A Large Deviations Analysis of the Transient of a Queue with Many Markov Fluid Inputs: Approximations and Fast Simulation

MICHEL MANDJES
Bell Laboratories/Lucent Technologies
and
AD RIDDER
Vrije Universiteit Amsterdam

This article analyzes the transient buffer content distribution of a queue fed by a large number of Markov fluid sources. We characterize the probability of overflow at time t, given the current buffer level and the number of sources in the on-state. After scaling buffer and bandwidth resources by the number of sources n, we can apply large deviations techniques. The transient overflow probability decays exponentially in n. In case of exponential on/off sources, we derive an expression for the decay rate of the rare event probability under consideration. For general, Markov fluid sources, we present a plausible conjecture. We also provide the "most likely path" from the initial state to overflow (at time t). Knowledge of the decay rate and the most likely path to overflow leads to (i) approximations of the transient overflow probability, and (ii) efficient simulation methods of the rare event of buffer overflow. The simulation methods, based on importance sampling, give a huge speed-up compared to straightforward simulations. The approximations are of low computational complexity, and accurate, as verified by means of simulation experiments.

Categories and Subject Descriptors: G.3 [**Probability and Statistics**]: *queuing theory*; I.6.1 [**Simulation and Modeling**]: Simulation Theory

General Terms: Algorithms, Performance

Additional Key Words and Phrases: ATM multiplexers, buffer overflow, calculus of variations, importance sampling simulations, IP routers, large deviations asymptotics, queuing theory, transient probabilities

1. INTRODUCTION

A characteristic feature of modern switches and routers is that typically a large number of flows are multiplexed. In an ATM (Asynchronous Transfer Mode)

Authors' present addresses: M. Mandjes, CWI, Amsterdam, The Netherlands, and the Faculty of Mathematical Sciences, University of Twente, The Netherlands; email: michel@cwi.nl; A. Ridder, Vrije Universiteit, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands; email: aridder@econ.vu.nl.

switch or an IP (Internet Protocol) router, the number of flows that share buffer and bandwidth resources can be in the order of many thousands. The key model for representing the stochastic behavior in the buffer of the switch or router is that of a large number of sources, alternating between bursts (the on state) and silences (the off state), feeding into a FIFO queue that is emptied at a constant rate. An accurate description of the stochastic properties of the buffer content of this queue are of utmost interest, as the loss due to overflow should be very rare. Particularly for highly loss-sensitive traffic allowed loss fractions in the order of $10^{-6} - 10^{-9}$ are typical.

Until now, the literature has been very much focused on *stationary* overflow probabilities, but one can argue that for specific applications *transient* overflow probabilities will be of great relevance. Such a transient overflow probability is defined by the probability of overflow at some time t, given the states of the modulating Markov chains and the queue length at time 0. For instance, for admission control purposes, it is essential to know the probability of overflow during the next time interval, given current system state.

The history of the above model goes back to the mid-Seventies—see Kosten [1974] and Cohen [1974]—and has inspired research in the teletraffic area significantly. Modelling the traffic as fluid has meant a great simplification, and is justified by the fact that information packets are typically small compared to the burst size. The fluid model analyzed by Anick et al. [1982] has become the fundamental reference for the analysis of packet-based switching or routing elements. It consists of a number of sources with exponentially distributed on and off times, where during the on-time traffic is sent at a constant rate. Anick et al. [1982] succeed in giving an explicit derivation of the steady state distribution of the buffer content.

Many generalizations followed. Kosten [1984] considered the case of sources that are driven by a Markov chain with more than two states, with which for instance Erlang or hypergeometric on and off times can be modelled; these more general sources are called *Markov fluid* in the sequel. He also solved the model with heterogeneous input: the sources do not share the same statistical properties, as is typically the case in a multiservice network. The solution of the buffer content distribution—and more specifically the buffer overflow probability—required the solution of a large eigensystem; its dimension is particularly high if the input is heterogeneous. To cope with this numerical problem, several types of asymptotic approximation techniques were proposed. One of them is the diffusion approach, valid in the regime of heavy traffic, as was proposed by Knessl and Morrison [1991]. However, in the present article, we focus on large-deviations-based asymptotics. Within this class, the two important regimes are the *large buffer asymptotics* and the *large system asymptotics*.

The large buffer asymptotics rely on the the exponential decay of the overflow probability as a function of the buffer size; it is asymptotically of the form $\alpha \exp[-\theta B]$, where amplitude α and decay rate θ are positive constants and B is the buffer size. It appears that θ is relatively easy to compute: it can be done without solving the above-mentioned eigensystem. The calculation of α, however, does require the entire solution of the eigensystem. Simulation

techniques have been proposed to (quickly) capture this amplitude, see Kesidis and Walrand [1993], Mandjes and Ridder[1995], and Ridder [1996]. Another approach is to replace α just by 1, but unfortunately this tends to be very inaccurate; in case many sources are multiplexed, the amplitude is orders of magnitude smaller than 1.

We believe that large system asymptotics have more practical interest than large buffer asymptotics. In particular, in switches or routers to be used by delay-sensitive traffic, buffer sizes tend to be small, whereas the number of inputs usually does grow large. As we mentioned, the amplitude of the large buffer asymptotic tends to be small for a large number of sources; this effect is taken care of explicitly by the large system asymptotics. Crucial is the scaling due to Weiss [1986]: in a system in which n homogeneous sources are present, buffer space B and link rate C are scaled with n, that is, $B = nb$ and $C = nc$. Notice that, as long as the mean input rate of a source is below c, then even the probability of a nonempty buffer is rare as n grows. Weiss succeeds in finding the asymptotics of the loss probability by using a *pathwise* large deviations approach: the decay rate of the overflow probability is the minimum of an action functional, were that minimum is taken over all paths that start off in the queue's equilibrium behavior, and that eventually arrive at buffer overflow. Interestingly, the optimizing path has the interpretation of most likely path: as n grows overflow becomes increasingly rare, but *if* it occurs, it does so according to this trajectory. Botvich and Duffield [1995] find—with different techniques—the decay rate for a much broader class of sources; related results can be found in Courcoubetis and Weber [1996] and Simonian and Guibert [1995]. Mandjes and Ridder [1999] and Wischik [2001] succeed in unifying both approaches, in that they explicitly find the most likely trajectories that give the decay rate of Botvich and Duffield [1995].

There are only few papers dedicated to the calculation of the transient behavior of a queue with Markov fluid input. The most notable contributions are by Kobayashi and Ren [1992] and Tanaka et al. [1995]. They succeed in finding the Laplace transform of the distribution of the queue length. Their approach had two obvious drawbacks: in the first place the methods are numerically demanding, as it requires both the solution of a (typically high-dimensional) eigensystem, and a numerical inversion of the Laplace transform. Both papers do not give explicit numerical results, nor evaluate numerical issues. In the second place, these methods do not give any insight into the system's behavior. For instance, we would like to know what the influence is of the initial queue length, or whether the system essentially returns to equilibrium before attaining the extreme value of buffer overflow at time t. A novel study on transient behavior is by Duffield [1998]. He only conditions on the states of the modulating chains and does not take into account the amount of traffic in the buffer.

This article aims at finding manageable and accurate asymptotics of the transient probability. The contribution of our study is twofold. In the first place, by using the "large system scaling" of resources proposed by Weiss [1986], we derive large deviations asymptotics of the transient buffer overflow probability. The calculations involved are relatively easy, as they only require the solution

of a low-dimensional optimization problem. As a by-product, we obtain the most likely trajectory to overflow, which sheds light on the question *how* overflow is reached. Typically, we will see that for small values of t, overflow is reached without the queue getting idle between 0 and t, whereas for larger t, the process first moves in the direction of its equilibrium behavior, and builds up the buffer during the last part of time interval $[0, t]$. Interestingly, a "bifurcation time" can be numerically evaluated. Our proofs strongly rely on the fundamental theorems provided by Shwartz and Weiss [1995, Section 13.6].

In the second place, we validate a number of approximations. This is done by a quick simulation method, based on importance sampling. Knowledge of the optimum path is used to change the underlying probability model such that the optimum path towards the rare event of buffer overflow becomes a frequently occurring event. The data is weighed by likelihood ratios, thus recovering an unbiased estimate. The novelty of the present article is that the underlying probability model has to be adapted continuously during the simulations, as the large system regime is considered. This is essentially different from earlier proposed importance sampling methods for the large buffer regime [Kesidis and Walrand 1993; Mandjes and Ridder 1995; Ridder 1996]. Earlier work on importance sampling with a changing alternative distribution can be found in Cottrell et al. [1983] and Kroese and Nicola [1998, 1999]. The simulation technique turns out to provide significant efficiency gains. Empirically, we show that the proposed approximations are accurate and in general conservative.

The structure of this article is as follows: Section 2 describes the model and presents a number of preliminaries. In Section 3, we formally find the decay rate of the transient buffer overflow probability for the important case of exponential on/off sources, and present a plausible conjecture for general Markov fluid sources. Section 4 is concerned with the performance of the corresponding large deviations approximation and quick simulation. In Section 5, conclusions are drawn.

2. MODEL AND PRELIMINARIES

This section first describes the model. Then, we define the transient probability that we attempt to approximate. We end up with stating a number of known results from large deviations theory that are needed in the analysis of Section 3.

2.1 Model and Notation

The model of this article can be described as a queue fed by a superposition of Markov fluid sources, an infinite buffer and a constant output rate.

