Stochastic discretization for the long-run average reward in fluid models

Stochastic discretization for the long-run average reward in fluid models

I.J.B.F. Adan[*], V.G. Kulkarni[†] and J.A.C. Resing[*]

November 15, 2002

Abstract

Stochastic discretization is a technique of representing a continuous random variable as a random sum of iid exponential random variables. In this paper we apply this technique to study the limiting behavior of a stochastic fluid model. Specifically, we consider an infinite capacity fluid buffer, where the net input of fluid is regulated by a finite-state irreducible continuous time Markov chain. Most long-run performance characteristics for such a fluid system can be expressed as the long-run average reward for a suitably chosen reward structure. In this paper we use stochastic discretization of the fluid content process to efficiently determine the long-run average reward. This method transforms the continuous-state Markov process describing the fluid model into a discrete-state Quasi-Birth-Death process. Hence, standard tools, such as the matrix-geometric approach, become available for the analysis of the fluid buffer. To demonstrate this approach we analyze the output of a buffer processing fluid from K sources on a FCFS basis.

1 Introduction

The model considered here is a fluid buffer with unlimited capacity. The net input of fluid is regulated by an irreducible continuous time Markov chain (CTMC) with state space $S = \{0, 1, \dots, M\}$ and generator Q. In state $i \in S$ the amount of fluid in the buffer changes at rate c_i units per time unit. That is, if $c_i > 0$ then fluid is added at a constant rate c_i when in state i; if $c_i < 0$ then fluid is removed at a constant rate $-c_i$ until the buffer is empty, at which point no fluid is removed. We call c_i the net input rate in state i. For stability we have to require that the mean net input rate is negative, so

$$\pi \cdot c < 0,$$

[*]Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

[†]Department of Operations Research, University of North Carolina, CB 3180, Chapel Hill, N.C. 27599

where $\pi = (\pi_0, \ldots, \pi_M)$ is the equilibrium distribution of the regulating Markov chain, and $c' = (c_0, \ldots, c_M)$. Most long-run performance characteristics can be expressed as the long-run average reward for a suitably defined reward structure. The aim of this paper is to develop an efficient method to determine the long-run average reward.

The method is based on stochastic discretization of the content of the buffer, i.e., we represent the fluid content as a random sum of iid exponential random variables. Each of these exponential random variables will be called a *stochastic quantum*. Thus we keep track of the number of stochastic quanta in the buffer rather than the actual fluid content. To guarantee that the quantum sizes are identically distributed we need to uniformize the generation of fluid quanta in an appropriate way. The result is that the original continuous-state Markov process describing the fluid buffer is transformed into a discrete-state Quasi-Birth-Death (QBD) process. Hence, for the analysis of the fluid buffer, we may use the powerful tools available for QBDs; for example, see Latouche and Ramaswami [9]. Ramaswami [13] was the first to analyze fluid models through QBDs and using matrix analytic methods, but his approach as well as the resulting QBD are different from the ones developed in the present paper.

The discretization technique has also been used in Adan et al. [1, 2]. The system studied in [2] is a two-level traffic shaper. This is a mechanism for shaping a cell input stream entering an ATM network. It consists of a cell buffer and a token buffer. The latter regulates the output rate of the shaper. The performance of the shaper is analyzed by using an approximative model based on a stochastic discretization of the content of the token buffer.

The paper is organized as follows. In Section 2 we first introduce the continuous-state fluid model and briefly present its standard analysis based on differential equations. The next section describes the transformation to a discrete-state model using stochastic discretization. The resulting QBD is presented in Section 4. The main result, stating that the long-run average rewards in the original fluid model and the QBD model are the same, is presented and proved in Section 5. The special case of a multi-class queue is treated in Section 6 and finally, Section 7 presents some numerical results.

2 Continuous-state fluid model

Let $X(t)$ be the total amount of fluid in the buffer at time t, and let $Z(t)$ be the state of the regulating Markov chain at time t. The (X, Z) process describes the fluid model, and its steady-state distribution

$$F_i(x) = \lim_{t\to\infty} P(X(t) \le x, Z(t) = i), \qquad x \ge 0,\ i = 0, 1, \ldots, M,$$

satisfies the balance equations

$$\frac{d}{dx}F(x) \cdot C = F(x)Q,$$

with boundary conditions

$$F(\infty) = \pi, \qquad F_i(0) = 0, \qquad i \in S_+,$$

where $F(x) = (F_0(x), \dots, F_M(x))$, and $C = diag(c)$. The set S_+ denotes the subset of S with positive net input rates; S_- will denote the subset with non-positive net input rates. It is well known that $F(x)$ can be expressed in terms of the generalized eigenvalues λ and eigenvectors ϕ of the system (Q, C) (that solve $(\lambda C - Q)\phi = 0$). See, e.g., Anick et al. [3] and Kulkarni [6].

