On the decay rates of buffers in continuous flow lines

D.P. Kroese (kroese@maths.uq.edu.au)
Department of Mathematics
The University of Queensland
Brisbane 4072
Australia

Abstract. Consider a tandem system of machines separated by infinitely large buffers. The machines process a continuous flow of products, possibly at different speeds. The life and repair times of the machines are assumed to be exponential. We claim that the overflow probability of each buffer has an exponential decay, and provide an algorithm to determine the exact decay rates in terms of the speeds and the failure and repair rates of the machines. These decay rates provide useful qualitative insight into the behaviour of the flow line. In the derivation of the algorithm we use the theory of Large Deviations.

Keywords: Continuous flow line, overflow probability, decay rate, methodology, dual flow line.

AMS Subject Classifications (1991): Primary 60K25, Secondary 90B22.

1. Introduction

A *flow line* (also called transfer line or production line) is a tandem system of machines separated by storage areas — which we will call buffers — through which a stream of items flows from one machine to the next. Flow lines are frequently encountered in manufacturing systems and other industrial processes, as well as in computer and communication applications. A typical flow line is depicted in Figure 1. For a comprehensive survey on flow lines we refer to (Dallery and Gershwin, 1992).

Figure 1. A flow line with three machines and two buffers (3-stage flow line).

Using the terminology of (Dallery and Gershwin, 1992), we consider a *continuous* flow line — in which we view the stream of products as a fluid flow — consisting of n machines and $n-1$ intermediate buffers, each buffer having infinite capacity. The life and repair times of the


© 2000 *Kluwer Academic Publishers. Printed in the Netherlands.*

machines are assumed to be independent of each other and exponentially distributed. Finally, the machines may have different processing speeds.

Important performance measures for this model are the *stationary distribution* of the content of a buffer, and the probability of a buffer *overflow* during a specified time interval. In this paper we focus on the overflow probability of a buffer, which we define as the probability that the buffer content exceeds a given threshold before returning to 0.

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 (Zimmern, 1956), 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 (Coillard and Proth, 1984). Although in (De Koster and Wijngaard, 1986) 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 (Kroese and Scheinhardt, 2001), where, amongst other fluid systems, a fluid tandem queue with on-off input is 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 is found and expressed in terms of integrals of modified Bessel functions. See also (Kella and Whitt, 1999) for a discussion on linear fluid networks.

Flow lines with more than two machines are usually analyzed via simulation and approximation methods. Approximations are often based on the principle of *decomposition* or *aggregation*. The idea is to decompose the flow line into a set of 2-stage subsystems which locally have the same behaviour as the original line. We refer again to (Dallery and Gershwin, 1992) for details.

Although the decomposition algorithms have proved to be very useful, we cannot expect them to yield accurate approximations of small overflow or steady-state probabilities. However, we claim that the overflow probability of a buffer has an *exponential decay*, i.e. the probability of reaching a high level L during a busy period, decreases exponentially with L. The purpose of this paper is to show how the corresponding decay rates for a general n-stage flow line can be determined. The exact knowledge of these decay rates provides useful qualitative insight into the behaviour of the flow line. As a by-product, we obtain information on the manner in which backlog builds up in a buffer.

We emphasize the *methodology* and the relative *simplicity* of the results. Related studies may be found in (De Veciana et al., 1993) and (Chang et al., 1994). In the latter article, an algorithm was proposed to efficiently simulate small buffer overflow probabilities in certain acyclic ATM queueing networks, using the Importance Sampling method. Although the "intree" network topology of (Chang et al., 1994) is more general than our series topology, the network components in (Chang et al., 1994) are completely reliable, whereas our flow line system allows for unreliable machines. Moreover, in (Chang et al., 1994) the concept of "effective bandwidth" has been used, which seems less natural in the context of flow lines. In fact, the method developed in the present paper seems simpler and is more in line with (De Veciana et al., 1993). For accessible accounts on Large Deviations, we refer to (Bucklew, 1990) and (Shwartz and Weiss, 1995).

The organization of the rest of the paper is as follows. In Section 2 we introduce the model and define the relevant parameters. Section 3, the main section of the paper, deals with the decay rates of (the overflow probabilities of) the buffers. First, we address the issue why the overflow probabilities should have an exponential decay. We then proceed to show how the corresponding decay rates can be determined via a minimization program. In Section 4 we illustrate the theory with a number of examples. Section 5 shows an alternative way, more algebraic in nature, to derive the decay rates. We conclude with some directions for future research.

2. The model

Consider a production line consisting of n machines in series and $n-1$ intermediate buffers. All buffers have infinite capacity. Each machine $i \in \{1, \ldots, n\}$ has a specific *machine speed* ν_i, which is the maximum rate at which it can transfer products from its upstream buffer to its downstream buffer. We view the flow of products as a fluid; in other words, we are dealing with a *continuous* flow line. The *lifetime* of machine i has an exponential distribution with parameter λ_i. The *repair* of machine i starts immediately after failure, and requires an exponential time with parameter μ_i. All life and repair times are assumed to be independent of each other. The first machine has an unlimited supply, and is therefore never "starved". The machines and buffers are numbered sequentially. For example, buffer i is situated between machines i and $i+1$, for $i \in \{1, \ldots, n-1\}$.

The *availability* a_i of machine i is the average fraction of time that the machine works (whether idle or not). By standard renewal theory,

we have

$$a_i = \frac{\mu_i}{\lambda_i + \mu_i},\ i = 1, \ldots, n.$$

The *Isolated Production Rate* (IPR) of a machine is defined as the product of the machine availability and speed. In other words, it is the average rate at which a machine would operate if it were "stand-alone" and had an unlimited supply of products.

The actual average amount of fluid that is processed by machine i, per unit of time, is called the (machine) *production rate*, p_i say. The production rate of the entire flowline, p say, is defined as the average amount of fluid that leaves the system per unit of time. Hence in our case, $p = p_n$. Moreover, it is not difficult to see that

$$p_k \ = \ \min_{i=1,\ldots,k} a_i\,\nu_i,\ \ k = 1, \ldots, n.$$

In particular, if the following *stationarity condition* holds,

$$a_1\,\nu_1 \ < \ a_i\,\nu_i,\ i = 2, \ldots, n, \tag{1}$$

none of the buffers will gradually build up to infinity and we have

$$p = p_1 = \cdots = p_n = a_1\,\nu_1.$$

This is the principle of *conservation of flow*, which states that in this case all machines have the same production rate (for a proof see the Appendix of (Dallery and Gershwin, 1992)).

