Efficient simulation of overflow probabilities in queues with breakdowns

Dirk P. Kroese[*] Victor F. Nicola[†]

Abstract

Efficient importance sampling methods are proposed for the simulation of a single server queue with server breakdowns. The server is assumed to alternate between the operational and failure states according to a continuous time Markov chain. Both, continuous (fluid flow) and discrete (single arrivals) sources are considered. In the fluid flow model, we consider Markov-modulated fluid sources and a constant output rate when the server is operational. In the discrete arrivals model, we consider Markov-modulated Poisson sources and exponential service time when the server is operational.

We show how known results on Markov additive processes may be applied to determine the optimal (exponentially tilted) change of measure for both models. The concept of effective bandwidth is used in models with multiple independent sources. Empirical studies demonstrate the effectiveness of the proposed change of measures when used in importance sampling simulations.

Keywords: Analysis methodology, rare event simulation, importance sampling, overflow probability, Markov-modulated rate processes, Markov additive processes, effective bandwidth.

1 Introduction

Many models for communication and manufacturing systems consider the behaviour of a reservoir (buffer or queue) which operates in a random environment. One usually distinguishes between *continuous* and *discrete* models. In the first case the content of the reservoir is viewed as a continuous fluid, in the second case the reservoir contains only discrete items, e.g., customers or packets. Examples of such models and their applications may be found in [10] and Chapter 6 of [9], respectively. Typically, in these models the net input 'rate' into the buffer depends (only) on the current state of a 'regulating' Markov process (which we refer to as the *environment* process). For example, in a fluid flow queue the net input rate into the buffer depends on the state of the input source (e.g., 'on' or 'off') and the state of the server (e.g., 'operational' or 'failed'). The stochastic process that describe the temporal evolution of the content of the buffer is sometimes called a *Markov-modulated rate process* (MMRP).

The literature abounds with exact analytical results for MMRPs whose environment processes have *finite* state spaces. It is often possible to derive exact formulas, e.g., for the steady-state distribution of the reservoir. On the other hand, the analysis is often complicated by various computational difficulties. For example, when the overall environment process is composed of a large number of sub-processes, the corresponding state space 'explodes,' leading to excessive memory requirements and very large computation times. Other numerical problems arise in connection with rare event probabilities. To compute these small probabilities, one typically has to solve a set of linear equations which are ill-conditioned, leading to unreliable answers. Based on the concept of *effective bandwidth*, a major reduction in computational effort for so-called *separable* MMRPs was reported in [4] and [10]. Basically, a MMRP is separable if it is composed of independent sub-processes. Usually, these sub-processes are modelled as reversible Markov processes with small state spaces.

Exact analysis of a *general* MMRP is often not possible, and one has to resort to either simulation or approximation techniques. However, since overflow

[*] Faculty of Mathematical Sciences, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands. e-mail: d.p.kroese@math.utwente.nl

[†] Faculty of Electrical Engineering, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands, e-mail: v.f.nicola@el.utwente.nl.

probabilities are typically small, standard simulation is very inefficient. One way to improve the efficiency of the simulation is to use *importance sampling*, which requires determining an appropriate change of measure (see, e.g., [3] for simulation of ATM intree networks.) In this paper we show how such a change of measure can be obtained for certain MMRPs that can be used to model queues with breakdowns. The concept of Markov additive processes serves as a unifying theory for the continuous and discrete flow models. The concept of effective bandwidth can be used to reduce the computational complexity for models with multiple independent sources.

Section 3 briefly outlines the relevant theory for the continuous fluid flow model, along with numerical procedures to determine the change of measure to be used in simulation. Similar treatment for the discrete queueing model is given in Section 4. Empirical results for examples of both models demonstrate the validity and effectiveness of our methodology.

2 Preliminaries

In this paper we focus on two related MMRP models, both describing a single server queue with server breakdowns, operating in a random environment. In the first model the content of the queue is viewed as a fluid, in the second model the queue contains only discrete items. We will refer to these queues as the *fluid queue* and the *discrete queue* respectively.

