International Journal
of Electronics
and Communications

Efficient Simulation of Backlogs in Fluid Flow Lines

Dirk P. Kroese and Victor F. Nicola

Abstract **An Importance Sampling procedure to efficiently estimate overflow probabilities in continuous flow lines is described. The corresponding change of measure is found by generalizing the procedure that yields the change of measure for an ordinary fluid queue. Empirical results demonstrate the validity and effectiveness of the approach.**

Keywords **Continuous flow line, production line, overflow probability, efficient simulation, importance sampling, uniformization.**

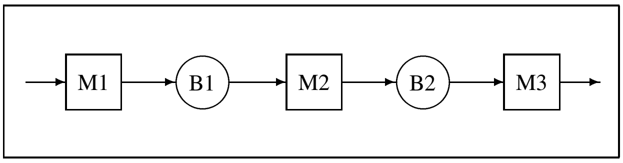

Fig. 1. A flow line with three machines and two buffers.

1. Introduction

A flow line is a tandem system of machines separated by buffers through which a stream of items/customers flows from one machine to the next. Flow lines are frequently encountered in manufacturing systems and other industrial processes, as well as in some computer and communication applications. A typical flow line is depicted in Fig. 1. We refer to a flow line with n machines and $n-1$ intermediate buffers as an *n-stage flow line*.

The design of a flow line involves the dimensioning of its buffers in order to keep the probability of overflow sufficiently low. Therefore, the study of backlogs in such systems is of particular interest. Among the factors that determine the backlog behaviour are the processing speed and reliability of each machine, as well as the size of the associated buffer. Two performance measures are of particular interest: the probability of a buffer overflow during a "busy cycle", and the stationary distribution of the content of a buffer. A comprehensive survey on flow lines may be found in [1]. Given the abundance of different flow lines, we remark that our model deals with a *continuous* flow line (i.e., the flow of products is modeled as a fluid flow) with *unreliable* machines (i.e., the machines are subject to service interruptions).

An exact analysis of flow lines is often not possible. Analytical results, mostly on stationary distributions, exist only for the most elementary systems. The 2-machine 1-buffer case (2-stage flow line) has been examined in [2], where the stationary distribution of the buffer content was found. For the 3-machine 2-buffer flow line with identical machines and (finite) buffers the joint stationary distribution of the buffers was found in [3]. Although in [4] more general 3-machine 2-buffer flow lines were considered, exact results were found only for a number of special cases which could be directly related to 2-stage flow lines. The fact that 3-stage flow lines are essentially more difficult to solve than 2-stage flow lines was demonstrated in [5], where a fluid tandem queue with on-off input was analyzed. This is basically a 3-stage flow line with one unreliable machine at the front of the line and two subsequent reliable machines. The joint stationary distribution of the content of the two buffers was expressed in terms of integrals of modified Bessel functions.

An alternative approach to study flow lines is simulation. However, since backlog is typically a rare event, standard simulation is very inefficient, i.e., excessively long simulation time is required to achieve an acceptable relative accuracy. Importance sampling has been used successfully to speed up simulations involving rare events. However, this involves determining an appropriate change of measure to be used in the importance sampling simulations. In relatively simple queueing and fluid flow systems, such a change of measure may be obtained from an asymptotic large deviations analysis (see, for example, [6] and [7]). We should point out that some care has to be taken here. As [8] shows, for some models, large deviations analysis may lead to inefficient changes of measure. In [9], a large deviation analysis for the study of backlogs in a fluid flow line with unreliable machines is considered. The results will be used in the present paper to speed up simulations of backlogs in a flow line using an importance sampling procedure. Preliminary empirical results demonstrate the validity and effectiveness of our approach.

The remainder of this paper is organized as follows. In Section 2 we introduce the model. We will restrict ourselves to the most simple (3-stage) flow line for which no exact results exist. In Section 3 we briefly review the theory on fluid queues; in particular we recall the importance sampling procedure for such queueing systems. The main contribution of this paper is in Section 4, where we pro-

Received November 11, 1997.

Dr. D. P. Kroese, Department of Applied Mathematics,
Dr. V. F. Nicola, Department of Electrical Engineering, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands, E-mail: d.p.kroese@math.utwente.nl., v.f.nicola@el.utwente.nl.

Int. J. Electron. Commun. (AEÜ) **52** (1998) No. 3, 1–7

pose a method to efficiently simulate a 3-stage flow line, using the analogy between fluid queues and flow lines. In Section 5 we give a flow line example to show how the IS change of measure is derived in practice. In Section 6 we discuss the uniformization approach and issues related to the implementation of importance sampling. And, in Section 7 we perform a simulation study of the flow line example using this change of measure. Finally, in Section 8 we give some conclusions and directions for further research.