A *Markov fluid* source is characterized by a generator and a traffic rate vector. The generator, say $\Lambda = (\lambda_{ij})_{i,j=1}^{d}$ governs a finite-state (dimension d) continuous-time Markov chain. Its state at time s is $X(s)$. Entry λ_{ij} (for $i \neq j$) denotes the transition rate from state i to state j. We follow the convention that $\lambda_{ii} := -\sum_{j \neq i} \lambda_{ij}$. The transient transition probabilities are denoted by $p_{ij}(s) := \mathrm{P}(X(s) = j \,|\, X(0) = i)$.

If the Markov chain is in state i, traffic is generated at a constant rate $r_i \geq 0$. One important type of Markov fluid source is the exponential on/off source, in which $d = 2$ and one of the traffic rates equals 0. In the analysis, we assume that the Markov chain is irreducible; consequently, there is a unique invariant vector $\pi = (\pi_1, \ldots, \pi_d)$, determined by the equation $\pi\Lambda = 0$.

The buffer is fed by n of these sources, and is emptied at a constant rate nc. We assume that the queuing system is stable: $\sum_i \pi_i r_i < c$.

Let us introduce a number of functions that will be used in the remainder of this article. First, define $A(t) := \int_0^t r_{X(s)}\,\mathrm{d}s$ as the total amount of fluid offered to the buffer by an arbitrary source during time interval $[0, t]$. $M(\theta; t) := \mathrm{E}[\exp(\theta A(t))]$, $\theta \in \mathbb{R}$, $t \geq 0$ is the moment generating function of $A(t)$. We also define the "conditional mgfs":

$$M_i(\theta; t) := \mathrm{E}[\exp(\theta A(t)) \mid X(0) = i].$$

Furthermore, define the matrix $(B(\theta; t))_{i,j=1}^d$ by

$$B_{ij}(\theta; t) := \mathrm{E}[\exp(\theta A(t))\mathbf{1}\{X(t) = j\} \mid X(0) = i], \quad \theta \in \mathbb{R}, t \geq 0.$$

The above-defined moment generating functions allow for explicit calculation, as follows. Given that the modulating chain X is in equilibrium at time 0, we may write

$$M(\theta; t) = \sum_{i=1}^{d} \pi_i M_i(\theta; t).$$

In Brandt and Brandt [1994] and Kesidis et al. [1993], it is proven that $B_{ij}(\theta; t) = (\exp((\Lambda + \theta R)t))_{ij}$, with $R := \mathrm{diag}\{r\}$. Consequently

$$M_i(\theta; t) = \sum_{j=1}^{d} B_{ij}(\theta; t).$$

2.2 Problem Description

While several previous papers dealt with the stationary buffer content distribution of this model, we focus on its transient. First, denote by $Q_n(t)$ the buffer content at time $t \geq 0$. The d-dimensional vector $F_n(t)$ denotes the distribution of the states of the n Markov chains at time t. Formally,

$$(F_n(t))_i := \frac{1}{n} \sum_{\ell=1}^{n} \mathbf{1}\{X_\ell(t) = i\}, \quad i = 1, 2, \ldots, d.$$

Obviously, for all t,

$$F_n(t) \in \left\{ x \mid x_i \geq 0, \sum_{i=1}^{d} x_i = 1 \right\}.$$

We are interested in the transient overflow probability, that is, the probability of reaching a particular buffer level at time t, given observations of queue length

$Q_n(0)$ and sources $F_n(0)$ at time 0:

$$\mathrm{P}(Q_n(t) \geq nb \mid Q_n(0) = nb_0 \text{ and } F_n(0) = f_0). \tag{1}$$

Notice that we use the same scaling as was introduced by Weiss [1986]: in the model with n sources, we scale buffer space and link rate by the same number.

As opposed to Kobayashi and Ren [1992] and Tanaka et al. [1995], we do not pursue an exact evaluation of (1). Instead, we focus on the derivation of its decay rate

$$I(b, t \mid b_0, f_0) := -\lim_{n\to\infty} \frac{1}{n} \log \mathrm{P}(Q_n(t) \geq nb \mid Q_n(0) = nb_0 \text{ and } F_n(0) = f_0). \tag{2}$$

ASSUMPTION 2.1. *Throughout this article, we assume the event $Q_n(t) \geq nb$, under conditions $Q_n(0) = nb_0$ and $F_n(0) = f_0$, to be rare, meaning that $I(b, t \mid b_0, f_0)$ is strictly positive and increases in b.*

Remark 2.2. In practice, one could be more interested in

$$\mathrm{P}\left(\sup_{s\in[0,t]} Q_n(s) \geq nb \mid Q_n(0) = nb_0 \text{ and } F_n(0) = f_0 \right)$$

rather than (1). Analogously to the reasoning in the proof of Theorem 1 of Botvich and Duffield [1995], one can show that the corresponding decay rate equals $\inf_{s\in[0,t]} I(b, s \mid b_0, f_0)$.

2.3 Preliminaries

In this section, we present the pathwise LDP for Markov processes. First, define the *local rate function* of Markov chains with generator $\Lambda = (\lambda_{ij})_{i,j}^d$:

$$\begin{aligned} I_x(y) &:= \sup_{\theta\in R^d} \left(\sum_{i=1}^d \theta_i y_i - \sum_{i=1}^d \sum_{\substack{j=1 \\ j\neq i}}^d x_i \lambda_{ij} (\exp(\theta_j - \theta_i) - 1) \right) \\ &= \sup_{\theta\in R^d} \left(\sum_{i=1}^d \theta_i y_i - \sum_{i,j=1}^d x_i \lambda_{ij} \frac{\exp(\theta_j)}{\exp(\theta_i)} \right) \\ &\qquad x \in \mathbb{R}_+^d, \quad \sum_{i=1}^d x_i = 1, \quad y \in \mathbb{R}^d, \quad \sum_{i=1}^d y_i = 0. \end{aligned} \tag{3}$$

The interpretation of this local rate function is the following: Consider a large number of Markov chains with generator Λ, and let the vector $x = (x_i)_{i=1}^d$ denote the empirical distribution of the Markov chains: a fraction x_i is in state i. Then, $I_x(y)$ is in fact the cost of the empirical distribution moving into direction y. This heuristically justifies the following theorem, rigorously proven by Shwartz and Weiss [1995, Theorem 13.37].

THEOREM 2.3 [SHWARTZ AND WEISS: LDP FOR MARKOV PROCESSES]. *For $d = 2$, and a set S of absolutely continuous functions on the interval $[0, t]$,*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathrm{P}(F_n(s) \in S, s \in [0,t] \mid F_n(0) = f_0) = - \inf_{f \in S: f(0)=f_0} J_t(f). \qquad (4)$$

Here $J_t(f) := \int_0^t I_{f(s)}(f'(s))\,ds$ is called the action functional.

Notice that, in Theorem 5.1 of Shwartz and Weiss [1995], this theorem is proved for general dimension d, but under the assumption that the logarithm of the transition rates is bounded. In our model, the rate of jumping from i to j is $m\lambda_{ij}$ if m sources are in state i. As m can attain value zero, the logarithm of this rate is not bounded, and Theorem 5.1 does not hold. However, for the special case $d = 2$, Shwartz and Weiss [1995] show in Theorem 13.37 that the process $F_n(\cdot)$ satisfies a large deviations principle. For higher dimension d, the LDP is not formally proven yet.

The optimizing f, say f^*, of variational problem (4) has an interesting interpretation. Given that the rare event under consideration occurs, with overwhelming probability it does so with $F_n(\cdot)$ following a path that lies close to f^* (where $n \to \infty$). A formal treatment of this concept is found in Chapter 6 of Shwartz and Weiss [1995], in particular Theorem 6.15. Clearly, this optimal path gives much insight into the system conditional on overflow, and appears to be useful in developing efficient simulation methods, see Section 4.

3. ANALYSIS

This section finds the decay rate (2) for the case of exponential on/off sources, and gives support to a conjecture for the decay rate for sources of dimension d larger than 2.

In Section 3.1, we find a variational problem corresponding to the decay rate for $d = 2$ (Proposition 3.1), which we can simplify (Lemma 3.2). Then we show in Proposition 3.3 that the variational problem is solved by an optimal path, which is the unique solution of a specific equation (Euler equation). Section 3.2 focuses on the intuition behind the variational problem. We then conjecture the optimal path (on the basis of probabilistic arguments) in Heuristic 3.4, and finally we show that the conjectured path solves the Euler equation (Proposition 3.5); both Heuristic 3.4 and Proposition 3.5 hold for general dimension d. Combining all results, in Section 3.3, we find decay rate and optimal path for $d = 2$ in Theorem 3.6 and Corollary 3.7.

If the dimension d equals 2, the paths are essentially one dimensional, as evidently the fraction of sources in the on-state and the fraction of sources in the off-state sum to 1; we therefore use one dimensional paths "$f = f_{\text{on}}$". For $d = 2$, for reasons of simplicity, we assume (without loss of generality) that the traffic rate in the on-state equals 1: "$r_{\text{on}} = 1$". This notation holds for Sections 3.1 and 3.3. In Section 3.2, we assume general dimension d, and therefore we use multidimensional paths and we have no tacit assumption regarding the traffic rates.

3.1 Derivation of the Variational Problem

In this section, we focus on exponential on–off sources. In Proposition 3.1, the decay rate under consideration is written as a variational problem, by invoking Theorem 2.3. Then, in Lemma 3.2, we prove that the set over which the action functional is minimized can be reduced considerably. Finally, we show in Proposition 3.3 that there is a unique minimizer.

Let $\{q_f(s), s \in [0,t]\}$ denote the scaled queue length, when given $q_f(0) = b_0$ and the path of the fraction of sources in the on state is $\{f(s), s \in [0,t]\}$.