In both models, the input (continuous or discrete) to the queue is 'modulated' by a continuous time Markov chain (CTMC) (I_t), with finite state space E and infinitesimal generator (Q-matrix) Q. How the input 'rate' to the queue depends on this modulating chain will be specified later. The server is assumed to alternate between the operational and failure state according to an alternating renewal process (M_t). The *environment process* is defined to be the stochastic process $(J_t) := (I_t, M_t)$. Finally, the content of the queue at time t will be denoted by X_t.

Two performance measures are of particular interest: the probability of a buffer overflow and the stationary distribution of the content of a buffer. In this paper we concentrate on the probability of overflow, starting from a certain buffer level and a given state for the environment process, before the buffer empties again. We will denote the overflow level by K. For the fluid queue the overflow probability, $p(K)$ say, has an *exponential decay*, i.e. we have

$$\lim_{K\to\infty} \frac{\log p(K)}{K} = -\bar{\theta}.$$

For the discrete queue we have, similarly, a *geometric decay*, such that

$$\lim_{K\to\infty} \frac{\log p(K)}{K} = \log \bar{z}.$$

We call $\bar{\theta}$ and $\bar{z}$ the *(assymptotic) decay rates* for the fluid and discrete queue, respectively. Notice that by definition $\bar{\theta} > 0$ and $0 \leq \bar{z} \leq 1$.

Notation Throughout this paper we will use the notation $\Delta(\mathbf{a})$ to denote the diagonal matrix derived from a vector $\mathbf{a}$.

3 Fluid queues

Consider a fluid queue in which the reservoir (*buffer*) is filled at rates which vary according to the current state of the CTMC (I_t) defined in Section 2. Specifically, the *input rate* is $r_i \geq 0$ whenever state $i \in E$ is visited. The buffer has a constant output rate c (when not empty). Let $\mathbf{r}$ be the vector of input rates, and let $R := \Delta(\mathbf{r})$ denote the corresponding diagonal matrix. Finally, let $I_+ := \{i \in E : r_i > c\}$, $I_- := \{i \in E : r_i < c\}$ and $I_0 := \{i \in E : r_i = c\}$. We assume that $|I_+| > 0$ (otherwise, an overflow can never happen.)

Remark 1 Notice that we can easily include *breakdown* of the server into the above (standard) model. We simply take the entire environment process (J_t) as our regulating process – instead of just (I_t) – and adapt the input rates; this of course *provided* that (J_t) is a CTMC. In particular, (M_t) needs to have a Markov structure. A server with exponential failure and repair times (independent of everything else) would expand the state space of the regulating process by a factor of 2. Note that the notion of server breakdown could be included in the framework of [8].

3.1 Overflow probabilities

Next, we consider the overflow probabilities as defined in Section 2. Specifically, starting at level x and the environment Markov chain in state i, let $p_i(x), i \in E$, be the probability that the buffer reaches overflow level K before it becomes empty. The reader may verify that the vector $\mathbf{p}(x)$ of overflow probabilities satisfies the following differential equation:

$$(R - c\,I)\,\mathbf{p}'(x) = -Q\,\mathbf{p}(x), \quad 0 < x \leq K. \qquad (1)$$

Note that $p_i(x), i \in I_0$, can be expressed in terms of $p_i(x), i \in \{I_+ + I_-\}$, which are obtained by solving a reduced system of differential equations, with the boundary conditions:

$$p_i(0+) = 0,\ i \in I_-, \text{ and } p_i(K) = 1,\ i \in I_+.$$

In order to find $\mathbf{p}(x)$, we need to specify $p_i(0)$ for $i \in I_+$; these can be determined by setting $x = K$ in the reduced system of (1). This yields a system of $|I_+ + I_-|$ linear equations, exactly $|I_+|$ of which have the left-hand-side equal to 1, thus giving just enough equations to be able to determine the unknowns boundary probabilities.