2. The Model

We consider a continuous flow line with three machines and two buffers, as in Fig. 1. Each machine is subject to failures and repairs. The life and repair times have exponential distributions and are all independent of each other. We denote the corresponding *failure* and *repair rates* by λ_i and μ_i, respectively, $i = 1, 2, 3$. Each machine i has a specific *machine speed* ν_i, which is the maximum rate at which products can be processed. Finally, the capacities of the two buffers are given by C_1 and C_2, respectively. When a buffer has reached its capacity any excess input stream of products is lost.

Remark 1 Notice that the *actual* production speed of a machine i at a certain time may be lower than ν_i. Consider for example the following situation. Assume $\nu_1 > \nu_2 > \nu_3$. Suppose machine 1 has failed and machines 2 and 3 are working. If the second buffer is not empty, machine 3 works at rate ν_3. If the second buffer is empty but the first buffer is not, then machine 3 works at rate ν_2. The machine does not process any products when both buffers are empty. □

The state of the system at time t is determined by the state of each machine and the content of each buffer. Let $M_i(t)$ be the state of machine i at time t, $i = 1, 2, 3$ (Working = 1, Failed = 0) and let $Z_i(t)$ be the content of buffer i at time t, $i = 1, 2$. Also define $\mathbf{M}(t) = (M_1(t), M_2(t), M_3(t))$ and $\mathbf{Z}(t) = (Z_1(t), Z_2(t))$.

Obviously, the stochastic process $(\mathbf{M}(t), \mathbf{Z}(t))$ is *regenerative*. A regeneration point is recognized when all buffers are empty, all machines are operational, and the input Markov modulating process (into the first buffer) enters a specified state. We assume that the process is *stable*, by which we mean that the expected length of a regeneration cycle is finite. In this case $(\mathbf{M}(t), \mathbf{Z}(t))$ converges in distribution to some random vector $(\mathbf{M}, \mathbf{Z})$. The distribution of this vector is a major importance measure for the backlog behaviour of the system. However, our main concern will be the estimation of overflow probabilities of the *second*[1] buffer.

For a given buffer in the flow line, define a *regeneration overflow probability* as the probability that, starting from a regeneration point, the buffer content exceeds a given threshold before returning to level 0. We are primarily interested in the regeneration overflow probability of the second buffer. Large deviation results in [9] are used to derive a change of measure, which we implement in an importance sampling procedure to efficiently estimate this rare event probability.

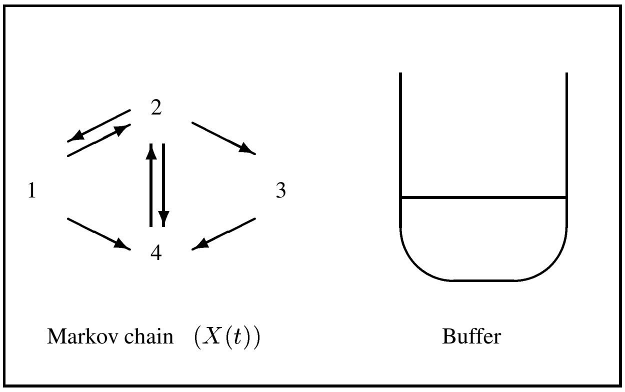

Fig. 2. A fluid queue consists of a buffer and a regulating Markov chain.

3. Efficient Simulation of Fluid Queues

In this section we briefly review the theory on fluid queues. These systems are closely related to flow lines. In fact, we may view a flow line as a tandem system of fluid queues which are subject to service interruptions.

A fluid queue is a queueing system (see Fig. 2) in which a reservoir is filled and depleted at rates which vary according to the current state of a regulating Markov chain. More precisely, let $(X(t))$ be a continuous-time Markov chain with a finite state space, say $E := \{1, \ldots, m\}$. This chain regulates the buffer content in such a way that net input rate is r_i whenever state $i \in E$ is visited, provided that the buffer is non-empty. A fluid queue is therefore defined by two quantities: the generator (Q-matrix), $Q = (q_{ij})$ say, of $(X(t))$ and the set $\{r_i\}$ of net input rates. It will be convenient to assemble the r_i's in a diagonal matrix $R := \mathrm{diag}\,(r_1, \ldots, r_m)$.

In principle, the joint stationary distribution of the Markov chain and the content of the fluid reservoir can be determined analytically, see e.g., [10] and [11]. Overflow probabilities may be established in the same way. However, if the state space is large, then the numerical calculations needed to determine the overflow probabilities become rather involved[2] and simulation becomes a valid option. However, since overflow is typically a rare event, efficient simulation requires the use of an appropriate change of measure in an importance sampling procedure.

