

A multi-layer fluid queue with boundary phase transitions and its application to the analysis of multi-type queues with general customer impatience

Gábor Horváth

Department of Telecommunications
Budapest University of Technology and Economics
Email: ghorvath@hit.bme.hu

Benny Van Houdt

Department of Mathematics and Computer Science
University of Antwerp - IBBT
Email: benny.vanhoudt@ua.ac.be

Abstract—Consider a Markov modulated fluid queue with multiple layers separated by a finite number of boundaries, where each layer is characterized by its own set of matrices. In the past, matrix analytic methods have been devised to determine the stationary behavior of such a fluid queue for no-resistance, sticky and repellent boundaries. In this paper we extend this approach by allowing general phase transitions at the boundaries.

As an application, we analyze the MMAP[K]/PH[K]/1 queue with general, customer type dependent impatience, where customers remain impatient while being served. We show that the steady state distribution of the age process of this queue can be expressed via the steady state distribution of a multi-layered fluid queue with phase transitions at the boundary. Based on the analysis of the age process, expressions for the sojourn time distribution and for the probability of abandonment are presented.

Keywords-fluid queue, age process, impatient customers

I. INTRODUCTION

For several decades, Markov modulated fluid queues have been successful modeling tools in the area of communication networks (e.g., [1], [2],[3]). Since their introduction in the field of performance evaluation, the basic model has been extended in several ways. Such extensions are for instance limiting the capacity of the fluid storage [4], making the fluid rates and the behavior of the background process fluid level dependent [5], replacing the linear fluid accumulation by a second order process [6], etc. In spite of the large number of model variants, robust numerical methods for fluid queues with a large number of states appeared only more recently ([7], [4]).

Fluid queues are also useful when analyzing processes which exhibit a mixture of smooth behavior and occasional jumps, controlled by continuous time Markovian processes on a finite state space, as the steady state distribution of such a jump process can be expressed via the steady state distribution of a fluid queue [8]. A generalization of this idea was used in [9] to analyze the workload process of a class of multi-type queues with Markovian arrival and service processes.

In this paper we consider a multi-layer fluid model as in [10], in which the background process can have phase

transitions when hitting the boundaries. In fact, multi-layer fluid queues that allow such phase transitions have been analyzed before in [11], but the authors focused on the transient solution only. Here we focus on the steady state solution of the system, furthermore, we extend the matrix analytical method of [10] to handle our extended fluid model as well. The numerical procedure developed is able to cope with a large number of states and a large number of layers.

As a possible application, we investigate the multi-type MMAP[K]/PH[K]/1 queue with general customer impatience, where the customers remain impatient while in service. The steady state distribution of the age process of this queueing system can be obtained from a fluid queue that is constructed in a similar manner as in [9], which also considered the same class of queues, but customers were only impatient while waiting. From the steady state distribution of the age process we are able to obtain the distribution of the sojourn time and the probability of service abandonment. Single type queueing systems in which customers remain impatient during their service have been studied in [12] and [13]. They are typically suitable in systems where the service time of a customer is unknown in advance while there are deadlines involved.

The rest of the paper is organized as follows. Section II gives a formal description of the extended fluid queue. The differential equations and the boundary conditions describing the steady state behavior are derived in Section III. Section IV presents the numerical method to solve the steady state distribution of the fluid level. Section V describes the analysis of the queueing system with impatient customers. Section VI presents several numerical examples to demonstrate the possible applications of the results of the paper. Finally, Section VII concludes the paper.

II. MODEL DESCRIPTION

Markov modulated fluid queues are characterized by a two-dimensional Markov process $\{X(t), Z(t), t \geq 0\}$, where $X(t)$ corresponds to the fluid level in the queue and $Z(t)$ is the underlying continuous time Markov chain (CTMC) that modulates the rate at which fluid is accumulated (referred to as the background process in this paper).

The fluid queue described in this paper makes use of $r \geq 0$ thresholds d_1, \dots, d_r such that $0 = d_0 < d_1 < \dots < d_r < d_{r+1} = \infty$. The queue is said to be in layer k if $X(t) \in (d_{k-1}, d_k)$. The transition rates of the underlying CTMC depend on the current layer the fluid level is residing in. To avoid weird and difficult to define behavior, we assume that the fluid rate matrix is the same in all layers (however, this restriction can be relaxed easily).

A. Behavior between the thresholds

The evolution of the fluid level between the thresholds is given by

$$\frac{d}{dt}X(t) = \rho_{Z(t)}, \quad \text{if } X(t) \neq d_k, \quad k = 0, \dots, r,$$

where ρ_i is the fluid rate associated with phase i of the background process. Let \mathcal{S}_+ , \mathcal{S}_- and \mathcal{S}_0 denote the set of phases with positive, negative and zero fluid rate, respectively. The diagonal matrix $\mathbf{R} = \text{diag}\langle \rho_i \rangle$ is constituted by the fluid rates belonging to different phases. The generator matrix of the background process in layer k is denoted by $\mathbf{F}^{(k)} = \{f_{ij}^{(k)}\}$, and it is of size N .

To simplify the description, we assume that the phases are ordered in such a way that phases with positive fluid rates are followed by phases with negative fluid rates followed by phases with zero fluid rates throughout the paper. The number of phases having positive, negative, zero fluid rates are denoted by N_+ , N_- , N_0 , respectively. Due to this ordering $\mathbf{F}^{(k)}$ and \mathbf{R} can be partitioned as follows:

$$\mathbf{F}^{(k)} = \begin{bmatrix} \mathbf{F}_{++}^{(k)} & \mathbf{F}_{+-}^{(k)} & \mathbf{F}_{+0}^{(k)} \\ \mathbf{F}_{-+}^{(k)} & \mathbf{F}_{--}^{(k)} & \mathbf{F}_{-0}^{(k)} \\ \mathbf{F}_{0+}^{(k)} & \mathbf{F}_{0-}^{(k)} & \mathbf{F}_{00}^{(k)} \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} \mathbf{R}_+ & & \\ & \mathbf{R}_- & \\ & & \mathbf{0} \end{bmatrix}.$$

For convenience, we introduce a separate notation for the absolute value of the rate matrix, $\mathbf{C} = |\mathbf{R}|$, thus $\mathbf{C} = \text{diag}\langle |\rho_i| \rangle$. We denote the transient joint distribution function by $\Pi_i^{(k)}(t, x)$ defined as

$$\Pi_i(t, x) = P(X(t) \leq x, Z(t) = i), \quad x \geq 0, i = 1, \dots, N.$$

The row vector of the transient joint distributions is $\Pi(t, x) = [\Pi_1(t, x) \quad \dots \quad \Pi_N(t, x)]$. The corresponding joint densities are defined by

$$\pi_i(t, x) = \frac{\partial}{\partial x} \Pi_i(t, x), \quad x \in (d_{k-1}, d_k), \quad k = 1, \dots, r+1.$$

Similar to the joint distributions, we introduce a vector notation of $\pi(t, x) = [\pi_1(t, x) \quad \dots \quad \pi_N(t, x)]$. By taking the limit we get the stationary joint distributions $\Pi_i(x) = \lim_{t \rightarrow \infty} \Pi_i(t, x)$, for $x \geq 0$, and the densities $\pi_i(x) = \lim_{t \rightarrow \infty} \pi_i(t, x)$, for $x \neq d_k$ and $k = 0, \dots, r$.