PROPOSITION 3.1

$$I(b,t \mid b_0, f_0) = \inf_{f \in S_0' \cup S_1'} \int_0^t I_{f(s)}(f'(s))\, ds,$$

with

$$S_0' := \left\{ f \mid f(0) = f_0, q_f(u) > 0 \text{ for all } u \text{ in } [0,t], \text{ and } \int_0^t f(s)\, ds = b - b_0 + ct \right\}$$

and

$$S_1' := \left\{ f \mid f(0) = f_0, \exists\, u \in [0,t] \text{ with } q_f(u) = 0, q_f(s) > 0 \text{ for all } s \in (u,t], \right.$$
$$\left. \text{and } \int_u^t f(s)\, ds = b + c(t-u) \right\}.$$

PROOF. We invoke the LDP of Theorem 2.3. The set over which the actional functional has to be minimized is

$$S := \{f \mid f(0) = f_0 \text{ and } q_f(t) \geq b\}.$$

Because the decay rate increases in b—see Assumption 2.1—S can be replaced by S':

$$S' := \{f \mid f(0) = f_0 \text{ and } q_f(t) = b\}.$$

The theorem is proven by showing that $S' = S_0' \cup S_1'$. This is done as follows:

Obviously, every trajectory of the buffer content has either zero or a positive number of idle periods. Let S_0' consist of all paths for which the buffer is always nonempty in $[0,t]$; then the amount of fluid to be built up in $[0,t]$ is at least $b - b_0 + ct$. Furthermore, S_1' is the set of paths that do yield an idle period; if u is the last epoch of a zero buffer content, then the amount of traffic to be built up after u equals at least $b + c(t-u)$. Therefore, the decay rate of the transient probability equals the minimum of the action functional on the union of the sets S_0' and S_1'. □

We then define two variational problems. The first variational problem corresponds to the decay rate of the probability of generating $b - b_0 + ct$ fluid in

the interval $[0, t]$:

$$I_0 := \inf_{f \in S_0} \int_0^t I_{f(s)}(f'(s))\, \mathrm{d}s,$$
$$S_0 := \left\{ f \mid f(0) = f_0 \text{ and } \int_0^t f(s)\, \mathrm{d}s = b - b_0 + ct \right\}. \tag{5}$$

The second variational problem reflects the decay rate of the probability that somewhere in $[0, t]$ the input rate equals output rate c, and from then $b + c(t - u)$ traffic is fed into the system:

$$I_1 := \inf_{f \in S_1} \int_0^t I_{f(s)}(f'(s))\, \mathrm{d}s,$$

where

$$S_1 := \left\{ f \mid f(0) = f_0, \exists\, u \in [0, t] \text{ with } f(u) = c, \text{ and } \int_u^t f(s)\, \mathrm{d}s = b + c(t - u) \right\}.$$

LEMMA 3.2. *Decay rate* (2) *equals the minimum of* I_0 *and* I_1.

PROOF. Trivially, $S_0' \subset S_0$. Noticing that $q_f(u) = 0$ and $q_f(s) > 0$ for all $s \in (u, t]$ implies that $f(u) = c$, we get that $S_1' \subset S_1$.

Consider the minimum of I_0 and I_1, that is, the minimum of the action functional over $S_0 \cup S_1$. Say that the minimum is reached for a path $f^*(\cdot)$. This optimizing path $f^*(\cdot)$ lies in $S_0' \cup S_1'$. That can be proven as follows:

—Suppose $f^*(\cdot)$ lies in $S_0 \setminus S_0'$. So consequently $I_0 \leq I_1$. Then there is a u such that $\int_0^u f^*(s)\, \mathrm{d}s < -b_0 + cu$ and $f^*(u) = c$. Consequently, $\int_u^t f^*(s)\, \mathrm{d}s = b' + c(t - u) > b + c(t - u)$. In accordance with Proposition 3.2 of Simonian and Guibert [1995], the decay rate of the probability exceeding b after $t - u$ time increases as a function of b. In other words, f^* is *more expensive* than a path $g(\cdot)$ in S_1 with $g(u) = c$ and $\int_u^t g(s)\, \mathrm{d}s = b + c(t - u)$. But then would hold that $I_1 < I_0$. Contradiction.

—Suppose $f^*(\cdot)$ lies in $S_1 \setminus S_1'$, and therefore $I_1 \leq I_0$. Consequently, $\{q_{f^*}(s), s \in [0, t]\}$ has no idle periods. So the buffer contents at time u is positive: $q_{f^*}(u) > 0$. But then the buffer contents at time t equals $q_{f^*}(t) = q_{f^*}(u) + b > b$. This is *more expensive* than a path $g(\cdot)$ in S_0 that exactly reaches b: $q_g(t) = b$ (again essentially equivalent to Proposition 3.2 of Simonian and Guibert [1995] and using Assumption 2.1). But then $I_0 < I_1$. Contradiction.

Trivially, $\arg\inf_{x \in G} y(x) \in H$ and $H \subset G$ imply that $\inf_{x \in G} y(x) = \inf_{x \in H} y(x)$. Noticing that we observed that the optimal path exactly hits level b, we are done. □

Now define the following variational problems, as required in the next theorem. Notice that their sum, optimized over $u \in [0, t]$, equals I_1.

$$I_1^{\mathrm{A}}(u) = \inf_{f \in S_1^{\mathrm{A}}(u)} \int_0^u I_{f(s)}(f'(s))\, \mathrm{d}s,$$
$$S_1^{\mathrm{A}}(u) := \{ f \mid f(0) = f_0 \text{ and } f(u) = c \}, \tag{6}$$

$$I_1^{\mathrm{B}}(u) = \inf_{f \in S_1^{\mathrm{B}}(u)} \int_u^t I_{f(s)}(f'(s))\,\mathrm{d}s, \tag{7}$$

$$\text{where} \quad S_1^{\mathrm{B}}(u) := \left\{ f \mid f(u) = c \text{ and } \int_u^t f(s)\,\mathrm{d}s = b + c(t-u) \right\}.$$

Notice that problems (5) and (7) are contraint variational problems. We recall from the theory of calculus of variations [Gelfand and Fomin 1963; Shwartz and Weiss 1995] the first order necessary conditions for an optimal (absolute continuous) f of an unconstraint problem

$$\inf \int_{t_1}^{t_2} I_{f(s)}(f'(s))\,\mathrm{d}s.$$

The conditions are known as the Euler equations, and say that for all $s \in [t_1, t_2]$ it must hold that

$$\frac{\partial}{\partial f} I_{f(s)}(f'(s)) = \frac{\mathrm{d}}{\mathrm{d}s} \frac{\partial}{\partial f'} I_{f(s)}(f'(s)).$$

In the next proposition, we prove that our constraint variational problems can be solved by solving the associated Euler equations.

PROPOSITION 3.3. *Variational problems* (5), (6), *and* (7) *lead to Euler equations with a unique solution.*

PROOF. Problems (5) and (7) are constraint variational problems, and therefore we apply the Euler equations to the Lagrangians. We first consider problem (5). The Lagrangian problem reads as

$$I_{0,K} := \inf_{f: f(0) = f_0} \int_0^t I_{f(s)}(f'(s))\,\mathrm{d}s - K\left(\int_0^t f(s)\,\mathrm{d}s - (b - b_0 + ct) \right). \tag{8}$$

This problem is similar to Lagrangian problem

$$I'_{0,K} := \inf_{f: f(0) = c} \int_0^t I_{f(s)}(f'(s))\,\mathrm{d}s - K\left(\int_0^t f(s)\,\mathrm{d}s - (b + ct) \right). \tag{9}$$

The Euler equations are

$$\frac{\partial}{\partial f} h(f(s), f'(s)) = \frac{\mathrm{d}}{\mathrm{d}s} \frac{\partial}{\partial f'} h(f(s), f'(s)),$$

where

$$h(f(s), f'(s)) := I_{f(s)}(f'(s)) - K(f(s) - c).$$

Then, Theorem 13.43 in Shwartz and Weiss [1995] says that, for any buffer level b and time t (where $t(1-c) > b$), there exists a Lagrange multiplier K such that the solution of the variational problem (9) satisfies these Euler equations. Furthermore, in Section 13.2 of Shwartz and Weiss [1995], it is shown that the Euler equations corresponding to this variational problem have a *unique* solution. It is easy to see that consequently this same property also holds for (8), for all t and b such that $t(1-c) > b - b_0$.

Completely analogously to (5), we can show that (6) and (7) have a unique optimizing path. □

In the next section, we deal with higher dimensions. The Euler equations that go with (8) will then be given explicitly in the proof of Proposition 3.5.

3.2 Heuristic Derivation of the Decay Rate and Optimal Path

In this section, we give—for general dimension d—a heuristic derivation of the most likely path (see Heuristic 3.4). The type of argument we use can be applied to find the optimal path in many other variational problems, and has an entirely probabilistic nature. For this reason, we include its derivation.

For a path to be optimal, a necessary condition is that it solves the Euler conditions. In Proposition 3.5, we show that the path of Heuristic 3.4 satisfies these equations. Notice that for $d = 2$ this means that we have found the unique solution, based on Proposition 3.3.