Let $r(x, i)$ be the rate at which reward is earned when the (X, Z) process is in state (x, i). Then the long-run average reward g is given by

$$g = \sum_{i=0}^{M} \int_0^\infty r(x, i) dF_i(x).$$

Most long-run performance characteristics can be expressed as the long-run average reward for some suitably defined reward rate $r(x, i)$. For example, using $r(x, i) = x$ implies that g is the mean buffer content, and using

$$r(x, i) = \begin{cases} 1 & \text{if } x \leq y, i = j; \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

implies that g is the probability $F_j(y)$.

Clearly, for the computation of g we need $F(x)$, and thus the generalized eigenvalues and eigenvectors of (Q, C). The computation of eigenvalues and eigenvectors, however, may become numerically unstable when the state space of the regulating Markov chain becomes large. But for special cases, efficient and numerically stable procedures have been developed; for example, see Lenin and Parthasarathy [10] who consider a fluid queue driven by a truncated Birth-Death process. In the next section we will develop an alternative method to compute g by transforming the (X, Z) process into a discrete-state QBD. This QBD can be efficiently analyzed by matrix-geometric techniques, thus avoiding the computation of eigenvalues and eigenvectors.

3 Stochastic discretization and uniformization

As mentioned before, the idea behind stochastic discretization of a continuous random variable is to represent it as a sum of a random number of iid exponential random variables, called the stochastic quanta. We apply this idea to the content of the fluid buffer. The sizes of the quanta are iid exponential, which implies that, once we know the number of quanta in the buffer, we also know the distribution of the actual buffer content.

Let us first describe how the number of stochastic quanta in the buffer changes with time. Quanta are only added during periods when the regulating Markov chain Z is in a state i with $c_i > 0$. At the beginning of such a period the buffer receives a quantum (i.e., the number of quanta jumps by 1), the size of which is initially zero, but increases with rate c_i until this period ends, i.e., Z makes a transition to another state. Clearly the size of the new quantum added is exponentially distributed with mean $-c_i/q_{ii}$. This immediately reveals a complication: the sizes of the quanta are not identically distributed, but depend on the state in which they were generated. To overcome this problem we uniformize the generation of quanta, i.e., we introduce artificial events that occur with rate δ_i while Z is in state i. Then, the size of the new quantum increases with rate c_i until either Z jumps to another state or an artificial event occurs, whichever happens first. In the latter case a second quantum is added. Its size grows with rate c_i until, again, either Z leaves state i, or an artificial events occurs, and so on; this repeats until eventually Z jumps to another state. Clearly, the total volume added to the buffer while Z is in state i remains the same; the number of quanta added during this period is geometrically distributed with mean $1 - \delta_i/q_{ii}$ and the size of each quantum is exponentially distributed with mean $c_i/(\delta_i - q_{ii})$. To achieve that the sizes of all quanta are exponentially distributed with the same parameter α, we set $\delta_i = \alpha c_i + q_{ii}$ in state i with $c_i > 0$, where

$$\alpha = \max_{i:c_i>0} \frac{-q_{ii}}{c_i}. \tag{2}$$

Then it is readily verified that the mean size of each quantum is $1/\alpha$.

In states i with $c_i \leq 0$ the quanta are drained at rate $-c_i$, one quantum at a time, and removed as soon as they are empty. Hence, while Z is in state i, quanta are removed with rate δ_i, which is now defined as $\delta_i = -\alpha c_i$.

Let $N(t)$ denote the number of quanta in the buffer at time t. Then the fluid model can also be described by the (N, Z) process. For this process we want to determine its steady-state distribution

$$p(n,i) = \lim_{t\to\infty} P(N(t) = n, Z(t) = i).$$

Once this distribution is known, we can compute $F(x)$ and thus the long-run average reward g, since, given the number of quanta at an arbitrary point in time, their sizes are independent and exponentially distributed with mean $1/\alpha$.

However, there is a further complication: the (N, Z) process is *not Markovian*. The reason is that the history at time t exactly determines the sizes of the quanta currently present, and hence, influences the future behavior of the (N, Z) process. Therefore we act as follows: each time we finish filling a quantum, we immediately replace it by another one, the size of which is exponentially distributed with mean $1/\alpha$ and independent of the history. Clearly, this version of the (N, Z) process is Markovian, but its stochastic behavior is different from the original one. However, as we will show later on, its *limiting* behavior and associated *long-run* average rewards are exactly the same as the original process. In the sequel, the (N, Z) process will refer to its Markovian version. Some limitations of this translation to a Markovian process are explained in the concluding section.

4 The discrete-state QBD model

The (N, Z) process constitutes a Markov process with state space $\{(n, i), n \geq 0, i \in S\}$. The set S may be partitioned into S_-, and S_+, denoting the subsets with non-positive and positive net input rates. The transition rates from state (n, i) are depicted in figure 1; it shows that quanta are generated (removed) at rate δ_i if $c_i > (\leq) 0$, where (see Section 3)

$$\delta_i = \begin{cases} \alpha c_i + q_{ii} & \text{if } c_i > 0; \\ -\alpha c_i & \text{if } c_i \leq 0. \end{cases}$$

Note that in states $(0, i)$ with $i \in S_-$ downward transitions are not possible (because there are no quanta), and states $(0, i)$ with $i \in S_+$ are transient.