A quantity that will be of particular interest for the next section is the average net input rate into buffer i when there is an infinite amount of fluid in the buffer. We will denote this by r_i. More precisely,

$$r_i := p_i - a_{i+1}\nu_{i+1} = \min_{j=1,\ldots,i} a_j\,\nu_j - a_{i+1}\,\nu_{i+1}. \tag{2}$$

(Recall that buffer i lies between machines i and $i+1$.)

The content of the ith buffer at time t is denoted by $Z_i(t)$. Let $M_i(t)$ be state of the ith machine; $M_i(t) = 1$ if the ith machine works at time t, and $M_i(t) = 0$ else. Let $\boldsymbol{Z}(t)$ and $\boldsymbol{M}(t)$ denote the corresponding random vectors.

Notation. Throughout this paper we use boldface letters for n- dimensional *or* $(n-1)$-dimensional row-vectors.

Assumption. For the rest of the paper we assume that stationarity condition (1) holds for the flow line defined by the parameters

$\{\lambda_i, \mu_i, \nu_i\}$. However, other (auxiliary) flow lines will be considered for which the corresponding condition does not hold.

3. Decay rates

In this section we investigate the overflow probability of a buffer in a general flow line. Since all buffers have infinite capacity, the behaviour of a buffer does not depend on the behaviour of its *downstream* buffers. We will therefore concentrate on the last buffer in the flow line, i.e. on buffer $n-1$. Notice that the net input rate into the buffer depends on the state $X(t) := (\boldsymbol{M}(t), Z_1(t), \ldots, Z_{n-2}(t))$ of the n machines and the $n-2$ other buffers. Indeed, we may view the last buffer as the reservoir of a *fluid queue* which is driven by the Markov process[1] $(X(t))$.

We are interested in the probability γ_L that the process $(Z_{n-1}(t))$, starting from 0, exceeds some threshold L before hitting 0. Of course this probability depends on the "starting states" of the machines and the other buffers at the beginning of the busy cycle of $(Z_{n-1}(t))$. To avoid trivialities, we only consider starting states for which the last buffer initially fills up. We claim that

$$\lim_{L\to\infty} \frac{\log \gamma_L}{L} = -\beta, \tag{3}$$

for some *decay rate* $\beta > 0$.

To see why this holds, assume without loss of generality that the busy cycle of $(Z_{n-1}(t))$ starts at time 0. Let $\mathbb{P}_x$ denote the probability measure under which $(X(t))$ starts at x. By E_L we denote the event that the content of the buffer reaches level L during the busy cycle. In particular, $\gamma_L = \mathbb{P}_x(E_L)$. Let T_L denote the first hitting time of level L when E_L occurs, or else put $T_L = \infty$. It is plausible that the conditional distribution $\rho_L(dx) := \mathbb{P}_x(X(T_L) \in dx \,|\, E_L)$ converges vaguely to a limiting distribution ρ, as $L \to \infty$, where ρ is independent of x.

Now, let v be a level such that the "entrance distribution" ρ_w at any level $w \geq v$ is "close" to the limiting entrance distribution ρ. Moreover, suppose that the probability of reaching level $z = w + a$, $(a > 0)$, starting from level w and from the limiting entrance distribution, depends approximately only on a and not on w, and is given by $h(a)$, for some

[1] Note, however, that the driving process $(X(t))$ has a non-denumerable state space; hence the well-established theory on fluid queues, see e.g. (Rogers, 1994), cannot be directly applied.

strictly positive function h. We then have[2]

$$\gamma_z \sim \gamma_v\, h(z-v) \sim \gamma_v\, h(w-v)\, h(z-w),$$

which leads to $\gamma_L \sim \exp(-\beta L)$, as claimed.

Remark 1. For standard fluid queues, where the driving Markov process has a finite state space, it is easy to check that the overflow probabilities are of the form (3). The limiting entrance distribution ρ for such processes can sometimes be determined explicitly and be used in efficient simulation techniques, see for example (Garvels and Kroese, 1999).

In the following proposition we describe a method to determine the decay rate for an arbitrary buffer in a general flow line. We first need some definitions.

Let $\tilde{\lambda}_1$ and $\tilde{\lambda}_n$ be strictly positive real numbers, and let I be a subset of $\{2, \ldots, n-1\}$. For any such $\tilde{\lambda}_1$, $\tilde{\lambda}_n$ and I, and for any $i \in \{2, \ldots, n-1\}$, let

$$\tilde{\lambda}_i^I := \begin{cases} \lambda_i & \text{if} \quad i \in I, \\ \sqrt{\frac{\mu_i\,\lambda_i\,\nu_i}{\mu_1\,\lambda_1\,\nu_1}\,(\mu_1\,\lambda_1 + \tilde{\lambda}_1^2) - \mu_i\,\lambda_i} & \text{else.} \end{cases} \tag{4}$$

Moreover, let

$$\tilde{b}_i := \frac{\mu_i\,\lambda_i}{\mu_i\,\lambda_i + \tilde{\lambda}_i^2},\ i \in \{1, n\}. \tag{5}$$

We may interpret $\tilde{b}_i$ as the availability of a machine with failure rate $\tilde{\lambda}_i$ and repair rate $\mu_i\lambda_i/\tilde{\lambda}_i$. Next, for any subset I define the function g^I on the set

$$D^I := \{(\tilde{\lambda}_1, \tilde{\lambda}_n) \geq 0\ :\ \tilde{b}_1\,\nu_1\ >\ \tilde{b}_n\,\nu_n;\ \ \tilde{b}_1\,\nu_1\ \leq\ \min_{i\in I} a_i\,\nu_i\} \tag{6}$$

by

$$g^I(\tilde{\lambda}_1, \tilde{\lambda}_n) := \frac{\sum_{i=1}^n \mu_i(\tilde{\lambda}_i^I - \lambda_i)^2/(\mu_i\,\lambda_i + (\tilde{\lambda}_i^I)^2)}{\tilde{b}_1\,\nu_1 - \tilde{b}_n\,\nu_n}, \tag{7}$$

When D^I is non-empty, by convexity arguments, each function g^I has a global minimum g^{*I} on D^I, possibly on the boundary.

[2] We will use the abbreviation $g(x) \sim h(x)$ to indicate that $g(x) = k(x)\,h(x)$, for some function k such that $\lim_{x\to\infty} x^{-1} \log k(x) = 0$.

