Applied Probability Trust (28 April 2005)

EFFICIENT ALGORITHMS FOR TRANSIENT ANALYSIS OF STOCHASTIC FLUID FLOW MODELS

SOOHAN AHN,* *The University of Seoul*

V. RAMASWAMI,** *AT&T Labs*

Abstract

Several algorithms for the busy period distribution of the canonical Markovian fluid flow model are derived. One of them is similar to the Latouche-Ramaswami algorithm for Quasi Birth Death models and is shown to be quadratically convergent. These algorithms add significant efficiency to the matrix-geometric procedures developed earlier by the authors for the transient and steady state analysis of fluid flow models.

Keywords: Stochastic fluid flows; transient analysis; matrix-geometric method; algorithms; quadratic convergence

2000 Mathematics Subject Classification: Primary 60K25
Secondary 90B05;68M20

1. Introduction

The subject of this paper is the construction of efficient algorithms for the transient (time-dependent) analysis of the canonical Markov modulated fluid flow (MMFF) model. That model is obtained by assuming as given an irreducible, continuous time Markov chain (CTMC) $J(t)$ of "phases" with a finite state space $S = S_1 \cup S_2 \cup S_3$ and infinitesimal generator Q, such that: while the phase $i \in S_1$, the fluid level increases at rate $c_i > 0$; while $i \in S_2$, the fluid level decreases at rate $c_i > 0$; and while $i \in S_3$, the fluid level remains constant.

In [2], we derived the joint distribution of $(F(t), J(t))$, where $F(t)$ and $J(t)$ are respectively the fluid level and phase at time $t+$, given that $F(0) = 0$ and $J(0) = i,\ i \in S_1$; see Section 7 in [2] . The matrix of Laplace-Stieltjes transforms (LSTs) with elements

$$\mathcal{E}_{(0,i)}[e^{-sF(t)}\chi(J(t) = j)], \quad i \in S_1 \text{ and } j \in S,$$

where $\mathcal{E}_{(x,i)}$ denotes conditional expectation given the initial state (x, i) and χ is an indicator function, was characterized in terms of three LST matrices $\tilde{K}(s), \tilde{\Psi}(s)$ and $\tilde{\Theta}(s)$. It was shown that these matrices are readily obtained from the LST matrix $\Psi(s)$ of the busy period $\tau = inf\{t > 0 \ : \ F(t) = 0\}$ of the fluid flow model and defined by the elements

$$[\Psi(s)]_{ij} = \mathcal{E}_{(0,i)}[e^{-s\tau}\chi(J(\tau) = j)], \quad i \in S_1 \text{ and } j \in S_2. \tag{1}$$

[*] Postal address: Department of Statistics, The University of Seoul, 90 Jeonnong-dong, Dongdaemun-gu, Seoul 130-743, South Korea

[**] Postal address: 180 Park Avenue, E-233, Florham Park, NJ 07932, USA

In the literature on stochastic fluid flow models, much attention is focused on the steady state distribution; see [3], [4], [7], [16]. As for time dependent distributions, past approaches have been based on Wiener-Hopf factorizations or on partial differential equations. Noteworthy are the work of Sericola [17] based on an ad-hoc series expansion and that of [9] based on spectral methods. Ramaswami [15] is the first systematic use of matrix-analytic methods in the context of fluid flow models and provided a highly efficient algorithm for computing the stationary distribution; Ahn and Ramaswami [1] demonstrated that approach to be based on stochastic coupling to a matrix-geometric queue. A continuation of that work in [2] characterized the time dependent distributions exactly in terms of the busy period transform of the fluid model and provided an accurate algorithm to evaluate them. The methods of [2] require repeated evaluations of the busy period transform, and therefore the quadratically convergent algorithm of this paper improves the efficiency of those methods very significantly. Note that a characterization of the busy period distribution has been obtained earlier by Asmussen [5] who also noted its importance as a fundamental quantity and provided a linearly convergent iterative scheme for its computation; being quadratically convergent, the new algorithm developed here is much faster.

The discrete state space analogue of the stochastic fluid flow is the quasi-birth-and-death (QBD) process for which matrix-geometric methods apply; see [12], [13]. The analogue of the transform $\Psi(s)$ in the QBD model is the matrix G of the latter for which an efficient quadratically convergent algorithm has been obtained by Latouche and Ramaswami [11] using probabilistic arguments. The work here provides an algorithm for the fluid model that is similar in spirit to the Latouche-Ramaswami algorithm for QBDs and has quadratic convergence.

The construction of the algorithms in this paper and the determination of their properties are achieved through the consideration of a closely related queue with interarrival dependent service times (see Section 5) and a probabilistic analysis of that queue. That coupled queue is quite different from those in [1] and [2] which involved service times independently distributed of interarrival times. Thus, this work may also be interpreted as an extension of the algorithms in [11] to queues with interarrival dependent service times.

Throughout this paper, I will denote an identity matrix and $\mathbf{1}$ a column vector of 1's both of whose dimensions will be determined by the context in which they appear. Where it is necessary to indicate the dimension explicitly, we will write I_n to denote the $n \times n$ identity matrix. For later use, we define the diagonal matrices

$$C_j = diag\{c_i, \quad i \in S_j\}, \quad j = 1, 2, 3, \tag{2}$$

where we set $c_i = 1$ for all $i \in S_3$, and let $C = diag(C_1, C_2, C_3)$. We partition the states of the Markov chain in conformity with the three sets S_i identified above and denote its infinitesimal generator in partitioned form as

$$Q = \begin{pmatrix} Q_{11} & Q_{12} & Q_{13} \\ Q_{21} & Q_{22} & Q_{23} \\ Q_{31} & Q_{32} & Q_{33} \end{pmatrix}. \tag{3}$$

Finally, to avoid confusion between submatrices in a partitioned structure and elements of a matrix, the (i, j)-th element of a matrix A will always be denoted by $[A]_{i,j}$, $[A]_{ij}$ or as $A(i, j)$ instead of as A_{ij} as is often customary.

2. Spatial Uniformization

A key step in the analysis of [2] is a procedure called spatial uniformization which we recall below.

A spatial uniformization (for the fluid flow) is effected by modeling the Markov process of phases as a Markov renewal process (MRP) with exponential sojourn times such that potential changes to the fluid level between epochs of that MRP are identically distributed. To that end, we let $\{(J_n, t_n) : n \geq 0\}$ be such an MRP, with successive states $J_n \in S$, transition epochs $0 = t_0 < t_1 < t_2 < \cdots$, and with semi-Markov kernel $H(\cdot)$ defined such that $H(i, j; t)$, the (i, j)th element of $H(t)$, is given by

$$H(i,j;t) = \mathcal{P}\{J_{n+1} = j, t_{n+1} - t_n \leq t | J_n = i\} = \left(1 - e^{-\lambda c_i t}\right) [P_\lambda]_{ij}, \tag{4}$$

where

$$P_\lambda = \lambda^{-1} C^{-1} Q + I, \text{ and } \lambda \geq \max_{i \in S} \left\{-[C^{-1}Q]_{ii}\right\}. \tag{5}$$

The associated semi-Markov process (SMP) $\mathcal{J} = \{J(t) : t \geq 0\}$ is specified such that it takes the value J_n in the interval $t_n \leq t < t_{n+1}$. The following result shows that $\mathcal{J}$ is indeed a realization of the phase process; for a proof, we refer to [2].

Theorem 1. *The process $\mathcal{J} = \{J(t), t \geq 0\}$ is a CTMC with infinitesimal generator Q.*

A sojourn interval of the SMP in $i \in S_1$ being distributed as $exp(\lambda c_i)$ with fluid accumulation at rate c_i per unit time, the additional fluid accumulation in that interval is distributed as $exp(\lambda)$. Similarly, for a state in S_2, given adequate fluid exists at the start of the interval, the potential decrease to the fluid level that could be effected is distributed as $exp(\lambda)$. This underlies our reason for using the nomenclature "spatial uniformization." Throughout the rest of the paper, we shall view the phase process, the CTMC $J(\cdot)$, as being specified by the above construction.

3. Busy Period

For $x > 0$, $i, j \in S$, and $Re(s) > 0$ let $[\hat{G}(s,x)]_{i,j}$ denote the LST

$$[\hat{G}(s,x)]_{i,j} \equiv \mathcal{E}_{(x,i)} \left[e^{-s\tau} \chi(J(\tau) = j)\right]. \tag{6}$$

We assume that the matrix $\hat{G}(s,x)$ of elements $[\hat{G}(s,x)]_{ij}$ is also partitioned according to the sets S_i, $i = 1, 2, 3$. Thus, for instance, the submatrix $\hat{G}_{12}(s,x)$ is the matrix of elements $[\hat{G}(s,x)]_{ij}$ as i varies over S_1 and j varies over S_2.

In our model, fluid gets depleted only in S_2. Thus, all busy periods must end in a state of S_2. From this, the following result is trivial.

Theorem 2. *The matrices $\hat{G}(s,x)$ have the structure*

$$\hat{G}(s,x) = \begin{pmatrix} 0 & \hat{G}_{12}(s,x) & 0 \\ 0 & \hat{G}_{22}(s,x) & 0 \\ 0 & \hat{G}_{32}(s,x) & 0 \end{pmatrix}. \tag{7}$$

We now proceed to determine the submatrices in the second column of the partitioned structure above.