B. Behavior at the thresholds

Whenever the fluid level hits a threshold d_k , an immediate phase transition occurs in the background process. The transition probability matrix describing the phase transition probabilities when crossing threshold d_k is denoted by $\mathbf{P}^{(k)} = \{p_{ij}^{(k)}\}$. As level zero can be hit in phases having negative fluid rates only, we have that the size of $\mathbf{P}^{(0)}$ is $N_- \times N$. The other thresholds can be hit in phases with negative and positive fluid rates (from above and below the threshold, respectively), thus matrices $\mathbf{P}^{(k)}$, $k = 1, \dots, r$ are of size $(N_+ + N_-) \times N$. For further use, we partition $\mathbf{P}^{(k)}$ and $\mathbf{P}^{(0)}$ as follows:

$$\mathbf{P}^{(k)} = \begin{bmatrix} \mathbf{P}_{++}^{(k)} & \mathbf{P}_{+-}^{(k)} & \mathbf{P}_{+0}^{(k)} \\ \mathbf{P}_{-+}^{(k)} & \mathbf{P}_{--}^{(k)} & \mathbf{P}_{-0}^{(k)} \end{bmatrix},$$

$$\mathbf{P}^{(0)} = \begin{bmatrix} \mathbf{P}_{-+}^{(0)} & \mathbf{P}_{--}^{(0)} & \mathbf{P}_{-0}^{(0)} \end{bmatrix}.$$

Observe that probability mass can be accumulated at phases having zero fluid rate at threshold k whenever $\mathbf{P}_{+0}^{(k)} \neq \mathbf{0}$ or $\mathbf{P}_{-0}^{(k)} \neq \mathbf{0}$ for some k . At the zero level the probability mass exists obviously also in the phases with negative fluid rates. The generator of the background process is defined by a separate matrix $\mathbf{Q}^{(k)} = [\mathbf{Q}_{0+}^{(k)} \quad \mathbf{Q}_{0-}^{(k)} \quad \mathbf{Q}_{00}^{(k)}]$ at threshold k . As level zero is left via one of the phases having a positive fluid rate, the size of $\mathbf{Q}^{(0)}$ is $(N_- + N_0) \times N$ and we partition it as follows:

$$\mathbf{Q}^{(0)} = \begin{bmatrix} \mathbf{Q}_{0+}^{(0)} & \mathbf{Q}_{0-}^{(0)} & \mathbf{Q}_{00}^{(0)} \\ \mathbf{Q}_{0+}^{(0)} & \mathbf{Q}_{0-}^{(0)} & \mathbf{Q}_{00}^{(0)} \end{bmatrix}.$$

The probability mass at time t at threshold k is defined as

$$c_i^{(k)}(t) = \begin{cases} P(X(t) = 0, Z(t) = i), & t \geq 0, k = 0, i \in \mathcal{S}^0 \cup \mathcal{S}^-, \\ P(X(t) = d_k, Z(t) = i), & t \geq 0, k = 1, \dots, r, i \in \mathcal{S}^0. \end{cases}$$

and the corresponding vectors are denoted by $c^{(0)}(t) = \{c_i^{(0)}(t), i \in \mathcal{S}^0 \cup \mathcal{S}^-\}$ and $c^{(k)}(t) = \{c_i^{(k)}(t), i \in \mathcal{S}^0\}$ for $k = 1, \dots, r$. The stationary distribution of the probability mass is given by $c^{(k)} = \lim_{t \rightarrow \infty} c^{(k)}(t)$.

Due to the presence of the potential masses at d_k , the left and right limits $\lim_{x \rightarrow d_k, x < d_k} \pi(t, x)$ and $\lim_{x \rightarrow d_k, x > d_k} \pi(t, x)$ may not coincide. As such we will denote these as $\pi(t, d_k^-)$ and $\pi(t, d_k^+)$, respectively.

III. STEADY STATE ANALYSIS

The differential equations and the corresponding boundary conditions describing the system in steady state are provided by the following theorem.

Theorem 1: The stationary joint densities satisfy the following differential equations

$$\frac{d}{dx} \pi(x) \mathbf{R} = \pi(x) \mathbf{F}^{(k)}, \quad x \in (d_{k-1}, d_k), \quad k = 1 \dots r+1. \quad (1)$$

The boundary conditions for level zero are as follows:

$$\begin{aligned} \pi_i(0+)\rho_i &= - \sum_{j \in \mathcal{S}^-} p_{ji}^{(0)} \pi_j(0+)\rho_j \\ &+ \sum_{j \in \mathcal{S}^0 \cup \mathcal{S}^-} c_j^{(0)} q_{ji}^{(0)}, \quad \forall i \in \mathcal{S}^+, \end{aligned} \quad (2)$$

$$\begin{aligned} 0 &= - \sum_{j \in \mathcal{S}^-} p_{ji}^{(0)} \pi_j(0+)\rho_j \\ &+ \sum_{j \in \mathcal{S}^0 \cup \mathcal{S}^-} c_j^{(0)} q_{ji}^{(0)}, \quad \forall i \in \mathcal{S}^0 \cup \mathcal{S}^-. \end{aligned} \quad (3)$$

For thresholds $d_k, k = 1, \dots, r$ we have the following linear system of equations:

$$\begin{aligned} \pi_i(d_k+)\rho_i &= \sum_{j \in \mathcal{S}^+} p_{ji}^{(k)} \pi_j(d_k-)\rho_j - \sum_{j \in \mathcal{S}^-} p_{ji}^{(k)} \pi_j(d_k+)\rho_j \\ &+ \sum_{j \in \mathcal{S}^0} c_j^{(k)} q_{ji}^{(k)}, \quad \forall i \in \mathcal{S}^+, \end{aligned} \quad (4)$$

$$\begin{aligned} -\pi_i(d_k-)\rho_i &= \sum_{j \in \mathcal{S}^+} p_{ji}^{(k)} \pi_j(d_k-)\rho_j - \sum_{j \in \mathcal{S}^-} p_{ji}^{(k)} \pi_j(d_k+)\rho_j \\ &+ \sum_{j \in \mathcal{S}^0} c_j^{(k)} q_{ji}^{(k)}, \quad \forall i \in \mathcal{S}^-, \end{aligned} \quad (5)$$

$$\begin{aligned} 0 &= \sum_{j \in \mathcal{S}^+} p_{ji}^{(k)} \pi_j(d_k-)\rho_j - \sum_{j \in \mathcal{S}^-} p_{ji}^{(k)} \pi_j(d_k+)\rho_j \\ &+ \sum_{j \in \mathcal{S}^0} c_j^{(k)} q_{ji}^{(k)}, \quad \forall i \in \mathcal{S}^0. \end{aligned} \quad (6)$$

Proof: Since the behavior between the thresholds does not differ from the classical definition of the fluid queues, the differential equation describing the evolution of the fluid level (1) is identical to the well known result presented for example in [14].

To prove (4), let us consider state i for which $\rho_i > 0$ holds. First we express the probability that the fluid level is between d_k and $d_k + \delta$ at time $t + \Delta$ as

$$\begin{aligned} P(d_k < X(t + \Delta) < d_k + \delta, Z(t) = i) &= \\ &\sum_{j \in \mathcal{S}^-} P(d_k - \rho_j \Delta + \delta \cdot \rho_j / \rho_i < X(t) < d_k - \rho_j \Delta, Z(t) = j) p_{ji}^{(k)} \\ &+ \sum_{j \in \mathcal{S}^+} P(d_k - \rho_j \Delta < X(t) < d_k - \rho_j \Delta + \delta \cdot \rho_j / \rho_i, Z(t) = j) p_{ji}^{(k)} \\ &+ \sum_{j \in \mathcal{S}^0} P(X(t) = d_k, Z(t) = j) q_{ji}^{(k)} \Delta \frac{\delta / \rho_i}{\Delta} \\ &+ \Theta(\delta) O(\Delta). \end{aligned} \quad (7)$$

By choosing δ such that $\delta \ll \max\{|\rho_i| \Delta\}$, we have that the fluid level can be in $(d_k, d_k + \delta)$ only in the following cases:

1) The state of the system is $(x, j), x > d_k, \rho_j < 0$ before hitting the threshold. To have the fluid level in the

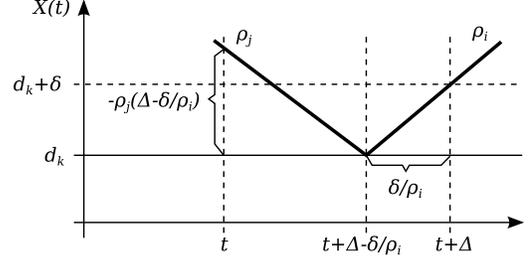


Figure 1. State transition when hitting the threshold

desired interval, the threshold must be hit in $(t + \Delta - \delta / \rho_i, t + \Delta)$ (see Figure 1). At rate ρ_j this can be achieved when the fluid level in time t falls into $(d_k - \rho_j \Delta + \delta \cdot \rho_j / \rho_i, d_k - \rho_j \Delta)$. This case is described by the first term of (7).

