Ann Oper Res (2008) 160: 215–225
DOI 10.1007/s10479-008-0309-2

The erlangization method for Markovian fluid flows

V. Ramaswami · Douglas G. Woolford · David A. Stanford

Published online: 24 January 2008

Abstract For applications of stochastic fluid models, such as those related to wildfire spread and containment, one wants a fast method to compute time dependent probabilities. Erlangization is an approximation method that replaces various distributions at a time t by the corresponding ones at a random time with Erlang distribution having mean t. Here, we develop an efficient version of that algorithm for various first passage time distributions of a fluid flow, exploiting recent results on fluid flows, probabilistic underpinnings, and some special structures. Some connections with a familiar Laplace transform inversion algorithm due to Jagerman are also noted up front.

Keywords Erlangization · Fluid flow · Markov modulation · Markov process

1 Introduction

The subject of this paper is the development of an efficient version of the "Erlangization" procedure for the canonical Markov modulated fluid flow (MMFF) model, called by some authors as a "fluid queue." This procedure involves approximating the joint distribution of the fluid level and the environment (otherwise referred to as "phase") at time t by the corresponding distribution at an independent, Erlang-distributed horizon with mean t. In addition to the computation of the distribution at time t, it also applies to the evaluation of a variety of first passage time distributions. The method appears to have been first used by Asmussen et al. (2002) to approximate finite-time ruin probabilities in certain insurance risk models. It has also been used in other contexts like models of fire spread and containment; see Stanford et al. (2005b).

V. Ramaswami
AT&T Labs Research, 180 Park Avenue, Florham Park, NJ 07932, USA
e-mail: vramaswami@att.com

D.G. Woolford (✉) · D.A. Stanford
Department of Statistical & Actuarial Sciences, The University of Western Ontario, London, ON, Canada
e-mail: doug.woolford@utoronto.ca

An inversion formula, based on a theorem due to Widder, see Hirchman and Widder (1955),

$$f_n(t) = \frac{(-1)^n}{n!} s^{n+1} \frac{d^n \tilde{f}(s)}{ds^n}\bigg|_{s=(n+1)/t}$$

providing an approximate inversion for a function f on $[0, \infty)$ from its Laplace transform $\tilde{f}$ has been provided by Jagerman (1978, 1982). Noting that

$$\tilde{f}(s) = \int_0^\infty e^{-st} f(t)\, dt,$$

and differentiating inside the integral on the right side repeatedly, we can rewrite the Jagerman inversion formula as

$$f_n(t) = \frac{(\frac{n+1}{t})^{n+1}}{\Gamma(n+1)} \int_0^\infty e^{-\frac{n+1}{t}u} u^n f(u)\, du,$$

from which it is easy to see that this inversion is none other than Erlangization with an order $(n+1)$ Erlang distribution. To the best of our knowledge, this fact has not been recognized in the literature. As an important consequence of this, from Jagerman (1978, 1982) we get not only that $f_n(t) \to f(t)$ as $n \to \infty$ pointwise, but also that the convergence is uniform over compact intervals, and furthermore that $f_n(\cdot)$ would inherit many properties of $f(\cdot)$ such as monotonicity, absolute monotonicity, convexity, and log-convexity. Thus, not only does Erlangization provide a quick way to compute transient results for the fluid model, but it is also attractive as an approximation preserving some important properties of the function approximated. We refer the reader to Jagerman (1978, 1982) for applicable error bounds, an acceleration scheme, etc., related to this method.

For fluid models, we will show below that the method entails essentially the solution of one non-linear problem, namely, the determination of the matrix Ψ_0 governing the phase transition in a busy period that ends before the expiry of the first stage of the Erlang process, and that all other quantities are determined from it through a solution of linear equations. This is quite unlike the general Laplace inversion methods proposed by Ahn and Ramaswami (2004, 2005, 2006), or Bean et al. (2005), which require the evaluation of the busy period transform at a multiple set of complex values of its argument. Note that, in the insurance context, it has been noted already in the literature (e.g., Asmussen et al. 2002; Stanford et al. 2005a) that with even very small values of n such as 1 or 2, Erlangization provides estimates of the required probabilities within 20% of the actual value, with the error exceeding 10% only rarely. Thus, the method is fast and has reasonable accuracy for many applications. While noting these nice properties of the method, we must, however, also note its major disadvantage, namely that its errors have been shown by Jagerman to be of order $O(1/n)$ and therefore go to zero very slowly with n. This fact has been partially mollified in Asmussen et al. (2002) and Stanford et al. (2005a) by resorting to a Richardson extrapolation. Nonetheless, in practice, this method is only useful when not more than a few decimal places of accuracy are needed in the computed probabilities; one must use more sophisticated methods as in Ahn and Ramaswami (2005, 2006), or Ramaswami (2006), if one needs greater levels of accuracy.