3.2 Efficient simulation

Although (1) and the corresponding boundary conditions give us complete knowledge of the overflow vector $\mathbf{p}(x)$, in practice we may run quickly into numerical problems. For example, when K grows large, the system of $|I_+|$ linear equations to determine $\mathbf{p}(0)$ becomes ill-conditioned, leading to unreliable numerical results. The dimension of the state space may cause other numerical problems. In such cases, simulation may be a valid option. But, since overflow is typically a rare event, we need an efficient simulation procedure, such as those based on importance sampling.

The appropriate exponential change of measure to be used in importance sampling follows from Markov additive theory. Below we just describe the basics; for details we refer to [2].

The decay rate $\bar{\theta}$, as defined in Section 2, is the smallest strictly positive eigenvalue of the eigenvalue equation

$$-Q\,\mathbf{w} = \theta\,(R - c\,I)\,\mathbf{w}. \tag{2}$$

Let $\overline{\mathbf{w}}$ denote the corresponding right-eigenvector. Define the *conjugate* Q-matrix of the fluid queue, $\bar{Q} = (\bar{q}_{ij})$ by putting

$$\bar{q}_{ij} = q_{ij}\,\frac{\overline{w}_j}{\overline{w}_i},\ i \neq j,$$

where $\overline{w}_i$ is the ith element of $\overline{\mathbf{w}}$. Importance sampling involves simulating the system with $\bar{Q}$ instead of Q, and weighing the simulated events by the corresponding likelihood ratios.

3.3 Decay rate

The decay rate $\bar{\theta}$ has been the subject of numerous studies, not only because of its relevance for efficient simulation based on importance sampling, but also because it gives important information about asymptotics of the overflow probabilities and steady-state distributions. We describe two numerically efficient methods to find the decay rate and the corresponding right-eigenvector in (2).

Power method

The first method is useful for general MMRPs (also non-separable, e.g., those representing models with single or multiple/correlated Markov modulated sources), particularly when the matrix Q is sparse. The decay rate $\bar{\theta}$ of (2) can be found by power iteration as follows. For $\varepsilon > 0$, let $A := Q + \varepsilon\,(R - c\,I)$. Next, put $\mathbf{w}^{(0)} := \mathbf{1}$ (vector of 1's) and, for $n \geq 0$, define

$$\begin{aligned} \mathbf{w}^{(n+1)} &:= \frac{A\mathbf{w}^{(n)}}{||\,A\mathbf{w}^{(n)}\,||}, \\ y^{(n+1)} &:= ||\,A\,\mathbf{w}^{(n+1)}\,||. \end{aligned}$$

When ε is 'close enough' to $\bar{\theta}$, the sequence $\{y^{(n)}\}$ converges to $\varepsilon - \bar{\theta}$, and $\{\mathbf{w}^{(n)}\}$ to the corresponding eigenvector $\overline{\mathbf{w}}$. An obvious difficulty is that we need to choose the 'shift' ε properly.

Effective bandwidth method

The second method uses the concept of effective bandwidth and applies to separable MMRPs (e.g., those representing models with multiple independent Markov modulated sources.) The basis of the method was laid in [6] and [8] and generalizations were made in [4]. The effective bandwidth of a fluid source characterized by (Q, R) is the function g such that $g(\theta)$ is the output capacity required to give the overflow probabilities a decay rate θ, for any initial environment state i and a starting level x. In particuler, it can be shown that $g(\theta)$ is the maximal real eigenvalue of the matrix

$$\frac{1}{\theta}\,Q + R.$$

We note that for a two-state source, an expression for the effective bandwidth as a function of the decay rate is given in [4]. (In this paper we refer to the decay rate as a strictly positive quantity.)