Lemma 3.2 implicitly says that if I_0 is the smaller, then the queuing trajectory corresponding to $f^*(\cdot)$ reaches level b at time t, without an idle period in between. Based on this observation, it can be expected that the transient overflow probability roughly equals

$$\mathrm{P}\left(\sum_{\ell=1}^{n} A_\ell(t) \approx n(b - b_0 + ct) \mid F_n(0) = f_0\right). \tag{10}$$

If on the other hand I_1 is the smaller, then the queuing trajectory corresponding to $f^*(\cdot)$ reaches level b at time t, with some idle time in between. This gives rise to the following rough characterization of the overflow probability:

$$\sup_{u,\, f(u)} \mathrm{P}(F_n(u) \approx f(u) \mid F_n(0) = f_0) \cdot \mathrm{P}\left(\sum_{\ell=1}^{n} A_\ell(t-u) \approx n(b + c(t-u)) \mid F_n(u) = f(u)\right), \tag{11}$$

where the optimization is over $u \in [0, t]$ and all $f(u)$ such that $\sum_{i=1}^{d} r_i f_i(u) = c$.

We see that the minimization over $S_0 \cup S_1$ determines the choice between a direct path, and a path with one or more idle periods. The advantage of a path in which the queue is empty for some time is that the traffic to be generated by the sources is lower than $b - b_0 + ct$, but its drawback is that $b + c(t - u)$ has to be built up quite quickly. Consequently, for small values of t, the direct path will be optimal. For larger t, the process first goes more or less to equilibrium, and builds up a buffer b during the last part of $[0, t]$. Interestingly, there is an epoch t_b that may be called the "bifurcation time": for $t < t_b$, the direct path is more likely; for $t > t_b$, the queue will have idle time before overflow. Based on these observations, we develop the following Heuristic.

HEURISTIC 3.4. *A heuristic derivation of the optimal paths of variational problems* (5), (6), *and* (7) *is given as follows:*

—First consider problem (5). As suggested above, the transient probability of our interest asymptotically equals (10). Notice that—with Cramér's

theorem—the exponential decay rate of this probability equals

$$\sup_{\theta}\left(\theta(b-b_0+ct)-\sum_{i=1}^{d} f_{i,0}\log M_i(\theta;t)\right); \tag{12}$$

let θ^* be the supremizing argument. Expression (10) provides information on the state of the sources at time 0, as well as the amount of fluid that entered the system in the interval $[0,t]$. However, by invoking Laplace's principle [Dupuis and Ellis 1997], implicitly the entire most likely trajectory of the distribution of the sources in the interval $[0,t]$ is given. This can be explained in the following three steps.

(1) First, we introduce the matrix $G_n(s)$: its (i,j)th entry is the fraction of the sources that were in state i at time 0 that is in state j at time s (of course, the rowsums of $G(s)$ equal 1). We condition probability (10) to all possible values of $G(s)$:

$$\int \mathrm{P}\left(\sum_{\ell=1}^{n} A_\ell(t)\approx n(b-b_0+ct)\mid F_n(0)=f_0, G_n(s)\approx g(s)\right) \cdot \mathrm{P}(G_n(s)=g(s)\mid F_n(0)=f_0)\,\mathrm{d}g(s). \tag{13}$$

(2) Now, consider both probabilities in the integrand of (13). Both of them can be evaluated asymptotically by means of the standard Large Deviations theorems of Cramér and Sanov [Dupuis and Ellis 1997; Shwartz and Weiss 1995]. Denote by α and β the exponential decay rates of both probabilities; then

$$\alpha = \sup_{\theta}\left(\theta(b-b_0+ct)-\sum_{i,j=1}^{d} f_{0,i}g_{ij}(s)\log\left(\frac{B_{ij}(\theta;s)M_j(\theta;t-s)}{p_{ij}(s)}\right)\right), \tag{14}$$

$$\beta = \sum_{i,j=1}^{d} f_{0,i}g_{ij}(s)\log\left(\frac{f_{0,i}g_{ij}(s)}{p_{ij}(s)}\right). \tag{15}$$

(3) Now we apply Laplace's principle, which states that the decay rate of an integral equals (under specific conditions) the decay rate of the maximum of the integrand. Heuristically, this principle says that asymptotically all probability mass is concentrated on one point, and this point can be regarded as "most likely." In this case, we minimize the sum $\alpha+\beta$ of (14) and (15) with respect to $g(s)$. Tedious calculations yield

$$g^*_{ij}(s)=\frac{B_{ij}(\theta^*;s)M_j(\theta^*;t-s)}{\sum_{k=1}^{d} B_{ik}(\theta^*;s)M_k(\theta^*;t-s)},$$

to be interpreted as the the most likely fraction (of the $f_{0,i}$ sources that were in state i at time 0) that are in state j at time s. Therefore, our conjecture

for the most likely fraction of sources in state j at time s is $\sum_{i=1}^{d} f_{0,i} g_{ij}^*(s)$:

$$\hat{f}_j(s) = \sum_{i=1}^{d} f_{0,i} \frac{B_{ij}(\theta^*;s) M_j(\theta^*;t-s)}{\sum_{k=1}^{d} B_{ik}(\theta^*;s) M_k(\theta^*;t-s)} = \sum_{i=1}^{d} f_{0,i} \frac{B_{ij}(\theta^*;s) M_j(\theta^*;t-s)}{M_i(\theta^*;t)}. \tag{16}$$

—Now consider (6). Again on the basis of the remark of the beginning of this subsection, if $I_1 < I_0$ the transient probability of our interest asymptotically equals (11), where at the optimizing u it holds that the distribution $f(u)$ is such that $\sum_{i=1}^{d} r_i f_i(u) = c$. The trajectory on the interval $[0,u]$ can be found analogously to the three-step recipe presented above (conditioning, large deviations on the individual terms, and "Laplace") (see Mandjes [1999]). After calculations, we find

$$\hat{f}_j(s) = \sum_{i,k=1}^{d} x_i p_{ij}(s) p_{jk}(u-s) y_k, \quad s \in [0,u], \tag{17}$$

where x and y are such that

$$\sum_{k=1}^{d} x_i p_{ik}(u) y_k = f_{i,0} \quad \text{and} \quad \sum_{i=1}^{d} x_i p_{ik}(u) y_k = f_k(u). \tag{18}$$

As was extensively treated in Mandjes [1999], the resulting decay rate is

$$\sum_{i=1}^{d} (f_{0,i} \log x_i + f_i(u) \log y_i), \tag{19}$$

where it is emphasized that x and y depend on u, as they are determined by (18).

—Finally, consider (7). Based on (11), the path on $[u,t]$ is heuristically derived analogously to (5), and is given by

$$\hat{f}_j(s) = \sum_{i=1}^{d} f_i(u) \frac{B_{ij}(\theta^*;s) M_j(\theta^*;t-s)}{M_i(\theta^*;t)}, \quad s \in [u,t], \tag{20}$$

where θ^* is optimizing argument in

$$\sup_{\theta} \left(\theta(b + c(t-u)) - \sum_{i=1}^{d} f_i(u) \log M_i(\theta;t) \right). \tag{21}$$

□

PROPOSITION 3.5. *The paths $\hat{f}(\cdot)$—given by* (16), (17), *and* (20)*—satisfy the Euler equations that correspond to variational problems* (5), (6), *and* (7), *respectively.*

PROOF. We again distinguish between the three cases. As the proof of (20) is completely analogous to (16), we omit this case.

—First, consider path (16). We first have to find an expression for the local rate function (3) along the above conjectured trajectory (16). This requires the

knowledge of the optimizing vector θ in the definition of (3). We conjecture that this so-called "twist" along path (16) is given by

$$\hat{\theta}_i(s) = \log M_i(\theta^*; t-s) + k(s),$$

for some function $k(s)$ (constant on $\{1, \ldots, d\}$). This is proved analogously to the proof in Section 3.4 of Mandjes and Ridder [1999]. We summarize the required steps and refer to Mandjes and Ridder [1999] for the details.

(1) Find the derivative of path (16). After some algebra:

$$\hat{f}'_j(s) = \sum_{i \neq j} \hat{f}_i \lambda_{ij} \frac{M_j(\theta^*; t-s)}{M_i(\theta^*; t-s)} - \sum_{i \neq j} \hat{f}_j \lambda_{ji} \frac{M_i(\theta^*; t-s)}{M_j(\theta^*; t-s)}.$$

(2) Solve (3) via the first order conditions. After algebra, $\hat{\theta}(s)$ must satisfy

$$y_j = \sum_{i \neq j} x_i \lambda_{ij} \frac{\exp(\hat{\theta}_j(s))}{\exp(\hat{\theta}_i(s))} - \sum_{i \neq j} x_j \lambda_{ji} \frac{\exp(\hat{\theta}_i(s))}{\exp(\hat{\theta}_j(s))}.$$

(3) Substitute in the last expression $x = \hat{f}(s)$, $y = \hat{f}'$ and $\hat{\theta}_i(s) = \log M_i(\theta^*; t-s) + k(s)$.

We then show that path $\hat{f}$ solves the Euler equations associated with problem (5). To this end, first define—similarly to Proposition 3.3—

$$h(x, y) := I_x(y) - K \sum_{i=1}^{d} x_i (r_i - c),$$

where K is the associated Lagrange multiplier. So, for all $s \in [0, t]$ and $i = 1, 2, \ldots, d$ it must hold that

$$\frac{\partial}{\partial x_i} h(x, y)\Big|_{x=\hat{f}(s), y=\hat{f}'(s)} = \frac{\mathrm{d}}{\mathrm{d}s} \left(\frac{\partial}{\partial y_i} h(x, y)\Big|_{x=\hat{f}(s), y=\hat{f}'(s)} \right) = \hat{\theta}'_i(s),$$

where the last equality is due to Theorem C.2 of Shwartz and Weiss [1995]. The expression of $I_x(y)$ can be worked out using the exact expressions for $y = \hat{f}'(s)$ and $\hat{\theta}(s)$ given above. After algebra, similarly as in Section 3.5 of Mandjes and Ridder [1999], we verify the Euler equations. We find that they are solved for $K = \theta^*$ (being the optimizing argument in (12)) and $k(\cdot)$ being a constant function.