[1] The overflow behaviour of the first buffer is well known from the theory of 2-stage flow lines.

[2] Basically, all the eigenvalues of the matrix $R^{-1}Q$ have to be calculated.

The following procedure establishes a change of measure which is optimal[3] in the class of exponentially twisted measures. Details can be found in [12, Section 17], [7, Chapter 3] or [13].

Firstly, determine the pair $(\alpha, \mathbf{w})$, satisfying the eigenvalue equation

$$Q\mathbf{w} = \alpha R\mathbf{w},$$

with α the largest negative eigenvalue (i.e., the negative eigenvalue which is closest to 0). Then, define a new Q-matrix $Q^* = (q^*_{ij})$ by putting

$$q^*_{ij} = q_{ij}\frac{w_j}{w_i}, \quad i \neq j,$$

where w_i is the ith element of $\mathbf{w}$. Q^* is called the *conjugate* Q-matrix of the fluid queue. The interpretation of Q^* is roughly the following: during an overflow period of the buffer, the driving Markov chain behaves like a chain with generator Q^* instead of Q. We now perform the simulation with Q^* instead of Q, and weigh the simulated events by the corresponding likelihood ratios. For example, if during a simulation run the driving chain has visited consecutively the states $i_0, i_1, \ldots, i_n$ with sojourn times $s_0, s_1, \ldots, s_n$, then the likelihood ratio of such a run is given by

$$\frac{\prod_{k=0}^{n-1} q_{i_k i_{k+1}} \exp\left(-\sum_{k=0}^{n} q_i s_i\right)}{\prod_{k=0}^{n-1} q^*_{i_k i_{k+1}} \exp\left(-\sum_{k=0}^{n} q^*_i s_i\right)},$$

where $q_i = \sum_{j\neq i} q_{ij}$ and $q^*_i = \sum_{j \neq i} q^*_{ij}$.

4. Efficient Simulation of a Flow Line

In this section we establish the change of measure to be used for the efficient simulation of the flow line in Fig. 1. To simplify the analysis somewhat, we only consider flow lines with 3 machines and 2 buffers in which the last machine is *perfect* (cannot fail). Without loss of generality, we assume that $\nu_1 > \nu_2 > \nu_3$ and the flow into machine 1 is continuous at a fixed rate higher than ν_1.

The key observation to make is that the second buffer may be viewed as the reservoir of a fluid queue that is regulated by the Markov process $(X(t)) := (M_1(t), M_2(t), Z_1(t))$ (notice that $M_3(t) = 1$). This driving process has a non-denumerable state space, so the theory of the previous section cannot be directly applied. However, we may reason analogously. For details we refer to [9].

The Markov process $(X(t))$ is completely specified by its infinitesimal generator, $\mathcal{Q}$ say. $\mathcal{Q}$ is an operator which operates on vector-valued test-functions $\mathbf{h} = (h_{00}, h_{01}, h_{10}, h_{11})^{\mathrm{T}}$, where the $h_{ij} : \mathsf{R} \to \mathsf{R}$ are differentiable functions of x, the content of the first buffer. Basically, $\mathcal{Q}$ maps a function $\mathbf{h}$ to the function $\mathcal{Q}\mathbf{h}$ for which the ij-th element has the following value at x:

$$\lim_{t \downarrow 0} \frac{\mathrm{E}_{ij,x}[h_{M_1(t)M_2(t)}(Z_1(t))] - h_{ij}(x)}{t},$$

where $\mathrm{E}_{ij,x}$ denotes the conditional expectation with respect to the event $\{M_1(0) = i, M_2(0) = j, Z_1(0) = x\}$.

We have in particular,

$$(\mathcal{Q}\mathbf{h})(x) = Q\mathbf{h}(x) + \begin{cases} (\mathcal{B}_0\mathbf{h})(0), & x = 0 \\ (\mathcal{B}\mathbf{h})(x), & 0 < x < C_1 \\ (\mathcal{B}_{C_1}\mathbf{h})(C_1), & x = C_1, \end{cases} \tag{1}$$

where $Q = (q_{ij,kl})$ is the Q-matrix[4] of $(M_1(t), M_2(t))$,

$$\mathcal{B}_0 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \nu_1\frac{\mathrm{d}}{\mathrm{d}x} & 0 \\ 0 & 0 & 0 & (\nu_1-\nu_2)\frac{\mathrm{d}}{\mathrm{d}x} \end{bmatrix},$$

$$\mathcal{B} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & -\nu_2\frac{\mathrm{d}}{\mathrm{d}x} & 0 & 0 \\ 0 & 0 & \nu_1\frac{\mathrm{d}}{\mathrm{d}x} & 0 \\ 0 & 0 & 0 & (\nu_1-\nu_2)\frac{\mathrm{d}}{\mathrm{d}x} \end{bmatrix},$$

$$\mathcal{B}_{C_1} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & -\nu_2\frac{\mathrm{d}}{\mathrm{d}x} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