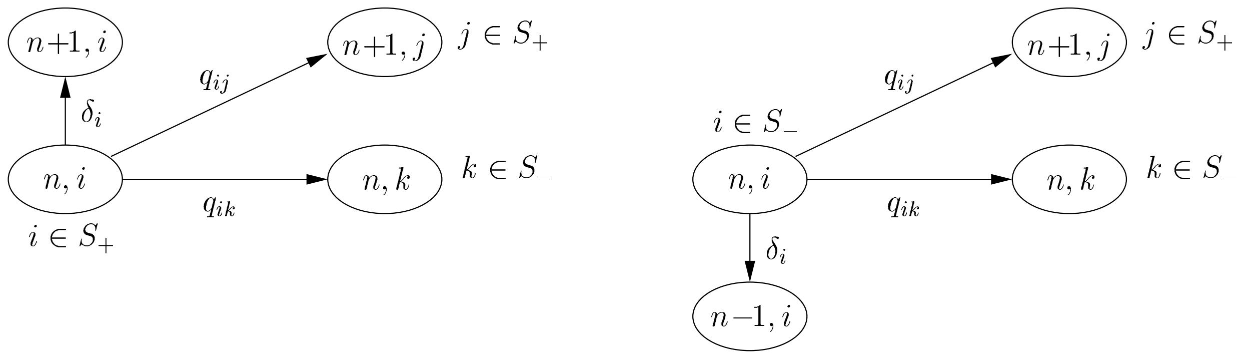

Figure 1: Transition rate diagram of the (N, Z) process.

By partitioning the state space into levels n, where level n is the set $\{(n, i), i \in S\}$, the generator of the (N, Z) process can be put in the form

$$\begin{pmatrix} B_1 & A_0 & 0 & 0 & 0 & \dots \\ A_2 & A_1 & A_0 & 0 & 0 & \dots \\ 0 & A_2 & A_1 & A_0 & 0 & \dots \\ 0 & 0 & A_2 & A_1 & A_0 & \dots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix}$$

where A_0, A_1, A_2 and B_1 are square matrices of order $M + 1$; the nonzero elements of these matrices are given by

$$A_0(i, j) = \begin{cases} q_{ij}, & i \neq j, i \in S, j \in S_+, \\ \delta_i, & i = j, i \in S_+; \end{cases} \qquad A_1(i, j) = \begin{cases} q_{ij}, & i \neq j, i \in S, j \in S_-, \\ q_{ii} - \delta_i, & i = j, i \in S; \end{cases}$$

$$A_2(i,j) = \delta_i, \; i = j, i \in S_-, \qquad\qquad B_1(i,j) = \begin{cases} q_{ij}, & i \neq j, i \in S, j \in S_-, \\ q_{ii} - \delta_i, & i = j, i \in S_+; \\ q_{ii}, & i = j, i \in S_-. \end{cases}$$

Hence, the (N, Z) process is a QBD. Also note that $A_0 + A_1 + A_2 = Q$. By partitioning the steady-state probability vector p into the sequence of vectors $p_0, p_1, \ldots$, where p_n is the probability vector of level n, we conclude from Neuts [12] that

$$p_n = p_0 R^n, \qquad n = 0, 1, 2, \ldots \tag{3}$$

where R is the minimal nonnegative solution of the matrix-quadratic equation

$$A_0 + RA_1 + R^2 A_2 = 0. \tag{4}$$

The vector p_0 follows from the boundary conditions and the normalization equation, i.e.,

$$p_0(B_1 + RA_2) = 0, \qquad p_0(I - R)^{-1} e = 1,$$

where I is the identity matrix and e the column vector with ones. Since states $(0, i)$ with $i \in S_+$ are transient, it follows that the corresponding $p(0, i)$'s are zero.

5 Long-run average rewards

In this section we define the reward rates for the QBD, and show that the long-run average reward of the QBD is exactly the same as the one of the continuous-state fluid model.

Let $E(T_{n,i})$ be the expected sojourn time in state (n, i) and $E(R_{n,i})$ the expected total reward earned during a visit to (n, i). The total reward is computed as the integral of rewards $r(X(t), i)dt$ over the whole duration of the visit, where the actual volume of the buffer at the beginning of this visit is distributed as the sum of n exponentials with mean $1/\alpha$ if $c_i \leq 0$, and as the sum of $n - 1$ exponentials if $c_i > 0$ (the new quantum is empty at the start). Now we define the reward rate $r(n, i)$ in state (n, i) as

$$r(n, i) = \frac{E(R_{n,i})}{E(T_{n,i})};$$

this may be interpreted as the average reward earned per time unit while the (N, Z) process is in state (n, i). By using the memoryless property of the exponential distribution we can derive a simpler expression for $r(n, i)$ as follows: Given that at an arbitrary point in time the (N, Z) process is in state (n, i), the sizes of all quanta in the buffer, including the one being emptied or filled, are exponentially distributed with mean $1/\alpha$. Hence,

$$r(n, i) = E(r(X, i)), \tag{5}$$

where the random variable X is the sum of n exponentials with mean $1/\alpha$.