—Let us concentrate on (17). Analogously to the three steps above (with details in Section 2.2 of Mandjes [1999]), we find that the twist along the first part of the conjectured path is

$$\hat{\theta}_i(s) = \log \left(\sum_{k=1}^{d} p_{ik}(u-s) y_k \right).$$

The Euler equations associated with (unconstrained) problem (6) are

$$\frac{\partial}{\partial x_i} I_x(y)\Big|_{x=\hat{f}(s), y=\hat{f}'(s)} = \frac{\mathrm{d}}{\mathrm{d}s} \left(\frac{\partial}{\partial y_i} I_x(y)\Big|_{x=\hat{f}(s), y=\hat{f}'(s)} \right),$$

for $s \in [0, u]$ and $i = 1, 2, \ldots, d$. The rest is algebra, since we have again the exact expressions of $\hat{f}'(s)$ and $\hat{\theta}(s)$. □

3.3 Calculation of the Decay Rate and Optimal Path

Combining the results of the previous two sections, we derive the decay rate and optimal path for the case $d = 2$.

THEOREM 3.6. *For $d = 2$, decay rate (2) is given by*

$$\min\left\{I_0, \min_{u\in[0,t]}\left\{I_1^{\mathrm{A}}(u) + I_1^{\mathrm{B}}(u)\right\}\right\}. \tag{22}$$

Here I_0 is given by (12), $I_1^{\mathrm{A}}(u)$ by (19), and $I_1^{\mathrm{B}}(u)$ by (21).

PROOF. The proof is a combination of the previous results of this section. First, Lemma 3.2 states that decay rate (2) is given by the minimum of I_0 and I_1, where the latter obviously equals $\min_{u\in[0,t]}\{I_1^{\mathrm{A}}(u) + I_1^{\mathrm{B}}(u)\}$. I_0 can be calculated by inserting the $\hat{f}(\cdot)$ of (16) into $J_t(\cdot)$, justified by Propositions 3.3 and 3.5. We get (12). Similarly, we find that $I_1^{\mathrm{A}}(u)$ and $I_1^{\mathrm{B}}(u)$ equal (19) and (21), respectively. □

COROLLARY 3.7. *For $d = 2$ and for any $\delta > 0$,*

$$\lim_{n\to\infty} \mathrm{P}\left(\sup_{v\in[0,t]} |F_n(v) - \hat{f}(v)| < \delta \mid Q_n(0) = nb_0, F_n(0) = f_0, Q_n(t) \geq nb\right) = 1.$$

If (22) is given by I_0, then $\hat{f}(\cdot)$ is given by (16). Else, $\hat{f}(\cdot)$ is given by the concatenation of (17) and (20), where

$$u = u^* := \arg\min_{u\in[0,t]}\left\{I_1^{\mathrm{A}}(u) + I_1^{\mathrm{B}}(u)\right\}. \tag{23}$$

In other words, for $d = 2$, we have $f^*(\cdot) = \hat{f}(\cdot)$.

PROOF. The statement that we have found the optimal path, follows from our previous results: Proposition 3.5 says that $\hat{f}$ solves the Euler equations, and Proposition 3.3 says that the Euler equations have a unique solution.

Given the rare event, the sample paths lie close to the optimal path almost surely. Under certain conditions, this holds more generally (see, e.g., Theorem 6.15 in Shwartz and Weiss [1995]). For the Markov fluid case, we refer also to Theorem 13.41 of Shwartz and Weiss [1995]. □

4. FAST SIMULATION

In this section, we consider the possibility of getting quick estimates of the transient overflow probability (1) by simulation. The goal is twofold. In the first place, such an efficient simulation technique is, of course, interesting on its own right. In the second place, we use it to verify possible approximations of the transient overflow probability.

Section 4.1 explains why variance reduction techniques are required to efficiently estimate probability (1). We advocate the use of *importance sampling*; this technique entails simulation under a probability model that differs from

the actual one, such that rare behavior becomes more frequent. We emphasize that the novelty of our method is that this alternative probability model is adapted during the simulation continuously, in contrast with all earlier proposed methods. In Section 4.2, we show how the knowledge of the optimal path (Section 3) enables to find this new probability model. Then Section 4.3 briefly sketches implementation details of our simulation method. In Section 4.4, we give simulation results, and comment on approximations.

4.1 Rare Event Simulation—Importance Sampling

Importance sampling has the potential to speed up simulations significantly, given that some global knowledge of the system is available (see, e.g., Asmussen and Rubinstein [1995] and Heidelberger [1995]). Here we briefly review its goal, some relevant concepts, and the status of the literature.

Rare Event Simulation Requires Variance Reduction. As we assumed decay rate (2) to be positive, the event under consideration is *rare*: as scaling parameter n grows, probability (1) decays exponentially fast. Denoting the event under consideration by A_n, in the standard simulation procedure—"crude Monte Carlo" (CMC)—one draws independent samples and estimates $\mathrm{P}(A_n)$ as the fraction of samples that lie in A_n. The accuracy of the estimate is measured by its *relative error* (RE), that is, the standard deviation divided by the expectation of the estimator. Let N be the number of simulation experiments, then $\mathrm{RE} \sim 1/\sqrt{\mathrm{P}(A_n)N}$ (see, e.g. Asmussen and Rubinstein [1995] and Heidelberger [1995]). This means that to achieve some fixed accurate estimate (i.e., 10% RE) the required sample size N is proportional to $1/\mathrm{P}(A_n)$. In our case, the transient overflow probability decays exponentially in n, and, hence, the required number of runs grows exponentially with n. Clearly, this justifies the need for a *variance reduction technique*.

Importance Sampling—Major Concepts. In simulations with importance sampling Monte Carlo (ISMC), the samples are drawn independently according to another probability Q. The simulation data of sample path ω (viz. "1" if the sample is in A_n and "0" else) are multiplied by a *likelihood ratio* $L(\omega) := (\mathrm{dP}/\mathrm{dQ})(\omega)$ to keep unbiasedness. The average of the obtained numbers is the importance sampling estimate.

A crucial issue in this approach is to find a good alternative measure Q; chosen wrongly, the procedure might even lead to variance increase. To measure the quality of the new probability model, one analyzes again the RE and the required sample size N (as explained above) of the estimate as functions of the parameter n (see, e.g., Asmussen and Rubinstein [1995] and Heidelberger [1995]).

—Ideal would be an estimate with bounded RE in which case the required N may be chosen fixed. A situation where this happens is as follows: From (2), one might suggest

$$P(A_n) \sim k_1 \exp(-nI),\ n \to \infty, \tag{24}$$

where $f(n) \sim g(n)$, $n \to \infty$ means $\lim_{n\to\infty} f(n)/g(n) = 1$, and where k_1 is some constant, and I is the decay rate in (2). Now *assume* that the second moment of the IS-estimate satisfies the asymptotic

$$\mathrm{E_Q}[L^2 1(A_n)] \sim k_2 \exp(-2In),\ n \to \infty, \tag{25}$$

$\mathrm{E_Q}[\cdot]$, denoting expectation under Q, and $1(\cdot)$ being the indicator function. Then it is easy to see that $\mathrm{RE} \sim \sqrt{(k_2 - k_1^2)/(Nk_1^2)}$ is bounded (in n) and the required sample size N to obtain $\mathrm{RE} = \epsilon$ is fixed $N \approx (k_2 - k_1^2)/(\epsilon^2 k_1^2)$.

—The situation sketched above is very unlikely. More generally, it holds that the constants k_1 and k_2 in the asymptotics (24) and (25) are functions of n, preferably $k_1(n)$ and $k_2(n)$ are polynomials. Then a number of properties hold that are easy to verify:

(i) The RE is polynomial.

(ii) The required N is polynomial.

(iii) The new probability model is *asymptotically optimal*, meaning

$$\lim_{n\to\infty} \frac{\log \mathrm{E_Q}[L^2 1(A_n)]}{\log \mathrm{P}(A_n)} = 2. \tag{26}$$

—Without assuming the asymptotics for the second moment $\mathrm{E_Q}[L^2 1(A_n)]$, one calculates the ratio (26) from the simulation data. By Jensen's inequality and because the denominator is negative, the ratio is at most 2. The closer it is to 2, the better the implemented Q. In the original probability model the ratio equals 1, when we assume that $\mathrm{P}(A_n)$ decays exponentially.

Notice that we did not take into account the time complexity in this analysis, that is, the computing time in the simulation experiments. For a discussion on this matter, we refer to Section 17.2.2 in Asmussen and Rubinstein [1995].

Use of Large Deviations—Optimal Paths. As said above, given some global knowledge on the system's behavior, ISMC can improve the estimator's variance properties considerably. This global knowledge involves the characteristics of the system during its path towards the unlikely state or event. The idea is that Q is chosen such that the average sample path during a simulation experiment mimics the (rare) optimal path to overflow. Therefore, large deviations techniques have proven to be useful for finding good Q.