- 2) The state of the system is $(x, j), x < d_k, \rho_j > 0$ before hitting the threshold. Considerations similar to the previous case leads to the second term of (7).
- 3) The state of the system is $(x, j), x = d_k, \rho_j = 0$ and a state transition occurs to state i in $(t + \Delta - \delta / \rho_i, t + \Delta)$. The probability that a state transition happens in $(t, t + \Delta)$ is $q_{ji}^{(k)} \Delta$, and the probability that the state transition falls into a given δ / ρ_i long interval is $\frac{\delta / \rho_i}{\Delta}$. This case is expressed by the third term.
- 4) The fourth term covers the cases when two events occur in $(t, t + \Delta)$. This can happen if both the threshold is hit and the state of the background process changes, or when the fluid level resides at the threshold in a zero state and two state transitions occur in the background process. This term vanishes when taking the limits below since $\lim_{\delta \rightarrow 0} \frac{\Theta(\delta)}{\delta}$ is constant and $\lim_{\Delta \rightarrow 0} O(\Delta) = 0$.

Dividing both sides by δ and taking the limit $\delta \rightarrow 0$ gives

$$\begin{aligned} \pi_i(t + \Delta, d_k+) &= \\ &\sum_{j \in \mathcal{S}^-} \pi_j(t, d_k - \rho_j \Delta) (-1) \frac{\rho_j}{\rho_i} p_{ji}^{(k)} \\ &+ \sum_{j \in \mathcal{S}^+} \pi_j(t, d_k - \rho_j \Delta) \frac{\rho_j}{\rho_i} p_{ji}^{(k)} \\ &+ \sum_{j \in \mathcal{S}^0} c_j^{(k)}(t) q_{ji}^{(k)} \frac{1}{\rho_i} \\ &+ O(\Delta). \end{aligned}$$

Note that in the first sum we have $\rho_j < 0$ thus $d_k - \rho_j \Delta$ falls above d_k . Multiplying both sides by ρ_i and taking the limits $\Delta \rightarrow 0, t \rightarrow \infty$ provides (4). The equality in (5) can be obtained similarly, but the probability that the fluid level is between $d_k - \delta$ and d_k has to be investigated.

Boundary condition (6) is proved by expressing the probability that the fluid level is d_k at time $t + \Delta$. This can happen in the following three cases:

- 1) the state of the system at time t is (x, i) , $x = d_k$, $\rho_i = 0$ and the background process stays in the same state in $(t, t + \Delta)$,
- 2) the state of the system at time t is (x, j) , $x = d_k$, $\rho_j = 0$ and the background process moves from state j to state i , $\rho_i = 0$,
- 3) the fluid level reached the threshold in $(t, t + \Delta)$ either from above or below and the state transition moved the background process to a zero state.

Thus we have

$$\begin{aligned}
P(X(t + \Delta) = d_k, Z(t + \Delta) = i) &= \\
&P(X(t) = d_k, Z(t) = i)(1 + q_{ii}^{(k)} \Delta) \\
&+ \sum_{j \in \mathcal{S}^0, j \neq i} P(X(t) = d_k, Z(t) = j) q_{ji}^{(k)} \Delta \\
&+ \sum_{j \in \mathcal{S}^+} P(d_k - \rho_j \Delta < X(t) < d_k, Z(t) = j) p_{ji}^{(k)} \\
&+ \sum_{j \in \mathcal{S}^-} P(d_k < X(t) < d_k + \rho_j \Delta, Z(t) = j) p_{ji}^{(k)} + o(\Delta).
\end{aligned}$$

Now we subtract $P(X(t) = d_k, Z(t) = i)$ from both sides, divide by Δ and take the limit of $\Delta \rightarrow 0$ that yields

$$\begin{aligned}
\frac{d}{dt} c_i^{(k)}(t) &= \sum_{j \in \mathcal{S}^0} c_j^{(k)}(t) q_{ji}^{(k)} + \\
&\sum_{j \in \mathcal{S}^+} \pi_j(t, d_k -) \rho_j p_{ji}^{(k)} - \sum_{j \in \mathcal{S}^-} \pi_j(t, d_k +) \rho_j p_{ji}^{(k)}.
\end{aligned}$$

Finally, $t \rightarrow \infty$ proves (6). The boundary equations for the zero level (2) and (3) can be proven similarly. \blacksquare

The boundary conditions can be re-stated in a more compact vector-matrix form if we introduce the following matrices:

$$\begin{aligned}
P_-^{(k)} &= \begin{bmatrix} I & 0 & 0 \\ P_{-+}^{(k)} & P_{--}^{(k)} & P_{-0}^{(k)} \\ 0 & 0 & I \end{bmatrix}, \\
P_+^{(k)} &= \begin{bmatrix} P_{++}^{(k)} & P_{+-}^{(k)} & P_{+0}^{(k)} \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}.
\end{aligned}$$

Corollary 1: The boundary conditions belonging to level zero (2) and (3) can be expressed in a matrix form as

$$\pi(0+) \mathbf{R} P_-^{(0)} = c^{(0)} \mathbf{Q}^{(0)},$$

and for $k = 1, \dots, r$ (4), (5) and (6) can be expressed as

$$\pi(d_k+) \mathbf{R} P_-^{(k)} - \pi(d_k-) \mathbf{R} P_+^{(k)} = c^{(k)} \mathbf{Q}^{(k)}$$

IV. AN EFFICIENT NUMERICAL METHOD

In this section we present an efficient matrix analytic approach to solve the set of equations given by (2)-(6). We start by defining the process $\{\tilde{X}(t), \tilde{Z}(t), t \geq 0\}$ by observing the process $\{X(t), Z(t), t \geq 0\}$ only when it resides in the set

of states $\{(x, i), i \in \mathcal{S}^+\} \cup \{(x, i), x > 0, i \in \mathcal{S}^-\}$. This process has no masses at the boundaries d_0, d_1, \dots, d_r , while its steady state densities $\tilde{\pi}(d_k+) = (\tilde{\pi}_+(d_k+), \tilde{\pi}_-(d_k+))$, for $k \geq 0$, and $\tilde{\pi}(d_k-) = (\tilde{\pi}_+(d_k-), \tilde{\pi}_-(d_k-))$, for $k \geq 1$, can be written as $\tilde{\pi}(d_k+) = \tilde{\eta} \pi(d_k+)$ and $\tilde{\pi}(d_k-) = \tilde{\eta} \pi(d_k-)$, with $\tilde{\eta}$ a normalization constant.