Our main contribution here is an extension of the Erlangization idea to fluid models with arbitrary rates; prior use of this method appears to have been restricted mainly to fluid models with unit rates of (absolute) change. In addition, we have also provided a recursive scheme for computing the blocks of a key matrix (the busy period matrix Ψ of the "enlarged

model" below) and given an interesting probabilistic interpretation of its sub-blocks in terms of the stages in the Erlangization process. A key contribution here is the identification of the block structure of Ψ that leads to obvious savings in computational complexity. Our derivation is also entirely probabilistic and very transparent, unlike the Feynman-Kac approach presented in Asmussen et al. (2002).

In our analysis we draw on the approach to fluid flows initiated by Ramaswami (1999) based on the matrix geometric method and level crossing arguments. Building on that work, da Silva Soares and Latouche (2002) gave an alternate QBD representation besides the one identified by Ramaswami and provided a probabilistic interpretation. Later, a more elegant interpretation in terms of stochastic coupling of the fluid model to a queue was obtained by Ahn and Ramaswami (2003), who also used it to provide a detailed transient analysis of the fluid model in a rigorous manner (Ahn and Ramaswami 2004). Ahn & Ramaswami also obtained the transient results in a somewhat easier manner via differential equations in Ahn and Ramaswami (2006), a paper in which they derived the transient distribution for an arbitrary initial state. This work will draw on several results from Ramaswami (1999, 2006) as well as Ahn and Ramaswami (2005) without re-deriving those results. These papers provide some of the background enabling an efficient implementation of Erlangization.

An attractive feature of the matrix-geometric approach is the ability it provides to relate the MMFF to a Quasi Birth-and-Death (QBD) process and thereby to compute the stationary distribution of the fluid level in terms of the familiar G matrix characterizing a busy period of the QBD. For computing G, powerful algorithms like the Latouche-Ramaswami algorithm (Latouche and Ramaswami 1993) exist. For general background on QBDs and the matrix-geometric method, the interested reader is directed to Neuts (1981) and Latouche and Ramaswami (1999).

2 Markov modulated fluid flows and their busy period

Consider a fluid flow $F(\cdot)$ modulated by a finite state Markov chain $J(\cdot)$. Let the Markov chain's infinitesimal generator be Q and fluid rates be given by a vector $\mathbf{r}$. Here, the component r_i is the instantaneous rate of change of the fluid when the Markov chain is in state i. We shall assume the state space of the Markov chain to be partitioned into subsets Ω_+ and Ω_- such that for $i \in \Omega_+$, we have $r_i > 0$, and for $i \in \Omega_-$, we have $r_i < 0$. (The restriction that all $r_i \neq 0$ is not germane to the analysis, and the more general case where some $r_i = 0$ can be handled in a straightforward manner; see the papers of Ahn and Ramaswami (2003, 2004).)

In the analysis, it is customary to partition various matrices that appear according to the partitioning of the state space given above; thus, for instance the infinitesimal generator Q gets partitioned as

$$Q = \begin{bmatrix} Q_{++} & Q_{+-} \\ Q_{-+} & Q_{--} \end{bmatrix}. \tag{1}$$

In the above, the indexing by pairs of symbols $+$ and $-$ denote the appropriate subsets of rows and columns appearing in the sub-matrix, and that convention shall be adopted throughout for all partitioned matrices.

For later purposes, we define the matrices C_{++} and C_{--} as the diagonal matrices whose diagonals are respectively the vector of absolute rates $|r_i|$ for states in Ω_+ and for states in Ω_- respectively. We also define the matrix S such that $S = (\mathrm{diag}(C_{++}, C_{--}))^{-1} Q$ with attending partitions.

For the fluid model, a fundamental quantity of interest is the transform matrix $\Psi(s)$ of order $|\Omega_+| \times |\Omega_-|$ and defined for $Re(s) \geq 0$. Let τ denote a return time to fluid level 0 by the MMFF and let $\chi(\cdot)$ denote an indicator function. Now, the element $[\Psi(s)]_{ij}$ is the Laplace transform of the random variable $\tau\,\chi(J(\tau) = j)$ given $F(0) = 0, J(0) = i$. The matrix $\Psi(s)$ is called "the busy period transform" as a carry over from queueing theory. By an abuse of notation, we shall let $\Psi = \Psi(0)$; a similar convention will be adopted with respect to all transforms of interest.