That the long-run average rewards of the (N, Z) and (X, Z) processes are the same can be seen as follows: Let us suppose that at $t = 0$ the fluid buffer is empty and Z is in state i_0; denote the time of the k-th event (transition) in the (N, Z) process by T_k and define R_k as the total reward earned in the interval $[T_{k-1}, T_k]$ (where $T_0 = 0$). Then the long-run average reward of the (N, Z) process can be computed as

$$\lim_{k\to\infty} \frac{E(R_1) + \cdots + E(R_k)}{E(T_k)}.$$

The important observation is that, when we *embed* the (N, Z) process on the event times $T_0, T_1, T_2, \ldots$ (so we do not observe the lengths $T_k - T_{k-1}$ of the inter-event times), then it is stochastically identical to the original non-Markovian (N, Z) process; embedded at event times, the sizes of the quanta are unknown and iid exponentially distributed. So, for each k, the distribution of $(N(T_k), Z(T_k))$ is the same in both versions of the (N, Z) process. Thus $E(R_k)$ and $E(T_k - T_{k-1})$ are the same, and hence, their long-run average rewards. To complete the argument, we note that the long-run average reward of the non-Markovian (N, Z) process is, of course, identical to the one of the (X, Z) process. Our findings our summarized in the following theorem.

Theorem 5.1 *The fluid process (X, Z) and the QBD process (N, Z) have the same long-run average reward g, given by*

$$g = \sum_{i=0}^{M} \int_0^\infty r(x, i) dF_i(x) = \sum_{(n,i)} p(n, i) r(n, i),$$

where $r(n, i)$ is given by Equation (5).

We now apply the above result to determine the mean and distribution of the buffer content. If we take $r(x, i) = x$, then g yields the mean buffer content $E(X)$. From (5) it follows that the corresponding reward rate in the discrete model is equal to $r(n, i) = n/\alpha$. Using Theorem 5.1 and substituting the matrix-geometric form (3) of $p(n, i)$ into the expression for g gives

$$E(X) = \frac{1}{\alpha} \sum_{n=0}^{\infty} p_0 n R^n e = \frac{1}{\alpha} p_0 (I - R)^{-2} Re.$$

This expression can be further simplified by using the identity

$$p_0 (I - R)^{-1} = \pi. \tag{6}$$

Corollary 5.2 *The mean buffer content is given by*

$$E(X) = \frac{1}{\alpha} \pi (I - R)^{-1} Re.$$

Next we compute the limiting distribution of the buffer content by using the reward rate given in Equation (1). By (5) the reward rate $r(n,i)$ corresponding to this reward rate structure is given by

$$r(n,i) = \begin{cases} 1 - e^{-\alpha y}\sum_{k=0}^{n-1}(\alpha y)^k/k! & \text{if } n \geq 0, i = j; \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

This immediately leads to the following expression for the steady-state distribution of the buffer content.

Corollary 5.3 *For all $y \geq 0$ it holds that*

$$F(y) = \pi\left(I - Re^{-\alpha(I-R)y}\right).$$

Proof: Substituting the reward rate (7) and identity (6) we obtain

$$\begin{aligned} F_j(y) &= \sum_{(n,i)} p(n,i)r(n,i) = \sum_{n=0}^{\infty} p_0 R^n e_j \left(1 - e^{-\alpha y}\sum_{k=0}^{n-1}\frac{(\alpha y)^k}{k!}\right) \\ &= p_0(I-R)^{-1}e_j - p_0\sum_{k=0}^{\infty} e^{-\alpha y}\frac{(\alpha y)^k}{k!}\sum_{n=k+1}^{\infty} R^n e_j \\ &= p_0(I-R)^{-1}e_j - p_0(I-R)^{-1}R\sum_{k=0}^{\infty} e^{-\alpha y}\frac{(\alpha R y)^k}{k!}e_j \\ &= \pi e_j - \pi R e^{-\alpha(I-R)y}e_j. \end{aligned}$$

6 Multi-class queue

In this section we consider the situation that the incoming fluid is produced by K sources, numbered $1, \ldots, K$. The input rates of these sources are regulated by the Markov process Z; in state $i \in S$ the input rate of source k is v_{ik}. The fluid of the K sources accumulates in the buffer, where it is removed at constant rate v. Thus the net input of fluid in state i is equal to

$$c_i = \sum_{k=1}^{K} v_{ik} - v.$$

The fluid is processed on a First-Come First-Served basis: Fluid arriving at time t will be removed from the buffer only when all fluid that arrived before time t has been removed. Typically, the buffer will be simultaneously processing fluids of many sources, and its output process may be rather complicated. This is illustrated for the special case of two independent on-off sources in figure 2.