This process $\{\tilde{X}(t), \tilde{Z}(t), t \geq 0\}$ is fully characterized by

$$\tilde{F}^{(k)} = \begin{bmatrix} F_{++}^{(k)} & F_{+-}^{(k)} \\ F_{-+}^{(k)} & F_{--}^{(k)} \end{bmatrix} - \begin{bmatrix} F_{+0}^{(k)} \\ F_{-0}^{(k)} \end{bmatrix} (F_{00}^{(k)})^{-1} \begin{bmatrix} F_{0+}^{(k)} & F_{0-}^{(k)} \end{bmatrix},$$

for $k = 1, \dots, r + 1$,

$$\tilde{P}^{(k)} = \begin{bmatrix} P_{++}^{(k)} & P_{+-}^{(k)} \\ P_{-+}^{(k)} & P_{--}^{(k)} \end{bmatrix} - \begin{bmatrix} P_{+0}^{(k)} \\ P_{-0}^{(k)} \end{bmatrix} (Q_{00}^{(k)})^{-1} \begin{bmatrix} Q_{0+}^{(k)} & Q_{0-}^{(k)} \end{bmatrix},$$

for $k = 1, \dots, r$ and

$$\tilde{P}^{(0)} = P_{-+}^{(0)} - \begin{bmatrix} P_{--}^{(0)} & P_{-0}^{(0)} \end{bmatrix} \begin{bmatrix} Q_{0-}^{(0)} & Q_{00}^{(0)} \\ Q_{0-}^{(0)} & Q_{00}^{(0)} \end{bmatrix}^{-1} \begin{bmatrix} Q_{0+}^{(0)} \\ Q_{0+}^{(0)} \end{bmatrix}.$$

The matrices $\tilde{F}^{(k)}$ and $\tilde{P}^{(k)}$ are also partitioned in the obvious manner. Equations (4)-(5) are also valid for the censored process, for which the masses $c_j^{(k)}$ are equal to zero. Hence,

$$\tilde{\pi}_+(d_k+) \mathbf{C}_+ = \tilde{\pi}_+(d_k-) \mathbf{C}_+ \tilde{P}_{++}^{(k)} + \tilde{\pi}_-(d_k+) \mathbf{C}_- \tilde{P}_{-+}^{(k)}, \quad (8)$$

$$\tilde{\pi}_-(d_k-) \mathbf{C}_- = \tilde{\pi}_-(d_k+) \mathbf{C}_- \tilde{P}_{--}^{(k)} + \tilde{\pi}_+(d_k-) \mathbf{C}_+ \tilde{P}_{+-}^{(k)}, \quad (9)$$

for $k = 1, \dots, r$, while (2) implies

$$\tilde{\pi}_+(d_0+) \mathbf{C}_+ = \tilde{\pi}_-(d_0+) \mathbf{C}_- \tilde{P}^{(0)}. \quad (10)$$

To express the $\tilde{\pi}_+(d_k-)$ and $\tilde{\pi}_-(d_k+)$ vectors in terms of $\tilde{\pi}_+(d_k+)$ and $\tilde{\pi}_-(d_k-)$, we define four sets of first passage probability matrices as in [4]:

- Entry (i, i') of $\hat{\Lambda}_{++}^{(k)}$, for $k = 1, \dots, r$, holds the probabilities that, starting from state (x, i) with $x = d_k$ and $i \in \mathcal{S}_+$, the next visit to the set of states $\{(x, j), x = d_k, j \in \mathcal{S}_-\} \cup \{(x, j), x = d_{k+1}, j \in \mathcal{S}_+\}$ occurs in state (d_{k+1}, i') .
- $\hat{\Psi}_{+-}^{(k)}$, for $k = 1, \dots, r+1$, holds the same probabilities as $\hat{\Lambda}_{++}^{(k)}$, except that the first visit occurs in state (d_k, i') .
- Entry (i, i') of $\hat{\Lambda}_{--}^{(k)}$, for $k = 1, \dots, r$, holds the probabilities that, starting from state (x, i) with $x = d_k$ and $i \in \mathcal{S}_-$, the next visit to the set of states $\{(x, j), x = d_k, j \in \mathcal{S}_+\} \cup \{(x, j), x = d_{k-1}, j \in \mathcal{S}_-\}$ occurs in state (d_{k-1}, i') .
- $\hat{\Psi}_{-+}^{(k)}$, for $k = 1, \dots, r$, holds the same probabilities as $\hat{\Lambda}_{--}^{(k)}$, except that the first visit occurs in state (d_k, i') .

Note, $(\hat{\Lambda}_{++}^{(k)} + \hat{\Psi}_{+-}^{(k)})$ and $(\hat{\Lambda}_{--}^{(k)} + \hat{\Psi}_{-+}^{(k)})$ are stochastic matrices, for $k = 1, \dots, r$. The matrix $\hat{\Psi}_{+-}^{(r+1)}$ is stochastic

if and only if the process $\{X(t), Z(t), t \geq 0\}$ is positive recurrent. The computation of these matrices is discussed in Appendix A.

Using the probabilistic interpretation of these matrices, we find

$$\begin{aligned}\tilde{\pi}_+(d_k-)C_+ &= \tilde{\pi}_+(d_{k-1}+)C_+\Lambda_{++}^{(k)} + \tilde{\pi}_-(d_k-)C_-\hat{\Psi}_{-+}^{(k)}, \\ \tilde{\pi}_-(d_k+)C_- &= \tilde{\pi}_-(d_{k+1}-)C_-\hat{\Lambda}_{--}^{(k)} + \tilde{\pi}_+(d_k+)C_+\Psi_{+-}^{(k)}, \\ \tilde{\pi}_-(d_r+)C_- &= \tilde{\pi}_+(d_r+)C_+\Psi_{+-}^{(r+1)},\end{aligned}$$

where $C_+ = R_+$ and $C_- = |R_-|$.

Define $\tilde{\pi} = (\tilde{\pi}_-(d_0+)C_-, \tilde{\pi}_-(d_1+)C_-, \tilde{\pi}_+(d_1-)C_+, \dots, \tilde{\pi}_-(d_r+)C_-, \tilde{\pi}_+(d_r-)C_+)$. We obtain a linear system of equations $\tilde{\pi} = \tilde{\pi}A$, by combining (8)-(10) with the above three equations, where

$$A = \begin{bmatrix} \mathbf{A}_{0,0} & \mathbf{A}_{0,1} & & & 0 \\ \mathbf{A}_{1,0} & \ddots & \ddots & & \\ & \ddots & \ddots & \mathbf{A}_{r-1,r} & \\ 0 & & \mathbf{A}_{r,r-1} & \mathbf{A}_{r,r} & \end{bmatrix},$$

with $\mathbf{A}_{0,0} = \tilde{P}^{(0)}$,

$$\mathbf{A}_{0,1} = \begin{bmatrix} 0 & \tilde{P}^{(0)}\Lambda_{++}^{(1)} \end{bmatrix}, \quad \mathbf{A}_{1,0} = \begin{bmatrix} \tilde{P}^{(1)}\hat{\Lambda}_{--}^{(1)} \\ \tilde{P}^{(1)}\hat{\Lambda}_{--}^{(1)} \end{bmatrix},$$

and

$$\mathbf{A}_{k,k} = \begin{bmatrix} \tilde{P}^{(k)}\Psi_{+-}^{(k+1)} & \tilde{P}^{(k)}\hat{\Psi}_{-+}^{(k)} \\ \tilde{P}^{(k)}\Psi_{+-}^{(k+1)} & \tilde{P}^{(k)}\hat{\Psi}_{-+}^{(k)} \end{bmatrix}$$

for $k = 1, \dots, r$,

$$\mathbf{A}_{k,k-1} = \begin{bmatrix} \tilde{P}^{(k)}\hat{\Lambda}_{--}^{(k)} & 0 \\ \tilde{P}^{(k)}\hat{\Lambda}_{--}^{(k)} & 0 \end{bmatrix}$$

for $k = 2, \dots, r$, and

$$\mathbf{A}_{k,k+1} = \begin{bmatrix} 0 & \tilde{P}^{(k)}\Lambda_{++}^{(k+1)} \\ 0 & \tilde{P}^{(k)}\Lambda_{++}^{(k+1)} \end{bmatrix}$$

for $k = 1, \dots, r-1$. As A is stochastic, the system $\tilde{\pi} = \tilde{\pi}A$ has a unique solution as we implicitly assumed that $\{X(t), Z(t), t \geq 0\}$ is irreducible. Due to the structure of A , we can compute $\tilde{\pi}$ in $O((N_+ + N_-)^3 r)$ time with $O((N_+ + N_-)^2 r)$ memory using for instance the linear level reduction algorithm in [15, Chapter 10]. The densities $\tilde{\pi}_+(d_k+)$ and $\tilde{\pi}_-(d_k-)$ can be computed from $\tilde{\pi}$ using (8)-(10).