Theorem 3. *For $x > 0$,*
(a) $\hat{G}_{12}(s,x) = \Psi(s)\hat{G}_{22}(s,x)$.
(b) $\hat{G}_{32}(s,x) = (sI - Q_{33})^{-1}Q_{31}\hat{G}_{12}(s,x) + (sI - Q_{33})^{-1}Q_{32}\hat{G}_{22}(s,x)$.
(c) With

$$H(s) = C_2^{-1}[Q_{22} - sI + Q_{23}(sI - Q_{33})^{-1}Q_{32} + \{Q_{21} + Q_{23}(sI - Q_{33})^{-1}Q_{31}\}\Psi(s)], \quad (8)$$

we have

$$\hat{G}_{22}(s,x) = e^{H(s)x}. \quad (9)$$

Proof. Part (a) follows easily by conditioning on the first return to level x in the set S_2. Part (b) is proved similarly by conditioning on the first epoch when the phase process escapes from S_3. To prove (c), consider the first epoch of spatial uniformization and note that, with δ_{ij} denoting the Kronecker delta, we can write

$$\begin{aligned}[\hat{G}_{22}(s,x)]_{i,j} &= \delta_{ij}e^{-\lambda c_i \frac{x}{c_i}}e^{-s\frac{x}{c_i}} \\ &\quad + \int_0^{x/c_i} \lambda c_i e^{-\lambda c_i t}e^{-st}\sum_{k\in S}[P]_{i,k}[\hat{G}(s,x-c_it)]_{k,j}\,dt \\ &= \delta_{ij}e^{-(\lambda+\frac{s}{c_i})x} + \lambda\int_0^x e^{-(\lambda+\frac{s}{c_i})(x-z)}\sum_{k\in S}[P]_{i,k}[\hat{G}(s,z)]_{k,j}\,dz.\end{aligned}$$

Multiplying with $e^{(\lambda+\frac{s}{c_i})x}$ and differentiating with respect to x, we obtain

$$(\lambda + \frac{s}{c_i})[\hat{G}_{22}(s,x)]_{i,j} + [\frac{\partial}{\partial x}\hat{G}_{22}(s,x)]_{i,j} = \lambda\sum_{k\in S}[P]_{i,k}[\hat{G}(s,x)]_{k,j}$$

which, after writing in matrix form, yields due to (a) and (b) the differential equation $\frac{\partial}{\partial x}\hat{G}_{22}(s,x) = H(s)\hat{G}_{22}(s,x)$, with the initial condition $\hat{G}_{22}(s,0) = I$. Therefore, $\hat{G}_{22}(s,x) = e^{H(s)x}$, and the proof is complete.

4. Random Initial Fluid

We wish to appeal to matrix-geometric results which are essentially developed for queues. So, a tool we shall employ is to view the fluid model as derived from the work process of a suitably defined queue. To relate quantities of interest to the busy period of such a queue, it helps to consider a busy period started with an amount of fluid X that is exponentially distributed with mean λ^{-1}; in the context of the queue, X will be the amount of work brought in by the customer starting a busy period of the queue.

We thus consider the transform matrix $\tilde{\hat{G}}(s,\lambda)$ defined by the elements

$$[\tilde{\hat{G}}(s,\lambda)]_{i,j} \equiv \mathcal{E}\left[\mathcal{E}_{(X,i)}\left[e^{-s\tau}\chi(J(\tau)=j)\right]\right], \quad (10)$$

where the outer expectation is with respect to X. Then we can easily obtain from Theorem 3 that

$$\tilde{\hat{G}}_{22}(s,\lambda) = \int_0^\infty \lambda e^{-\lambda y}e^{H(s)y}\,dy = \lambda\{\lambda I - H(s)\}^{-1}, \quad (11)$$

$$\tilde{\hat{G}}_{12}(s,\lambda) = \Psi(s)\tilde{\hat{G}}_{22}(s,\lambda), \tag{12}$$

and

$$\tilde{\hat{G}}_{32}(s,\lambda) = (sI - Q_{33})^{-1}Q_{31}\tilde{\hat{G}}_{12}(s,\lambda) + (sI - Q_{33})^{-1}Q_{32}\tilde{\hat{G}}_{22}(s,\lambda). \tag{13}$$

The following result expresses $\Psi(s)$ in terms of the sub-matrices of $\tilde{\hat{G}}(s,\lambda)$.

Theorem 4.

$$\Psi(s) = (P_{11} - \frac{s}{\lambda}C_1^{-1})\tilde{\hat{G}}_{12}(s,\lambda) + P_{12}\tilde{\hat{G}}_{22}(s,\lambda) + P_{23}\tilde{\hat{G}}_{32}(s,\lambda). \tag{14}$$

Proof. If we consider the first epoch of spatial uniformization of the underlying Markov process, then for $i \in S_1, j \in S_2$,

$$\begin{aligned} [\Psi(s)]_{i,j} &= \int_0^\infty \lambda c_i e^{-\lambda c_i t} e^{-st} \sum_{k\in S} [P]_{i,k} \, [\hat{G}(s, c_i t)]_{k,j} \, dt \\ &= \lambda \int_0^\infty e^{-(\lambda + \frac{s}{c_i})y} \sum_{k \in S} [P]_{i,k} \, [\hat{G}(s,y)]_{k,j} \, dy, \end{aligned}$$

and we can rewrite this in matrix form as

$$\Psi(s) = \lambda \int_0^\infty e^{-(\lambda I + sC_1^{-1})y}[P_{11}\hat{G}_{12}(s,y) + P_{12}\hat{G}_{22}(s,y) + P_{13}\hat{G}_{32}(s,y)] \, dy.$$

Using the equation (13) and Theorem 3, we can rewrite $\Psi(s)$ as

$$\Psi(s) = \lambda \int_0^\infty e^{-sC_1^{-1}y} L(s) e^{-(\lambda I - H(s))y} \, dy,$$

where

$$L(s) = P_{11}\Psi(s) + P_{12} + P_{13}(sI - Q_{33})^{-1}Q_{31}\Psi(s) + P_{13}(sI - Q_{33})^{-1}Q_{32}.$$

Using integration by parts, we get

$$\Psi(s) = L(s)\lambda(\lambda I - H(s))^{-1} - \frac{s}{\lambda}C_1^{-1}\Psi(s)\lambda(\lambda I - H(s))^{-1}.$$

Substituting for $L(s)$ and using equations (11), (12) and (13) immediately yields equation (14).

The following is a key theorem for developing an algorithm for computing $\Psi(s)$ via $\tilde{\hat{G}}(s,\lambda)$.

Theorem 5. *The matrices $\tilde{\hat{G}}_{12}(s,\lambda)$, $\tilde{\hat{G}}_{22}(s,\lambda)$, $\tilde{\hat{G}}_{32}(s,\lambda)$ satisfy the following equations.*

$$\tilde{\hat{G}}_{12}(s,\lambda) = [P_{11} - \frac{s}{\lambda}C_1^{-1}]\tilde{\hat{G}}_{12}(s,\lambda)\tilde{\hat{G}}_{22}(s,\lambda) + P_{12}\tilde{\hat{G}}_{22}^{2}(s,\lambda) + P_{13}\tilde{\hat{G}}_{32}(s,\lambda)\tilde{\hat{G}}_{22}(s,\lambda),$$

$$\tilde{\hat{G}}_{22}(s,\lambda) = \lambda C_2(sI + 2\lambda C_2)^{-1}\left[I + P_{21}\tilde{\hat{G}}_{12}(s,\lambda) + P_{22}\tilde{\hat{G}}_{22}(s,\lambda) + P_{23}\tilde{\hat{G}}_{32}(s,\lambda)\right],$$

$$\tilde{\hat{G}}_{32}(s,\lambda) = \frac{\lambda}{s+\lambda}P_{31}\tilde{\hat{G}}_{12}(s,\lambda) + \frac{\lambda}{s+\lambda}P_{32}\tilde{\hat{G}}_{22}(s,\lambda) + \frac{\lambda}{s+\lambda}P_{33}\tilde{\hat{G}}_{32}(s,\lambda).$$

Proof. The first equation is clear from equations (12) and (14). The third equation can be obtained easily by conditioning on the first epoch of spatial uniformization. Thus, we need to prove only the second equation.

By considering the first epoch of spatial uniformization, we can write

$$[\hat{G}_{22}(s,x)]_{i,j} = e^{-\lambda c_i \frac{x}{c_i}} e^{-s\frac{x}{c_i}} + \int_0^{x/c_i} \lambda c_i e^{-\lambda c_i t} e^{-st} \sum_{k\in S} [P]_{i,k}\, [\hat{G}(s, x - c_i t)]_{k,j}\, dt.$$

Multiplying this by $\lambda e^{-\lambda x}$ and integrating over x, it is easy now to get the formula

$$[\tilde{\hat{G}}_{22}(s,\lambda)]_{i,j} = \lambda \left(2\lambda + \frac{s}{c_i}\right)^{-1} + \lambda \left(2\lambda + \frac{s}{c_i}\right)^{-1} \sum_{k\in S} [P]_{i,k}\, [\tilde{\hat{G}}(s,\lambda)]_{k,j}.$$

This written in matrix form yields the required formula for $\tilde{\hat{G}}_{22}(s,\lambda)$.