PROPOSITION 1. The decay rate β of the stationary distribution of the $(n-1)$st buffer is given by

$$\beta = \min g^{*I}, \tag{8}$$

where the minimum is taken over all subsets $I \subset \{2, \ldots, n-1\}$ for which the set D^I is not empty.

Proof. For a heuristic proof we use the theory of Large Deviations along the lines of (De Veciana et al., 1993).

Let L be the overflow threshold and let E_L (as before) denote the event that an overflow[3] of level L occurs in the $(n-1)$st buffer, during a busy cycle of this buffer. Let $S := \{0,1\}^n$ be the state space of the Markov process $(\boldsymbol{M}(t))$. The generator of $(\boldsymbol{M}(t))$ is denoted by $Q := (q_{ij})$.

We assume the following is true.

1. The most likely path to an overflow is a *straight* line. In other words, overflows of the last buffer are due to a steady buildup of fluid.

2. In order to establish an overflow, $(\boldsymbol{M}(t))$ has to behave like a Markov process with a *different* generator $\tilde{Q} = (\tilde{q}_{ij})$ during an extended period of time.

For a justification of these assumptions we refer to (Anantharam, 1988), (De Veciana et al., 1993) and (Kesidis and Walrand, 1993).

By (Kesidis and Walrand, 1993), the probability that $(\boldsymbol{M}(t))$ behaves like a Markov process with generator $\tilde{Q}$ during an interval $[0,T]$ (T large) is approximately

$$\mathrm{e}^{-T\,H(\tilde{Q}\,||\,Q)},$$

where

$$H(\tilde{Q}\,||\,Q) = \sum_{i\in S} \pi_{\tilde{Q}}(i) \left(\sum_{j\neq i} \tilde{q}_{i,j} \log \frac{\tilde{q}_{i,j}}{q_{i,j}} + q_{i,j} - \tilde{q}_{i,j} \right) \tag{9}$$

is called the *relative entropy* of $\tilde{Q}$ with respect to Q. Here $\pi_{\tilde{Q}}$ denotes the stationary distribution of $(\boldsymbol{M}(t))$ under $\tilde{Q}$. Let $\tilde{r}$ denote the average net input rate into the $(n-1)$st buffer, when $(\boldsymbol{M}(t))$ has generator $\tilde{Q}$ and when the content of the buffer is infinite, as defined in (2). In other words, let

$$\tilde{r} := \min_{i=1,\ldots,n-1} \tilde{a}_i \nu_i - \tilde{a}_n \nu_n,$$

[3] The term *overflow* should not be taken too literally here, since all buffers have infinite capacity.

where $\tilde{a}_i := \tilde{\mu}_i/(\tilde{\mu}_i + \tilde{\lambda}_i)$ denotes the availability of machine i under $\tilde{Q}$.

Combining the results above, the most likely scenario leading to an overflow is the following: $(\boldsymbol{M}(t))$ behaves like a Markov process with generator $\tilde{Q}$, yielding a net input rate c into the last buffer, during an interval of length L/c, at the end of which level L is reached. We have asymptotically

$$\gamma_L \sim \sup_{c>0} \sup_{\{\tilde{Q}\,|\,\tilde{r}=c\}} \exp\left\{-\frac{L}{c}\,H(\tilde{Q}\,||\,Q)\right\} \sim \exp\left\{-L \inf_{\{\tilde{Q}\,|\,\tilde{r}>0\}} \frac{H(\tilde{Q}\,||\,Q)}{\tilde{r}}\right\},$$

which identifies β as

$$\beta = \min_{\{\tilde{Q}\,|\,\tilde{a}_i\nu_i > \tilde{a}_n\nu_n,\, i=1,\ldots,n-1\}} \frac{H(\tilde{Q}\,||\,Q)}{\min\limits_{i=1,\ldots,n-1} \tilde{a}_i\nu_i - \tilde{a}_n\nu_n,}. \tag{10}$$

Thus, β may be obtained by solving a complicated minimization program. Below, we show how to simplify this program.

First, we evaluate the entropy function, as given in (9). For the stationary probability $\pi_{\tilde{Q}}(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{M} = \boldsymbol{x})$, we have

$$\pi_{\tilde{Q}}(\boldsymbol{x}) = \prod_{i=1}^{n} \{(1-\tilde{a}_i)(1-x_i) + x_i\tilde{a}_i\}, \;\; \boldsymbol{x} \in S.$$

Consequently, the entropy function has the following form:

$$H(\tilde{Q}\,||\,Q) = \sum_{i=1}^{n} \{\phi_i(1-\tilde{a}_i) + \psi_i\tilde{a}_i\}, \tag{11}$$

with

$$\phi_i = \tilde{\mu}_i \log\left(\frac{\tilde{\mu}_i}{\mu_i}\right) + \mu_i - \tilde{\mu}_i, \;\; i = 1, \ldots, n,$$

and

$$\psi_i = \tilde{\lambda}_i \log\left(\frac{\tilde{\lambda}_i}{\lambda_i}\right) + \lambda_i - \tilde{\lambda}_i, \;\; i = 1, \ldots, n.$$

Next, we consider the Lagrangian

$$L(\tilde{\mathbf{a}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\mu}}, \mathbf{K}) := w(\tilde{\mathbf{a}}) \sum_{i=1}^{n} \{\phi_i(1-\tilde{a}_i) + \psi_i\tilde{a}_i\} + \sum_{i=1}^{n} K_i \left(\frac{\tilde{\mu}_i}{\tilde{\mu}_i + \tilde{\lambda}_i} - \tilde{a}_i\right),$$

with $w(\tilde{\mathbf{a}}) := 1/(\min_{i=1,\ldots,n-1} \tilde{a}_i\nu_i - \tilde{a}_n\nu_n)$. Note that minimizing L over all $(\tilde{\mathbf{a}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\mu}}, \mathbf{K})$ such that $w(\tilde{\mathbf{a}}) > 0$, solves (10).