We may now determine the optimal exponential change of measure in (2) as follows. First we determine the decay rate $\bar{\theta}$ by solving

$$g(\theta) = c.$$

Then, we determine the right-eigenvalue $\overline{\mathbf{w}}$ of the matrix $Q/\bar{\theta} + R$ corresponding to the eigenvalue $g(\bar{\theta}) = c$.

The power of the effective bandwidth concept lies in the fact that a similar procedure can be followed when dealing with a separable MMRP. Specifically, suppose we have an input source that consists of many independent sub-sources; the kth sub-source is defined by the matrices $Q^{(k)}$ and $R^{(k)}$, respectively, and has effective bandwidth $g^{(k)}$. The decay rate $\bar{\theta}$ of the total system is the unique $\theta > 0$ satisfying

$$\sum_k g^{(k)}(\theta) = c, \tag{3}$$

where the summation is over all sources. The conjugate transition rates for each source are given by

$$\bar{q}_{ij}^{(k)} = q_{ij}^{(k)} \frac{\overline{w}_j^{(k)}}{\overline{w}_i^{(k)}},\ i \neq j,$$

where $\overline{\mathbf{w}}^{(k)}$ is the right-eigenvector of $Q^{(k)}/\bar{\theta} + R^{(k)}$ corresponding to the eigenvalue $g^{(k)}(\bar{\theta})$.

Remark 2 Notice that we may view a server with exponential on and off times (independent of everything else) as a two-state source with a input rate 0 when the server is operational, and input rate c when the server is failed. See Figure 1 for an illustration.

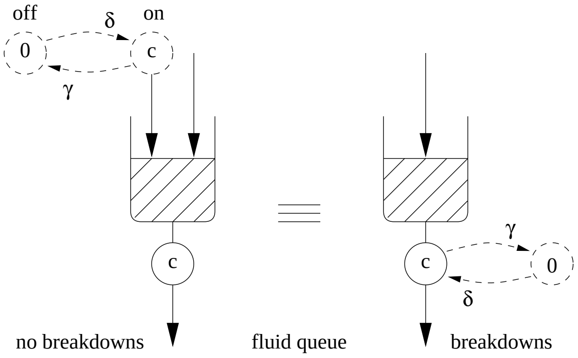

Figure 1: *A fluid queue with server breadowns*

3.4 A fluid queue with breakdowns

As an illustration of validity of the approach described above, we consider a fluid queue consisting of 10 independent on-off sources, 5 of Type 1 and 5 of Type 2, and an unreliable server. Sources of Type i ($i = 1, 2$) have exponential on- and off-times with parameters α_i and β_i, respectively. When a source of type i is active ('on') it sends fluid to the buffer at rate r_i. The up- and down-times of the server have exponential distributions with parameters γ and δ respectively, independent of the input process. The capacity of the buffer is denoted by c.

For the simulation we have taken the following parameters: $\alpha_1 = 3$, $\alpha_2 = 1$, $\beta_1 = 2$, $\beta_2 = 4$, $r_1 = 3$, $r_2 = 6$, $\gamma = 3$, $\delta = 4$, $c = 100$. Using the separability of the system and the effective bandwidth method, we easily find that the decay rate $\bar{\theta}$ is 0.20243. Already for an overflow level of $K = 5$ it is difficult to solve equation (1) using standard (numerical) methods due to badly conditioned matrices. Instead, we use importance sampling in a simulation procedure to estimate the overflow probabilities, starting from the following system state: the buffer is empty, the server is 'down', one source of Type 2 is 'on' and the rest of the sources are 'off.'