Now, if we define the matrices

$$A_2(s,\lambda) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \lambda C_2(sI + 2\lambda C_2)^{-1} & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{15}$$

$$A_1(s,\lambda) = \Lambda C(sI + \Lambda C)^{-1} \begin{pmatrix} 0 & 0 & 0 \\ \frac{1}{2}P_{21} & \frac{1}{2}P_{22} & \frac{1}{2}P_{23} \\ P_{31} & P_{32} & P_{33} \end{pmatrix}, \tag{16}$$

$$A_0(s,\lambda) = \begin{pmatrix} P_{11} - \frac{s}{\lambda}C_1^{-1} & P_{12} & P_{13} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{17}$$

where $\Lambda = diag(\lambda I, 2\lambda I, \lambda I)$, then the set of equations given by Theorem 5 can be written simply as

$$\tilde{\hat{G}}(s,\lambda) = A_2(s,\lambda) + A_1(s,\lambda)\tilde{\hat{G}}(s,\lambda) + A_0(s,\lambda)[\tilde{\hat{G}}(s,\lambda)]^2, \tag{18}$$

reminiscent of the equation for the G-matrix of a QBD; see [12], [13]. That then also yields the following corollary.

Corollary 1. *If we define*

$$\tilde{\hat{U}}(s,\lambda) = A_1(s,\lambda) + A_0(s,\lambda)\tilde{\hat{G}}(s,\lambda), \tag{19}$$

then for $Re(s) > 0$, we have

$$\tilde{\hat{G}}(s,\lambda) \quad = \quad (I - \tilde{\hat{U}}(s,\lambda))^{-1} A_2(s,\lambda). \tag{20}$$

Proof. We note first of all that the blocks $\tilde{\hat{G}}_{ij}(s,\lambda)$ of the matrix $\tilde{\hat{G}}(s,\lambda)$ are zero except for $j = 2$; this is so because fluid gets depleted only when the phase is in S_2

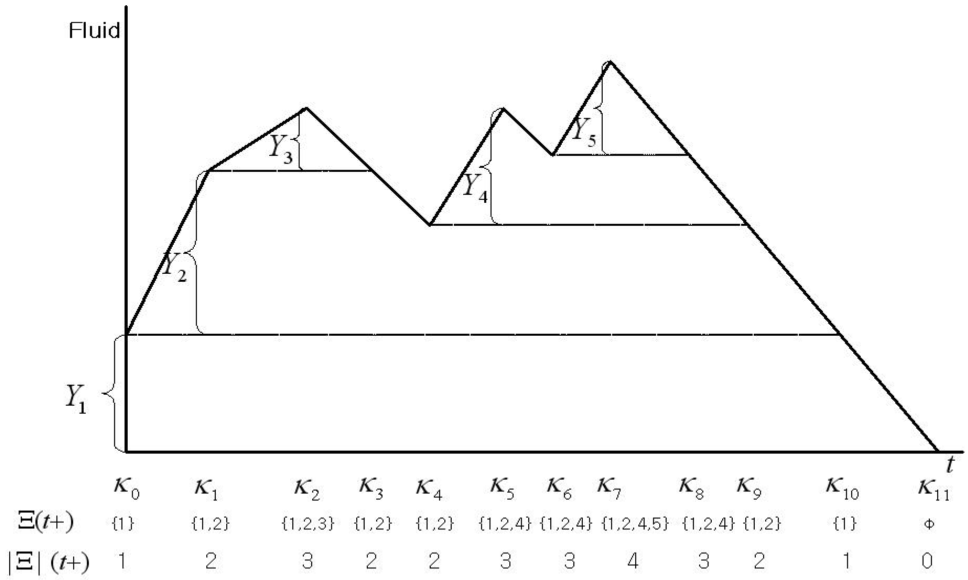
Fluid
Y_1
Y_2
Y_3
Y_4
Y_5
t
κ_0 κ_1 κ_2 κ_3 κ_4 κ_5 κ_6 κ_7 κ_8 κ_9 κ_{10} κ_{11}
$\Xi(t+)$ {1} {1,2} {1,2,3} {1,2} {1,2} {1,2,4} {1,2,4} {1,2,4,5} {1,2,4} {1,2} {1} Φ
$|\Xi|(t+)$ 1 2 3 2 2 3 3 4 3 2 1 0

FIGURE 1: Illustration

and therefore the busy period must also end in S_2. If we now let $g_2(s,\lambda) = 2\lambda C_2(sI + 2\lambda C_2)^{-1}$ and $g_3(s,\lambda) = \lambda(s+\lambda)^{-1}$, then it follows from evaluating the right side of (19) in partitioned form using Theorem 5 and (14)-(17) that

$$\tilde{\tilde{U}}(s,\lambda) = \begin{pmatrix} 0 & \Psi(s) & 0 \\ g_2(s,\lambda)\frac{1}{2}P_{21} & g_2(s,\lambda)\frac{1}{2}P_{22} & g_2(s,\lambda)\frac{1}{2}P_{23} \\ g_3(s,\lambda)P_{31} & g_3(s,\lambda)P_{32} & g_3(s,\lambda)P_{33} \end{pmatrix}. \tag{21}$$

This shows that the matrix $\tilde{\tilde{U}}(s,\lambda)$ is a matrix of LSTs which is strictly substochastic for all $s > 0$ and has all eigenvalues less than 1 in absolute value. For s complex with $Re(s) > 0$, the matrix of absolute values of this transform matrix is bounded above by a strictly substochastic matrix, and therefore all its eigenvalues are also less than 1 in absolute value; see [14], Paragraph 2.4.9. Thus, the inverse in (20) exists, and we can get that equation from (18) using (19).

Note that equations (19), (20) are similar to those obtained by Latouche [10] (see also [12], Chapter 8.1, 8.2) for the G-matrix of the QBD and suggest an iterative procedure. Later in the paper, we shall consider such an iterative scheme (Algorithm 2) and demonstrate that it converges to the required matrices.

5. An algorithm for $\tilde{\tilde{G}}(s,\lambda)$ through a queue

5.1. Notations and definitions

We assume in what follows that the MMFF process $(\mathcal{F},\mathcal{J}) = \{(F(t),J(t)), t \geq 0\}$ operates under a LIFO scheme; that is, the most recently arrived fluid is purged first. This does not affect the distribution of the busy period, the quantity of interest to us.

Our approach rests on constructing a closely related queue whose work process yields the fluid flow.

To facilitate the discussion, we introduce some notations and definitions below. It is helpful to consider the underlying ideas with the illustration shown in Figure 1 which depicts a path of the MMFF process; for simplicity, we have in that illustration (alone) assumed that S_3 is empty. Note that if we replace each of the upward linear segments of the path of the MMFF during the sojourn in an exponentially distributed interval resulting in the spatial uniformization by a jump that occurs at the termination of that segment, we could interpret the resulting path as that of the work in a LIFO queue. In that queue, at the end of each sojourn of the phase process in states of S_1, a new customer with an exponentially distributed amount of work with mean λ^{-1} joins the queue, and work gets depleted only while the phase process is in S_2 and at rate c_j per unit time when in state $j \in S_2$. The queue is quite complex in that the amount of work brought by a customer is directly proportional to the interval of sojourn in S_1 whose end marks the arrival epoch of that customer. What is, however, noteworthy is that with this correspondence between the MMFF and the queue, the busy period of the MMFF process is identical to the corresponding busy period of the queue. When such a busy period of the MMFF starts with an exponentially distributed amount of fluid, the distribution of the busy period of the MMFF becomes identical to the busy period of the queue initiated by an arrival to an empty system. Many of the terms we introduce will be interpreted in terms of that queue also, and this is what would allow us to draw freely from the matrix-geometric literature.

Here are some of the notations to be used by us.

(1) Let $\{0 = \tau_0 < \tau_1 < \cdots\}$ denote the set of successive spatial uniformization epochs of $\mathcal{J}$.

(2) Let Y_1 denote the initial amount of fluid at time 0, which we assume to be exponentially distributed with mean $1/\lambda$. Let $Y_{n+1},\ n \geq 1$ denote the amount of fluid incoming during the n-th sojourn in S_1 of the (spatially uniformized) phase process. Note that, by the spatial uniformization, $Y_n, n \geq 2$ are all also exponentially distributed with mean $1/\lambda$.

(3) Let $n \geq 1$ and assume $\tau_{r(n)}$ is the epoch of the n-th visit to S_1. (If $J(0) \in S_1$, then we treat 0 as the epoch of the first visit to S_1.) Now, for $t \geq \tau_{r(n)+1}$, define $Y_{n+1}(t)$ to be the remaining amount of fluid out of Y_{n+1} which remains in the system at time t. Note that $Y_{n+1}(\tau_{r(n)+1}) = Y_{n+1}$. Similarly, let us denote by $Y_1(t)$ the remaining amount of fluid out of Y_1 which is not depleted by time $t > 0$. The epoch when a $Y_n(t)$ attains the value 0, is clearly a departure epoch for the LIFO queue; in the illustration in Figure 1, such epochs correspond to the right side terminal points where the lines drawn parallel to the t-axis meet the path of the MMFF.