Suppose now that L is minimal at $\theta^* = (\mathbf{a}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*, \mathbf{K}^*)$. Then the requirements $\frac{\partial L(\theta^*)}{\partial \tilde{\mu}_i} = 0$ and $\frac{\partial L(\theta^*)}{\partial \tilde{\lambda}_i} = 0$, $i = 1, \ldots, n$ lead to the equations

$$w(\mathbf{a}^*)\,(1 - a_i^*) \log\left(\frac{\mu_i^*}{\mu_i}\right) + K_i^* \frac{1 - a_i^*}{\mu_i^* + \lambda_i^*} = 0$$

and

$$w(\mathbf{a}^*)\, a_i^* \log\left(\frac{\lambda_i^*}{\lambda_i}\right) - K_i^* \frac{a_i^*}{\mu_i^* + \lambda_i^*} = 0,$$

from which it follows that

$$\mu_i^*\, \lambda_i^* = \mu_i \lambda_i,\ i = 1, \ldots, n. \tag{12}$$

Consequently, we may replace $\tilde{\mu}_i$ with $\lambda_i\, \mu_i / \tilde{\lambda}_i$ in (10) and (11). In particular, the right-hand side of (11) becomes $\sum_{i=1}^n \mu_i (\tilde{\lambda}_i - \lambda_i)^2 / (\mu_i\, \lambda_i + \tilde{\lambda}_i^2)$, where the log-terms cancel. By replacing also $\tilde{a}_i$ with $\tilde{b}_i := \mu_i\, \lambda_i / (\mu_i\, \lambda_i + \tilde{\lambda_i}^2)$ in (10) and (11), we are left with the minimization of a function g defined by

$$g(\tilde{\boldsymbol{\lambda}}) := \frac{\sum_{i=1}^n \mu_i (\tilde{\lambda}_i - \lambda_i)^2 / (\mu_i\, \lambda_i + \tilde{\lambda}_i^2)}{\min\limits_{i=1,\ldots,n-1} \tilde{b}_i\, \nu_i - \tilde{b}_n\, \nu_n}$$

on the set $A := \{\tilde{\boldsymbol{\lambda}} \geq 0 \,:\, \tilde{b}_i\, \nu_i > \tilde{b}_n\, \nu_n,\ i = 1, \ldots, n-1\}$.

Let us call the flow line for which g is minimal the *dual* flow line. We denote the failure and repair rates of machine i in the dual system by λ_i^* and $\mu_i^* = \mu_i \lambda_i / \lambda_i^*$, respectively; and we denote the corresponding availability by b_i^*, $i = 1, \ldots, n$. In particular, we cannot find any vector $\boldsymbol{\lambda}$ on A such that $g(\boldsymbol{\lambda}) < g(\boldsymbol{\lambda}^*)$.

Suppose that machine $k \in \{1, \ldots, n-1\}$ has the smallest Isolated Production Rate $b_k^*\, \nu_k$ in the dual flow line. Hence, for any machine $i \in \{1, \ldots, n-1\}$ either $b_i^*\, \nu_i > b_k^*\, \nu_k$ or $b_i^*\, \nu_i = b_k^*\, \nu_k$. For the first case we must have

$$\lambda_i^* = \lambda_i,$$

because any other choice for λ_i^* would give a higher value for (the numerator of) g. For the second case we have by definition

$$\lambda_i^* = \sqrt{\frac{\mu_i\, \lambda_i\, \nu_i}{\mu_k\, \lambda_k\, \nu_k} (\mu_k\, \lambda_k + \lambda_k^{*2}) - \mu_i\, \lambda_i}.$$

Moreover, we may assume that machine 1 has the smallest IPR in the dual flow line. For, suppose this is not the case. Then the arguments

above show that there exists a $k \neq 1$ such that $b_k^* \nu_k < b_1^* \nu_1 = a_1 \nu_1$. Define the vector $\boldsymbol{\lambda}'$ to be equal to $\boldsymbol{\lambda}^*$ except for the kth entry, λ_k^*, which is replaced with λ_k. The numerator of $g(\boldsymbol{\lambda}')$ is obviously larger than that of $g(\boldsymbol{\lambda}^*)$. Moreover, the denominator of $g(\boldsymbol{\lambda}')$ is smaller than that of $g(\boldsymbol{\lambda}^*)$, because $b_k^* \nu_k < a_1 \nu_1 \leq a_k \nu_k$ (the last inequality follows from the stability assumption (1)). Thus $g(\boldsymbol{\lambda}^*)$ is not minimal, which is a contradiction.

This shows that

$$\beta = \min_{I \subset \{2,\ldots,n-1\}} \min_{(\tilde{\lambda}_1, \tilde{\lambda}_n) \in D^I} \frac{\sum_{i=1}^n \mu_i (\tilde{\lambda}_i^I - \lambda_i)^2 / (\mu_i \lambda_i + \left(\tilde{\lambda}_i^I\right)^2)}{\tilde{b}_1 \nu_1 - \tilde{b}_n \nu_n},$$

where $\tilde{b}_1$ and $\tilde{b}_n$ are given in (5) and, λ_i^I in (4).

The latter minimization program can be easily divided into 2^{n-2} simpler minimization programs, leading to formulas (4), (7) and (8), and finally to Proposition 1.

Remark 2. We may interpret the results in the proof of Proposition 1 in the following way. In the dual flow line — i.e. during an overflow period of the last buffer — the first machine has the smallest IPR. Any other machine in the flow line either has the same IPR as the first machine or has its original failure and repair rate (as in the original system). In the latter case, the IPR of the machine is larger than that of the first machine. This case typically occurs when a machine has a much higher availability than the first machine. In the case where all dual rates differ from the original rates, the net input rates into buffers 1 through $n-2$ are 0, while the average net input rate into the last buffer is strictly positive. This "balancing property", perhaps holds for more general network topologies, making it easier to identify the decay rates is such networks.

Remark 3. We now indicate how the above method could be generalized to general life and repair time distributions. Consider a arbitrary machine, where the lifetimes $X_1, X_2, \ldots$ have density f and the repair times $Y_1, Y_2, \ldots$ have density g. Assume that the repair and lifetimes are independent of each other, and that the moment generating functions

$$M_X(s) := \mathbb{E}\mathrm{e}^{sX_i} \quad \text{and} \quad M_Y(s) := \mathbb{E}\mathrm{e}^{sY_i}$$

exist. Note that $X_1, Y_1, X_2, Y_2, \ldots$ is a so-called *alternating renewal process.*

Suppose that the life and repair times have different densities during the interval $[0, T]$. Specifically, suppose that we have the *exponential*

change of measures

$$\tilde{f}(x) = \frac{\mathrm{e}^{\theta} f(x)}{M_X(\theta)} \quad \text{and} \quad \tilde{g}(x) = \frac{\mathrm{e}^{\eta} g(x)}{M_Y(\eta)}.$$

