[6] B.R. Davis, An improved importance sampling method for digital communication system simulations, IEEE Trans. Commun. COM-34 (1986).

[7] M.R. Frater, T.M. Lennon and B.D.O. Anderson, Optimally efficient estimation of the statistics rare events in queueing networks, IEEE Trans. Auto. Contr. AC-36 (1991) 1395–405.

[8] M.R. Frater, J. Walrand and B.D.O. Anderson, Optimally efficient simulation of buffer overflows in queues with deterministic service times, Australian Telecommun. Res. 24 (1990) 1–8.

[9] S.R.S. Varadhan, *Large Deviations and Applications* (Society for Industrial and Applied Mathematics, Philadelphia, PA, 1984).