(4) Now, consider the spatial uniformization epochs $\{\tau_n\}$ along with the departure epochs identified above, and denote the resulting set of ordered epochs by κ_n; we have,

$$0 = \kappa_0 < \kappa_1 < \kappa_2 < \cdots, \quad a.s.$$

(5) Let $\Xi(t)$ denote the set of indices n's such that $Y_n(t) > 0, n = 1, 2, \cdots$. We define

$|\Xi|(t)$ as the total number of elements in $\Xi(t)$, and we represent the set $\Xi(t)$ as

$$\Xi(t) = \{n_1(t), \cdots, n_{|\Xi|(t)}(t)\}$$

where $n_j(t)$ denotes the j-th largest index among the indices in the set $\Xi(t)$. Defined thus, the set $\Xi(t)$ gives the identities of the customers still present in the queue and $|\Xi|(t)$ the total number of customers present at time t. In the illustration of Figure 1, we have shown the epochs κ_n in a busy period along with the corresponding sets Ξ and their cardinalities at those epochs. Note that a.s., $|\Xi|(t) = 0$ iff $F(t) = 0$.

Armed with these notations, we are now ready to state the following important result which is a direct consequence of the memoryless property of the exponential distribution and can be established by mathematical induction; we omit the proof.

Theorem 6. *Given $|\Xi|(\kappa_n) = m$, the fluid level $F(\kappa_n)$ (or equivalently, the total amount of work in the queue at κ_n) is distributed as the sum of m independent, identically distributed (*iid*) random variables with common distribution $exp(\lambda)$. Furthermore, the residual amounts of work for the m customers in the queue at κ_n are* iid *with distribution $exp(\lambda)$.*

Remark 1. The above result shows that if the fluid level (work in the sytem) at κ_0 is distributed as $exp(\lambda)$, then knowing the number of customers present at κ_n determines the law of the MMFF process in the interval $[\kappa_n,\ \infty)$. In fact, if N_n denotes the level at κ_n+, then (κ_n, N_n), $n \geq 0$ form a semi-regenerative sequence ([6], Chapter 10.6) for the MMFF process as well as for the work in the queue.

For the queue, we now introduce a set of random variables similar to those introduced by Latouche [10] in the computation of G in a QBD. Given $|\Xi|(\kappa_m) = n \geq 1$, let κ_U denote the first epoch in $\{\kappa_j\ :\ j \geq m+1\}$ for which $|\Xi|(\kappa_j) = n$. Also, let κ_G denote the first epoch in $\{\kappa_j\ :\ j \geq m+1\}$ for which $|\Xi|(\kappa_j) = n-1$. Thus, if by "level" n, we denote the set of states with queue length n, then $\kappa_U - \kappa_m$ is a return time to level n avoiding lower levels, and $\kappa_G - \kappa_m$ is the first passage time to the immediately lower level $n-1$ given that one starts in level n. Finally, we denote the amount of remaining work of the customer in service at κ_m by X_m.

Now, let $\hat{U}(k, s, x)$ denote the transform matrix such that

$$\begin{aligned}[\hat{U}(k,s,x)]_{i,j} \quad = & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (22)\\ & \mathcal{E}(e^{-s(\kappa_U-\kappa_m)}\chi\{J(\kappa_U)=j,\ n \leq |\Xi|(t) < n+k\ \forall\ t \in [\kappa_m, \kappa_U)\} \\ & \qquad\qquad \mid\ |\Xi|(\kappa_m)=n,\ J(\kappa_m)=i,\ X_m=x).\end{aligned}$$

Also, let

$$\begin{aligned}[\hat{G}(k,s,x)]_{i,j} \quad = & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (23)\\ & \mathcal{E}(e^{-s(\kappa_G-\kappa_m)}\chi\{J(\kappa_G)=j,\ n \leq |\Xi|(t) < n+k\ \forall\ t \in [\kappa_m, \kappa_G)\} \\ & \qquad\qquad \mid\ |\Xi|(\kappa_m)=n,\ J(\kappa_m)=i,\ X_m=x).\end{aligned}$$

From the structure of the process under consideration, it is clear that the above transform matrices do not depend on n. Note that these matrices, of course, depend on the uniformization parameter λ, but we have suppressed that fact to simplify notations.

We also define the matrices $\tilde{\hat{U}}(k,s,\lambda)$ and $\tilde{\hat{G}}(k,s,\lambda)$ as

$$\tilde{\hat{U}}(k,s,\lambda) = \int_0^\infty \lambda e^{-\lambda x}\hat{U}(k,s,x)\,dx, \text{ and } \tilde{\hat{G}}(k,s,\lambda) = \int_0^\infty \lambda e^{-\lambda x}\hat{G}(k,s,x)dx. \tag{24}$$

From the definitions, we can get the following results.

Lemma 1.
(a) If we let $g_2(s,\lambda) = 2\lambda C_2(sI + 2\lambda C_2)^{-1}$ and $g_3(s,\lambda) = \lambda(s+\lambda)^{-1}$, then $\tilde{\hat{U}}(k,s,\lambda)$ has the following form.

$$\tilde{\hat{U}}(k,s,\lambda) = \begin{pmatrix} 0 & \tilde{\hat{U}}_{12}(k,s,\lambda) & 0 \\ g_2(s,\lambda)\frac{1}{2}P_{21} & g_2(s,\lambda)\frac{1}{2}P_{22} & g_2(s,\lambda)\frac{1}{2}P_{23} \\ g_3(s,\lambda)P_{31} & g_3(s,\lambda)P_{32} & g_3(s,\lambda)P_{33} \end{pmatrix}. \tag{25}$$

(b) For $Re(s) > 0$, $\tilde{\hat{U}}_{12}(k,s)$ converges as k increases, and

$$\lim_{k\to\infty} \tilde{\hat{U}}_{12}(k,s,\lambda) = \Psi(s). \tag{26}$$

Furthermore, for $s \geq 0$, the matrices $\tilde{\hat{U}}_{12}(k,s,\lambda)$ are (entry-wise) monotonically non-decreasing. That is,

$$\tilde{\hat{U}}(k,s,\lambda) \to \tilde{\hat{U}}(s,\lambda) \text{ as } k \to \infty, \tag{27}$$

and the convergence is monotone for $s \geq 0$.

Proof. In Part (a), the submatrices in the second and third rows are obtained by noting that the first return to a given level avoiding lower levels occurs at the first step of the spatial uniformization iff the time to that step is less than the amount of time to serve the customer in service at time 0. The zero elements of the first row follow from the fact that when a return to a given level occurs from S_1, the phase visited must be in S_2. All the other results follow immediately upon noting that the set of paths that go up to level $n+k-1$ in a return to level n avoiding lower levels form a non-decreasing set converging to the set of all paths returning to level n avoiding lower levels.

Lemma 2. *Let $Re(s) > 0$,*
(a) $\hat{G}_{12}(k,s,\lambda) = \tilde{\hat{U}}_{12}(k,s,\lambda)\hat{G}_{22}(k,s,\lambda)$.
(b) $\hat{G}_{32}(k,s,\lambda) = (sI - Q_{33})^{-1}Q_{31}\hat{G}_{12}(k,s,\lambda) + (sI - Q_{33})^{-1}Q_{32}\hat{G}_{22}(k,s,\lambda)$.
(c) If we let

$$H(k,s) = C_2^{-1}[Q_{22} - sI + Q_{23}(sI - Q_{33})^{-1}Q_{32} + \{Q_{21} + Q_{23}(sI - Q_{33})^{-1}Q_{31}\}\tilde{\hat{U}}_{12}(k,s,\lambda)],$$

then

$$\hat{G}_{22}(k,s,x) = e^{H(k,s)x}. \tag{28}$$

(d) $\tilde{\hat{G}}_{22}(k,s,\lambda) = \lambda(\lambda I - H(k,s))^{-1}$.
(e) For $Re(s) > 0$, $H(k,s) \to H(s)$ and $\tilde{\hat{G}}(k,s,\lambda) \to \tilde{\hat{G}}(s,\lambda)$ as $k \uparrow \infty$. Furthermore, for $s \geq 0$, $\tilde{\hat{G}}(k,s,\lambda) \uparrow \tilde{\hat{G}}(s,\lambda)$ as $k \uparrow \infty$.

Proof. Part (a) follows by conditioning on the first epoch of return to level 1 before the end of the busy period. Part (b) follows by conditioning on the first exit time from S_3. Part (c) is proven along the same lines as Theorem 3c, and (c) immediately yields (d). The proof of (e) follows by the simple observation that as $k \to \infty$, the set of paths yielding a first passage from level 1 to level 0 avoiding level $n+k$ form an increasing set converging to the set of all paths yielding a first passage from level 1 to level 0.

5.2. Relation between $\tilde{\hat{U}}$ and $\tilde{\hat{G}}$

For the analysis, we introduce the operator [8] *vec* defined on matrices $A = (a_{ij})$ of order $m \times n$ by

$$vec(A) = (a_{11} \ \cdots \ a_{m1} \ \cdots \ a_{1n} \ \cdots \ a_{mn})^t \tag{29}$$

where $(\cdot)^t$ denotes the transpose operator. Then, the operator *vec* can be seen to satisfy the following lemma.