K	$\hat{p}$ (IS)	RE (IS)	$\hat{p}$ (SS)	RE (SS)
5	1.888e-01	5.0e-03	1.87e-01	2.1e-02
10	6.220e-02	5.5e-03	6.25e-02	3.9e-02
20	7.853e-03	5.7e-03	8.60e-03	1.1e-01
40	1.368e-04	5.7e-03	2.00e-04	7.1e-01
80	4.143e-08	5.9e-03	–	–
160	3.815e-15	5.8e-03	–	–
320	3.279e-29	5.9e-03	–	–
640	2.397e-57	5.9e-03	–	–

Table 1: *Estimation of overflow probabilities for a fluid queue with breakdowns. Importance sampling (IS) versus standard simulation (SS) results. RE denotes the relative error of the estimate, i.e., standard deviation/mean.*

For each estimate we perform a simulation of 10000 independent replications, all starting from the same system state, as described above. Each replication ends when either the overflow level is reached or the buffer empties. Each simulation (to obtain one estimate) lasted less than two minutes. Table 1 lists the estimates of the overflow probabilities and their relative errors for different overflow levels, K.

The results clearly indicate the efficiency of importance sampling (IS) which maintains a bounded relative error for estimates of extremely small values of overflow probabilities. (Compare with standard simulation (SS).)

4 A single server queue in a random environment

Next, we consider a discrete flow queueing model. Let (I_t) be the CTMC of Section 2. This chain regulates the arrivals to, and departures from, an ordinary queue in such a way that when (I_t) is in state

$i \in E$ customers arrive according to a Poisson process at rate λ_i, and are served singly (for an exponential time) at rate μ_i. Let $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ denote the vectors of arrival rates and service rates, respectively. The content of the queue at time t is again denoted by X_t. Obviously, the joint process (I_t, X_t) is a Markov process with an infinite block diagonal Q-matrix. The diagonal elements of each $(|E| \times |E|)$ block are given by the vector $-\boldsymbol{\tau} = -(\mathbf{q} + \boldsymbol{\mu} + \boldsymbol{\lambda})$, where $-\mathbf{q}$ is the vector of diagonal elements of Q.

4.1 Overflow probabilities

We wish to calculate (or estimate) the probability of overflow of some level K, as defined in Section 2. Starting with the buffer at level $X_0 = x$ and the regulating process in state $I_0 = i$, define T as the first time that the queue either hits level K or becomes empty. We are interested in the probabilities

$$\mathbb{P}(J_T = j,\, X_T = K \,|\, J_0 = i,\, X_0 = x),\ i, j \in E,$$

which we collect (in the obvious way) into a matrix $P(x)$. To find $P(x)$ we first define the matrices $S_k, k = 1, 2, \ldots$, such that the (i, j)th element of S_k is the probability of entering level $k+1$ at environment state j before reaching level 0, starting from level k at environment state i. Let S_0 be the 0-matrix, the reader my verify that for $k = 1, 2, \ldots$

$$\Delta(\boldsymbol{\tau})\, S_k = \Delta(\boldsymbol{\lambda}) + (Q + \Delta(\mathbf{q}))\, S_k + \Delta(\boldsymbol{\mu})\, S_{k-1} S_k,$$

It follows that $S_k,\ k = 1, 2, \ldots$, can be determined recursively from

$$S_k = (\Delta(\boldsymbol{\lambda} + \boldsymbol{\mu}) - Q - \Delta(\boldsymbol{\mu})\, S_{k-1})^{-1}\ \Delta(\boldsymbol{\lambda}). \quad (4)$$

Moreover, the matrix $P(x)$ is given by

$$P(x) = S_x\, S_{x+1} \cdots S_{K-1}. \quad (5)$$

4.2 Efficient simulation

Solving $P(x)$ from (4) and (5) may in practice be difficult due to numerical problems similar to those discussed in the previous section. Here too, we may use importance sampling to avoid these problems. As in the fluid queue case, the appropriate change of measure follows from the general theory of Markov additive processes. Next we give the main results; for details we refer to Section 17.5.2 of [1].