This procedure has been pursued successfully in systems where the rarity was due to a large buffer rather than a large number of sources. The crucial feature there is that the most likely path to overflow is essentially a *straight line* (cf. Anantharam [1988]). Measure Q is chosen such that the resulting average path coincides with this straight line. The alternative measure induces constant probability distributions of the random variables involved. For instance, in GI/G/1 queues, the interarrival times and the service times have given fixed new laws. Notice that under the new measure Q the queue becomes unstable. References for (networks of) GI/G/m queues are Parekh and Walrand [1989] and Sadowsky [1991]; for Markov fluid driven queues, we mention Kesidis and Walrand [1993], and Mandjes and Ridder [1995], and Ridder [1996]. In many

cases, asymptotic optimality has been proven [Asmussen and Rubinstein 1995; Heidelberger 1995].

The model of the present article is essentially different: the rarity is due to a large number of sources. The optimal path to the rare event is nonlinear, and therefore a constant change of measure cannot apply. As a consequence, the transition rates that correspond to the alternative measure have to be updated *during the simulation run.* Related procedures were considered in Kroese and Nicola [1998] and Ridder [1999] and in a more abstract context by Cottrell et al. [1983]. We have not succeeded in formally proving asymptotic optimality; empirically, we show in Section 4.4 that a significant speed-up is achieved.

4.2 The New Transition Rates

As motivated in the previous section, measure $\mathbb{Q}$ has to be chosen such that the process behaves under $\mathbb{Q}$ on average according to the paths from Heuristic 3.4. We will show below that under $\mathbb{Q}$ the original time-homogeneous modulating Markov chains are replaced by time-inhomogeneous Markov chains (with rate matrix $\Lambda(s) = (\lambda_{ij}(s))_{i,j=1}^d$, for $s \in [0,t]$).

The average path under $\Lambda(s)$ is given by the distributions $p(s) = (p_i(s))_{i=1}^d$ that satisfy the forward Kolmogorov equations $p'(s) = p(s)\Lambda(s)$. Or, in detail:

$$p_j'(s) = \sum_{i:i\neq j} p_i(s)\lambda_{ij}(s) - \sum_{i:i\neq j} p_j(s)\lambda_{ji}(s), \quad j = 1,2,\ldots,d; \quad s \in [0,t]. \tag{27}$$

As our objective was to mimic the optimal path, we have to find rate $\Lambda(s)$ such that $p(\cdot) \equiv \hat{f}(\cdot)$. The following proposition gives such a generator, which lies within the class of exponential twists [Asmussen and Rubinstein 1995; Mandjes and Ridder 1995; Ridder 1996].

PROPOSITION 4.1. *Transition rates* $\Lambda(s)$ *that solve* (27) *are the following:*

(i) *If* $I_0 < I_1$*: let* θ^* *defined as an optimizer of* (12)*. Then, for* $s \in [0,t]$,

$$\lambda_{ij}(s) = \lambda_{ij}\,\frac{M_j(\theta^*;t-s)}{M_i(\theta^*;t-s)}\;(i \neq j). \tag{28}$$

(ii) *If* $I_0 \geq I_1$*: let* u^* *as in* (23) *and* θ^* *as an optimizer of* (21)*. Then, for* $s \in [0,u^*]$,

$$\lambda_{ij}(s) = \lambda_{ij}\,\frac{\sum_{k=1}^d p_{jk}(u^*-s)y_k}{\sum_{k=1}^d p_{ik}(u^*-s)y_k}\;(i \neq j), \tag{29}$$

and for $s \in [u^*,t]$,

$$\lambda_{ij}(s) = \lambda_{ij}\,\frac{M_j(\theta^*;t-s)}{M_i(\theta^*;t-s)}\;(i \neq j).$$

The diagonal elements $\lambda_{ii}(s)$ *are given by* $-\sum_{j:j\neq i}\lambda_{ij}(s)$.

PROOF. First consider the case $I_0 < I_1$. Rewrite the expression (16) of the optimal path

$$\hat{f}_j(s) = \sum_{i=1}^d \frac{f_{0,i}}{M_i(\theta^*;t)}\;\sum_{k=1}^d B_{ij}(\theta^*;s)B_{jk}(\theta^*;t-s),$$

where θ^* stems from (12). Take the derivative with respect to s:

$$\hat{f}'_j(s) = \sum_{i=1}^{d} \frac{f_{0,i}}{M_i(\theta^*;t)} \sum_{k=1}^{d} \left(\frac{\partial}{\partial s} B_{ij}(\theta^*;s) B_{jk}(\theta^*;t-s) + B_{ij}(\theta^*;s) \frac{\partial}{\partial s} B_{jk}(\theta^*;t-s) \right). \tag{30}$$

The partial derivative of the matrix element $B_{ij}(\theta;s)$ follows from the definition $B(\theta;s) := \exp((\Lambda + \theta R)s)$:

$$\begin{aligned} \frac{\partial}{\partial s} B_{ij}(\theta;s) &= \sum_{k=1}^{d} B_{ik}(\theta;s)(\Lambda + \theta R)_{kj} \\ \frac{\partial}{\partial s} B_{ij}(\theta;s) &= \sum_{k=1}^{d} (\Lambda + \theta R)_{ik} B_{kj}(\theta;s), \end{aligned}$$

where $R := \mathrm{diag}\{r\}$. Apply the first variant to the first term in (30) and the second variant to the second term. Then, after some obvious manipulations, we obtain that the right hand side of (30) is indeed

$$\sum_{i:i\neq j} \hat{f}_i(s)\lambda_{ij}(s) - \sum_{i:i\neq j} \hat{f}_j(s)\lambda_{ji}(s),$$

when the transition rates are given by (28).

For the second case, the line of reasoning is similar as above. First, consider the $[0, u^*]$ part. The expression of the optimal path is given in (17). Take the derivative with respect to s, and apply the property that transition matrix $P(s)$ satisfies the differential equations

$$P'(s) = P(s)\Lambda \quad \text{and} \quad P'(s) = \Lambda P(s).$$

After manipulations, we get

$$\hat{f}'_j(s) = \sum_{i:i\neq j} \hat{f}_i(s)\lambda_{ij}(s) - \sum_{i:i\neq j} \hat{f}_j(s)\lambda_{ji}(s),$$

when the transition rates are given by (29). For the $[u^*, t]$ part, the optimal path is given in (20). Taking the derivative goes similarly as above. Again, we conclude that the optimal path $\hat{f}(\cdot)$ satisfies Eq. (27). □

4.3 Implementation Issues

In the previous section, we have found the time-inhomogeneous Markov chain that should be used as probability measure in the ISMC simulation. This section presents details on the implementation.

First, we introduce the d-dimensional vector $Y(s)$; its ith component records the number of modulating Markov chains in state i ($i \in \{1, \ldots, d\}$). We have simulated the jumps of this process by applying *uniformization* (see, e.g., Tijms [1994, p. 154]). This is done as follows. Choose γ such that

$$\gamma \geq n \max_{s\in[0,t]} \max_{i\in\{1,\ldots,d\}} \left\{ \sum_{j:j\neq i} \lambda_{ij}(s) \right\}. \tag{31}$$

The simulation of the time-inhomogeneous Markov chain is done as follows. In the simulation, we realize so-called *jump epochs* according to a Poisson(γ) process. Only at these epochs the process $Y(\cdot)$ can change state. If y_i denotes the number of sources in state i at that jump, then the probability of a source moving from i to j is $\lambda_{ij}\, y_i/\gamma$ (note that there is a possibly positive probability of a self-transition).

The above implementation would require on-line calculation of the rates $\Lambda(s)$, and consequently $M_i(\theta^*; t-s)$ or $p_{ij}(s)$. To avoid this, we propose the following alternative. Divide interval $[0, t]$ in K subintervals. Say that $[t_k, t_{k+1}]$ is the kth subinterval. Then, we let all transition rates in this subinterval obey, the same rule, in that there is in that interval a fixed change of measure. For $s \in [t_k, t_{k+1}]$, the approximation for $\lambda_{ij}(s)$ is

$$\hat{\lambda}_{ij}(s) := \frac{1}{2}(\lambda_{ij}(t_k) + \lambda_{ij}(t_{k+1})).$$

The advantage is that the alternative transition rates can be computed off-line, once for all simulation runs. Experiments show that the loss of efficiency in variance reduction is marginal, but the gain in simulation times is considerable [Ridder 1999]. We let the number of subintervals depend on the overflow time t, keeping the widths small enough. In the experiments, we took the uniformization constant γ in (31) as small as possible, and the widths of the subintervals about $1/50$ of the overflow time t.

4.4 Simulation Study

In this section, we consider the following model. We assume that sources are of the exponential on/off type; the mean time in the on-state is 0.5 seconds and the mean off-time is 1.0 seconds. This model is commonly used for voice [Schwartz 1996, p. 26]. While in the on-state a source transmits at a rate of 100 per second. The scaled link rate is $c = 50$, the scaled initial content $b_0 = 0.25$, the scaled target buffer value is $b = 1.0$, and the initial fraction in the on-state is $f_0 = 0.55$.

The goal of this section is twofold. First, we comment on a number of approximations of the transient overflow probability, and then we assess the quality of the proposed simulation approach.