Lemma 3. *Given $m \times m$ matrix A, $n \times n$ matrix B and $m \times n$ matrix Y, then*

$$vec(AYB) = (B^t \otimes A) vec(Y) \tag{30}$$

where $\otimes$ denotes the Kronecker Product of matrices.

Now, we can can get the following result establishing a relationship $\tilde{\hat{U}}$ and $\tilde{\hat{G}}$.

Lemma 4.
(a) $\tilde{\hat{G}}(k,s,\lambda) = (I - \tilde{\hat{U}}(k,s,\lambda))^{-1} A_2(s,\lambda)$.
(b) The submatrices $\tilde{\hat{U}}_{12}(k,s,\lambda)$ in (25) are such that $\tilde{\hat{U}}_{12}(1,s,\lambda) = 0$, and for $k \geq 2$,

$$\begin{aligned}\tilde{\hat{U}}_{12}(k,s,\lambda) \quad = \quad & P_{11}\tilde{\hat{G}}_{12}(k-1,s,\lambda) + P_{12}\tilde{\hat{G}}_{22}(k-1,s,\lambda) \\ & + P_{13}\tilde{\hat{G}}_{32}(k-1,s,\lambda) - \frac{s}{\lambda} C_1^{-1} \tilde{\hat{U}}_{12}(k,s,\lambda)\tilde{\hat{G}}_{22}(k-1,s,\lambda).\end{aligned} \tag{31}$$

(c)

$$\begin{aligned}vec(\tilde{\hat{U}}_{12}(k,s,\lambda)) = \big(I + \frac{s}{\lambda}\tilde{\hat{G}}_{22}^t(k-1,s,\lambda) \otimes C_1^{-1}\big)^{-1} \times \\ vec\big(P_{11}\tilde{\hat{G}}_{12}(k-1,s,\lambda) + P_{12}\tilde{\hat{G}}_{22}(k-1,s,\lambda) + P_{13}\tilde{\hat{G}}_{32}(k-1,s,\lambda)\big).\end{aligned} \tag{32}$$

Proof. Recall the definition of κ_U and κ_G in Section 5.1. Given the state $J(\kappa_U)$, it is clear that κ_G and $\kappa_G - \kappa_U$ are independent, and the distribution of $\kappa_G - \kappa_U$ given $J(\kappa_U) = j$ is identical to the distribution of κ_G given $J(0) = j$. Therefore,

$$\tilde{\hat{G}}(k,s,\lambda) = A_2(s,\lambda) + \tilde{\hat{U}}(k,s,\lambda)\tilde{\hat{G}}(k,s,\lambda),$$

which completes the proof of (a). Part (c) is a direct consequence of (b) and Lemma 3, and we only need to prove (b). Now, because of Lemma 2(b), we can, without loss of generality, assume $S = S_1 \cup S_2$ for the proof. By its definition, it is trivial that $\tilde{\hat{U}}_{12}(1,s,\lambda) = 0$, for, given an initial state in S_1, the level increases in the very first step of the spatial uniformization. Thus, (b) holds for $k = 1$. Assume as induction hypothesis that it holds for $k-1$, for some $k \geq 2$. Now, consider k, and assume that $i \in S_1$ and $j \in S_2$. Then

$$\begin{aligned}
[\tilde{\hat{U}}_{12}(k,s,\lambda)]_{i,j} &= \int_0^\infty \lambda c_i e^{-\lambda c_i t} e^{-st} \sum_{l \in S} [P]_{i,l}\, [\hat{G}]_{l,j}(k-1,s,c_i t)\, dt \\
&= \int_0^\infty \lambda e^{-\lambda y} e^{-\frac{s}{c_i} y} \sum_{l \in S} [P]_{i,l}\, [\hat{G}]_{l,j}(k-1,s,y)\, dy,
\end{aligned}$$

and it follows from the induction assumption that

$$\begin{aligned}
\tilde{\hat{U}}_{12}(k,s,\lambda) &= \lambda \int_0^\infty e^{-\lambda y} e^{-sC_1^{-1}y} P_{11} \hat{G}_{12}(k-1,s,y)\, dy \\
&\quad + \lambda \int_0^\infty e^{-\lambda y} e^{-sC_1^{-1}y} P_{12} \hat{G}_{22}(k-1,s,y)\, dy \\
&= \lambda \int_0^\infty e^{-sC_1^{-1}y} [P_{11} \tilde{\hat{U}}_{12}(k-1,s,\lambda) + P_{12}] e^{-(\lambda I - H(k-1,s))y}\, dy.
\end{aligned}$$

Using integration by parts in the expression above and the results in Lemma 2, we can now get

$$\tilde{\hat{U}}_{12}(k,s,\lambda) = P_{11} \tilde{\hat{G}}_{12}(k-1,s,\lambda) + P_{12} \tilde{\hat{G}}_{22}(k-1,s,\lambda) - \frac{s}{\lambda} C_1^{-1} \tilde{\hat{U}}_{12}(k,s,\lambda) \tilde{\hat{G}}_{22}(k-1,s,\lambda),$$

and the proof is complete by mathematical induction.

From Lemma 4, we can now construct the following iterative scheme.

5.3. Algorithm 1

```
Let $Re(s) > 0$.

Fix $\epsilon > 0$. Let $k = 1$ and diff $= 100$.
Determine $\lambda > 0$ such that $\lambda \geq \max_{i \in S} \left\{ -[C^{-1}Q]_{ii} \right\}$. Initialize as
$\tilde{\hat{U}}(1,s,\lambda) = A_1(s,\lambda)$ and $\tilde{\hat{G}}(1,s,\lambda) = [I - \tilde{\hat{U}}(1,s,\lambda)]^{-1} A_2(s,\lambda)$.

Do while ( diff $> \epsilon$ )
     $k = k + 1$;
     $\tilde{\hat{U}}(k,s,\lambda) = A_1(s,\lambda)$;
     $M = \left(I + \frac{s}{\lambda} \tilde{\hat{G}}^t_{22}(k,s,\lambda) \otimes C_1^{-1}\right)^{-1}$;
     $N = vec\big(P_{11} \tilde{\hat{G}}_{12}(k,s,\lambda) + P_{12} \tilde{\hat{G}}_{22}(k,s,\lambda) + P_{13} \tilde{\hat{G}}_{32}(k,s,\lambda)\big)$;
    $vec((\tilde{\hat{U}})_{12}(k,s,\lambda) = MN$
     $\tilde{\hat{G}}(k,s,\lambda) = (I - \tilde{\hat{U}}(k,s,\lambda))^{-1} A_2(s,\lambda)$;
     diff $= \max_{i,j \in S} \mid [\tilde{\hat{G}}(k,s,\lambda)]_{i,j} - [\tilde{\hat{G}}(k-1,s,\lambda)]_{i,j} \mid$;
end

$\Psi(s) \approxeq \tilde{\hat{U}}_{12}(k,s,\lambda)$;  $\tilde{\hat{G}}(s,\lambda) \approxeq \tilde{\hat{G}}(k,s,\lambda)$.
```

Due to Lemma 4, the iterates in Algorithm 1 are such that the k-th iterates $\tilde{\tilde{U}}(k, s, \lambda)$ and $\tilde{\tilde{G}}(k, s, \lambda)$ are respectively the quantities in (24) defined using the taboo paths of the MMFF, and therefore converge as $k \uparrow \infty$ to the required matrices $\Psi(s)$ and $\tilde{\tilde{G}}(s, \lambda)$ as shown in Lemmas 1 and 2. We have shown already that for $s \geq 0$, the convergence is (entry-wise) monotonic. Furthermore, the convergence is linear since each additional iteration obtains paths that go up by one more level during a busy period; this is similar to the linear algorithm of Latouche [10]. Indeed, by the results in [12], Chapter 8, the difference between the limit values and the k-th iterates are asymptotically $O([\eta(s)]^k)$ as $k \to \infty$, where $0 < \eta(s) < 1$ is the minimal solution in $(0, 1)$ of the equation $\eta(s) = sp\left(A_0(s) + \eta(s)A_1(s) + \{\eta(s)\}^2 A_2(s)\right)$, where $sp(A)$ denotes the spectral radius of the matrix A.

Remark 2. Although, in principle, the iterates of Algorithm 1 converge as required, when implemented on a computer (a finite arithmetic machine), we have found it to misbehave due to round offs and truncations. A thorough numerical analysis of the iterative schemes given in this paper has not been made. However, it is easy to show that if we choose λ using a more stringent criterion, viz., that in addition to the condition $\lambda \geq \max_{i \in S}\left\{-[C^{-1}Q]_{ii}\right\}$ of spatial uniformization, if we also require that

$$\max_{i \in S}\left[\frac{Re(s)}{\lambda}C^{-1}\right]_{i,i} \leq \delta < 1, \text{ and } 0 < \max_{i \in S}[P_\lambda - \frac{Re(s)}{\lambda}C^{-1}]_{ii},$$

then the iterates remain within a bounded region of the complex plane and behave well. This becomes obvious from the easily verified fact that under these conditions, the matrices $A_i(Re(s), \lambda)$, $i = 0, 1, 2$ are nonnegative and strictly substochastic and sum to a strictly substochastic matrix. Thus, each of the matrices $\tilde{\tilde{U}}(k, Re(s), \lambda)$ and $\tilde{\tilde{G}}(k, Re(s), \lambda)$ remains strictly substochastic. We therefore recommend implementing all the algorithms in this paper using this more stringent scheme so that numerical stability is maintained in the presence of round offs and truncations.