The net input rate matrix R of the previous section is also replaced by an operator, namely $\mathcal{R}$, with

$$(\mathcal{R}\mathbf{h})(x) = \begin{cases} R_0\,\mathbf{h}(0), & x = 0 \\ R\,\mathbf{h}(x), & x > 0, \end{cases}$$

where

$$\begin{aligned} R_0 &= \mathrm{diag}\,(-\nu_3, -\nu_3, -\nu_3, \nu_1 - \nu_3), \\ R &= \mathrm{diag}\,(-\nu_3, \nu_2 - \nu_3, -\nu_3, \nu_2 - \nu_3). \end{aligned}$$

By analogy with the previous section we determine the optimal change of measure of the flow line via the following procedure. (Empirical results support our claim of asymptotic efficiency, however, a formal proof is yet to be established.) Firstly, determine the pair $(\alpha, \mathbf{w})$, satisfying the eigenvalue equation

$$\mathcal{Q}\mathbf{w}(x) = \alpha\mathcal{R}\mathbf{w}(x), \quad 0 \le x \le C_1, \tag{2}$$

where α is the largest negative eigenvalue. Next, define the *conjugate transition rates* as

$$q^*_{ij,kl}(x) = q_{ij,kl}\frac{w_{kl}(x)}{w_{ij}(x)}, \quad i,j,k,l \in \{0,1\}, \tag{3}$$

[3] By "optimality" we mean asymptotic efficiency (see, e.g., Glasserman and Wang [8] and others) which implies that the relative error asymptotically grows less than exponential with the level of buffer overflow.

[4] The order on the state space of $(M_1(t), M_2(t))$ is lexicographic: 00,01,10,11.

where $w_{ij}(x)$ is the ij-th element of $\mathbf{w}(x)$. Now perform the simulation with $q^*_{ij,kl}(x)$ instead of $q_{ij,kl}$, and weigh the simulated events by the corresponding likelihood ratio.

Notice that the conjugate rates depend on x. The interpretation is that in the conjugate flow line (that is, during an overflow period of the second buffer) the failure and repair rates depend on the content of the first buffer.

Two questions remain. How to find the dominant eigenvalue α and the corresponding function $\mathbf{w}$, and how to update the likelihood ratio. The first issue is perhaps easiest to address in an example.

5. Flow Line Example

Consider the flow line of Fig. 1 with the following parameters. $\lambda_1 = 5$, $\lambda_2 = 2$, $\mu_1 = 1$, $\mu_2 = 1$, $\nu_1 = 3$, $\nu_2 = 2$, $\nu_3 = 1$. The third machine is perfectly reliable. The buffer capacities are $C_1 = 1$ and $C_2 = \infty$.

Notice that with any pair $(\alpha, \mathbf{w})$ satisfying (2), the pair $(\alpha, c\mathbf{w})$, where c is a constant, satisfies the eigenvalue equation as well. We may therefore normalize $\mathbf{w}$ such that $w_{00}(0) = 1$. Also, we define $\kappa := w_{01}(0)$.

For every α, (2) represents a set of linear (differential) equations. The reader may verify that (2) is equivalent to the following:

$$\begin{aligned} w_{10}(0) &= 2 - \alpha - \kappa, \\ w_{11}(0) &= -2 + (3-\alpha)\kappa. \\ w_{00}(x) &= \frac{w_{01}(x) + w_{10}(x)}{2-\alpha}, \quad 0 \le x \le 1, \end{aligned}$$

and, for $0 < x < 1$,

$$\tilde{\mathbf{w}}'(x) = \begin{bmatrix} -\frac{-4+\alpha+\alpha^2}{2(-2+\alpha)} & \frac{1}{2-\alpha} & \frac{1}{2} \\ \frac{5}{3(-2+\alpha)} & \frac{7-8\alpha+\alpha^2}{3(2-\alpha)} & -\frac{1}{3} \\ -5 & -2 & 7+\alpha \end{bmatrix} \tilde{\mathbf{w}}(x), \quad (4)$$

where $\tilde{\mathbf{w}}(x) := (w_{01}(x), w_{10}(x), w_{11}(x))^{\mathrm{T}}$, and finally,

$$\begin{aligned} 5\,w_{00}(1) - 6\,w_{10}(1) + w_{11}(1) &= -\alpha w_{10}(1), \\ 5\,w_{01}(1) + 2\,w_{10}(1) - 7\,w_{11}(1) &= \alpha w_{11}(1). \end{aligned}$$

Thus, $\mathbf{w}(1)$ is completely determined by κ and α; and these parameters can be found through the last two boundary conditions above.