Consider a path of the on-off (alternating renewal) process with jumps at $t_1, t_2, \ldots$. The likelihood of this path over $[0,T]$ under the new measure satisfies

$$\begin{aligned} L &= \frac{f(t_1)}{\tilde{f}(t_1)} \frac{g(t_2-t_1)}{\tilde{g}(t_2-t_1)} \frac{f(t_3-t_2)}{\tilde{f}(t_3-t_2)} \cdots \\ &\approx \mathrm{e}^{-\theta T - \eta T} \left(M_X(\theta) M_Y(\eta)\right)^N, \end{aligned}$$

where N is the total number of repairs before time T. Under the change of measure

$$N \approx \frac{T}{\tilde{\mathbb{E}}X + \tilde{\mathbb{E}}Y} = \frac{T}{(\log M_X)'(\theta) + (\log M_Y)'(\eta)}.$$

This suggests that the relative entropy of one on-off source (machine) is

$$-\theta - \eta + \frac{\log M_X(\theta) + \log M_Y(\eta)}{(\log M_X)'(\theta) + (\log M_Y)'(\eta)}.$$

Extending this to n independent machines gives an relative entropy

$$H = \sum_{i=1}^{n} \left(-\theta_i - \eta_i + \frac{\log M_X^{(i)}(\theta_i) + \log M_Y^{(i)}(\eta_i)}{(\log M_X^{(i)})'(\theta_i) + (\log M_Y^{(i)})'(\eta_i)} \right),$$

where the $M_X^{(i)}$ ($M_Y^{(i)}$) is the generating function of a lifetime (repair-time) of the ith machine.

The optimization, as in (10) now needs to be performed over the θ_i and η_i. Notice that for a given reliability $\tilde{a}_i$ of machine i, θ_i and η_i are related by

$$(1-\tilde{a}_i)\,(\log M_X^{(i)})'(\theta_i) = \tilde{a}_i\,(\log M_Y^{(i)})'(\eta_i).$$

Remark 4. It should be noted that the actual number of subsets I for which $D^I \neq \emptyset$ may be far smaller than 2^{n-2}. Also, if for a certain I, g^{*I} exists, then $g^{*I} \leq g^{*J}$, for $J \subset I$, so that we do not have to evaluate all possibilities. Observe that $g^{\emptyset}$ always has a solution.

Remark 5. If the last machine is *perfect* ($\lambda_n = 0$), the optimization program (8) should be slightly changed. Specifically, the functions g^I in (8) are not defined by (7), but instead by

$$g^I(\tilde{\lambda}_1) = \frac{\sum_{i=1}^{n-1} \mu_i (\tilde{\lambda}_i^I - \lambda_i)^2 / (\mu_i\,\lambda_i + (\tilde{\lambda}_i^I)^2)}{\tilde{b}_1\,\nu_1 - \nu_n},$$

on the set $D^I := \{\tilde{\lambda}_1 > 0 : \tilde{b}_1\,\nu_1 \;>\; \nu_n;\; \tilde{b}_1\,\nu_1 \;\leq\; \min_{i\in I}\; a_i\,\nu_i\}$, where $\tilde{b}_1$ and $\tilde{\lambda}_i$ are defined in (5) and (4).

Remark 6. It is tempting to use the "optimal parameters" $\lambda_1^*, \ldots, \mu_n^*$ to efficiently estimate overflow probabilities via the technique of *Importance Sampling*, as in (Chang et al., 1994). This would be valid if during the buildup of backlog in the last buffer the failure and repair rates would be *independent* of the contents of the upstream buffers. It turns out, however, see (Kroese and Nicola, 1998), that this is not the case. The actual, so-called *conjugate* rates *do* depend on the sizes of the upstream buffers. However, for large buffer contents, the machine rates are approximately constant and close to the optimal parameters above. It will now be clear why we chose the term *dual* flow line above (despite the overuse of this adjective) instead of *conjugate* flow line. See also Example 2.

Remark 7. In many fluid queueing models, the decay rate of a certain buffer overflow probability often coincides with the decay rate of the corresponding steady state distribution of the buffer. However, for the present model this is presumably not the case, mainly because the buffers have unlimited capacity. An analogy can be made with an ordinary 2-node Jackson tandem queue, where the decay rate of the overflow probability of the second buffer is not always equal to the decay rate of the stationary distribution of the second buffer, which is simply λ/μ_2, where λ is the arrival rate of customers and μ_2 is the service rate in the second queue.

4. Examples

In this section we illustrate the theory with a number of examples.

Example 1. Consider a flow line with three machines and two intermediate buffers. The corresponding machine speeds, failure and repair rates are given in the second, third and fourth column of Table I, respectively. We wish to determine the decay rate of the second buffer.

In view of (8), we have to minimize the functions $g^{\{2\}}$ and $g^{\emptyset}$ over the sets $D^{\{2\}} := \{(x,y) \in \mathbb{R}^2 : 2 \leq x < y\sqrt{2}\}$ and $D^{\emptyset} := \{(x,y) \in \mathbb{R}^2 : 0 < x < y\sqrt{2}\}$, respectively. In the first case we have

$$g^{\{2\}}(x,y) = \frac{x^2(-6+4y-3y^2) + x(32+8y^2) - 16(5-2y+2y^2)}{2(x^2-2y^2)}.$$

By straightforward algebra we find that $g^{\{2\}}(x,y)$ is minimal for $x=2$ and $y=2+\sqrt{2}$, with minimum $(7+4\sqrt{2})/(1+\sqrt{2}) \approx 5.24$.

For the second case we have

$$g^{\emptyset}(x,y) := \frac{x^2(-5+2y-2y^2)+x(24+6y^2)-48+16y-18y^2}{x^2-2y^2}.$$

The minimum $(55+7\sqrt{97})/24 \approx 5.164$ is attained in $((31-\sqrt{97})/12, (17+\sqrt{97})/8)$. Since this is smaller than the previous minimum, we have found the solution to the minimization program (8); and thus

$$\beta = \frac{55+7\sqrt{97}}{24}.$$

The original and dual rates are given in the table below. Notice that in the dual flow line the first buffer is "balanced", such that the average net input rate into the buffer is 0:

$$\nu_1 \frac{\mu_1^*}{\mu_1^*+\lambda_1^*} - \nu_2 \frac{\mu_2^*}{\mu_2^*+\lambda_2^*} = 0$$

Also, the average net input rate into the second buffer during an overflow period is

$$\nu_2 \frac{\mu_2^*}{\mu_2^*+\lambda_2^*} - \nu_3 \frac{\mu_3^*}{\mu_3^*+\lambda_3^*} = \frac{4753+12863\sqrt{97}}{286847} \approx 0.458219.$$