Having obtained all the boundary densities of the process $\{X(t), Z(t), t \geq 0\}$ using the censored process $\{\tilde{X}(t), \tilde{Z}(t), t \geq 0\}$ (up to a normalizing constant $\tilde{\eta}$), we are now in a position to obtain the required probability masses. Due to (6), one finds

$$c_0^{(k)} = \left(\pi_+(d_k-)C_+P_{+0}^{(k)} + \pi_-(d_k+)C_-P_{-0}^{(k)} \right) (-Q_{00})^{-1},$$

while (3) yields

$$(c_-^{(k)}, c_0^{(k)}) = \pi_-(d_0+)C_- \begin{bmatrix} P_{--}^{(0)} & P_{-0}^{(0)} \end{bmatrix} \begin{bmatrix} Q_{-0}^{(0)} & Q_{-0}^{(0)} \\ Q_{0-}^{(0)} & Q_{00}^{(0)} \end{bmatrix}^{-1}.$$

The normalization constant and the solution of (1) in terms of the boundary densities is briefly discussed in Appendix A as we can make direct use of on the approach developed in [4].

V. APPLICATION: QUEUES WITH IMPATIENT CUSTOMERS

A. Queueing model description

In this section we calculate the sojourn time distribution of the customers in a MMAP[K]/PH[K]/1 queue with impatience. This is a multi-type queue in which the arrival process is a marked Markovian arrival process (MMAP, [16]), and the service time is phase-type distributed (different types of customers have different service time distributions). The customers are impatient: when their sojourn time exceeds a limit (that is given by a distribution), they leave the system. In this system customers can leave even during their service. This case has several possible practical applications, including systems where customers are not aware of that they are in service and leave as soon as their deadline expires.

Let us denote the number of customer types by K . The matrices characterizing the MMAP[K] of the arrivals are denoted by $D_k, k = 0, \dots, K$, with D_0 describing the transition rates not accompanied by an arrival and D_k the ones accompanied by the arrival of a type k customer. The mean arrival rate of type k customers is λ_k that is calculated as $\lambda_k = aD_k$ with vector a being the unique solution of $a(\sum_{k=0}^K D_k) = 0, a\mathbb{1} = 1$. The PH distributed service time of a type k customer is given by initial vector and transient generator $\alpha_k, S_k, k = 1, \dots, K$, respectively.

The distribution of the impatience of a type k customer is given by a step function consisting of r steps. Without loss of generality we may assume that the points of the step function are the same for all customer types, denoted by $0 = d_0 < d_1 < \dots < d_r < d_{r+1} = \infty$. The distribution of the impatience is then described by probabilities $a_{i,k}$ defined as

$$a_{i,k} = P(\text{patience of a type } k \text{ customer} \geq d_i). \quad (11)$$

B. Analysis based on the age process

The distribution of the waiting time and/or the sojourn time of multi-type queues is often analyzed by using either the workload process or the age process. The workload process is defined as the remaining busy period of the queue, while the age process tracks the age of the customer under service [17].

In [9] a queueing model similar to our system is considered and its analysis is based on the workload process. The arrival process of that queue is more general than the one

we have, but impatience in [9] applies on the waiting time only, while in our model customers remain impatient during their service as well. While this may appear as a relatively minor difference in the behavior, it prevents us from using a standard multi-layered fluid queue to model the workload process and the age process. The fluid model introduced in Section II, however, is able to describe the behavior of both the workload and the age process.

The age of the customer in service increases linearly between the service instants and jumps downwards when the customer leaves the server. The length of the downward jump is equal to the inter-arrival time between the customer leaving the system and the next customer who is about to enter the server (see Figure 2), unless the server becomes idle for a while. There are various ways to deal with idle periods when using an age process: (i) negative values could be allowed for the age, the absolute value of which is the time until the server becomes busy again (ii) these idle periods could be skipped or (iii) the age is said to equal zero until the server becomes occupied again. We will make use of the latter approach and define the age process $\{A(t), Z(t), t \geq 0\}$ as follows.

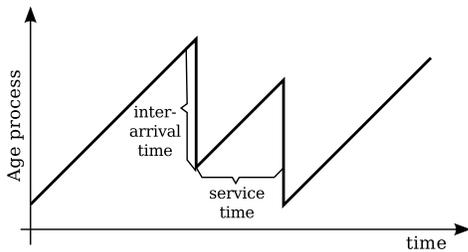


Figure 2. Evolution of the age process

$A(t)$ represents the age of the customer in service at time t , that is, $t - A(t)$ represents the time of arrival of the customer in service, it is equal to zero in case the server is idle. If $A(t) > 0$, then $Z(t)$ keeps track of (a) the type of the customer in service, (b) the current phase of the service and (c) the state of the MMAP[K] process at time $t - A(t)$. If $A(t) = 0$, then $Z(t)$ simply reflects the state of the MMAP[K] process at time t .

The direct analysis of the age process (exhibiting a skip free to the right behavior) seems hard due to the jumps. Observe that in our model the size of the jumps is not arbitrary, it is governed by the MMAP[K] process generating the arrivals, which can be exploited to develop an efficient analysis method. We use the approach taken in [9], which was a generalization of [8]. The basic idea is to construct a fluid queue that is skip-free in both directions and to derive the steady state distribution of the age process from the steady state distribution of this fluid queue.

The background process of the fluid queue has two set of phases according to the following considerations:

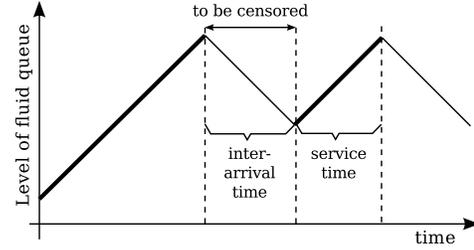


Figure 3. Fluid queue model for the age process

- A first set of phases corresponds to the evolution of the age process when $A(t)$ increases. The fluid rates corresponding to this set of phases are $\rho_i = 1$, since the age of the customer in the server increases according to a slope of one.
- The second set of phases is used whenever a customer leaves the system (for whatever reason). In this case the age of the customer in service has to be decreased by the inter-arrival time between the customer leaving and the one who is about to enter the server. According to the definition of the age process, this decrease is immediate: it is a jump. As the inter-arrival time follows a MMAP[K], the same amount of decrease of the age process can be achieved in an alternative way as well. Let us set the fluid rate to -1 and start the evolution of the MMAP[K] where it has been stopped before. When the MMAP[K] generates an arrival with a sufficient amount of patience, the queue level representing the age process has been decreased appropriately, so the MMAP[K] can be frozen again and the fluid queue can go back to the first set of phases corresponding to the service periods.