In the output of type k fluid (i.e., fluid originating from source k) we may distinguish on and off periods; the output is off when no type k fluid leaves the buffer, and on otherwise.

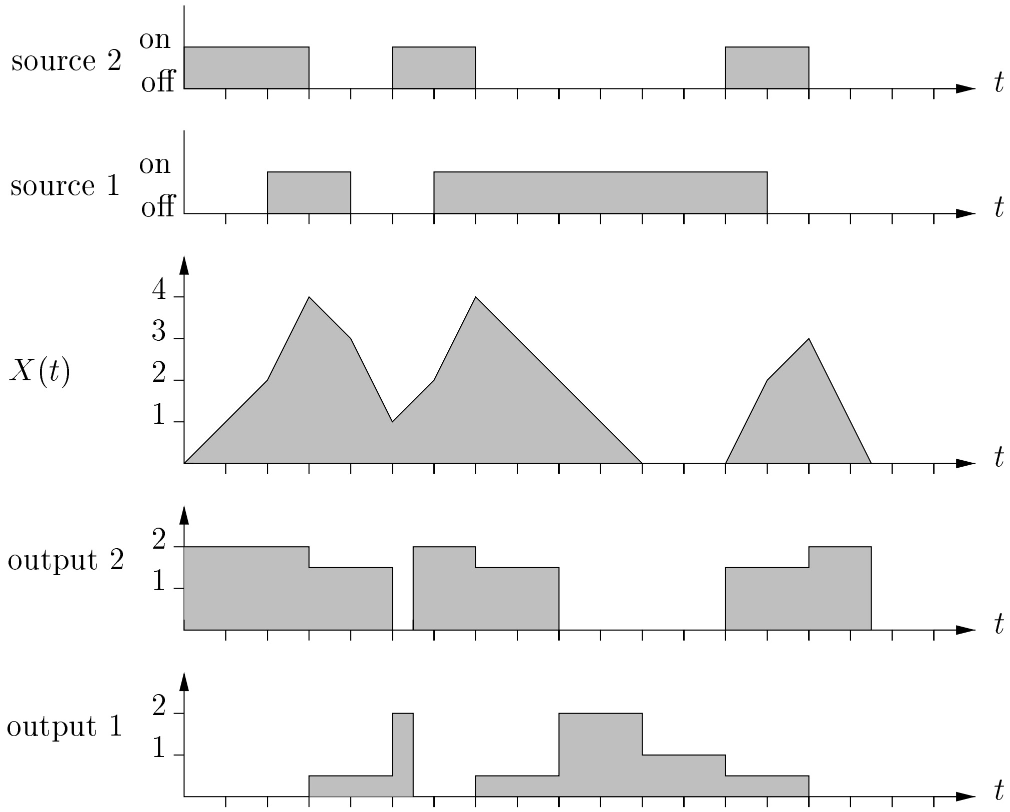

Figure 2: Sample paths of input, buffer content $X(t)$ and output for two independent on-off sources, where the input rate of source 1 (resp. source 2) during the on-time is 1 (resp. 3) and $v = 2$.

However, the rate at which type k fluid leaves the buffer during an on-period can change, as can be seen in figure 2.

Below we indicate how various performance characteristics of the buffer and its output process can be expressed as the long-run average reward for suitably defined reward structures. In data network applications fluid models are used to describe the flow of (small) data cells through the network; the input rate of a source can be interpreted as the number of cells generated per time unit. Then an important performance characteristic is the delay experienced by cells.

Mean cell delay of type k fluid: The mean time a type k cell arriving in state (i, n) spends in the buffer is $n/\alpha \cdot 1/v$; the input rate, or number of cells arriving per time unit, in state (i, n) is v_{ik}. Hence, if we treat delays as rewards, and define

$$r(n, i) = v_{ik} \cdot \frac{n}{\alpha v},$$

then g is equal to the mean total delay 'earned' per time unit; dividing g by the mean number of type k cells arriving per time unit yields the mean cell delay $E(D_k)$ of type k

fluid. The mean input rate of type k fluid is πv_k where v_k denotes the column vector of input rates v_{ik}. So

$$E(D_k) = \frac{1}{\pi v_k} \sum_{n=0}^{\infty} p_n v_k \frac{n}{\alpha v} = \frac{1}{\pi v_k} \cdot \frac{1}{\alpha v} \cdot \pi (I - R)^{-1} R v_k.$$

Distribution of the cell delay of type k fluid: The time a type k cell arriving in state (n, i) spends in the buffer is distributed as the sum of n exponentials with mean $1/(\alpha v)$. Hence, to find the probability $P(D_k \le y)$, we define

$$r(n,i) = v_{ik} \cdot \left(1 - e^{-\alpha v y} \sum_{l=0}^{n-1} \frac{(\alpha v y)^l}{l!}\right).$$