We find[5] $\alpha = -3.80464$ and $\kappa = 2.38007$. The 3×3-matrix in (4) has one real eigenvalue $\gamma = 3.64902$ and two complex conjugate eigenvalues $\zeta \pm \mathrm{i}\,\xi$, with $\zeta = 1.55101$ and $\xi = 1.11015$. It follows that, for $i, j \in \{0, 1\}$, the functions w_{ij} (defined in (2)) are of the form

$$w_{ij}(x) = a_{ij}\,\mathrm{e}^{\gamma x} + \mathrm{e}^{\zeta x}(b_{ij}\cos(\xi x) + c_{ij}\sin(\xi x)).$$

The numerical values of the constants are given in Table 1.

The conjugate intensities now follow from (3), and are depicted in Fig. 3. Notice that in the conjugate system the machine and repair rates depend on the state of the machines and the content of the first buffer.

[5] All constants in this section have been rounded off.

Table 1. The constants for $w_{ij}(x)$.

ij	a_{ij}	b_{ij}	c_{ij}
00	$9.5455 \cdot 10^{-3}$	$9.9045 \cdot 10^{-1}$	$7.3397 \cdot 10^{-1}$
01	$-1.7181 \cdot 10^{-2}$	2.3973	4.8641
10	$7.2589 \cdot 10^{-2}$	3.3520	$-6.0363 \cdot 10^{-1}$
11	$-1.3066 \cdot 10^{-1}$	$1.4326 \cdot 10^{1}$	4.3841

6. Uniformization and Implementation Issues

In Section 4 the conjugate transition rates (to be used in importance sampling) are given as continuous functions of the content of the first buffer. That is, according to this change of measure, the failure and repair rates of machines are no longer constants but functions of the content of the first buffer. This suggests a suitable implementation of importance sampling using uniformization [14]. In this section we give a brief review of uniformization as a method for sampling from non-homogeneous Poisson processes. We also consider some implementation details in relation to our flow line problem.

Uniformization is a simple technique for sampling (i.e., simulating) the times of certain stochastic processes including non-homogeneous Poisson processes, renewal processes, or Markovian processes in continuous time on either discrete or continuous state spaces (see, e.g. [15]–[18]). We describe it in the case of a non-homogeneous Poisson process. Let $(\lambda(t),\, t \ge 0)$ denote the intensity rate function of a non-homogeneous Poisson process $(N(t),\, t \ge 0)$ and assume that $\lambda(t) \le \beta$ for some finite constant β. Let T_n denote the time of the n-th event in a time homogeneous Poisson process $(N_\beta(t),\, t \ge 0)$ with a constant rate β. Then T_n is included (accepted) as an event in $(N(t))$ with probability $\lambda(T_n)/\beta$, otherwise the point is not included (rejected). Rejected events are sometimes called pseudo events. Renewal processes can be simulated using uniformization as described above provided $\lambda(t)$ is the hazard rate of the inter-event time distribution at time t. Uniformization can be generalized to cases in which the process being thinned is not a time homogeneous Poisson process (see [17]). For example, at time T_{n-1}, we can let $T_n = T_{n-1} + E_n$, where E_n has an exponential distribution with rate β_n. The point T_n is then accepted with probability $\lambda(T_n)/\beta_n$. This requires only that $\lambda(t) \le \beta_n$ for all $t \ge T_{n-1}$.

As in Section 4, let us consider the flow line with 3 machines and two buffers in which the third machine is perfectly reliable. The transition rate matrix $Q = (q_{ij,kl})$ is the generator matrix of $(M_1(t), M_2(t))$, where $i, j, k, l \in \{0, 1\}$. Let $q_{ij} = \sum_{kl \neq ij} q_{ij,kl}$, i.e., q_{ij} is the total transition rate in state ij.