Consider the matrix

$$G(z) = Q + \Delta((1/z - 1)\boldsymbol{\lambda} + (z - 1)\boldsymbol{\mu}).$$

The decay rate $\bar{z}$, as defined in Section 2, is the largest $z \in (0, 1)$ such that $|G(z)| = 0$. Let $\overline{\mathbf{w}}$ denote the eigenvector of $G(\bar{z})$ corresponding to the eigenvalue 0. As before, define the conjugate Q-matrix $\bar{Q} = (\bar{q}_{ij})$ by setting

$$\bar{q}_{ij} = q_{ij}\, \frac{\overline{w}_j}{\overline{w}_i},\ i \neq j.$$

Moreover, define conjugate arrival and service rates

$$\bar{\lambda}_i := \lambda_i / \bar{z},\ \ \bar{\mu}_i := \mu_i\, \bar{z},\ \ i \in E.$$

Importance sampling involves simulating the system with $\bar{Q}, \bar{\boldsymbol{\lambda}}$ and $\bar{\boldsymbol{\mu}}$ instead of the original parameters and weighing the simulated events by the corresponding likelihood ratios.

4.3 Decay rate

The decay rate $\bar{z}$ may be viewed as a *dominant eigenvalue* (Perron-Frobenius eigenvalue) of a certain matrix. Consider the definition of S_k, and suppose that $S_k \to S$ as $k \to \infty$, for some fixed matrix S. For large K, equation (5) suggests that the overflow probabilities decay geometrically, at a rate which is determined by the dominant eigenvalue of S. Moreover, from (4), S should satisfy

$$\Delta(\boldsymbol{\mu})\, S^2 + (Q - \Delta(\boldsymbol{\lambda} + \boldsymbol{\mu}))\, S + \Delta(\boldsymbol{\lambda}) = 0. \quad (6)$$

It turns out that $\bar{z}$ is indeed the dominant eigenvalue of S, and that $\overline{\mathbf{w}}$ is the corresponding right-eigenvector. Notice that S is closely related (but is not identical) to the *rate matrix* R of [9].

Power method

We may determine $S, \bar{z}$ and $\overline{\mathbf{w}}$ by the following iterative (power) method. Put $S^{(0)} := \Delta(\mathbf{0})$ (0-matrix) and $\mathbf{w}^{(0)} := \mathbf{1}$. For $n = 0, 1, \ldots$, define the recursions

$$\begin{aligned} S^{(n+1)} &:= (\Delta(\boldsymbol{\lambda} + \boldsymbol{\mu}) - Q)^{-1} \left\{ \Delta(\boldsymbol{\mu})\, (S^{(n)})^2 + \Delta(\boldsymbol{\lambda}) \right\}, \\ \mathbf{w}^{(n+1)} &:= \frac{S^{(n+1)}\, \mathbf{w}^{(n)}}{||S^{(n+1)}\, \mathbf{w}^{(n)}||}, \end{aligned}$$

and let $z^{(n)} := ||S^{(n)}\, \mathbf{w}^{(n)}||$. Then the sequences $\{z^{(n)}\}$ and $\{\mathbf{w}^{(n)}\}$ converge to $\bar{z}$ and the corresponding eigenvector $\overline{\mathbf{w}}$, respectively.

Effective bandwidth method

As in the fluid queue case we may determine $\bar{z}$ for separable MMRPs using the effective bandwidth concept. In Section 7 of [4], the effective bandwidth of a single MMPP (Markov-modulated Poisson process) source is given *only* for the case in which the service requirement is exponentially distributed and the server is reliable with a constant capacity or rate (i.e., independent of the environment.) Clearly, if we include server breakdowns in the model, then the service rate depends on the state of the server, and thus it is no longer a constant. However, we still can use the concept of effective bandwidth by considering the workload process (amount of work in the system), rather than the queue length process (number in the system.) In particular, consider the workload process in an MMPP/G/1 queue. Work arrives according to a MMPP and is released at constant rate 1. Let L_G denote the Laplace Stieltjes transform (LST) of the service time (i.e., the size of the workload increment at an arrival epoch) distribution G. In Section 17.5.2 of [1], Markov additive process theory is used to determine the decay rate, say $\bar{s}$, of the stationary distribution of the workload in the MMPP/G/1 system.