4.4.1 *Approximations.* First, we derive from Cramér's theorem the *most likely overflow time*, and denote it by t^*. It is the time t that minimizes the exponential decay rate (12) of the overflow probability:

$$t^* := \arg\inf_{t>0} \sup_{\theta} \left(\theta(b - b_0 + ct) - \sum_{i=1}^{d} f_{i,0} \log M_i(\theta; t) \right).$$

Furthermore, we denote the bifurcation time by t_b. This is the time epoch where the large deviations rate function switches from I_0 to I_1. In the example, we found numerically the most likely overflow time $t^* = 0.2181$, and the bifurcation time $t_b = 1.4894$. So, in a natural way, we discriminate between three cases for the overflow time t: (i) $t < t^*$, (ii) $t^* < t < t_b$, and (iii) $t > t_b$. These three cases

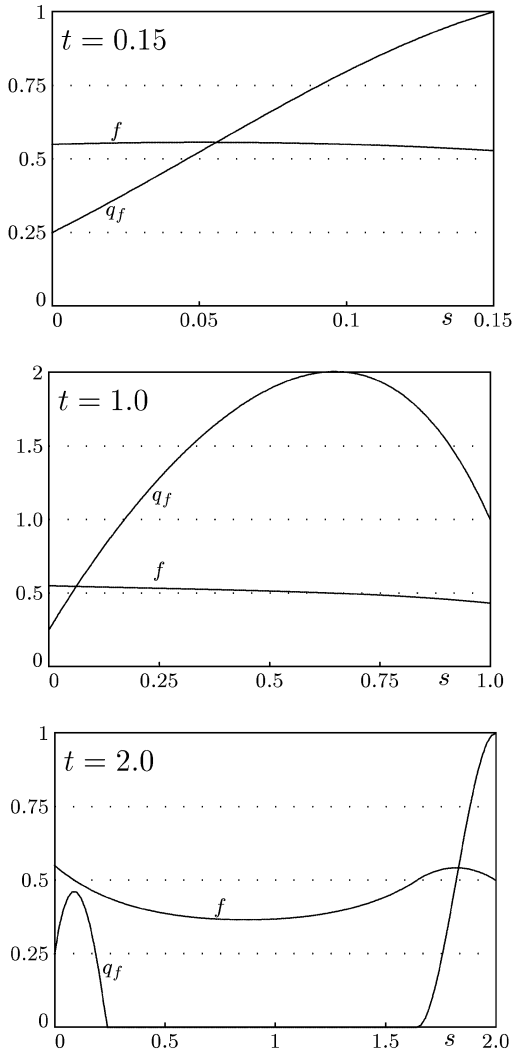

Fig. 1. Optimal paths $f(s)$ and buffers $q_f(s)$, where $f(s) := f_{\text{on}}(s)$, the fraction of sources in the on-state.

show essentially different overflow behavior, as demonstrated by the optimal paths below. Figure 1 shows the optimal paths of the fraction of sources in the on-state and the buffer contents. These are numerically determined from the expressions (16), (17), and (20). Remarkably, we see that, in the case of $t^* < t_b$, the buffer content reaches higher levels before dropping down to b at time t.

For the three regimes identified above, a number of possible approximations are the following:

—The first approximation is simply based on (2):

$$\mathrm{P}(Q_n(t) \geq nb \mid Q_n(0) = nb_0 \text{ and } F_n(0) = f_0) \approx \exp[-nI(b, t \mid b_0, f_0)].$$

—The second approximation applies the Bahadur–Rao refinement of Cramér's theorem [Bahadur and Rao 1960].

$$\mathrm{P}(Q_n(t) \geq nb \mid Q_n(0) = nb_0 \text{ and } F_n(0) = f_0) \approx \frac{1}{\sqrt{2\pi n}\,\sigma\theta^*} \exp[-nI(b, t \mid b_0, f_0)],$$

where θ^* solves Cramér's equation (12) and σ^2 is the variance of the total input of a source generated during $[0, t]$ given the distribution of initial state of the source, and assuming that the on/off times of the source are θ^*-exponentially tilted. σ^2 is computed by

$$\sigma^2 = \sum_{i=1}^{d} f_{i,0} \left[\frac{M_i''(\theta^*; t)}{M_i(\theta^*; t)} - \left(\frac{M_i'(\theta^*; t)}{M_i(\theta^*; t)} \right)^2 \right].$$

This approximation is only possible when Cramér's theorem is applied for determining the large deviations decay rate (see Section 3.2). That is, when the overflow time t is smaller than the bifurcation time t_b.

We use simulation to validate these approximations of the transient overflow probabibility. Figure 2 shows the estimates and approximations of the transient overflow probabilities for the three cases $t = 0.15, 1.0, 2.0$ for varying number of n sources. The simulations were executed until the relative half width of the 95%-confidence interval of the estimate is 15% to both sides of the estimate: we call this (95%, 15%)-efficiency. The order of the probability estimates goes down to 10^{-10}, which can be seen from the plots: the values are given in ^{10}log scale. Clearly, the Bahadur–Rao approximations perform very well (Note: This approximation is possible only for $t < t_b \approx 1.5$). The large deviations approximations form an upper bound and exceed the estimates by a factor 10. As a more detailed comparison, we give in Table I these probability estimates for values larger than 10^{-6}. The differences are due to (i) we performed each experiment one time, and (ii) only in 95% of the cases we are sure that the true probability differs at most 15% of both the CMC and the ISMC values. Also, in Table I, we give the number of simulations runs that were executed to obtain these estimates.

4.4.2 *Comparison of Simulation Methods.* We consider three ways of comparing the CMC and ISMC simulation methods.

—The first possibility is by keeping the number of simulation runs constant and calculating the relative error (RE) of the estimates. A smaller RE means variance reduction. A simulation run starts at time 0 with a given number of sources n, the given initial buffer content nb_0 and fraction f_0. A run ends at time t.

Table II shows these relative errors of the probability estimates for increasing number of sources n, when the number of runs is fixed: 200000 in CMC and 10000 in ISMC. From the numbers, we conjecture that the CMC relative error increases exponentially (as a function of the number of sources n), while the ISMC relative error increases linearly. This indicates a huge acceleration of the simulations. There are no overflow observations in the CMC

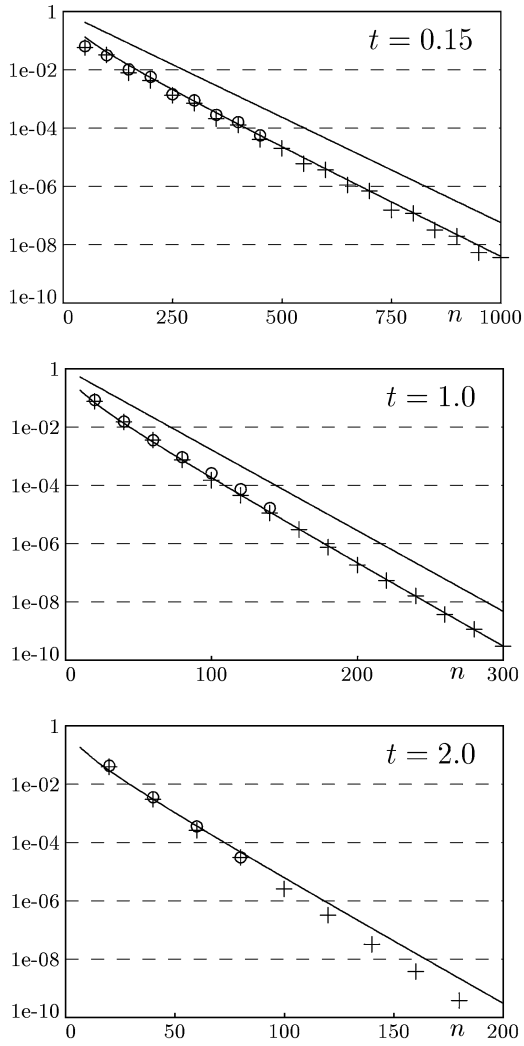

Fig. 2. Simulation estimates of the overflow probability (○: CMC, +: ISMC) and approximations (top: Cramér, bottom: Bahadur–Rao). No B-R for $t = 2$.

simulations for large n. In these cases, we give (between parentheses) the estimated relative error using the estimated probability $\hat{p}$ from the ISMC simulations: as mentioned above, RE $\sim 1/\sqrt{\hat{p}N}$, where N is the number of runs.

—In an alternative comparison, we require that the relative width of the 95%-confidence interval of the estimate is 15% to both sides of the estimate, that is, a (95%, 15%)-efficiency. The relative error of the estimate is approximately 0.0765 in such cases. Then we compare the number of simulation runs that is required before this happens. In Table I, we have listed these figures (below the corresponding estimates). We conclude that the number of runs increases roughly linearly under ISMC and exponentially under CMC, and hence there is significant variance reduction.