Our goal in this section is to derive an equation satisfied by $\Psi(s)$ that will play an important role in what follows. A particular case of that equation for $s = 0$ in the case when $|r_i| = 1$ can be found in the literature (see e.g. Asmussen et al. 2002) and has been derived using a Feynman-Kac type equation. Our derivation is for the general case, is probabilistic, and hopefully more transparent. The required equation is actually equivalent to an equation derived in Ahn and Ramaswami (2005) using similar arguments as given below, but went unrecognized as such. We state our result as a theorem.

Theorem 1 *The matrix $\Psi(s)$ satisfies the equation*

$$S_{++}\Psi(s) + S_{+-} + \Psi(s)(S_{--} - sC_{--}^{-1}) + \Psi(s)S_{-+}\Psi(s) - sC_{++}^{-1}\Psi(s) = 0. \tag{2}$$

Proof Choose a positive number ξ such that $\xi \geq \max_i |S(i,i)|$. Define the nonnegative stochastic matrix $P = \xi^{-1}S + I$, where I is an identity matrix. Consider a Markov renewal process defined by a discrete time Markov chain with successive states evolving according to the transition matrix P, whose steps are separated by time intervals which are such that a transition interval out of state i has an exponential distribution with parameter $\xi|r_i|$ independent of all else. It has been shown in Theorem 1 of Ahn and Ramaswami (2004) that the resulting semi Markov process (SMP) is indeed a version of the continuous time Markov chain with generator Q.

We can now consider the fluid flow as being modulated by the foregoing semi-Markov process, and the advantage resulting thereby is what has been called in the literature as "spatial uniformization," namely that the potential change to the fluid in any interval of transition of the SMP is i.i.d. as an exponential variable with parameter ξ.

Now, conditioning on the first step of this SMP and the amount of increment x for the fluid level that results in it, we can write

$$[\Psi(s)]_{ij} = \int_0^\infty \xi e^{-\xi x} e^{-s(x/r_i)} \left[\sum_{k \in S} [P]_{ik} [G(s,x)]_{kj} \right] dx \tag{3}$$

where $G(s,x)$ is the matrix of transforms of first visit to fluid level 0 from fluid level x; in the above, the term $\exp(-s(x/r_i))$ is the conditional transform of the time to the first transition of the SMP given that in it we have an increment x for the fluid level. In partitioned matrix form, we can write the above equations as

$$\Psi(s) = \int_0^\infty e^{-(\xi I + sC_{++}^{-1})x} [P_{++}G_{+-}(s,x) + P_{+-}G_{--}(s,x)]\,dx. \tag{4}$$

It is well-known (see e.g. Theorem 3(c) in Ahn and Ramaswami 2005) that

$$G_{--}(s,x) = e^{H(s)x}, \quad \text{and} \quad G_{+-}(s,x) = \Psi(s)G_{--}(s,x),$$

where $H(s) = S_{--} + S_{-+}\Psi(s) - sC_{--}^{-1}$. Substituting the above in (4) yields

$$\Psi(s) = \int_0^\infty e^{-sxC_{++}^{-1}}[P_{++}\Psi(s) + P_{+-}]e^{-(\xi I - H(s))x}\,dx. \tag{5}$$

Post-multiplying both sides of the above by $\xi I - H(s)$ and integrating the right side by parts, one obtains

$$\Psi(s)[\xi I - H(s)] = \xi[P_{++}\Psi(s) + P_{+-}] - sC_{++}^{-1}\Psi(s). \tag{6}$$

The asserted equation (2) is immediately obtained by substituting for $H(s)$ and simplifying the above. □

Corollary 2 *The matrix* Ψ *satisfies the equation*

$$\Psi S_{-+}\Psi + \Psi S_{--} + S_{++}\Psi + S_{+-} = 0. \tag{7}$$

3 Erlangization of the busy period

Our starting point is a fluid flow modulated by the finite state, continuous time Markov chain $\{\varphi_t : t \geq 0\}$ with infinitesimal generator Λ and (nonzero) fluid rates $\mathbf{r}$. As before, we shall denote the fluid level at time t by X_t but the state of the underlying Markov chain by φ_t. We shall let Φ denote the state space of this Markov chain, and use a partitioning similar to that introduced for the general fluid flow with respect to states with positive and negative rates. Our goal is to determine the first passage time distribution of the fluid level to zero within an Erlang horizon; more precisely, we determine the probability that level zero has been reached prior to a random time η which has an Erlang distribution of order L and rate parameter λ. Here, η is thus the sum of L independent exponential random variables each with mean $1/\lambda$. When $\lambda = L/t$, the computed distributions at η provide an Erlangized approximation for the distribution at time t. For later use we denote by R_{++} and R_{--} the diagonal matrices formed by the fluid rates for states in Φ_+ and absolute fluid rates for states in Φ_-.