Henceforth, we will assume that for each s, an appropriate $\lambda(s)$ meeting the stringent criteria established above is being used. However, to simplify the notations, we shall simply write λ suppressing the dependence of λ on s.

6. Quadratically Convergent Algorithm

An iterative procedure is said to have linear convergence if the error in the k-th iterate is of order $O(\eta^k)$ as $k \to \infty$ and to have quadratic convergence if that error is of the order $O(\eta^{2^k})$ for some $0 < \eta < 1$. Having obtained a linear algorithm for $\Psi(s)$, we now examine an algorithm resulting from Corollary 1 and then an accelerated version thereof. The accelerated version will be shown to have quadratic convergence.

The following is an iterative scheme obtained by bootstrapping in the equations of Corollary 1.

6.1. Linear Algorithm

Algorithm 2

Fix $\epsilon > 0$ and set diff $= 100$;
Initialize $U^*(1, s, \lambda) = A_1(s, \lambda), G^*(1, s, \lambda) = (I - U^*(1, s, \lambda))^{-1} A_2(s, \lambda)$;

```
Do while ( diff $> \epsilon$ )
    $k = k + 1$;
    $U^*(k,s,\lambda) = A_1(s,\lambda) + A_0(s,\lambda)G^*(k-1,s,\lambda)$;
    $G^*(k,s,\lambda) = (I - U^*(k,s,\lambda))^{-1}A_2(s,\lambda)$;
    diff $= \max_{i,j \in S} \mid [G^*(k,s,\lambda)]_{i,j} - [G^*(k-1,s,\lambda)]_{i,j} \mid$;
end
```

$\Psi(s) \approxeq U^*_{12}(k,s,\lambda)$; $\tilde{\tilde{G}}(s,\lambda) \approxeq G^*(k,s,\lambda)$.

Comparing the matrices $U^*(k,s,\lambda)$ and $G^*(k,s,\lambda)$ of Algorithm 2 respectively with $\tilde{\tilde{U}}(k,s,\lambda)$, $\tilde{\tilde{G}}(k,s,\lambda)$ of Algorithm 1, we can see that (a) $U^*(1,s,\lambda) = \tilde{\tilde{U}}(1,s,\lambda)$ and $G^*(1,s,\lambda) = \tilde{\tilde{G}}(1,s,\lambda)$; (b) $U^*_{lm}(k,s,\lambda) = \tilde{\tilde{U}}_{lm}(k,s,\lambda)$ for $l = 2,3$, $m = 1,2,3$ and $k = 1,2,\cdots$; (b) Furthermore, for all $k = 1,2,\cdots$,

$$G^*(k,s,\lambda) = (I - U^*(k,s,\lambda))^{-1}A_2(s,\lambda), \quad \tilde{\tilde{G}}(k,s,\lambda) = (I - \tilde{\tilde{U}}(k,s,\lambda))^{-1}A_2(s,\lambda). \tag{33}$$

Thus, the difference in the two algorithms arises from the difference in the iterates $U^*_{12}(k,s,\lambda)$ and $\tilde{\tilde{U}}_{12}(k,s,\lambda)$. As we can see in Lemma 4(b), $\tilde{\tilde{U}}_{12}(k,s,\lambda)$ satisfies

$$\begin{aligned} \tilde{\tilde{U}}_{12}(k,s,\lambda) &= P_{11}\tilde{\tilde{G}}_{12}(k-1,s,\lambda) + P_{12}\tilde{\tilde{G}}_{22}(k-1,s,\lambda) + P_{13}\tilde{\tilde{G}}_{32}(k-1,s,\lambda) \\ &\quad - \frac{s}{\lambda}C_1^{-1}\tilde{\tilde{U}}_{12}(k,s,\lambda)\tilde{\tilde{G}}_{22}(k-1,s,\lambda). \end{aligned} \tag{34}$$

but, $U^*_{12}(k,s,\lambda)$ in Algorithm 2 satisfies

$$\begin{aligned} & U^*_{12}(k,s,\lambda) \\ &= \left(P_{11} - \frac{s}{\lambda}C_1^{-1}\right) G^*_{12}(k-1,s,\lambda) + P_{12}G^*_{22}(k-1,s,\lambda) + P_{13}G^*_{32}(k-1,s,\lambda) \\ &= P_{11}G^*_{12}(k-1,s,\lambda) + P_{12}G^*_{22}(k-1,s,\lambda) + P_{13}G^*_{32}(k-1,s,\lambda) \\ &\quad - \frac{s}{\lambda}C_1^{-1}U^*_{12}(k-1,s,\lambda)G^*_{22}(k-1,s,\lambda). \end{aligned} \tag{35}$$

We draw particular attention to the indices of the U-matrices arising in (34) and (35). Note that the first one yields a linear equation for the unknown that needs to be solved, while the second one is a true recursion. A result we will establish soon is that despite these differences, the iterates in both algorithms converge to the same matrices.

We begin with the following result whose proof by mathematical induction is quite straightforward and therefore omitted.

Lemma 5.
(a) For $s > 0$, the matrices $U^*(k,s,\lambda)$ and $G^*(k,s,\lambda)$ are monotonely non-decreasing as k increases.
(b) For all $k \geq 1$, $s > 0$, the matrices $U^*(k,s,\lambda)$ and $G^*(k,s,\lambda)$ are nonnegative and strictly substochastic.

Let $U^*(s,\lambda) = \lim_{k\to\infty} U^*(k,s,\lambda)$ and $G^*(s,\lambda) = \lim_{k\to\infty} G^*(k,s,\lambda)$, for $Re(s) > 0$. Our next result shows that for $s > 0$, $U^*(s,\lambda) = \tilde{\tilde{U}}(s,\lambda)$ and $G^*(s,\lambda) = \tilde{\tilde{G}}(s,\lambda)$ so that the iterative schemes in Algorithm 1 and Algorithm 2 both yield the same results for $s > 0$.

Lemma 6. *Let $s > 0$.*
(a) For all $k \geq 1$, $U^*(k,s,\lambda) \geq \tilde{\tilde{U}}(k,s,\lambda)$ and $G^*(k,s,\lambda) \geq \tilde{\tilde{G}}(k,s,\lambda)$.
(b) $U^*_{12}(s,\lambda) = \Psi(s)$.
(c) $U^*(s,\lambda) = \tilde{\tilde{U}}(s,\lambda)$, and $G^*(s,\lambda) = \tilde{\tilde{G}}(s,\lambda)$

Proof. (a) We will prove this part by induction. Note that $U^*(1,s,\lambda) = \tilde{\tilde{U}}(1,s,\lambda)$ and $G^*(1,s,\lambda) = \tilde{\tilde{G}}(1,s,\lambda)$. From the equations (34) and (35), we can see that

$$\begin{aligned}\tilde{\tilde{U}}_{12}(k,s,\lambda) &= P_{11}\tilde{\tilde{G}}_{12}(k-1,s,\lambda) + P_{12}\tilde{\tilde{G}}_{22}(k-1,s,\lambda) + P_{13}\tilde{\tilde{G}}_{32}(k-1,s,\lambda)\\ &\quad - \frac{s}{\lambda}C_1^{-1}\tilde{\tilde{U}}_{12}(k-1,s,\lambda)\tilde{\tilde{G}}_{22}(k-1,s,\lambda)\\ &\quad - \frac{s}{\lambda}C_1^{-1}[\tilde{\tilde{U}}_{12}(k,s,\lambda) - \tilde{\tilde{U}}_{12}(k-1,s,\lambda)]\tilde{\tilde{G}}_{22}(k-1,s,\lambda).\end{aligned}$$

and

$$\begin{aligned}&U^*(2,s,\lambda) - \tilde{\tilde{U}}(2,s,\lambda)\\ &\quad= [A_1(s,\lambda) + A_0(s,\lambda)G^*(1,s,\lambda)] - [A_1(s,\lambda) + A_0(s,\lambda)\tilde{\tilde{G}}(1,s,\lambda)\\ &\qquad - \frac{s}{\lambda}[\tilde{\tilde{U}}(2,s,\lambda) - \tilde{\tilde{U}}(1,s,\lambda)]\tilde{\tilde{G}}(1,s,\lambda)]\\ &\quad= \frac{s}{\lambda}[\tilde{\tilde{U}}(2,s,\lambda) - \tilde{\tilde{U}}(1,s,\lambda)]\tilde{\tilde{G}}(1,s,\lambda) \geq 0.\end{aligned}$$