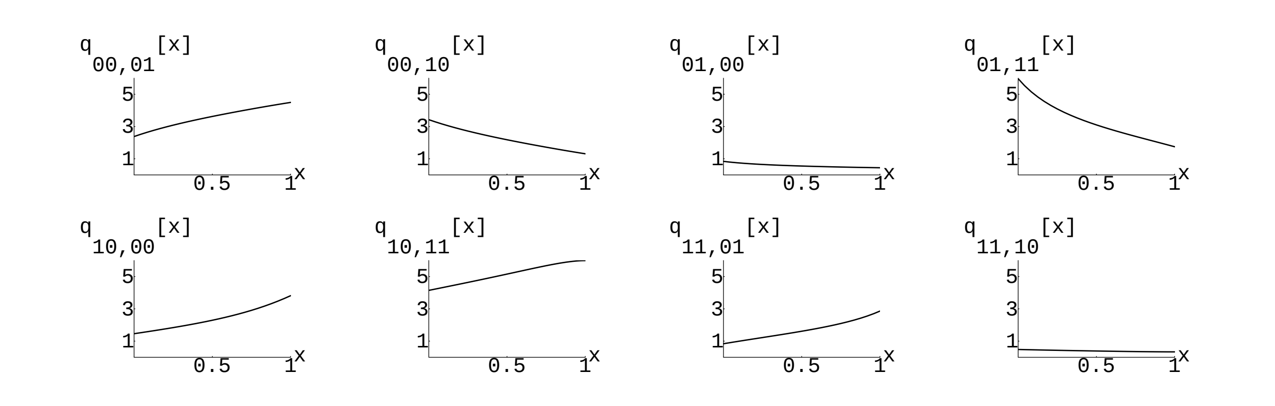

Fig. 3. Conjugate transition rates as functions of the buffer content x.

Uniformization can be used in flow line simulation as follows. Choose a constant uniformization rate β, such that $\beta > \max_{i,j\in\{0,1\}}\{q_{ij}\}$. Generate the times between uniformization events from an exponential distribution with a mean β^{-1}. In state ij $(i,j \in \{0,1\})$, a uniformization event is accepted as a real transition to state $kl \neq ij$ with probability $q_{ij,kl}/\beta$. It follows that the uniformization event is rejected (i.e., it is a pseudo event) with probability $1 - \sum_{kl\neq ij} q_{ij,kl}/\beta = 1 - q_{ij}/\beta$.

First, let us consider standard simulation to estimate the regeneration overflow probability p for the second buffer (as defined in Section 2). Generate n sample paths $\omega_i, i = 1,2,...,n$, each starting with a pure regeneration (all machines are operational and all buffers are empty) and ending with the content of the second buffer hitting overflow (a rare event) or hitting zero (a typical event), whichever occurs first. For each sample path ω_i, evaluate the indicator function $I_i = I(\omega_i)$, with $I(\omega) = 1$, if the sample path ω includes a buffer overflow, otherwise, $I(\omega) = 0$. An unbiased estimator for p is given by $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} I_i$. The relative error of this estimator is given by

$$\mathrm{RE}\,(\hat{p}) = \frac{\sqrt{\mathrm{VAR}\,(\hat{p})}}{E(\hat{p})} = \sqrt{\frac{1-p}{np}}.$$

It follows that $\mathrm{RE}\,(\hat{p}) \to \infty$ as $p \to 0$, which explains why standard simulation is not efficient for estimating very small values of p.

With importance sampling, the conjugate transition rates $q^*_{ij,kl}(x)$ (as determined from (3)) are used instead of $q_{ij,kl}$ to generate n samples of $I(\omega)L(\omega)$, where $L(\omega)$ is the likelihood ratio associated with a sample path ω. An unbiased estimator of p is given by $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} I_i L_i$. For the above change of measure, given by $q^*_{ij,kl}(x), i,j,k,l \in \{0,1\}$, it can be shown that the relative error of this estimator remains bounded as $p \to 0$.

At the beginning of a sample path (pure regeneration), the likelihood ratio L is set to one, and importance sampling is turned on. The likelihood ratio is updated with each uniformization (real or pseudo) event in that sample path as follows. In state ij, if an event is accepted as a transition to state $kl \neq ij$, then

$$L := L \times \frac{q_{ij,kl}}{q^*_{ij,kl}(x)}.$$

Otherwise, if the event is rejected (pseudo event), then

$$L := L \times \frac{1 - q_{ij}/\beta}{1 - q^*_{ij}(x)/\beta},$$

where q_{ij} and $q^*_{ij}(x)$ are the total transition rates out of state ij in the original and the conjugate matrices, Q and Q^*, respectively. Importance sampling is turned off at the occurrence of either an overflow or an empty (second) buffer, whichever occurs first.

There is a considerable freedom in choosing the uniformization rate β, provided it is higher than the maximum total transition rate in the current state, according to both the original and the conjugate transition matrices, Q and Q^*, respectively. (Note that, with importance sampling, the total transition rate in a given state changes with the content of the first buffer.) An appropriate uniformization rate should be low enough to limit excessive generation of pseudo events, yet high enough to yield less noisy estimates of the likelihood ratio. Experiments have shown that a good choice of the uniformization rate is 5 to 10 times the maximum total transition rate in the current state. The uniformization rate can also be set independent of the state, provided it is sufficiently high; however, this may lead to excessive generation of pseudo events in states with low total transition rates.