Then g yields the mean number of type k cells per time unit that experience a delay less or equal to y; to obtain the desired probability, we divide g by the mean input rate of type k fluid. Thus

$$P(D_k \le y) = \frac{1}{\pi v_k} \sum_{n=0}^{\infty} p_n v_k \left(1 - e^{-\alpha v y} \sum_{l=0}^{n-1} \frac{(\alpha v y)^l}{l!}\right) = \frac{1}{\pi v_k} \cdot \pi \left(I - R e^{-\alpha v (I - R) y}\right) v_k.$$

Mean amount of type k fluid in the buffer: The mean amount $E(X_k)$ of type k fluid cannot be obtained directly from the limiting distribution of the (N, Z) process, since we do not keep track of the type of fluid in the buffer. However, $E(X_k)$ can be obtained from the mean cell delay of type k fluid and application of Little's law. This yields

$$E(X_k) = \pi v_k E(D_k) = \frac{1}{\alpha v} \cdot \pi (I - R)^{-1} R v_k.$$

The following performance characteristics are related to the output process of the fluid buffer. The mean durations of on- and off-periods of type k fluid leaving the buffer are of special interest; they may be used to approximate the output process of type k fluid by an exponential on-off source, see Kulkarni [7].

Fraction of time type k fluid leaves the buffer: If at time t the (N, Z) process is in state $(0, i)$, the fluid that arrives in $(t, t + dt)$ will be immediately transmitted (i.e., at the same rate as it enters the buffer); in state (n, i) with $n > 0$ the incoming fluid will first stay in the buffer for a while before being removed (at rate v). The time needed to remove it is equal to $\sum_{l=1}^{K} v_{il} dt / v$. Thus, if we define the reward rates as

$$r(n,i) = \begin{cases} 1 & \text{if } n = 0, v_{ik} > 0; \\ \sum_{l=1}^{K} v_{il}/v & \text{if } n > 0, v_{ik} > 0; \\ 0 & \text{otherwise,} \end{cases}$$

then the long-run average reward yields the fraction of time type k fluid leaves the buffer.

Mean number of type k on-periods per time unit: For simplicity we now assume that the regulating Markov chain Z has exactly one state in which all sources are off; this is state 0. If in state (n, i) with $v_{ik} = 0$ and $\sum_{l=1}^{K} v_{il} > 0$ the Z process makes a transition from i to j with $v_{jk} > 0$, then this transition generates a new type k on-period. It will start as soon as the current buffer content has been removed. In case Z makes a transition to 0, then this transition only generates a new type k on-period if the next transition of Z is directed to j with $v_{jk} > 0$. Further, if in state (n, i) with $v_{ik} > 0$ the Z process makes a transition to 0, then this transition generates a new type k on-period, provided the next transition of Z is directed to state j with $v_{jk} > 0$ and this transition takes place after the buffer has been emptied. Summarizing, to obtain the mean number of type k on-periods per time unit, we should define the reward rates as follows.

$$r(n,i) = \begin{cases} \sum_{j:v_{jk}>0} q_{ij} - q_{i0}q_{0j}/q_{00} & \text{if } v_{ik} = 0, \sum_{l=1}^{K} v_{il} > 0; \\ \sum_{j:v_{jk}>0} -q_{i0}\beta^n q_{0j}/q_{00} & \text{if } v_{ik} > 0; \\ 0 & \text{otherwise,} \end{cases}$$

where β is the probability that a quanta is emptied before Z leaves state 0, so

$$\beta = \frac{\alpha v}{\alpha v - q_{00}}.$$

Note that the reward rates are of the form

$$r(n,i) = a_i + b_i \beta^n.$$

Hence, if n_k denotes the mean number of type k on-periods per time unit and a and b denote the column vectors of a_i and b_i, then we get

$$\begin{aligned} n_k &= \sum_{(n,i)} p(n,i) r(n,i) \;=\; = \sum_{n=0}^{\infty} p_0 R^n (a + b\beta^n) \\ &= p_0 (I-R)^{-1} a + p_0 (I - \beta R)^{-1} b \;=\; \pi a + \pi (I-R)(I-\beta R)^{-1} b. \end{aligned}$$

Mean duration of type k on- and off-periods: Let us denote the mean on- and off-period of type k output by $1/\mu_k^o$ and $1/\lambda_k^o$. Further, let v_k^o be the mean output rate during type k on-periods, ρ_k the fraction of time type k fluid leaves the buffer and n_k the mean number of type k on-periods per time unit. Then it readily follows that

$$\frac{1}{\mu_k^o} = \frac{\rho_k}{n_k}, \qquad \frac{1}{\lambda_k^o} = \frac{1-\rho_k}{n_k}, \qquad v_k^o = \frac{\pi v_k}{\rho_k}.$$