Table I. Original and dual parameters of the flow line of Example 1.

i	ν_i	λ_i	μ_i	λ_i^*	μ_i^*
1	1	4	2	$\frac{31-\sqrt{97}}{12}$	$\frac{96}{31-\sqrt{97}}$
2	1	1	2	$\frac{31-\sqrt{97}}{24}$	$\frac{48}{31-\sqrt{97}}$
3	1	1	4	$\frac{17+\sqrt{97}}{8}$	$\frac{32}{17+\sqrt{97}}$

Example 2. Consider a 3-stage flow line with parameters given in Table II. Notice that the last machine is perfect. In view of Remark 5 we have to minimize the functions $g^{\{2\}}$ and $g^{\emptyset}$ given by

$$g^{\{2\}}(x) = \frac{(x-5)^2}{10-x^2}$$

and

$$g^{\emptyset}(x) = \frac{75-20x+4x^2-2\sqrt{30}\sqrt{-5+2x^2}}{20-2x^2},$$

over the sets

$$D^{\{2\}} := \{x < \sqrt{10};\ \frac{5/x}{x + 5/x}\, 3 \leq \frac{2}{3}\}$$

and $D^{\emptyset} := \{x < \sqrt{10}\}$, respectively. However, $D^{\{2\}}$ is empty; and since $g^{\emptyset}$ attains its minimum at $x^* := 2.271275438\ldots$, we have

$$\beta = g^{\emptyset}(x^*) \approx 2.576668739.$$

The reader may check that x^* is a zero of the polynomial

$$-10000 + 23000\, x - 11825\, x^2 - 6900\, x^3 + 5750\, x^4 - 920\, x^5 + 16\, x^6.$$

Also, as in the previous example, the average net input rate into the first buffer is 0 in the dual system.

Table II. Original and dual parameters of the flow line of Example 2.

i	ν_i	λ_i	μ_i	λ_i^*	μ_i^*
1	3	5	1	2.271275438	2.201406274
2	2	2	1	0.842012211	2.375262464
3	1	0	∞	0	∞

In (Kroese and Nicola, 1998) a more detailed study of this very system is given. It turns out that the buildup of backlog in the second buffer does not quite happen in the way that is suggested in Section 3. During an overflow period, the machine rates (life and repair rates) do not remain constant, but depend on the content of the first buffer. Specifically, these so-called *conjugate* rates are of the form

$$q^*_{ij,kl}(x) = \frac{c_1 + c_2\, x}{d_1 + d_2\, x},\ x \geq 0,\ i,j,k,l \in \{0,1\}.$$

The connection with the dual rates is the following. As $x \to \infty$ the conjugate rates converge to the corresponding dual rates. For example, $q^*_{00,01}(x) \to \mu_2^*$, as $x \to \infty$. Thus μ_2^* could be interpreted as the limiting repair rate of the second machine in the conjugate system. Similar interpretations hold for λ_1^*, λ_2^* and μ_1^*.

Example 3. Consider the 3-stage system with parameters given in Table III. Copying the procedure of Example 2, we first have to minimize

$$g^{\{2\}}(x) := \frac{(x-3)^2}{6 - x^2}$$

on the set $D^{\{2\}} = \{x < \sqrt{6}\}$. This corresponds to the situation where in the dual system the second machine retains its original repair and failure rate. On $D^{\{2\}}$, $g^{\{2\}}$ is minimal at $x^* := 2$. Since x^* lies in the interior of $D^{\{2\}}$, we do not need to evaluate $g^{\emptyset}$, and consequently, $\beta = g^{\{2\}}(2) = 1/2$. The dual rates are given in Table III.

Table III. Original and dual parameters of the flow line of Example 3.

i	ν_i	λ_i	μ_i	λ_i^*	μ_i^*
1	3	3	1	2	3/2
2	2	1/3	4	1/3	4
3	1	0	∞	0	∞

This time, in contrast to the two previous examples, the first buffer in the dual system is not balanced.

5. An alternative method to find β

For an ordinary fluid queue driven by a finite state Markov process, the decay rate, α say, of the buffer can in general be found via two methods. The first method involves an optimization program similar in nature to the one described in the previous section. In the second method $-\alpha$ is identified as the largest strictly negative eigenvalue of the eigenvalue equation

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

where Q is the Q-matrix of the driving process and R denotes the diagonal matrix of net input rates. For more details, see for example Chapter 3 of (Mandjes, 1996). The eigenvalue equation is closely related to the Kolmogorov Forward Equations of the joint driving and the buffer content process.

Without going into details, we indicate an alternative way to derive the decay rates for some flow lines. A more detailed study of the idea, for a 3-stage flow line, is given in (Kroese and Nicola, 1998). The method goes as follows. For any binary vector $\boldsymbol{x} = (x_1, \ldots, x_n) \in S = \{0,1\}^n$, let $\boldsymbol{x}_i$ denote the vector $(x_1, \ldots, 1 - x_i, \ldots, x_n)$, and let $f_{\boldsymbol{x}}$ be the "density" defined by

$$f_{\boldsymbol{x}}(u_1, \ldots, u_{n-1}) := \frac{\partial^{n-1}\, \mathbb{P}(\boldsymbol{M} = \boldsymbol{x}, Z_1 \le u_1, \ldots, Z_{n-1} \le u_{n-1})}{\partial u_1 \cdots \partial u_{n-1}},$$

$u_1, \ldots, u_{n-1} > 0$. The 2^n densities $f_{\boldsymbol{x}}$ satisfy the following set of partial differential equations (these are in fact the time-stationary Kolmogorov Forward Equations of the process $(\boldsymbol{M}(t), \boldsymbol{Z}(t))$, see e.g. (Zimmern, 1956)):

$$\sum_{i=1}^{n-1}(x_i\nu_i - x_{i+1}\nu_{i+1})\frac{\partial f_{\boldsymbol{x}}}{\partial u_i} =$$
$$((1-x_i)\lambda_i + x_i\mu_i)\ f_{\boldsymbol{x}_i} - \sum_{i=1}^{n-1}(x_i\lambda_i + (1-x_i)\mu_i)\ f_{\boldsymbol{x}}.$$