7. Simulation Results

In this section we perform a number of simulation experiments to demonstrate the validity of our approach. In particular, we use importance sampling to estimate the probability of overflow of the second buffer in the 3-stage flow line (described in Section 5) for several overflow levels, k. The optimal change of measure used in the IS method is a continuous function of the content of the first buffer, x (see (3)), which follows directly from the constants in Table 1. Using an implementation based on uniformization (as described in the previous section) we compare standard simulation (SS) with importance sampling (IS) simulation. In each simulation run, we collect a fixed number of observations, each starting with a "pure" regeneration (all machines are operational and all buffers are empty) and ending with the second buffer either reaching overflow or getting empty, whichever occurs first.

Let p be the probability that the content of the second buffer reaches the overflow level k, say, before hitting 0, starting with $M_1(0) = M_2(0) = 1$ and $Z_1(0) = Z_2(0) = 0$.

In order to appropriately set the uniformization level β, a number of preliminary IS simulation runs are performed, in which the relative error $\mathrm{RE}\,(\hat{p})$ is determined for increasing values of β (see Table 2). Each simulation run is based on 10^4 observations. In Table 2, we see that initially the relative error $\mathrm{RE}\,(\hat{p})$ decreases sharply with an increase in β. Improvement in accuracy quickly levels off with further increase in β. A uniformization level $\beta > 70$ yields a relative error close to the (asymptotic) minimum. However, an "optimal" β may be determined such that $\beta \times (\mathrm{RE}\,(\hat{p}))^2$ is minimal; this yields $\beta = 80$ (i.e., about eight times the maximum total transition rate in any state of the original and conjugate processes.) A uniformization level higher than $\beta = 80$ is not efficient, since it increases the simulation time without significantly improving the accuracy. Among other factors, the "optimal" uniformization rate depends on the (conjugate) hazard rate functions, which are used in importance sampling.

Table 2. Relative error $\mathrm{RE}\,(\hat{p})$ for increasing values of β.

β	$\mathrm{RE}\,(\hat{p})$	β	$\mathrm{RE}\,(\hat{p})$	β	$\mathrm{RE}\,(\hat{p})$
10	$27.6 \cdot 10^{-2}$	40	$3.31 \cdot 10^{-2}$	70	$2.35 \cdot 10^{-2}$
20	$5.29 \cdot 10^{-2}$	50	$2.65 \cdot 10^{-2}$	80	$2.14 \cdot 10^{-2}$
30	$3.81 \cdot 10^{-2}$	60	$2.75 \cdot 10^{-2}$	90	$2.08 \cdot 10^{-2}$
				100	$2.07 \cdot 10^{-2}$

In Table 3 we list the estimates $\hat{p}$ of p, using standard simulation and importance sampling, for various overflow levels k. For each estimate the half-width of the 95 %-confidence interval is included, expressed as a percentage of the estimate. All estimates are based on 10^6 observations, with a uniformization level $\beta = 80$. The results in Table 3 demonstrate a significant reduction of variance when using importance sampling, compared with standard simulation (for $k = 4$, no overflow events were observed with SS). Note that SS could be implemented more efficiently without uniformization; however, also in this case, experimental results show that IS (with uniformization) yields significant variance reduction for overflow levels as low as $k = 2$.

Notice also that the relative error increases slightly (at a linear rate) with the overflow level k. Numerical experiments (not reported here) suggest that such dependence does not occur when we use a sufficiently large β. Finally, it is interesting to note that

$$p \approx 0.7\ \mathrm{e}^{\,\alpha k},$$

for large k, where $\alpha = -3.80464$, as determined in Section 5.

Table 3. Estimates of overflow probabilities using standard simulation (SS) and importance sampling (IS).

k	$\hat{p}$ (SS)	$\hat{p}$ (IS)
1	$1.53 \cdot 10^{-2} \pm 1.32\,\%$	$1.515 \cdot 10^{-2} \pm 0.32\,\%$
2	$3.58 \cdot 10^{-4} \pm 8.69\,\%$	$3.529 \cdot 10^{-4} \pm 0.37\,\%$
3	$6.00 \cdot 10^{-6} \pm 78.0\,\%$	$7.809 \cdot 10^{-6} \pm 0.43\,\%$
4	—	$1.739 \cdot 10^{-7} \pm 0.48\,\%$

8. Conclusions and Further Research

The backlog behaviour of a given buffer in a flow line with unreliable machines is studied by viewing the buffer as a reservoir of a fluid queue driven by a Markov process. This modulating process depends on the state of the succeeding machine as well as all the preceding machine-buffer pairs. In general, this modulating process has a non-denumerable state space, with the number of continuous state variables equal to the number of the preceding buffers in the flow line.