To obtain the age process, the second set of phases has to be censored out whenever the fluid $X(t) > 0$. If we also censor out this set of states when $X(t) = 0$, we would in fact skip any time intervals where the server is idle, which was option (ii) in our earlier discussion. Consequently, the age process will be analyzed using a fluid queue: in the first set of phases the fluid rates are $+1$, in the second set of phases they are -1 . The corresponding phase space partitions are denoted by \mathcal{S}^- and \mathcal{S}^+ , respectively. The first (second) set of phases will be referred to as positive (negative) phases in the sequel.

Since the arrival rates of customers depend on the fluid level (through the distribution of the impatience $a_{i,k}$), the generator of the background process is level dependent as well. The generator corresponding to layer i is partitioned as

$$F^{(i)} = \begin{bmatrix} F_{++}^{(i)} & F_{+-}^{(i)} \\ F_{-+}^{(i)} & F_{--}^{(i)} \end{bmatrix},$$

for $i = 1, \dots, r+1$. In the positive phases (belonging to the intervals between jumps of the age process) the evolution

of the background process is determined by the evolution of the PH distribution corresponding to the service time. Hence we have

$$\mathbf{F}_{++}^{(i)} = \begin{bmatrix} \mathbf{S}_1 & & 0 \\ & \ddots & \\ 0 & & \mathbf{S}_K \end{bmatrix} \otimes \mathbf{I}.$$

Note that besides keeping track of the type of the customer in service and the current phase of the service process, we also ensure that the state of the MMAP[K] is maintained. This state corresponds to the state of the MMAP[K] arrival process at time $t - X(t)$, where t is the current point in time and $X(t)$ the age of the customer in service. Hence, the MMAP[K] state remains frozen while the same customer remains in service.

When the service of a customer completes, a transition occurs to the negative phases, thus we get

$$\mathbf{F}_{+-}^{(i)} = \begin{bmatrix} s_1 \\ \vdots \\ s_K \end{bmatrix} \otimes \mathbf{I}.$$

In the negative phases the evolution of the background process is determined by the evolution of the MMAP[K]. Transitions not leaving \mathcal{S}^- are the transitions of the MMAP[K] that are not accompanied by an arrival, and arrivals that have less patience than d_k (these customers leave the system already before reaching the server). This yields

$$\mathbf{F}_{--}^{(i)} = \mathbf{D}_0 + \sum_{k=1}^K \mathbf{D}_k (1 - a_{i,k}).$$

When a type k inter-arrival time is generated (taking into account that the patience of the arriving customer exceeds his waiting time) a transition occurs to \mathcal{S}^+ and the corresponding type k PH distribution is initiated. We have

$$\mathbf{F}_{-+}^{(i)} = \sum_{k=1}^K [0 \quad \dots \quad 0 \quad \alpha_k \quad 0 \quad \dots \quad 0] \otimes \mathbf{D}_k a_{i,k}.$$

In our system the impatient customers can leave the system during their service as well. Observe customers in service can only run out of patience when their age is equal to one of the threshold values d_1, \dots, d_r as their patience distribution is a step function. The probability that a customer of type k loses patience when hitting threshold d_i is given by

$$\begin{aligned} & P(\text{leaves during service at } d_i) \\ &= P(\text{patience} < d_{i+1} \mid \text{patience} \geq d_i) \quad (12) \\ &= \frac{a_{i,k} - a_{i+1,k}}{a_{i,k}}. \end{aligned}$$

Whenever a customer loses patience during his service, the fluid model of the system returns to \mathcal{S}^- immediately. The immediate phase transitions of the background process

at threshold i are governed by the matrix $\mathbf{P}^{(i)}$ that is partitioned according to \mathcal{S}^- and \mathcal{S}^+ as well, yielding

$$\mathbf{P}^{(i)} = \begin{bmatrix} \mathbf{P}_{++}^{(i)} & \mathbf{P}_{+-}^{(i)} \\ \mathbf{P}_{-+}^{(i)} & \mathbf{P}_{--}^{(i)} \end{bmatrix}.$$

Due to (12) the immediate phase transitions from \mathcal{S}^+ to \mathcal{S}^- at threshold d_i are given by

$$\begin{aligned} \mathbf{P}_{+-}^{(i)} &= \begin{bmatrix} \left(\frac{a_{i,1} - a_{i+1,1}}{a_{i,1}} \right) \mathbb{1}_{m_1} \\ \vdots \\ \left(\frac{a_{i,K} - a_{i+1,K}}{a_{i,K}} \right) \mathbb{1}_{m_K} \end{bmatrix} \otimes \mathbf{I}, \\ \mathbf{P}_{-+}^{(i)} &= \begin{bmatrix} \frac{a_{i+1,1}}{a_{i,1}} \mathbf{I}_{m_1} & & \\ & \ddots & \\ & & \frac{a_{i+1,K}}{a_{i,K}} \mathbf{I}_{m_K} \end{bmatrix} \otimes \mathbf{I}. \end{aligned}$$

Since immediate phase transitions occur only in \mathcal{S}^+ causing the return to \mathcal{S}^- we have that

$$\mathbf{P}_{--}^{(i)} = \mathbf{I}, \quad \mathbf{P}_{-+}^{(i)} = \mathbf{0}.$$

It remains to discuss the matrices $\mathbf{Q}_{--}^{(0)}$ and $\mathbf{Q}_{-+}^{(0)}$ to complete the description of the fluid queue. Type k arrivals have zero patience with probability $1 - a_{i,k}$, thus even if they find the server idle upon arrival, they will only enter the server with probability $a_{1,k}$. Hence, $\mathbf{Q}_{--}^{(0)} = \mathbf{F}_{--}^{(1)}$ and $\mathbf{Q}_{-+}^{(0)} = \mathbf{F}_{-+}^{(1)}$.

The steady state distribution of the fluid queue $\pi(x)$ is calculated by the method presented in Section IV and Appendix A. The steady state distribution of the age process is obtained by censoring the results of the fluid queue on the positive phases. Let us denote the steady state joint density of the age process and the phase of the background process by $g_i(x)$ defined as

$$g_i(x) = \lim_{t \rightarrow \infty} \frac{d}{dx} P(A(t) < x, Z(t) = i),$$

for $x \geq 0$ and $i = 1, \dots, N$, and the corresponding vector by $g(x) = \{g_i(x), i \in \mathcal{S}_+\}$.

The density of the age process is obtained from the density of the fluid model as follows:

$$g(x) = \frac{\pi_+(x)}{c^{(0)} \mathbb{1} + \int_{y=0}^{\infty} \pi_+(y) \mathbb{1} dy}, \quad x > 0,$$

while

$$c_i^{(0)} = \lim_{t \rightarrow \infty} P(A(t) = 0, Z(t) = i),$$

for $i \in \mathcal{S}_-$.

C. Performance measures

There is a direct relation between the sojourn time of the customers and the age process at service instants. The probability that the sojourn time of a type k customer D_k is

less than x given that it did not leave before getting served is expressed by

$$P(D_k \leq x) = \frac{\int_{y=0}^x g(y) \sigma_k dy}{\int_{y=0}^{\infty} g(y) \sigma_k dy} \quad (13)$$

$$= \frac{\sum_{i=0}^{j-1} \int_{d_i}^{d_{i+1}} g(y) \sigma_k dy + \int_{d_j}^x g(y) \sigma_k dy}{\sum_{i=0}^r \int_{d_i}^{d_{i+1}} g(y) \sigma_k dy},$$

for $x \in (d_j, d_{j+1}]$, where σ_k denotes the vector of transition rates corresponding to service completions of type k customers:

$$\sigma_k = [0 \quad \dots \quad 0 \quad s_k^T \quad 0 \quad \dots \quad 0]^T.$$

As the steady state distribution of the fluid queue is calculated by the numerical procedure of Section IV and Appendix A, the integrals of $g(y)$ in (13) have a closed form solution.