Now, suppose that the system is composed of many independent input sources, where the kth source is characterized by $Q^{(k)}$ and $\Lambda^{(k)} := \Delta(\boldsymbol{\lambda}^{(k)})$. We define the effective bandwidth for the workload of the kth source as the real function $g^{(k)}$ on $\mathbb{R}_+$, such that $g^{(k)}(s)$ is the maximal real eigenvalue of the matrix

$$\frac{Q^{(k)}}{s} + \frac{(L_G(-s) - 1)}{s}\Lambda^{(k)}.$$

This is analogous to the definitions in [5] for the effective bandwidth of the workload in an M/G/1 queue. It turns out that the decay rate $\bar{s}$ of the workload overflow probability is identical to the decay rate of the stationary workload distribution, and satisfies

$$\sum_k g^{(k)}(\bar{s}) = 1. \qquad (7)$$

The conjugate transition rates are given by

$$\bar{q}_{ij}^{(k)} = q_{ij}^{(k)} \frac{\overline{w}_j^{(k)}}{\overline{w}_i^{(k)}},\ i \neq j,$$

where $\overline{\mathbf{w}}^{(k)}$ is the right-eigenvector of matrix $[Q^{(k)} + (L_G(-\bar{s}) - 1)\Lambda^{(k)}]/\bar{s}$ corresponding to the eigenvalue $g^{(k)}(\bar{s})$. Moreover, the conjugate arrival rates are given by

$$\bar{\lambda}_i^{(k)} := \lambda_i^{(k)}\, L_G(-\bar{s}),\ i \in E$$

and the conjugate service time distribution $\bar{G}^{(k)}$ has LST

$$L_{\bar{G}}^{(k)}(s) := \frac{L_G^{(k)}(\bar{s} - s)}{L_G^{(k)}(\bar{s})}.$$

Finally, the decay rate $\bar{z}$ of the queue length overflow probabilities (as defined in Section 1) is given by (see Section 1 of [11])

$$\bar{z} = 1/L_G(-\bar{s}). \qquad (8)$$

Remark 3 It remains to note that *active* server breakdowns (during a customer's service) may be included in the above model by considering another MMPP/G/1 system, in which the service time (workload increment) is modified to include the breakdown periods during a customer's service. For Poisson failures, the distribution of this modified service time (sometimes referred to as the completion time or virtual service time) can be easily derived as a special case of [7]. If the server breakdowns are *independent* (i.e., they may also occur when the server is idle), then, strictly speaking, the system can be viewed as an MMPP/G/1 system in which the first (virtual) service time in a busy period has a different distribution. However, it is interesting to note that the (asymptotic) decay rates are identical to those of the (same) system with active server breakdowns. See Figure 2 for an illustration.

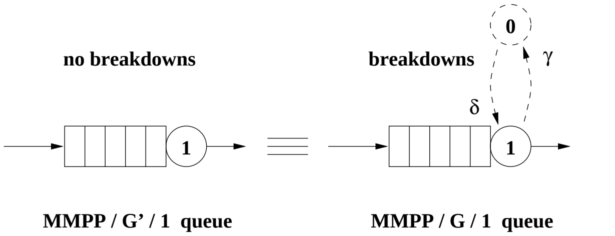

Figure 2: *A discrete queue with server breadowns*

4.4 An MMPP/M/1 queue with breakdowns

We illustrate the validity of our approach by considering a single server in a random environment which is identical to that considered in Section 3.4. That is, the server has a failure rate $\gamma = 3$ and a repair rate $\delta = 4$. The arrival process is the aggregate of 10 independent on-off sources, 5 of Type 1 (with $\alpha_1 = 3$ and $\beta_1 = 2$), and 5 of Type 2 (with $\alpha_2 = 1$ and $\beta_2 = 4$). When a source of is active ('on') it sends customers to the buffer according to a Poisson process at rate $\lambda_1 = 3$ (for Type 1 source) or $\lambda_2 = 6$ (for Type 2 source). The service time is exponentially distributed with a parameter $\mu = 100$. It follows that the LST of the modified service time distribution (including server breakdowns) is given by [7]

$$L_G(s) = \frac{\mu\,(\delta + s)}{(\mu + s)\,(\delta + s) + \gamma s}.$$

Using the effective bandwidth method described in Section 4.3, we find that the (geometric) decay rate $\bar{z}$ is 0.836993.