It then follows that

$$\begin{aligned}&G^*(2,s,\lambda) - \tilde{\tilde{G}}(2,s,\lambda)\\ &\quad= [I - U^*(2,s,\lambda)]^{-1}A_2(s,\lambda) - [I - \tilde{\tilde{U}}(2,s,\lambda)]^{-1}A_2(s,\lambda)\\ &\quad= \sum_{i=1}^{\infty}[U^*(2,s,\lambda)^i - \tilde{\tilde{U}}(2,s,\lambda)^i]A_2(s,\lambda) \geq 0,\end{aligned}$$

because $U^*(2,s,\lambda) \geq \tilde{\tilde{U}}(2,s,\lambda)$. For $k \geq 2$, if we assume $U^*(k,s,\lambda) - \tilde{\tilde{U}}(k,s,\lambda) \geq 0$ and $G^*(k,s,\lambda) - \tilde{\tilde{G}}(k,s,\lambda) \geq 0$, then

$$\begin{aligned}&U^*(k+1,s,\lambda) - \tilde{\tilde{U}}(k+1,s,\lambda)\\ &\quad= [A_1(s,\lambda) + A_0(s,\lambda)G^*(k,s,\lambda)] - [A_1(s,\lambda) + A_0(s,\lambda)\tilde{\tilde{G}}(k,s,\lambda)\\ &\qquad - \frac{s}{\lambda}[\tilde{\tilde{U}}(k+1,s,\lambda) - \tilde{\tilde{U}}(k,s,\lambda)]\tilde{\tilde{G}}(k,s,\lambda)]\\ &\quad= A_0(s)[G^*(k,s,\lambda) - \tilde{\tilde{G}}(k,s,\lambda)]\\ &\qquad + \frac{s}{\lambda}[\tilde{\tilde{U}}(k+1,s,\lambda) - \tilde{\tilde{U}}(k,s,\lambda)]\tilde{\tilde{G}}(k,s,\lambda) \geq 0,\end{aligned}$$

and

$$\begin{aligned}&G^*(k+1,s,\lambda) - \tilde{\tilde{G}}(k+1,s,\lambda)\\ &\quad= [I - U^*(k+1,s,\lambda)]^{-1}A_2(s,\lambda) - [I - \tilde{\tilde{U}}(k+1,s,\lambda)]^{-1}A_2(s,\lambda)\\ &\quad= \sum_{i=1}^{\infty}[U^*(k+1,s,\lambda)^i - \tilde{\tilde{U}}(k+1,s,\lambda)^i]A_2(s,\lambda) \geq 0.\end{aligned}$$

This completes the induction proof of Part (a). Note that (c) is a direct consequence of (b) and Corollary 1; so we need to prove only (b). From Part (a) of this lemma and (b) of Lemma 1, it is enough to show that

$$U_{12}^*(s,\lambda) \leq \Psi(s) = \tilde{\hat{U}}_{12}(s,\lambda) \tag{36}$$

and we will prove this by induction. First, we can easily see from Corollary 1 that

$$U^*(1,s,\lambda) = A_1 \leq \tilde{\hat{U}}(s,\lambda)$$

and this implies that

$$G^*(1,s,\lambda) = (I - U^*(1,s,\lambda))^{-1}A_2 \leq \tilde{\hat{G}}(s,\lambda) = (I - \tilde{\hat{U}}(s,\lambda))^{-1}A_2.$$

Now, assume that it holds for $k \geq 1$ that

$$U^*(k,s,\lambda) \leq \tilde{\hat{U}}(s,\lambda) \text{ and } G^*(k,s,\lambda) \leq \tilde{\hat{G}}(s,\lambda).$$

Then, from the Algorithm 2,

$$U^*(k+1,s,\lambda) = A_1 + A_0G^*(k,s,\lambda) \leq A_1 + A_0\tilde{\hat{G}}(s,\lambda) = \tilde{\hat{U}}(s,\lambda)$$

and this implies

$$G^*(k+1,s,\lambda) = (I - U^*(k+1,s,\lambda))^{-1}A_2 \leq (I - \tilde{\hat{U}}(s,\lambda))^{-1}A_2 = \tilde{\hat{G}}(s,\lambda).$$

Therefore,

$$\lim_{k\to\infty} U^*(k,s,\lambda) \leq \tilde{\hat{U}}(s,\lambda) \text{ and } \lim_{k\to\infty} G^*(k,s,\lambda) \leq \tilde{\hat{G}}(s,\lambda).$$

Having established the above lemma, it is now a trivial matter to see that for all $Re(s) > 0$, the two iterative schemes converge and yield the same transforms. One may prove this by appealing to an analytic continuation argument or directly using the Dominated Convergence Theorem by comparing the matrices of absolute of values of the iterates. We omit the details but in light of its importance, we state the result as a theorem.

Theorem 7. *Let $Re(s) > 0$. Then the iterative schemes in Algorithm 1 and Algorithm 2 yield in the limit as $k \to \infty$, the required matrices $\tilde{\hat{U}}$ and $\tilde{\hat{G}}$.*

6.2. A Quadratically Convergent Algorithm

We now present the following algorithmic scheme which is an accelerated version of Algorithm 2.

Algorithm 3

```
Fix $\epsilon > 0$ and set diff = 100;
$H^{**}(1,s,\lambda) = (I - A_1(s,\lambda))^{-1}A_0(s,\lambda)$;
$L^{**}(1,s,\lambda) = (I - A_1(s,\lambda))^{-1}A_2(s,\lambda)$;
$G^{**}(1,s,\lambda) = L^{**}(1,s,\lambda)$;
$T(1) = H^{**}(1,s,\lambda)$;
```

```
Do while ( diff $> \epsilon$ )
    $k = k + 1$;
    $U^{**}(k,s,\lambda) = H^{**}(k-1,s,\lambda)L^{**}(k-1,s,\lambda)$
                          $+ L^{**}(k-1,s,\lambda)H^{**}(k-1,s,\lambda)$;
    $M = (H^{**}(k-1,s,\lambda))^2$;
    $H^{**}(k,s,\lambda) = (I - U^{**}(k,s,\lambda))^{-1}M$;
    $M = (L^{**}(k-1,s,\lambda))^2$;
    $L^{**}(k,s,\lambda) = (I - U^{**}(k,s,\lambda))^{-1}M$;
    $G^{**}(k,s,\lambda) = G^{**}(k-1,s,\lambda) + T(k-1)L^{**}(k,s,\lambda)$;
    $T(k) = T(k-1)H^{**}(k,s,\lambda)$;
    diff $= \max_{i,j \in S} \mid [G^{**}(k,s,\lambda)]_{i,j} - [G^{**}(k-1,s,\lambda)]_{i,j} \mid$;
end
```

$\Psi(s) \approxeq G^{**}_{12}(k,s,\lambda)[G^{**}_{22}(k,s,\lambda)]^{-1}$; $\tilde{\tilde{G}}(s,\lambda) \approxeq G^{**}(k,s,\lambda)$.

Theorem 8.

$$\lim_{k \to \infty} G^{**}(k,s,\lambda) = G^{*}(s,\lambda) = \tilde{\tilde{G}}(s,\lambda), \tag{37}$$

and the convergence is quadratic.

Proof. As before, it suffices to prove the result for $s > 0$. In that case, the iterative scheme of Algorithm 2 is precisely the linear iteration scheme of Latouche ([12], p.170) and Algorithm 3 is the L-R algorithm (see [12], p.193) for the QBD defined by the nonnegative matrices $A_i(s,\lambda)$, $i = 0,1,2$. Thus, the results follow from [11]; see also Chapter 8 of [12] since the k-th iterate of Algorithm 3 is indeed the 2^k-th iterate of Algorithm 2. A direct proof can be given in terms of taboo paths that go up by at most 2^k levels and identifying the k-th iteration here as resulting from restrictions to such paths, but given those details in the cited references, we omit them here. Incidentally, the results in [12], Chapter 8 also show that the error in the k-th iterate is asymptotically $O([\eta(s)]^{2^k})$.

Remark 3. (a) Note that a comparison of Algorithm 2 with the linear scheme of Latouche at best only shows that it converges. The fact that it converges to $\Psi(s)$ has been established by us by comparing its iterates to those of Algorithm 1. Unfortunately, we have not been able to develop an argument leading to Algorithm 3 directly. (b) With $s = 0$, Algorithm 3 provides a powerful means to compute $\Psi(0)$ using which it is easy [1] to compute the steady state distribution of the fluid flow, when it exists.

7. Numerical Results

We have shown that all the iterative schemes developed by us converge to the quantities of interest. To facilitate their implementation, we have devised procedures that yield numerically stable schemes that do not appear to suffer from round off and truncation errors. While a careful error analysis of the algorithms has not been possible yet, our experimentation thus far has confirmed that the procedures developed by us are sound and work well. In this section, we report on a model class studied earlier by Sericola [17] and also used by us in [2], Section 8 for comparison.