As is well-known (Neuts 1981; Latouche and Ramaswami 1999), the Erlang distributed time period can be considered to be the distribution of the absorption time in a continuous time Markov chain which gets absorbed after spending time in a set of L transient phases in series. The time spent in each transient phase is i.i.d. exp(λ). The infinitesimal rates for the transient phases are governed by the $L \times L$ matrix

$$H = [\text{diag}(-\lambda), \text{superdiag}(\lambda)],$$

which is a matrix whose diagonal elements are $-\lambda$, principal super-diagonal elements are λ, and all other elements are zero.

To facilitate the study of the fluid until the expiry of the Erlang horizon, we construct an absorbing, composite continuous time Markov chain made up of the stages of the Erlang and the phase of the Markov chain modulating the fluid. Denoting by e_k a vector of 1's of order k and $h = He_L$, we can write the infinitesimal generator of this expanded Markov chain as

$$\tilde{Q} = \begin{bmatrix} 0 & 0 & 0 \\ h \otimes e_k & H \oplus \Lambda_{++} & I_L \otimes \Lambda_{+-} \\ h \otimes e_m & I_L \otimes \Lambda_{-+} & H \oplus \Lambda_{--} \end{bmatrix} \tag{8}$$

$$= \begin{bmatrix} 0 & 0 & 0 \\ h \otimes e_k & Q_{++} & Q_{+-} \\ h \otimes e_m & Q_{-+} & Q_{--} \end{bmatrix}, \quad \text{say,} \tag{9}$$

where $\otimes$ denotes the Kronecker product, $\oplus$ the Kronecker sum, and k and m are respectively the number of positive and negative rates r_i. A typical transient state of this Markov chain (p, i) with $1 \le p \le L,\ i \in \Phi$ denotes that the Markov process defining the Erlang horizon is in stage p and the phase of the original Markov chain modulating the fluid is in state i. Entry into the absorbing state 0 of this Markov chain denotes the end of the Erlang horizon, beyond which we have no interest. Now to view simultaneously the Markov chain defining the Erlang horizon and the evolution of the fluid up to the end of that horizon, we may consider a fluid model modulated by $\tilde{Q}$ whose fluid rate in the state (p, i) is r_i.

We note that for the purpose of determining the behavior of the fluid up to the Erlang horizon, it suffices to consider the fluid model defined on the (defective) infinitesimal generator

$$Q = \begin{bmatrix} Q_{++} & Q_{+-} \\ Q_{-+} & Q_{--} \end{bmatrix}. \tag{10}$$

The states of this (defective) Markov chain are of the form (p, i) with $1 \le p \le L$ and $i \in \Phi$, and escape from this chain marks the expiry of the Erlang horizon. This Markov chain and the associated fluid model will henceforth be called the "expanded fluid model."

Before we proceed further, it is convenient to consider the matrix S associated with Q. Recall from Sect. 2 that to obtain S, we must re-scale the rates in Q by the absolute values of fluid rates. Specifically, we have

$$S((\cdot, i), (\cdot, j)) = \frac{1}{|r_i|}\ Q((\cdot, i), (\cdot, j)),$$

and it is fairly easy to verify that we can write the matrix S in a partitioned form similar to the Q-matrix in (10) with

$$S_{ij} = (I_L \otimes R_{ii})^{-1} Q_{ij}, \quad i, j \in \{+, -\},$$

where the diagonal matrices R_{ii} are the diagonal R-matrices defined by the fluid model (X_t, φ_t). In the notation of Sect. 2, for the fluid model under consideration driven by the defective Markov chain, $C_{++} = I_L \otimes R_{++}$ and $C_{--} = I_L \otimes R_{--}$.

We can now prove the following structural result for the matrix Ψ associated with the expanded fluid model.