7 Numerical example

We consider a fluid queue where the input is generated by $K = 3$ independent on-off sources. The mean on- and off-times of source k, $k = 1, 2, 3$, are equal to $1/k$ and 1. The off-times are exponential and the on-times are Coxian-2 distributed. This means that the on-time goes through at most two exponential phases, with mean lengths $1/\mu_{1,k}$ and $1/\mu_{2,k}$. The on-time starts in phase 1. After this phase, it comes to an end with probability $1-p_{1,k}$, and otherwise it enters the second (and last) phase. We use the Coxian distribution to demonstrate the effect of variability in the input stream on the buffer content and output process. Denote the squared coefficient of variation of the on-time by cv^2 (which is the same for all sources). Provided $cv^2 \geq 0.5$, we can match the mean and coefficient of variation of the on-time of source k by choosing (cf. Marie [11])

$$\mu_{1,k} = 2k, \qquad p_{1,k} = 0.5/cv^2, \qquad \mu_{2,k} = \mu_{1,k} p_{1,k}.$$

The input rate of source k during the on-time is equal to $k + 1$, so the *mean* input rate is 1. Hence, the mean total input rate is 3 and the maximum input rate (when all sources are on) is 9.

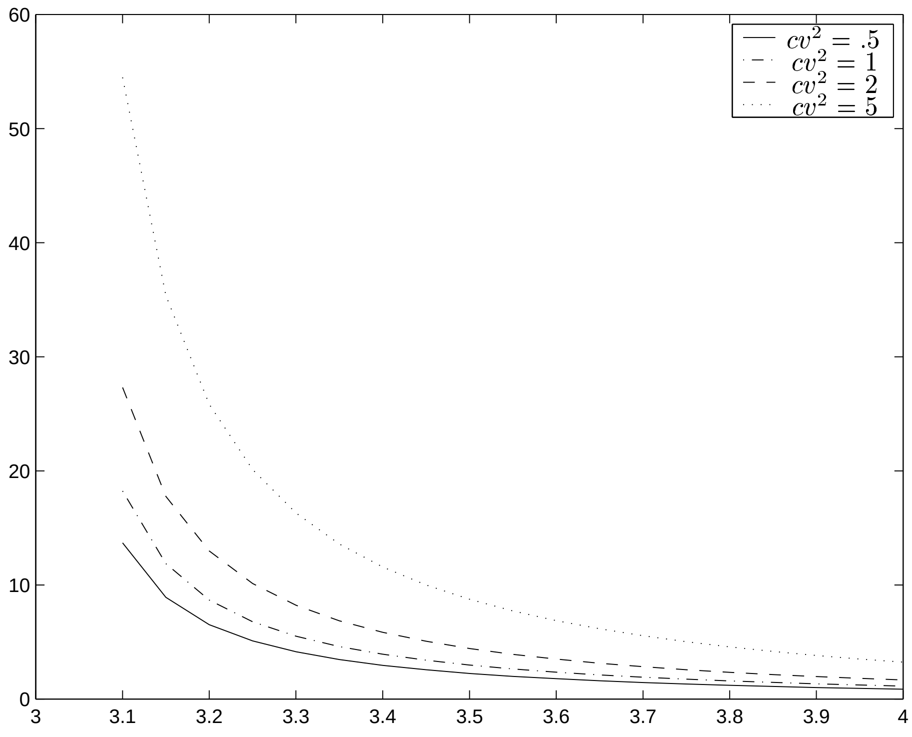

Figure 3: Mean buffer content as a function of the output rate v for $cv^2 = 0.5, 1, 2, 5$.

Note that the states of the regulating Markov chain Z can be described by the triples (m_1, m_2, m_3) where m_k indicates the state of source k: $m_k = 1$ means that source k is in state 1 of the off-time, $m_k = 2$ that it is in state 2 and $m_k = 3$ that source k in on. Hence, in this example the total number of states of Z is equal to $3^K = 27$.

In figure 3 we plot the mean buffer content as a function of the output rate v over $3.1 \leq v \leq 4$ for various values of cv^2. It shows that the mean buffer content decreases with v (from ∞ to 0) and increases with cv^2. Further, variability in the on-times seems to have a considerable effect on the buffer content process, especially when the output rate v is close to its minimum value 3.

We now consider the output process. In case of exactly one input source with exponential off-times, it is known that the mean duration of the on-period of the output process only depends on the mean on-time of the input source (not on higher moments); cf. Hirasawa [5] and Göbel [4]. In figure 4 we plot the mean duration of type 1 on-periods as a function of v for various values of cv^2. The results show that in the presence of multiple input sources the mean duration of on-periods is fairly insensitive to the variability in the on-times of the input sources.