We use importance sampling to estimate the overflow probabilities for different overflow levels K, starting from the state in which one Type 2 source has just sent a 'customer' to an empty queue, all other nine sources are 'off', and the server is down. As in Section 3.4, the method of replication is used for simulating the system. Each replication starts at the same initial state and ends when either the overflow level is reached or the buffer empties. Table 2 lists estimates of the overflow probabilities and their relative errors, for both importance sampling and standard simulation. The number of replications used in each simulation to obtain one estimate is 10000, and each simulation lasted only a few minutes.

Again, the results indicate the efficiency of importance sampling (IS) which maintains a bounded relative error for estimates of extremely small values of overflow probabilities. (Compare with standard simulation (SS).)

K	$\hat{p}$ (IS)	RE (IS)	$\hat{p}$ (SS)	RE (SS)
5	2.979e-01	5.9-e03	2.97e-01	1.5e-02
10	1.052e-01	7.3e-03	1.10e-01	2.9e-02
20	1.626e-02	7.2e-03	1.63e-02	7.8e-02
40	4.599e-04	7.5e-03	9.00e-04	3.3e-01
80	3.675e-07	7.5e-03	–	–
160	2.453e-13	7.3e-03	–	–
320	1.056e-25	7.4e-03	–	–
640	1.975e-50	7.2e-03	–	–

Table 2: *Estimation of overflow probabilities for an MMPP/M/1 queue with breakdowns. Importance sampling (IS) versus standard simulation (SS) results and their relative errors.*

References

[1] S. Asmussen and R.Y. Rubinstein (1995). Steady state rare events simulation in queueing models and its complexity properties. In *Advances in Queueing: Theory, Methods and Open problems.* J.H. Dshalalow (ed.), New York: CRC Press, 429–461.

[2] S. Asmussen (1995). Stationary distributions for fluid flow models with or without Brownian noise. *Stochastic Models* **11** (1) 21–49.

[3] C.S. Chang, P. Heidelberger, S. Juneja and P. Shahabuddin (1994). Effective bandwidth and fast simulation of ATM intree networks. *Performance Evaluation* **20** 45–65.

[4] A.I. Elwalid and D. Mitra (1993). Effective bandwidth of general Markovian traffic sources and admission control of high speed networks. *IEEE/ACM Trans. Networking* **1** (3) 329–343.

[5] F. Kelly (1991). Effective bandwidths at multi-class queues. *Queueing Systems* **9** 5–16.

[6] L. Kosten (1984). Stochastic theory of data-handling systems with groups of multiple sources. In *Performance of Computer-Communication Systems*, H. Ruding and W. Bux (eds.), Elsevier, Amsterdam, 321–331.

[7] V.G. Kulkarni, V.F. Nicola and K.S. Trivedi (1987). The completion time of a job on a multi-mode system. *Adv. Appl. Probab.* **19** 932–954.

[8] D. Mitra (1988). Stochastic theory of a fluid model of producers and consumers coupled by a buffer. *Adv. Appl. Probab.* **20** 646–676.

[9] M.F. Neuts (1981). *Matrix-geometric solutions in stochastic models*, John Hopkins University Press, Baltimore.

[10] T.E. Stern and A.I. Elwalid (1991). Analysis of separable Markov-modulated rate models for information-handling systems. *Adv. Appl. Prob.* **23** 105–139.

[11] W. Whitt (1993). Tail probabilities with statistical multiplexing and effective bandwidths in multi-class queues. *Telecommunication Systems* **2** 71–107.