Theorem 3 *The matrix Ψ associated with the expanded fluid model has the structure*

$$\Psi = \begin{bmatrix} \Psi_0 & \Psi_1 & \Psi_2 & \dots & \Psi_{L-1} \\ 0 & \Psi_0 & \Psi_1 & \dots & \Psi_{L-2} \\ 0 & 0 & \Psi_0 & \dots & \Psi_{L-3} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \Psi_0 \end{bmatrix}. \tag{11}$$

Furthermore, for $0 \le k \le L-1$, the block Ψ_k in the above form captures the probabilities of returning to fluid level 0 before expiry of the Erlang horizon and while the Erlang stage index has increased by k.

Proof The block upper triangular structure of Ψ arises due to the fact that the stages up to the absorption epoch in the Erlang process form a nondecreasing sequence. The repeating block structure of Ψ is a consequence of the fact that up to the absorption epoch, the rates of fluid change do not depend on the number of elapsed stages of the Erlang process. The interpretation of the blocks is a direct consequence of what the states mean: the stage (k, j) means that the Erlang process is in stage k at time t and $\varphi_t = j$. □

The next result is a precursor to obtaining a convenient set of recursions (see (14)) in computing the blocks of Ψ and leads to a full exploitation of the structural features of the model.

Theorem 4 *The matrix Ψ_0 is indeed the busy period transform associated with the original fluid model (X_t, φ_t) evaluated at $s = \lambda$. Furthermore, the matrices Ψ_k satisfy the equations*

$$\begin{aligned}&\Psi_0(R_{--}^{-1}\Lambda_{-+})\Psi_0 + R_{++}^{-1}(\Lambda_{++} - \lambda I_k)\Psi_0 \\ &\quad + \Psi_0 R_{--}^{-1}(\Lambda_{--} - \lambda I_m) + (R_{++}^{-1}\Lambda_{+-}) = 0; \end{aligned} \tag{12}$$

$$\begin{aligned}&\Psi_l R_{--}^{-1}(\Lambda_{--} - \lambda I_m) + \sum_{j=0}^{l} \Psi_j (R_{--}^{-1}\Lambda_{-+})\Psi_{l-j} + R_{++}^{-1}(\Lambda_{++} - \lambda I_k)\Psi_l \\ &\quad + \lambda R_{++}^{-1}\Psi_{l-1} + \lambda \Psi_{l-1} R_{--}^{-1} = 0; \quad l = 1, \ldots, L-1. \end{aligned} \tag{13}$$

Proof Clearly Ψ_0 contains the probabilities of the fluid process (X_t, φ_t) visiting fluid level 0 before a single Erlang stage is completed. For the fluid X_t to hit level 0 before such an epoch, its busy period length should be some t and no Poisson event in the Poisson process of rate λ defining the Erlang horizon could have occurred till then. Thus Ψ_0 is none other than the transform of the busy period of the process (X_t, φ_t) evaluated at $s = \lambda$.

The equations (12) and (13) satisfied by the sub-blocks of Ψ are obtained by writing out equation (7) using the structure of Ψ and the matrices S_{++} etc., appearing in it; we omit the details. □

Given a matrix A, we define the vector $\mathrm{vec}(A)$ as the vector obtained by concatenating successive columns of A to form a single vector. With this notation, it is well-known (see Graham 1981) that we can write for a conformable matrix product AXB that

$$\mathrm{vec}(AXB) = (B^t \otimes A)\mathrm{vec}(X),$$

where B^t denotes the transpose of the matrix B. We can use this to solve the linear equations (13), and obtain the following recursive solution for $\Psi_l, l = 1, \ldots, L-1$:

$$\begin{aligned}\mathrm{vec}\{\Psi_l\} = &\{[R_{--}^{-1}(\Lambda_{--} - \lambda I_m + \Lambda_{-+}\Psi_0)]^t \otimes I_k \\ &+ I_m \otimes [R_{++}^{-1}(\Lambda_{++} - \lambda I_k) + \Psi_0 R_{--}^{-1}\Lambda_{-+}]\}^{-1} \\ &\times \mathrm{vec}\left\{-\left[\lambda(R_{++}^{-1}\Psi_{l-1} + \Psi_{l-1}R_{--}^{-1}) + \sum_{j=1}^{l-1} \Psi_j R_{--}^{-1}\Lambda_{-+}\Psi_{l-j}\right]\right\}. \end{aligned} \tag{14}$$

This shows that the only nonlinear equation to solve is (12) for Ψ_0, but that matrix can be efficiently computed using the quadratic algorithm for the G-matrix of a QBD (Ramaswami