Table I. (95%, 15%)-Efficient Probability Estimates and the Corresponding Required Number of Simulation Runs (in parentheses)

$t = 0.15$			$t = 1.0$			$t = 2.0$		
n	CMC	ISMC	n	CMC	ISMC	n	CMC	ISMC
50	6.5735e-02 (2434)	5.8970e-02 (479)	20	8.6788e-02 (1809)	7.7276e-02 (1167)	20	4.3850e-02 (3740)	4.0011e-02 (2186)
100	3.2108e-02 (5170)	3.3204e-02 (684)	40	1.5803e-02 (10694)	1.5023e-02 (972)	40	3.6441e-03 (46925)	3.1320e-03 (4786)
150	1.0652e-02 (15865)	7.9807e-03 (937)	60	3.6214e-03 (47219)	3.6431e-03 (2110)	60	3.6417e-04 (469559)	2.6619e-04 (2028)
200	5.8968e-03 (28829)	4.3287e-03 (1322)	80	9.4180e-04 (181567)	7.6056e-04 (2815)	80	3.1311e-05 (5461423)	3.0851e-05 (4092)
250	1.4851e-03 (115145)	1.3477e-03 (1595)	100	2.6500e-04 (645283)	1.5055e-04 (2846)	100		2.5805e-06 (2881)
300	9.1527e-04 (186831)	7.2262e-04 (1756)	120	7.5268e-05 (2271873)	4.5558e-05 (5873)			
350	2.9039e-04 (588866)	2.1591e-04 (2111)	140	1.7016e-05 (10049560)	1.1509e-05 (3722)			
400	1.6495e-04 (1036699)	1.3100e-04 (3081)	160		3.0628e-06 (10364)			
450	5.7827e-05 (2957097)	4.1582e-05 (3787)						
500	2.2524e-05 (7591867)	2.0444e-05 (4176)						
550		6.0543e-06 (4646)						

Table II. Relative Errors (200000 Runs in CMC and 10000 Runs in ISMC)

$t = 0.15$			$t = 1.0$			$t = 2.0$		
n	CMC	ISMC	n	CMC	ISMC	n	CMC	ISMC
100	0.0117	0.0195	40	0.0174	0.0404	40	0.0365	0.0824
200	0.0304	0.0290	60	0.0362	0.475	50	0.0646	0.0660
300	0.0716	0.0368	80	0.0675	0.0408	60	0.1132	0.0813
400	0.1562	0.0444	100	0.1291	0.0675	70	0.2425	0.0501
500	0.5000	0.0958	120	0.2887	0.0565	80	0.4082	0.0616
600	(1.15)	0.0869	140	0.5773	0.0666	90	0.5773	0.0480
700	(2.64)	0.0712	200	(5.21)	0.1008	150	(21.7)	0.0795
800	(6.45)	0.0747	300	(127.7)	0.1487	200	(353.3)	0.0522
900	(15.80)	0.1120	400	(3332.6)	0.2084	250	(6241.7)	0.0677
1000	(36.62)	0.1023	500	(101241)	0.1656	300	(86771)	0.1117

—Finally, we checked whether the ratios of (26) converge to 2, which would indicate that the new model is asymptotically optimal. We found in the experiments of Table I that the ratios remain between 1.7 and 1.85, however slowly increasing when the number of sources gets larger.

5. CONCLUSION

In this article, we have studied the Markov fluid model with many sources and infinite buffer. Given initial conditions (states of the Markov sources and the buffer content at time 0), we were interested in the probability that the buffer content exceeds a certain level at a finite time t (which we call the transient

overflow probability). The emphasis of our study was on the determination of the decay rate of this probability as the number of sources increases; as a by-product, we gain insight into the typical statistical behavior of the sources and the buffer content during the trajectory to overflow. Asymptotically, that is, as the number of sources is growing large, this problem becomes a variational problem of which no solutions were known. For exponential on–off sources, we solved this variational problem; for general Markov fluid sources, we have proposed a plausible heuristic. This heuristic is based on the standard large deviations theorems of Cramér and Sanov and on Laplace's principle.

We have considered some interesting consequences of these large deviations results. Most importantly, we have described and implemented a technique for quick simulation of transient overflow probabilities, based on importance sampling. The idea is to use a change of measure, in such a way that the average statistical behavior under the new measure coincides with the deviant behavior of reaching buffer overflow at time t under the old measure. This approach led to a probability model in which the Markov sources have time dependent transition rates. The results showed that the importance sampling simulations yield strong variance reduction compared to standard simulations. We could not prove that the importance sampling simulation approach is asymptotically optimal. A second consequence of our large deviations results is the use of the asymptotic decay rate of the transient overflow probability in two approximations (of this probability) that perform very well; particularly, the Bahadur–Rao based approximation in the small t region is very accurate.

Further investigations are needed to develop more accurate approximations in the large t region. Also, further analysis is required to prove the large deviations principle to hold for higher dimensions ($d > 2$, general Markov fluid sources). Also, the asymptotic optimality of the simulation method may be verified formally.

ACKNOWLEDGMENTS

The authors wish to thank the referees for their assistance in improving the presentation of the article.

REFERENCES

ANANTHARAM, V. 1988. How large delays build up in a $GI/G/1$ queue. *Queue. Syst. 5*, 345–368.

ANICK, D., MITRA, D., AND SONDHI, M. 1982. Stochastic theory of a data-handling system with multiple sources. *Bell Syst. Tech. J. 61*, 1871–1894.

ASMUSSEN, S. AND RUBINSTEIN, R. 1995. Steady state rare event simulation in queueing models and its complexity properties. In *Advances in Queueing Theory, Theory, Methods and Open Problems*, J. Dshalalow, Ed. CRC Press, Boca Raton, USA, 429–461.

BAHADUR, R. AND RAO, R. R. 1960. On deviations of the sample mean. *Ann. Math. Stat. 31*, 1015–1027.

BOTVICH, D. AND DUFFIELD, N. 1995. Large deviations, the shape of the loss curve, and economies of scale in large multiplexers. *Queue. Syst. 20*, 293–320.

BRANDT, A. AND BRANDT, M. 1994. On the distribution of the number of packets in the fluid flow approximation of packet arrival streams. *Queue. Syst. 17*, 275–315.

COHEN, J. 1974. Superimposed renewal processes and storage with gradual input. *Stochast. Proc. Appl. 2*, 31–57.

COTTRELL, M., FORT, J.-C., AND MALGOUYRES, G. 1983. Large deviations and rare events in the study of stochastic algorithms. *IEEE Trans. Automat. Cont. 28*, 907–920.

COURCOUBETIS, C. AND WEBER, R. 1996. Buffer overflow asymptotics for a buffer handling many traffic sources. *J. Appl. Prob. 33*, 886–903.

DUFFIELD, N. 1998. Conditioned asymptotics for tail probabilities in large multiplexers. *Perform. Eval. 31*, 281–300.

DUPUIS, P. AND ELLIS, R. 1997. *A Weak Convergence Approach to the Theory of Large deviations*. Wiley, New York.

GELFAND, I. AND FOMIN, S. 1963. *Calculus of Variations*. Prentice-Hall, Englewood Cliffs, N.J.

HEIDELBERGER, P. 1995. Fast simulation of rare events in queueing and reliability models. *ACM Trans. Model. Comput. Simul. 5*, 43–85.

KESIDIS, G. AND WALRAND, J. 1993. Quick simulation of ATM buffers with on-off multiclass Markov fluid sources. *ACM Trans. Model. Comput. Simul. 3*, 269–276.

KESIDIS, G., WALRAND, J., AND CHANG, C.-S. 1993. Effective bandwidths for multiclass Markov fluids and other ATM sources. *IEEE/ACM Trans. Netw. 1*, 424–428.

KNESSL, C. AND MORRISON, J. 1991. Heavy traffic analysis of a data-handling system with many sources. *SIAM J. Appl. Math. 51*, 187–213.

KOBAYASHI, H. AND REN, Q. 1992. A mathematical theory for transient analysis of communication networks. *IEICE Trans. Commun. E75-B*, 1226–1276.

KOSTEN, L. 1974. Stochastic theory of a multi-entry buffer, part 1. *Delft Progress Report, Series F 1*.

KOSTEN, L. 1984. Stochastic theory of data-handling systems with groups of multiple sources. In *Performance of Computer-Communication Systems*, H. Rudin and W. Bux, Eds. Elsevier, Amsterdam, the Netherlands, 321–331.

KROESE, D. AND NICOLA, V. 1998. Efficient simulation of backlogs in fluid flow lines. *AEU International Journal on Electronics and Communications 52*, 165–171.

KROESE, D. AND NICOLA, V. 1999. Efficient simulation of a tandem jackson network. In *Proceedings of the Winter Simulation Conference* (Phoenix, Az.), 411–419.

MANDJES, M. 1999. Rare event of the state frequencies of a large number of markov chains. *Stochastic Models 15*, 577–592.

MANDJES, M. AND RIDDER, A. 1995. Finding the conjugate of Markov fluid processes. *Prob. Engin. Inf. Sci. 9*, 297–315.

MANDJES, M. AND RIDDER, A. 1999. Optimal trajectory to overflow in a queue fed by a large number of sources. *Queue. Syst. 31*, 137–170.

PAREKH, S. AND WALRAND, J. 1989. A quick simulation method for excessive backlogs in networks of queues. *IEEE Trans. Automat. Cont. 34*, 54–66.

RIDDER, A. 1996. Fast simulation of markov fluid models. *J. Appl. Prob. 33*, 786–803.

RIDDER, A. 1999. Efficient simulation of fluid queues with many sources. In *Proceedings of the 2nd International Workshop on Rare Event Simulation*. University of Twente, Netherlands, 19–27.

SADOWSKY, J. 1991. Large deviations and efficient simulation of excessive backlogs in a $GI/G/m$ queue. *IEEE Trans. Automat. Cont. 36*, 1383–1394.

SCHWARTZ, M. 1996. *Broadband Integrated Networks*. Prentice-Hall, Upper Saddle River, N.J.

SHWARTZ, A. AND WEISS, A. 1995. *Large Deviations for Performance Analysis, Queues, Communication, and Computing*. Chapman and Hall, New York.

SIMONIAN, A. AND GUIBERT, J. 1995. Large deviations approximation for fluid queues fed by a large number of on/off sources. *IEEE J. Sel. Areas Commun. 13*, 1017–1027.

TANAKA, T., HASHIDA, O., AND TAKAHASHI, Y. 1995. Transient analysis of fluid model for ATM statistical multiplexer. *Perform. Eval. 23*, 145–162.

TIJMS, H. 1994. *Stochastic Models. An Algorithmic Approach*. Wiley, New York.

WEISS, A. 1986. A new technique of analyzing large traffic systems. *Adv. Appl. Prob. 18*, 506–532.

WISCHIK, D. 2001. Sample path large deviations for queues with many inputs. *Ann. Appl. Prob.*

Received October 1999; accepted December 2001