Taking $((n-1)$-dimensional) Laplace transforms, we obtain[4] the algebraic equation

$$\sum_{i=1}^{n-1}(x_i\nu_i - x_{i+1}\nu_{i+1})s_i\,\tilde{f}_{\boldsymbol{x}} - c_{\boldsymbol{x}}) = \tag{13}$$
$$((1-x_i)\lambda_i + x_i\mu_i)\ \tilde{f}_{\boldsymbol{x}_i} - \sum_{i=1}^{n-1}(x_i\lambda_i + (1-x_i)\mu_i)\ \tilde{f}_{\boldsymbol{x}},$$

for some constants $c_{\boldsymbol{x}}, \boldsymbol{x} \in S$. Hence, if we gather the $\tilde{f}_{\boldsymbol{x}}$'s and $c_{\boldsymbol{x}}$'s into vectors $\tilde{\mathbf{f}}$ and $\mathbf{c}$ respectively, we obtain the matrix equation

$$A(s_1, \ldots, s_{n-1})\,\tilde{\mathbf{f}}(s_1, \ldots, s_{n-1}) = \mathbf{c}(s_1, \ldots, s_{n-1}),$$

where $A(s_1, \ldots, s_{n-1})$ follows from (13).

The polynomial $\det A(s_1, \ldots, s_{n-1})$ is of particular interest. In certain cases, in particular when $\beta = g^{*\emptyset}$, $-\beta$ is an extreme point of a closed subset of the set $\{(s_1, \ldots, s_{n-1}) : \det A(s_1, \ldots, s_{n-1}) = 0\}$. We will illustrate the idea with an example.

Example 4. Consider the flow line of Example 1. Arranging the $f_{\boldsymbol{x}}$ in lexicographical order (000,001,010, ... ,111) into a vector $\mathbf{f}$, we obtain the following matrix equation:

$$A\,\tilde{\mathbf{f}}(p, s) = \mathbf{c}, \tag{14}$$

[4] We denote the Laplace transform of a vector-valued function $\mathbf{h}$ by $\tilde{\mathbf{h}}$.

where

$$A = \begin{pmatrix} 8 & -1 & -1 & 0 & -4 & 0 & 0 & 0 \\ -4 & 5-s & 0 & -1 & 0 & -4 & 0 & 0 \\ -2 & 0 & 7-p+s & -1 & 0 & 0 & -4 & 0 \\ 0 & -2 & -4 & 4-p & 0 & 0 & 0 & -4 \\ -2 & 0 & 0 & 0 & 10+p & -1 & -1 & 0 \\ 0 & -2 & 0 & 0 & -4 & 7+p-s & 0 & -1 \\ 0 & 0 & -2 & 0 & -2 & 0 & 9+s & -1 \\ 0 & 0 & 0 & -2 & 0 & -2 & -4 & 6 \end{pmatrix}.$$

The determinant of $A =: A(p,s)$ is

$$\begin{aligned} \det A(p,s) &= -332640\,p - 108944\,p^2 + 10140\,p^3 + 1800\,p^4 - 133056\,s \\ &+ 171128\,p\,s - 12036\,p^2\,s - 5376\,p^3\,s - 184\,p^4\,s \\ &- 96848\,s^2 + 20160\,p\,s^2 + 8304\,p^2\,s^2 + 68\,p^3\,s^2 \\ &- 48\,p^4\,s^2 + 3744\,s^3 - 3720\,p\,s^3 + 420\,p^2\,s^3 \\ &+ 96\,p^3\,s^3 + 1152\,s^4 - 304\,p\,s^4 - 48\,p^2\,s^4. \end{aligned}$$

In Figure 2 a part of the set $\{(p,s) \in \mathbb{R}^2 : \det A(p,s) = 0\}$ is depicted. Let us call the left-most s-coordinate of the closed curve through (0,0) the *cut-off point*, s^* say. The reader may check that s^* is exactly $-\beta$.

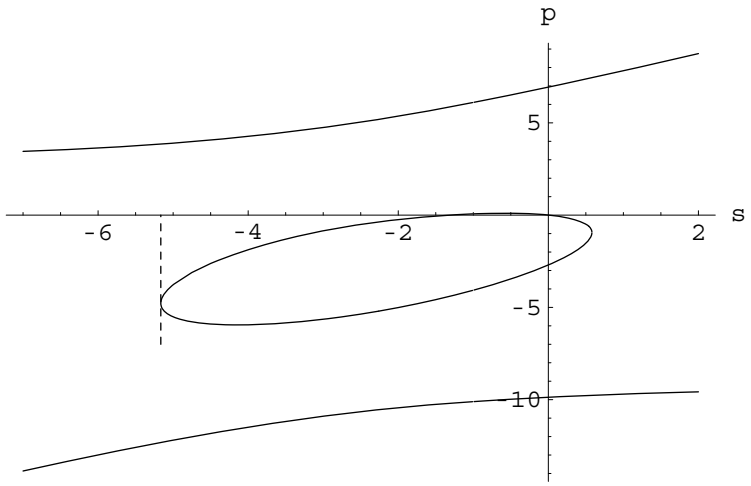

Figure 2. A subset of $\{(p,s) \in \mathbb{R}^2 : \det A(p,s) = 0\}$ forms a closed curve through (0,0). The left-most s-coordinate of this curve is $-\beta$.

For a *general* 3-stage flow line with *identical* machine speeds $\nu_i = 1, i = 1, 2, 3,$ we have similarly

$$A\,\tilde{\mathbf{f}}(p,s) := \begin{pmatrix} B & -\lambda_1 I_4 \\ -\mu_1 I_4 & B + (p + \lambda_1 - \mu_1) I_4 \end{pmatrix} \tilde{\mathbf{f}}(p,s) = \mathbf{c}, \qquad (15)$$

where

$$B := \begin{pmatrix} \mu_1+\mu_2+\mu_3 & -\lambda_3 & -\lambda_2 & 0 \\ -\mu_3 & -s+\mu_1+\mu_2+\lambda_3 & 0 & -\lambda_2 \\ -\mu_2 & 0 & -p+s+\mu_1+\lambda_2+\mu_3 & -\lambda_3 \\ 0 & -\mu_2 & -\mu_3 & -p+\mu_1+\lambda_2+\lambda_3 \end{pmatrix},$$

It is possible to identify the cut-off point s^* as a zero of a fifth-degree polynomial, in which the coefficients are rational functions of the parameters. The formula is rather lengthy (two pages of output) and we therefore omit it. However, for the non-trivial case where the availabilities of the first two machines are identical, $a_1 = a_2$, we have the simple expression

$$s^* = \frac{(\mu_1+\mu_2)(\lambda_3\,\mu_2 - \lambda_2\,\mu_3)(\lambda_1+\lambda_2+\lambda_3+\mu_1+\mu_2+\mu_3)}{\mu_2\,(\lambda_1+\lambda_2+\lambda_3)(\mu_1+\mu_2+\mu_3)}.$$

Finally, a note on the algebra involved in finding s^* for general 3-stage flow lines (not necessarily with identical speeds). All we have to do is to determine (p,s) satisfying the equations $\det A(p,s) = 0$ and $\frac{\partial}{\partial p} \det A(p,s) = 0$. We can eliminate the variable p from the equations by taking the *resultant* of the polynomials $\det A(p,s)$ and $\frac{\partial}{\partial p} \det A(p,s)$, with respect to p.