1999; Ahn and Ramaswami 2003). These steps constitute an extremely efficient computation of the matrix Ψ as opposed to computing Ψ directly as done in cited references. For example, when using an Erlang-L horizon, the upper triangular structure of Ψ is comprised of $L(L+1)/2$ blocks, but only L of these blocks are actually unique. If one wanted to solve for Ψ in its entirety, one could apply the quadratically convergent algorithm of Ahn and Ramaswami (2005), but in this case one would be solving for a matrix of dimension $L \times |\Omega_+| \times |\Omega_-|$. Using our method, we only need apply the same Ahn & Ramaswami algorithm to solve for the Ψ_0 block which is of a much smaller dimension, namely $|\Omega_+| \times |\Omega_-|$. Once this Ψ_0 block is determined, the remaining $L-1$ blocks that are unknown can be found recursively, each through a single application of (14), starting with Ψ_1.

Remark 1 A similar computation can be made of the transform $\Psi(s)$, but for our purposes it is not needed and hence we have restricted our comments to $\Psi = \Psi(0)$.

The foregoing discussion yields the following result for the busy period distribution of the fluid model (X_t, φ_t).

Theorem 5 *Assume that $X_0 = 0$ and the initial phase φ_0 has a distribution given by the vector α which is a probability distribution on Φ_+. With $\lambda = L/t$, the quantity $\alpha \sum_{k=0}^{L-1} \Psi_k e_m$ gives an Erlang order L approximation to the probability that a return to fluid level 0 occurs at or before time t.*

Proof This follows from the fact that by the interpretation of the matrices Ψ_k, the given expression is the probability that the fluid level returns to 0 before the expiry of an Erlang horizon with mean t and L transient stages. □

4 Some other first passage times

Ramaswami (2006) has provided a variety of important first passage time distributions in the canonical fluid model. Naturally, each one of them has a corresponding Erlangized approximation which is obtained by considering a first passage occurring before an Erlang horizon with mean t. Not to belabor the point much, we shall demonstrate how the approximation is constructed with respect to the downward passage times. This discussion is kept brief, and at a fairly intuitive level without unnecessary pedantic detail and notation. Other passage times are handled in a very similar manner, and we omit the details which are fairly straightforward. The interested reader is directed to Woolford (2007) where formulas for the Erlangized approximation to many first passage times are developed. Woolford (2007) also applies the Erlangization method for MMFFs in models for fire perimeter growth.

First passage to lower levels Since in the fluid model (X_t, φ_t), the first passage time distribution from fluid level $y+x$ to fluid level y, $x, y > 0$, is independent of y due to spatial homogeneity, it suffices to consider a first passage from x to 0. As noted by Ahn and Ramaswami (2005), the transform of the downward passage time from x to 0 is given in matrix form as $e^{U(s)x}$ where

$$U(s) = R_{--}^{-1}[\Lambda_{--} + \Lambda_{-+}\Psi(s)] \tag{15}$$

and $\Psi(s)$ is the busy period transform; the special case of the above for the unit rate case has appeared in several earlier papers: Asmussen (1995), Ramaswami (1999) and da Silva Soares and Latouche (2002).

Underlying the formula above is a method of cutting out the return times to various intermediate levels between x and 0 during a first passage from x to 0 and treating the evolution of the phase a function of the purged amount of fluid which is made to play the role of "time" in such a construction; see Ramaswami (1999), Theorem 3.2 for details of this procedure which is also used in an earlier paper of Asmussen (1995). This construction leads to the consideration of a Markov reward model (here reward is the transform of elapsed clock time) whose infinitesimal changes can be described by the matrix $U(s)$ in (15).

The first term in $U(s)$ keeps track of the local increment in clock time when the phase remains in the downward set of phases Φ_- while the second term keeps track of the clock time (forming a return time distributed as a busy period) that needs to be excised while treating the phase as a function of the fluid level. The appearance of the diagonal matrix R_{--}^{-1} in the above and similar other formulae can be explained by the intuition given in Ahn and Ramaswami (2006) as arising from a change of measure (Jacobian) that is needed to translate increments over small time intervals to increments over small fluid levels (in phase i a small fluid change dx translates to a small time change $r_i^{-1}dt$) and is rigorously established in Ahn and Ramaswami (2004) and Ahn and Ramaswami (2006); see also Sect. 3.5 in Ahn et al. (2007). Note that in creating this infinitesimal generator we start off by considering the phase process as a function of fluid levels.