Another important performance measure of a queue with customer impatience is the probability of abandonment $p_\ell^{(k)}$ of a type k customer, i.e., the probability that a type k customer leaves without starting or completing service. It is computed from the complementary event, that is the rate of successful service divided by the rate of arrivals, thus

$$p_\ell^{(k)} = 1 - \frac{1}{\lambda_k} \int_{y=0}^{\infty} g(y) \sigma_k dy.$$

VI. NUMERICAL EXAMPLES

In this section we provide numerical examples that demonstrate some interesting features and application possibilities of the studied impatient queue as well as the efficiency of the numerical method.

In all of the examples there are 2 types of customers. Type 1 customers are impatient, while type 2 customers are patient. The matrices of the MMAP[K] controlling the arrivals are:

$$D_0 = \begin{bmatrix} -0.971 & 0.001 \\ 0.01 & -4.06 \end{bmatrix}, \quad D_1 = \begin{bmatrix} 0.37 & 0 \\ 0 & 0.15 \end{bmatrix},$$

$$D_2 = \begin{bmatrix} 0.6 & 0 \\ 0 & 3.9 \end{bmatrix},$$

that correspond to arrival intensities $\lambda_1 = 0.35$, $\lambda_2 = 0.9$, respectively.

The service time is a PH distribution with two phases, constructed by the mean m_s and the squared coefficient of variance c_s^2 as

$$\alpha(m_s, c_s^2) = \left[\frac{1}{2c_s^2} \quad 1 - \frac{1}{2c_s^2} \right], \quad S(m_s, c_s^2) = \begin{bmatrix} -\frac{1}{m_s c_s^2} & \frac{1}{m_s c_s^2} \\ 0 & -\frac{2}{m_s} \end{bmatrix}.$$

The default parameters of the service times are given by the following table:

	Type 1	Type 2
m_s	0.666	0.222
c_s^2	3	1.5

	$m_s = 10$ (high load)		$m_s = 1$ (low load)	
	$E(D_1)$	$p_\ell^{(1)}$	$E(D_1)$	$p_\ell^{(1)}$
$c_s^2 = 0.5$	71.076	0.78058	1.6774	0.0677
$c_s^2 = 1.5$	64.768	0.76301	1.8617	0.0719
$c_s^2 = 3$	58.28	0.74119	2.0542	0.0767

Table I
MEAN TYPE 1 SOJOURN TIME AND $p_\ell^{(1)}$

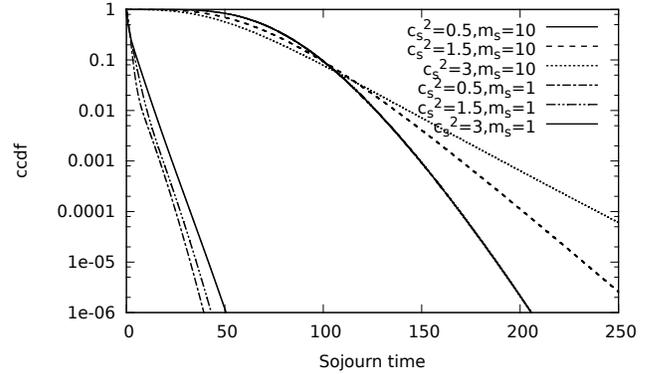


Figure 4. Ccdf of the sojourn time

The distribution of the patience of type 1 customers is a piece-wise approximation of the Weibull distribution X with parameters (k, λ) , i.e., $P[X \geq x] = 1 - e^{-(x/\lambda)^k}$, where λ is set such that the mean value is 50. The Weibull distribution has a heavy tail for $k < 1$. To ensure that the Weibull distributed impatience is approximated appropriately well with a step function in our queueing model, we use as many as 4096 steps. As only the type 1 customers are impatient, we will denote the probabilities $a_{i,1}$ as a_i . The thresholds are positioned in a linear manner such that

$$a_i = P(X \geq d_i) = \frac{r+1-i}{r+1}, \quad i = 1, \dots, r. \quad (14)$$

A. Impact of the burstiness of the service time

In this example we look at the impact of the coefficient of variation of the type 1 service time. The k parameter of the Weibull distribution has been set to 0.8. The results are depicted in Figure 4 and in Table I. Under low load the sojourn time increases with the c_s^2 , under high load the reverse happens. This is due to the customer impatience in the server: when the load is high and the service rate is more variable, only customers requiring a low amount of service tend to get fully served, which is beneficial for the mean sojourn time.

For each plot in Figure 4 about 38 seconds of computation time were required on a PC with an 1.8 GHz CPU and 2 GB of RAM, 30 seconds of which were spent calculating the matrices $\Psi, \hat{\Psi}$ for each of the 4096 layers and for solving the boundary equations. The remaining time was taken by the successive evaluation of the sojourn time distribution at

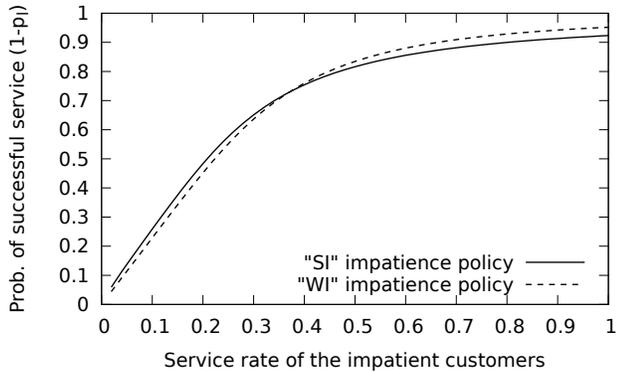


Figure 5. Probability of successful service vs. the service rate $1/m_s$

the necessary number of points. We did not experience any numerical issues when calculating the results neither in this, nor in the subsequent examples.

B. Comparison of two impatience policies

Now we investigate how the performance measures differ from the case where customers become patient when their service starts (this case is covered by [9]). To simplify the discussion we introduce the abbreviation SI for the case when the customers remain impatient in the server, and WI for the case when impatience applied in the waiting room only.

With regard to the ratio of customers leaving prematurely, one would think that the SI policy yields larger p_ℓ , since customers can leave not only during the waiting time, but during the service time as well. On the other hand, the workload decreases as more customers leave, so customers are more likely to start service, which reduces abandonments. Hence, there are two opposing forces at work here. This can also be noted by remarking that the SI system wastes some of its time on serving customers partially, but those that do receive full service experience a shorter mean service time.

To investigate how $p_\ell^{(1)}$ behaves, we varied the service rate of the type 1 customers over a wide range. The results are depicted in Figure 5. According to the figure, more customers leave prematurely with the SI policy when the utilization is low to moderate. In an overload situation, however, SI results in a lower $p_\ell^{(1)}$ value. Hence, it is good to drop customers that require lots of service under high load as perhaps two or more other customers requiring less service can be served instead. Figure 6 depicts the distribution of the sojourn time corresponding to the two studied impatience policies, with the mean service rate $1/m_s$ of the type 1 customers set to $1/4$. As expected, the sojourn time of the successfully served customers is larger in the "WI" case.