Given a Markov-modulated characterization of its input, the backlog behaviour of the considered buffer can be simulated using large deviations results in an importance sampling procedure. It is shown that the "optimal" change of measure is not fixed between discrete events (such as machine failure and repair events), but it must be varied continuously depending on the levels of all the preceding buffers. This suggests a suitable implementation of importance sampling using uniformization.

For a model with three machines and two buffers, empirical results demonstrate the validity and the effectiveness of the approach proposed in this paper. Further research is needed to investigate the feasibility of this and/or other approaches for flow lines consisting of more than two buffer-machine pairs. Extensions of the approach to (discrete) tandem queues and to non-Markovian failure/repair machine behaviour are of practical interest and are currently being investigated.

References

[1] Dallery, Y.; Gershwin, S.: Manufacturing flow line systems: A review of models and analytical results. Queueing Systems **12** (1992), 3–94.

[2] Zimmern, B.: Etudes de la propagation des arrêts aléatoires dans les chaines de production. Rev. Statist. Appl. **4** (1956), 85–104.

[3] Coillard, P.; Proth, J.-M.: Sur l'effet de stocks tampons dans une fabrication en ligne. Rev. Belge Statist. Inform. et Recherche Oper. **24** (1984), 3–27.

[4] De Koster, M. B. M.; Wijngaard, J.: A continuous flow model for three production units in series with buffers. Proc. DGOR Operations Research, 1986. Berlin: Springer-Verlag, 253–264.

[5] Kroese, D. P.: Analysis of a Markov-modulated fluid tandem queue. Memorandum 1367, Enschede, The Netherlands: University of Twente, Faculty of Applied Mathematics, 1997.

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

[7] Mandjes, M.: Rare event analysis of communication networks. Dissertation. Amsterdam, The Netherlands: Free University, 1996.

[8] Glasserman, P.; Wang, Y.: Counterexamples in importance sampling for large deviations probabilities. Ann. Appl. Probab. **7** (1997), 731–746.

[9] Kroese, D. P.: How backlog builds up in a manufacturing flow line. In preparation, 1998.

[10] Anick, D.; Mitra, D.; Sondhi, M. M.: Stochastic theory of a data-handling system with multiple sources. Bell Syst. Tech. J. **61** (1982), 1871–1894.

[11] Rogers, L. C. G.: Fluid models in queueing theory and Wiener-Hopf factorization of Markov chains. Ann. Appl. Probab. **4** (1994), 390–413.

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

[13] Ridder, A.: Fast simulation of Markov fluid models. J. Appl. Prob. **33** (1996), 786–803.

[14] Nicola, V. F.; Heidelberger, P.; Shahabuddin, P.: Uniformization and exponential transformation: Techniques for fast simulation of highly dependable non-Markovian systems. Proc. 23rd Inte. Symp. Fault-Tolerant Computing, 1992. New York: IEEE Computer Society Press, 130–139.

[15] Gross, D.; Miller, D. R.: The randomization technique as a modeling tool and solution procedure for transient Markov processes. Operations Research **32** (1984), 343–361.

[16] Jensen, A.: Markov chains as an aid in the study of Markov processes. Skand. Aktuarietidskr. **36** (1953), 87–91.

[17] Lewis, P. A. W.; Shedler, G. S.: Simulation of non-homogeneous Poisson processes by thinning. Naval Research Logistic Quarterly **26** (1979), 403–412.

[18] Shanthikumar, J. G.: Uniformization and hybrid simulation/analytic models of renewal processes. Operations Research **34** (1986), 573–580.

Dirk P. Kroese was born in Wapenveld, The Netherlands, in 1963. He received his masters degree (Ir.) in 1986 and his Ph.D. degree in 1991 from the University of Twente, The Netherlands. Since 1991 he is Lecturer in Probability and Statistics at the Faculty of Applied Mathematics at the University of Twente. His professional interests include queueing theory, reliability, stochastic processes, point processes, large deviations theory, heavy- and light-traffic approximations, polling systems, fluid queues and simulation. Currently he is working on the analysis and simulation of fluid flow lines with applications to telecommunication and manufacturing systems.

Victor F. Nicola holds the Ph.D. degree in computer science from Duke University, North Carolina, the B.S. and the M.S. degrees in electrical engineering from Cairo University, Egypt, and Eindhoven University of Technology, The Netherlands, respectively. From 1979, he held faculty and research staff positions at Eindhoven University and at Duke University. In 1987, he joined IBM Thomas J. Watson Research Center, Yorktown Heights, New York, as a Research Staff Member. Since 1993, he has been an Associate Professor at the Department of Electrical Engineering, University of Twente, The Netherlands. His research interests include performance and reliability modeling, fault-tolerance, queueing theory, analysis and simulation methodologies, with applications to computer systems and telecommunication networks.