Under Erlangization, we would approximate the downward passage time distribution at time t by an Erlang horizon with L phases and $\lambda = L/t$. Clearly, for the first passage to occur before the end of the Erlang horizon, all the periods in question, including all the excised periods in the above construction, should end before the Erlang horizon. Naturally, we need to keep track of the number of stages of the Erlang process elapsing in excised periods as well. A little thought will reveal that the matrix

$$Q_{--} + Q_{-+}\Psi \tag{16}$$

indeed describes the evolution of the stages of the Erlang process and the phase process of the original fluid model after the various return intervals have been excised. This is so because in these matrices there is an automatic accounting of the stage of the Erlang we are in as we make every transition, and also immediately after each excised interval. The Erlangized approximation for the first passage time from x to 0 would be e^{Ux} where

$$U = S_{--} + S_{-+}\Psi. \tag{17}$$

Once again the change from the Q-matrices to the S-matrices accounts for the change of measure that should be effected in moving from the phase process when it is described as a function of the fluid level in (16), as opposed to our need to describe the phase as a function of clock time which occurs when we use the matrix in (17). We can now determine the chance of the fluid reaching level zero prior to the expiry of the horizon.

Theorem 6 *Assume that, starting at fluid level $u > 0$, the system is in an increasing state with the various relative likelihoods given by initial probability vector $\boldsymbol{\alpha_0}$. The probability $\Psi(u, H)$ of reaching level 0 prior to the expiry of the horizon is*

$$\Psi(u, H) = (\alpha_0, \mathbf{0}, \ldots, \mathbf{0})\Psi e^{Uu} e. \tag{18}$$

First passage to higher levels First-passage times to higher levels require the matrix K defined by

$$K = S_{++} + \Psi S_{-+}. \tag{19}$$

As first shown in Ramaswami (1999), and once again exploiting spatial homogeneity, K is such that the element $[e^{Kx}]_{ij}$ represents the expected number of visits to level $u+x$ in increasing phase j, starting from increasing phase i, before returning to any initial level $u \geq 0$. (Here, such visits must also occur prior to expiry of the horizon, as must the various visits we describe below without explicit mention.) Consistent with Ramaswami (2006), we define the following two matrices: $R(x)$, where $[R(x)]_{ij}$ is the expected number of visits to level $x > 0$, in increasing phase j without visiting level 0, starting from level x in decreasing phase i, and the matrix $\Xi(x) = \Psi R(x)$. We interpret $[\Xi(x)]_{ij}$ as the expected number of return visits to level $x > 0$ in increasing phase j, without visiting level 0, starting from level x in increasing phase i. A transform result in Ramaswami (2006), Lemma 2, enables us to write $R(x)$ as

$$R(x) = \int_{y=0}^{x} e^{Uy} S_{-+} e^{Ky}\, dy. \tag{20}$$

Fairly direct application of Ramaswami (2006), Theorem 1 yields the following theorem and the consequent corollary:

Theorem 7 *Assume that the process starts from level $u \geq 0$ in increasing phase i. The probability $[{}_u f_{++}(u,x)]_{ij}$ of reaching level $x > u$ for the first time in an increasing phase j prior to reaching the horizon or returning to the initial level is given by the (i,j)th element of the matrix*

$${}_u f_{++}(u,x) = e^{K(x-u)}[I + \Xi(x-u)]^{-1}. \tag{21}$$

Applying the above result to the expanded model, we get the following result.

Corollary 8 *Assume that, starting at fluid level $u > 0$, the system (X_t, φ_t) is in an increasing phase with the various relative likelihoods given by initial probability vector α_0 on Φ_+. The probability of reaching level $x > u$ prior to the Erlang horizon is given by the quantity*

$$(\alpha_0, \mathbf{0}, \ldots, \mathbf{0}) \exp\{K(x-u)\}[I + \Xi(x-u)]^{-1} e. \tag{22}$$

Remark 2 Badescu et al. (2005) provides an infinite series for $R(x)$. However, an explicit, finite solution is in fact given in Ramaswami (2006), Lemma 2, by once again vectorizing a matrix. Applying this solution in our context, we obtain

$$\mathrm{vec}\{R(x)\} = [K^T \oplus U]^{-1} \mathrm{vec}\{\exp(Ux) S_{-+} \exp(Kx) - S_{-+}\}. \tag{23}$$

Remark 3 Other similar probabilities can be found via this line of reasoning, such as the chance of reaching level zero before expiry of the horizon, starting from a decreasing phase. These require us to exploit results pertaining to rate-reversed fluid flows found in Ramaswami (2006) and Ahn et al. (2007), so we direct the interested reader to these papers for the relevant exposition.