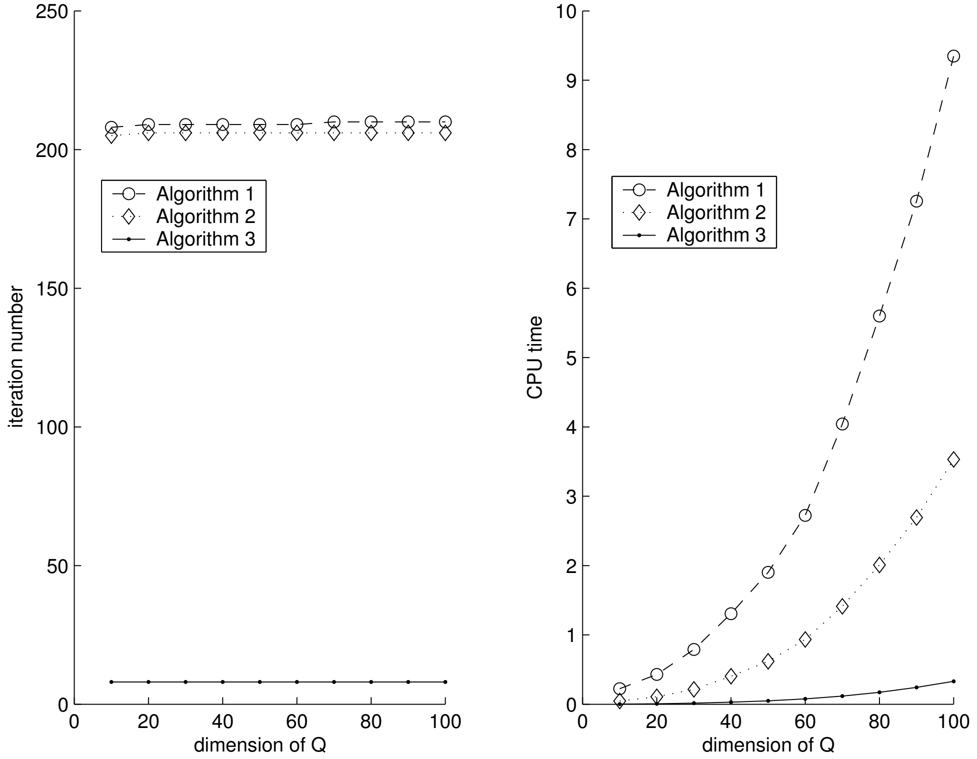

FIGURE 2: $s = 2$, $\rho = 0.9$.

The model class comprises of a fluid flow model wherein each of a set of m on-off sources provides fluid input at rate 1 while it is on, and the combined fluid is drained at a constant rate 0.8 per unit time. The means of the on and off periods are respectively 1 and $1/\gamma$, and the traffic intensity of the model is given by $\rho = (m\gamma)/[0.8(1+\gamma)]$. For a large number of cases, we used our algorithms to compute the transient distribution of the fluid flow and compared the results with those of Sericola [17], where the latter is available, obtaining extremely favorable comparisons for our methods.

For numerical experimentation with the algorithms, here is the procedure we adopted. For each problem, we considered several values of s, both real and complex. For each fixed s value, Algorithm 3 was used to compute $\Psi(s)$ with $\epsilon = 10^{-15}$. The resulting value of $\Psi(s)$ was used as the target value, and other algorithms were then used and iterative processes continued until iterates differed from $\Psi(s)$ by at most 10^{-10} in absolute value. Given below in the figures are the results obtained by us, and our emphasis here is only on providing a glimpse of the relative speeds of the various algorithms in computing the key transform $\Psi(s)$. A detailed complexity analysis has not been performed on them. For brevity, we have shown only two cases: $s = 2$ and $s = 2 + 3i$. Reported for them are the number of iterations and CPU times (in seconds) taken by each of the algorithms to reach the same level of accuracy for Ψ. All computations reported here were performed using MATLAB. Each example was run 20 times, and what is reported is the average CPU utilization per run.

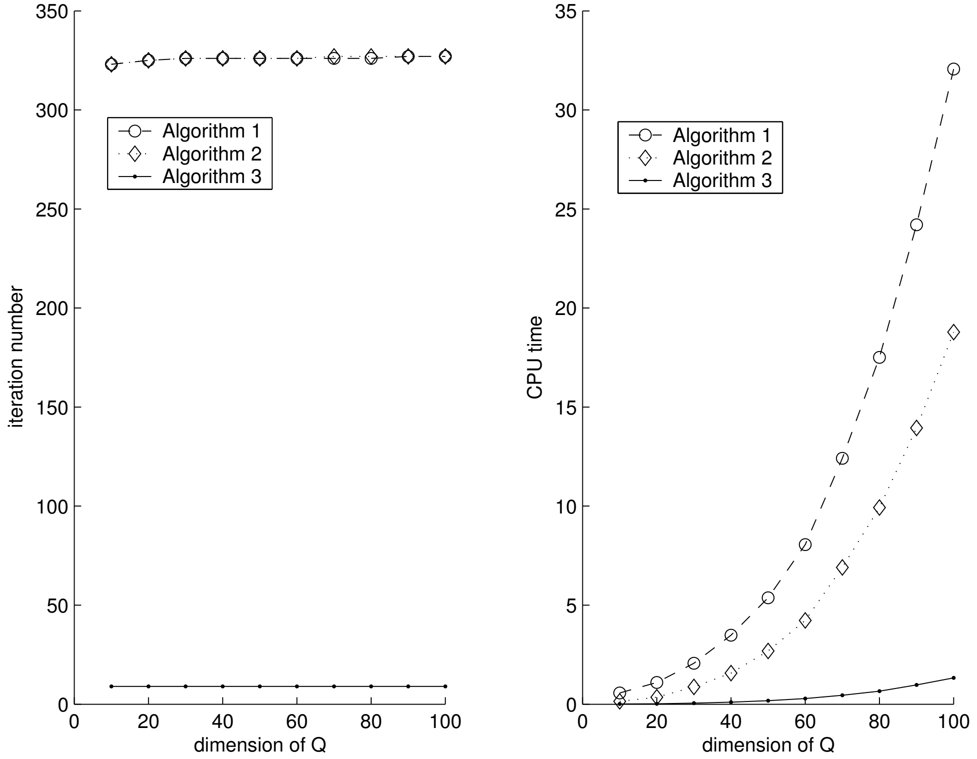

FIGURE 3: $s = 2 + 3i$, $\rho = 0.9$.

Based on these examples and the many others we have worked out, we can assert that we have an excellent algorithm for computing the transient results for stochastic fluid flow models.

References

[1] AHN, S & RAMASWAMI, V (2003). Fluid flow models & queues - A connection by stochastic coupling. *Stochastic Models,* **19(3)**, 325–348.

[2] AHN, S. & RAMASWAMI, V. (2004). Transient analysis of fluid flow models via stochastic coupling to a queue. *Stochastic Models,* **20(1)**, 71-101.

[3] ANICK, D., MITRA, D & SONDHI, M.M (1982). Stochastic theory of data handling system with multiple sources. *Bell System Tech. J.* **61**, 1871-1894.

[4] ASMUSSEN, S (1995). Stationary distributions for fluid flow models with or without Brownian noise. *Stochastic Models*, **11**, 1-20.

[5] ASMUSSEN, S (1994). Busy period analysis, rare events and transient behavior in fluid flow models. *J. Appl. Math. and Stoch. Anal.*, **7(3)**, 269-299.

[6] ÇINLAR, E. (1975). *Introduction to Stochastic Processes.* Prentice Hall, Englewood Cliffs, New Jersey.

[7] GAVER, D.P. & LEHOCZKY, J.P. (1982) Channels that cooperatively service a data stream and voice messages, *IEEE Trans. Com.*, **30**, 1153-1161.

[8] GRAHAM, A. (1981). *Kronecker Products and Matrix Calculus: with Applications.* John Wiley & Sons, NY.

[9] KOBAYASHI, H & REN, Q. (1992). A mathematical theory for transient analysis of communication networks, *IEICE Trans. Commun.*, **(12)**, 1266-1276.

[10] LATOUCHE, G. (1993). Algorithms for infinite Markov chains with repeating columns. in *Linear Algebra, Markov Chains and Queueing Models.* Meyer, C.D. and Plemmons, R.J (eds.), Springer Verlag, New York.

[11] LATOUCHE, G & RAMASWAMI, V. (1993). A logarithmic reduction algorithm for Quasi-Birth-and-Death processes, *J. Appl. Prob.,* **30**, 650-674.

[12] LATOUCHE, G & RAMASWAMI, V. (1999). *Introduction to Matrix Analytic Methods in Stochastic Modeling.* SIAM & ASA, Philadelphia.

[13] NEUTS, M.F. (1981). *Matrix-Geometric Solutions in Stochastic Models – An Algorithmic Approach.* The Johns Hopkins University Press, Baltimore, MD.

[14] ORTEGA, J.M. AND RHEINBOLDT, W.C. (1970). *Iterative Solution of Nonlinear Equations in Several Variables.* Academic Press, NY.

[15] RAMASWAMI, V. (1999). Matrix analytic methods for stochastic fluid flows. in *Teletraffic Engineering in a Competitive World - Proc. of the 16th International Teletraffic Congress.* D. Smith and P. Key (eds.), 1019–1030, Elsevier, NY.

[16] ROGERS, L.C. (1994) Fluid models in queueing theory and Wiener-Hopf factorization of Markov chains. *The Annals of Appl. Prob.*, **4(2)**, 390-413.

[17] SERICOLA, B. (1998). Transient analysis of stochastic fluid models. *Performance Evaluation.* **32**, 245-263.

Acknowledgements

This work was completed at the Institute of Mathematics & its Applications, University of Minnesota, and both authors thank the IMA for its hospitality. Soohan Ahn was supported partially by grant No. R01-2004-000-10284-0 from the Basic Research Program of the Korea Science & EngineeringFoundation He also thanks AT&T where this work was initiated during his teunure as a Post-Doctoral Researcher.