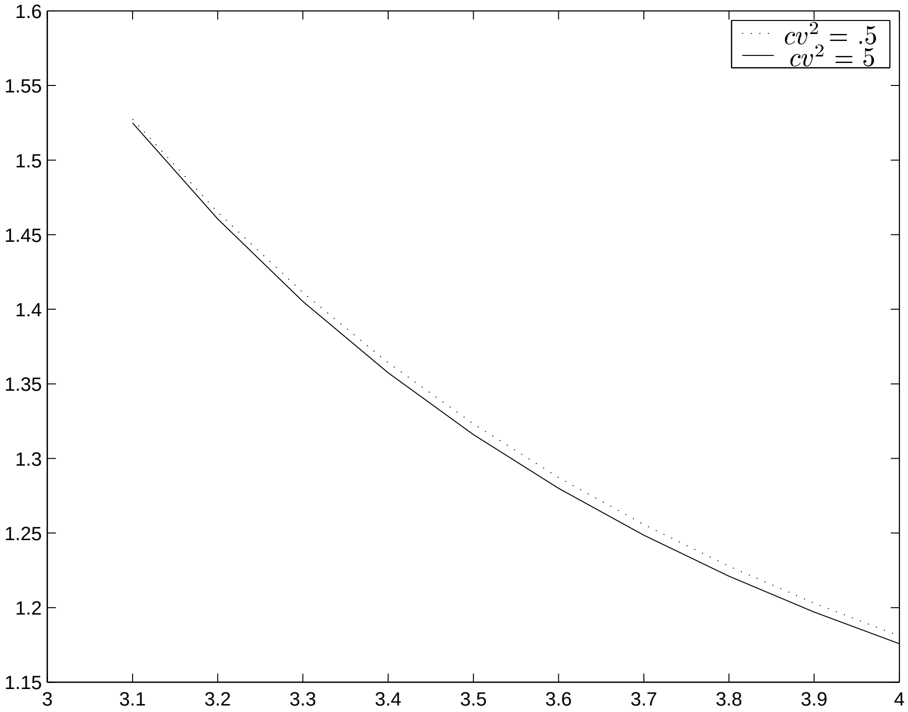

Figure 4: Mean duration of type 1 on-periods as a function of the output rate v for $cv^2 = 0.5, 5$.

Remark 7.1 *(Computation of R)* The rate matrix R can be determined from the quadratic equation (4) by an iterative algorithm (see, e.g., Latouche and Ramaswami [8]). One may expect that the number of iterations to accurately determine R depends on the mean size of the quanta; the number of iterations will be large if the quanta are very small, and this is indeed confirmed by numerical experiments. The quanta are small if one of the net input rates c_i is close to zero (cf. (2)).

8 Conclusion

In this paper we analyze a stochastic fluid model by using stochastic discretization. The determination of the long-run average reward in the (continuous-state) fluid model is reduced to the long-run average reward in a discrete-state QBD process. In doing so, powerful tools such as matrix-analytical methods become available for the steady-state analysis of the fluid model. The method of stochastic discretization, however, has also its limitations: some performance characteristics may not be expressed as long-run average rewards, and it does not provide transient results for the fluid model. For example, the QBD can be used to determine the *mean* duration of a period over which the fluid buffer is nonempty, but it fails to determine its distribution. In this case the dependency between the time needed to fill a quantum and the time needed to empty that quantum cannot be ignored.

References

[1] I.J.B.F. Adan, E.A. van Doorn, J.A.C. Resing and W.R.W. Scheinhardt (1998) Analysis of a single-server queue interacting with a fluid reservoir. *Queueing Systems*, **29**, 313-336.

[2] I.J.B.F. Adan and J.A.C. Resing (2000) A two-level traffic shaper for an on-off source. *Performance Evaluation*, **42**, 279-298.

[3] D. Anick, D. Mitra and M.M. Sondhi (1982) Stochastic theory of a data-handling system with multiple resources. *The Bell System Technical Journal*, **61**, 1871-1894.

[4] F. Göbel (1975) Queueing models involving buffers, Mathematical Centre Tract 60, Amsterdam.

[5] Y. Hirasawa (2000) Approximating traffic parameters in multiclass fluid networks. PhD thesis, Dept. of. Oper. Res., University of North Carolina, Chapel Hill, NC 27599-3180.

[6] V. G. Kulkarni (1997) Fluid models for single buffer systems, *Frontiers in Queueing: Models and Applications in Science and Engineering*, 321-338, Ed. J. H. Dshalalow, CRC Press.

[7] V.G. Kulkarni and K.D. Glazebrook (2002) Output analysis of a single-buffer multi-class queue: FCFS service. Technical Report No. UNC/OR/TR 01-5, under revision for J. Appl. Prob..

[8] G. Latouche and V. Ramaswami (1993) A logarithmic reduction algorithm for quasi-birth-death processes. *J. Appl. Prob.*, **30**, 650-674.

[9] G. Latouche and V. Ramaswami (1999) *Introduction to Matrix Analytic Methods in Stochastic Modeling.* SIAM.

[10] R.B. Lenin and P.R. Parthasarathy (2001) A computational approach for fluid queues driven by truncated birth-death processes. *Methodology and Computing in Applied Probability*, **2**, 373-392.

[11] R.A. Marie (1980) Calculating equilibrium probabilities for $\lambda(n)/C_k/1/N$ queue, in: *Proceedings Performance' 80*, Toronto, 117-125.

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

[13] V. Ramaswami (1999) Matrix analytic methods for stochastic fluid flows, in: *Teletraffic Engineering in a Competitive World, Proceedings of the International Teletraffic Congress, ITC-16*, 1019-1030, Eds. P. Key and D. Smith, Edinburgh, Elsevier.