For example, in the model of Example 2 we have found numerically $\beta \approx 2.576668739$; we wish to express β as a zero of a polynomial with integer coefficients. We have

$$\begin{aligned} \det A(p,s) = &-27\,p - 86\,p^2 - 12\,p^3 - 54\,s + 85\,p\,s \\ &+ 43\,p^2\,s + 6\,p^3\,s - 53\,s^2 - 27\,p\,s^2 + p^2\,s^2 + 2\,s^3 - 4\,p\,s^3 + s^4 \end{aligned}$$

and

$$\begin{aligned} \frac{\partial \det A(p,s)}{\partial p} = &-27 - 172\,p \\ &-36\,p^2 + 85\,s + 86\,p\,s + 18\,p^2\,s - 27\,s^2 + 2\,p\,s^2 - 4\,s^3. \end{aligned}$$

The resultant of the two polynomials above with respect to p is the polynomial

$$\begin{aligned} &-18(-2+s)(-45+19\,s+s^2)^2 \\ &\quad (732 - 22468\,s - 12793\,s^2 + 2706\,s^3 + 3163\,s^4 + 712\,s^5 + 48\,s^6). \end{aligned}$$

This identifies $-\beta$ as a zero of the polynomial

$$732 - 22468\,x - 12793\,x^2 + 2706\,x^3 + 3163\,x^4 + 712\,x^5 + 48\,x^6.$$

6. Conclusions and directions for future research

We have demonstrated how the decay rates of the buffer overflow probabilities in a general flow line can be determined via a minimization program. These decay rates provide useful qualitative insight into the behaviour of the flow line. The analysis reveals, as a by-product, information on the manner in which an overflow occurs. Specifically, we obtain the "dual" failure and repair rates of the machines. These quantities can be interpreted as asymptotic failure and repair rates during an overflow period of a buffer.

A possible direction for research is the efficient simulation of loss probabilities through Importance Sampling (IS), as in (Chang et al., 1994). The corresponding optimal change of measure turns out to be more complicated than the one in (Chang et al., 1994), but can still be determined in some cases. In particular, for a 3-stage flow line this has been done in (Kroese and Nicola, 1998). Empirical results using IS show several orders of magnitudes of variance reduction compared to standard simulation when estimating small loss probabilities.

Other topics include generalizations of our methodology to general networks, e.g. intree or fork-join networks. Also the failure and repair mechanism might be generalized by considering Phase-Type or even general distributions. It should be noted that the main reason why the overflow probabilities have an exponential decay is that the life and repair times have "thin-tailed" distributions. In the case of heavy-tailed distributions, buffer overflow is more likely to result from an exceptionally long repair or lifetime of a machine rather than from a steady buildup in the buffer.

The relationship between the decay rate of the steady-state distribution of the content of a buffer and the decay rate of the corresponding overflow probability should also be investigated.

As a starting point for a rigorous treatment we could view the content process of a buffer in the flow line as a (reflected version of a) Markov Additive Process (MAP). The main difficulty is that the driving Markov process has a *non-denumerable* state space and a *continuous* time parameter. Although large deviations for MAP's in *discrete* time have been considered in (Ney and Nummelin, 1987), it is not clear at this stage whether the theory in the latter article provides a convenient basis for a rigorous study of backlogs in flow lines.

References

Anantharam, V.: 1988, How large delays build up in a GI/G/1 queue. *Queueing Systems* **5** 345-368.

Bucklew, J: 1990, *Large Deviation Techniques in Decision, Simulation and Estimation*, Wiley, New York.

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

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

Dallery, Y. and Gershwin, S.: 1992, Manufacturing flow line systems: a review of models and analytical results. *Queueing systems* **12** 3 – 94.

De Veciana, G, Olivier, C. and Walrand, J.: 1993, Large deviations of birth death Markov fluids. *Probability in the Engineering and Informational Sciences* **7** 237-255

Garvels, M.J.J. and Kroese, D.P. : 1999, On the entrance distribution in RESTART simulation. *Proceedings of the 1999 Rare Event Simulation Workshop*, Enschede, The Netherlands, March 9–12.

Kesidis, G. and Walrand, J.: 1993, Relative entropy between two Markov transition rate matrices. *IEEE Transactions on Information Theory* **39** 1056-1057.

De Koster, M.B.M. and Wijngaard, J.: 1986, A continuous flow model for three production units in series with buffers. *Operations Research proceedings DGOR* (Berlin: Springer-Verlag) 253 - 264.

Kroese, D.P. and Scheinhardt W.R.W: 2001, Joint distributions for interacting fluid queues, *Queueing Systems: special issue on Stochastic Models with Tractable Steady State Characteristics*, To Appear, 46 pages.

Kroese, D.P. and Nicola, V.F.:1998, Efficient Simulation of Backlogs in Fluid Flow Lines. *AEÜ Int. J. Electron. Commun.*, **52**, 165–172.

Mandjes, M.: *Rare Event Analysis of Communication Networks.* PhD Thesis, Free University, Amsterdam, The Netherlands, 1996.

Ney, P. and Nummelin, E.: 1987, Markov additive processes II. Large deviations. *Ann. Prob.* **15** 593-609.

Rogers, L.C.G.: 1994, Fluid models in queueing theory and Wiener-Hopf factorization of Markov Chains. *Ann. Appl. Probab.* **4** (2) 390-413.

Shwartz, A. and Weiss, A. *Large Deviations for Performance Analysis: queues, communication, and computing*, Chapman and Hall, New York.

Kella, O. and Whitt: 1999, Linear stochastic fluid networks *J. Appl. Probab.* **36** (1) 244-260.

Zimmern, B.: 1956, Etudes de la propagation des arrêts aléatoires dans les chaines de production. *Rev. Statist. Appl.* **4** 85-104.