Acknowledgements The following funding sources are gratefully acknowledged: NSERC Discovery Grants have assisted the work of David Stanford and Douglas Woolford. OGSST scholarships, along with NPCDS and GEOIDE grants have provided additional support for Douglas Woolford's work. The authors thank the anonymous referees for their helpful suggestions which led to an improved presentation.

References

Ahn, S., & Ramaswami, V. (2003). Fluid flow models & queues—a connection by stochastic coupling. *Stochastic Models*, *19*(3), 325–348.

Ahn, S., & Ramaswami, V. (2004). Transient analysis of fluid flow models via stochastic coupling to a queue. *Stochastic Models*, *20*(1), 71–101.

Ahn, S., & Ramaswami, V. (2005). Efficient algorithms for transient analysis of stochastic fluid flow models. *Journal of Applied Probability*, *42*(2), 531–549.

Ahn, S., & Ramaswami, V. (2006). Transient analysis of fluid models via elementary level crossing arguments. *Stochastic Models*, *22*(1), 129–147.

Ahn, S., Badescu, A. L., & Ramaswami, V. (2007). Time dependent analysis of finite buffer fluid flows and risk models with a dividend barrier. *Queueing Systems: Theory and Applications*, *55*(4), 207–222.

Asmussen, S. (1995). Stationary distributions for fluid flow models with or without Brownian noise. *Stochastic Models*, *11*, 1–20.

Asmussen, S., Avram, F., & Usábel, M. (2002). Erlangian approximations for finite time ruin probabilities. *ASTIN Bulletin*, *32*, 267–281.

Badescu, A., Breuer, L., Drekic, S., Latouche, G., & Stanford, D. A. (2005). The surplus prior to ruin and the deficit at ruin for a correlated risk process. *Scandinavian Actuarial Journal*, *6*, 433–445.

Bean, N. G., O'Reilly, M., & Taylor, P. G. (2005). Hitting probabilities and hitting times for stochastic fluid flows. *Stochastic Processes and Their Applications*, *115*(9), 1530–1556.

da Silva Soares, A., & Latouche, G. (2002). Further results on the similarity between fluid queues and QBDs. In G. Latouche & P. Taylor (Eds.) *Proceedings of the 4th international conference on matrix-analytic methods* (pp. 89–106). River Edge: World Scientific.

Graham, A. (1981). *Kronecker products and matrix calculus with applications*. New York: Wiley.

Hirchman, I. I., & Widder, D. V. (1955). *The convolution transform*. Princeton: Princeton University Press.

Jagerman, D. L. (1978). An inversion technique for the Laplace transform with application to approximation. *Bell System Technical Journal*, *57*(3), 669–710.

Jagerman, D. L. (1982). An inversion technique for the Laplace transform. *Bell System Technical Journal*, *61*(8), 1995–2002.

Latouche, G., & Ramaswami, V. (1993). A logarithmic reduction algorithm for quasi-birth-death processes. *Journal of Applied Probability*, *30*(3), 650–674.

Latouche, G., & Ramaswami, V. (1999). *Introduction to matrix analytic methods in stochastic modeling*. Philadelphia: SIAM and ASA.

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

Ramaswami, V. (1999). Matrix analytic methods for stochastic fluid flows. In D. Smith & P. Hey (Eds.) *Teletraffic engineering in a competitive world (Proceedings of the 16th international teletraffic congress)* (pp. 1019–1030). Edinburgh: Elsevier Science B.V.

Ramaswami, V. (2006). Passage times in fluid models with application to risk processes. *Methodology and Computing in Applied Probability*, *8*, 497–515.

Stanford, D. A., Avram, F., Badescu, A., Breuer, L., da Silva Soares, A., & Latouche, G. (2005a). Phase-type approximations to finite-time ruin probabilities in the Sparre-Andersen and stationary renewal risk models. *ASTIN Bulletin*, *35*(1), 131–144.

Stanford, D. A., Latouche, G., Woolford, D. G., Boychuk, D., & Hunchak, A. (2005b). Erlangized fluid queues with application to uncontrolled fire perimeter. *Stochastic Models*, *21*(23), 631–642.

Woolford, D. G. (2007). *Stochastic models applied to lightning and wildfires in Ontario*. Ph.D. thesis, University of Western Ontario, London, Ontario, Canada.