VII. CONCLUSION

In this paper we introduced a multi-layer Markov modulated fluid queue that allows general phase transitions at

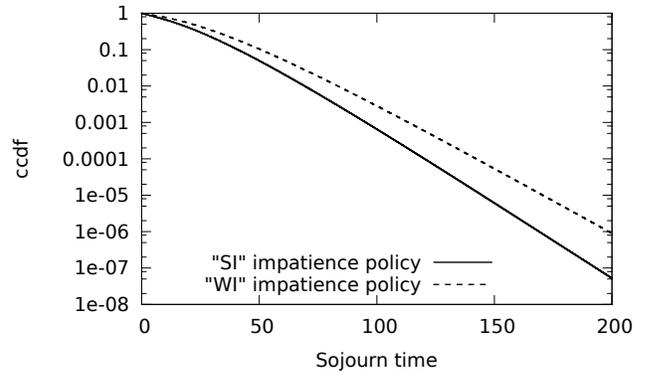


Figure 6. Ccdf of the sojourn time with "SI" and "WI" impatience policies

the boundaries. Together with the stationary solution of the system, we also provide an efficient numerical method based on the matrix analytic approach that is able to cope with a large number of layers.

The extended fluid queue allowed us to analyze the steady state of the age process of the MMAP[K]/PH[K]/1 queue with general customer impatience, where the customers remain impatient in the server as well.

Several numerical examples demonstrate the usefulness of the model, including a comparison between two impatience policies.

REFERENCES

- [1] A. I. Elwalid and D. Mitra, "Fluid models for the analysis and design of statistical multiplexing with loss priorities on multiple classes of bursty traffic," in *IEEE INFOCOM '92*. IEEE Computer Society Press, 1992, pp. 415–425.
- [2] M. Mandjes, D. Mitra, and W. Scheinhardt, "Models of network access using feedback fluid queues," *Queueing Syst. Theory Appl.*, vol. 44, pp. 365–398, August 2003.
- [3] H. E. Kankaya and N. Akar, "Exact analysis of single-wavelength optical buffers with feedback markov fluid queues," *IEEE/OSA Journal Optical Communication Networks*, vol. 1, no. 6, pp. 530–542, Nov 2009.
- [4] A. da Silva Soares and G. Latouche, "Matrix-analytic methods for fluid queues with finite buffers," *Perform. Eval.*, vol. 63, pp. 295–314, May 2006.
- [5] M. Gribaudo and M. Telek, "Stationary analysis of fluid level dependent bounded fluid models," *Perform. Eval.*, vol. 65, no. 3–4, pp. 241–261, Mar. 2008.
- [6] S. Asmussen, "Stationary distributions for fluid flow models with or without brownian noise," *Stochastic Models*, vol. 11, no. 1, pp. 21–49, 1995.
- [7] N. Akar and K. Sohraby, "Infinite- and finite-buffer markov fluid queues: a unified analysis," *J. Appl. Prob.*, vol. 41, pp. 557–569, 2004.

- [8] T. Dzial, L. Breuer, A. da Silva Soares, G. Latouche, and M. Remiche, "Fluid queues to solve jump processes," *Perform. Eval.*, vol. 62, pp. 132–146, October 2005.
- [9] B. Van Houdt, "Analysis of the adaptive MMAP[K]/PH[K]/1 queue: a multi-type queue with adaptive arrivals and general impatience," *European Journal of Operational Research*, vol. 220, no. 3, pp. 695–704, 2012.
- [10] A. da Silva Soares and G. Latouche, "Fluid queues with level dependent evolution," *European Journal of Operational Research*, vol. 196, pp. 1041–1048, 2009.
- [11] N. Bean and M. O'Reilly, "Performance measures of a multi-layer markovian fluid model," *Annals of Operations Research*, vol. 160, pp. 99–120, 2008.
- [12] F. Baccelli, P. Boyer, and G. Hebuterne, "Single-server queues with impatient customers," *Adv. in Appl. Prob.*, vol. 16, pp. 887–905, 1984.
- [13] J. Van Velthoven, B. Van Houdt, and C. Blondia, "Response time distribution in a D-MAP/PH/1 queue with general customer impatience," *Stochastic Models*, vol. 21, pp. 745–765, 2005.
- [14] V. G. Kulkarni, *Fluid models for single buffer systems*. Boca Raton, FL, USA: CRC Press, Inc., 1997, pp. 321–338.
- [15] G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods and stochastic modeling*. Philadelphia: SIAM, 1999.
- [16] Q.-M. He and M. F. Neuts, "Markov chains with marked transitions," *Stochastic Processes and their Applications*, vol. 74, no. 1, pp. 37 – 52, 1998.
- [17] Q. He, "Age process, workload process, sojourn times, and waiting times in a discrete-time SM[K]/PH[K]/1/FCFS queue," *Queueing Systems*, vol. 49, pp. 363–403, 2005.
- [18] C.-H. Guo, B. Iannazzo, and B. Meini, "On the doubling algorithm for a (shifted) nonsymmetric algebraic Riccati equation," *SIAM J. Matrix Anal. Appl.*, vol. 29, pp. 1083–1100, 2007.
- [19] W.-G. Wang, W.-C. Wang, and R.-C. Li, "ADDA: Alternating-Directional Doubling Algorithm for M-matrix algebraic Riccati equations," The University of Texas Arlington, Tech. Rep. 2011-04, 2011.
- [20] A. da Silva Soares, "Fluid queues: Building upon the analogy with QBD processes," Ph.D. dissertation, Université Libre de Bruxelles, 2005.

APPENDIX

For completeness we also indicate how the densities $\pi(x)$ for $x \in (d_{k-1}, d_k)$ for $k \geq 1$ can be obtained with the matrix analytic approach of [4], [10]. As in Section IV we first consider the censored process $\{\tilde{X}(t), \tilde{Z}(t), t \geq 0\}$. Define $\Psi^{(k)}$ and $\hat{\Psi}^{(k)}$, for $k \geq 1$, as the smallest non-negative solution of the algebraic Riccati equation

$$C_+^{-1} \tilde{F}_{+-}^{(k)} + \Psi^{(k)} C_-^{-1} \tilde{F}_{--}^{(k)} + C_+^{-1} \tilde{F}_{++}^{(k)} \Psi^{(k)} + \Psi^{(k)} C_-^{-1} \tilde{F}_{-+}^{(k)} \Psi^{(k)} = 0,$$

and its dual (i.e., for the level-reversed queue)

$$C_-^{-1} \tilde{F}_{-+}^{(k)} + \hat{\Psi}^{(k)} C_+^{-1} \tilde{F}_{++}^{(k)} + C_-^{-1} \tilde{F}_{--}^{(k)} \hat{\Psi}^{(k)} + \hat{\Psi}^{(k)} C_+^{-1} \tilde{F}_{+-}^{(k)} \hat{\Psi}^{(k)} = 0.$$

The solutions $\Psi^{(k)}$ and $\hat{\Psi}^{(k)}$ can be computed simultaneously with the SDA [18] or ADDA [19] algorithm and the first passage probability matrices defined in Section IV can be expressed in terms of $\Psi^{(k)}$ and $\hat{\Psi}^{(k)}$ (see [4, Corollary 5.3]). Define

$$K^{(k)} = C_+^{-1} \tilde{F}_{++}^{(k)} + \Psi^{(k)} C_-^{-1} \tilde{F}_{--}^{(k)},$$

and

$$\hat{K}^{(k)} = C_-^{-1} \tilde{F}_{--}^{(k)} + \hat{\Psi}^{(k)} C_+^{-1} \tilde{F}_{++}^{(k)},$$

for $k \geq 1$. Then, $\tilde{\pi}(x) = \tilde{\eta}\pi(x)$ can be expressed as

$$\tilde{\pi}(x)C = \tilde{\pi}_+(d_r+)C_+e^{K^{(r+1)}(x-d_r)}[I, \Psi^{(r+1)}],$$

for $x > d_r$ and

$$\tilde{\pi}(x)C = \nu_1^{(k)} e^{K^{(k)}(x-d_{k-1})}[I, \Psi^{(k)}] + \nu_2^{(k)} e^{\hat{K}^{(k)}(d_k-x)}[\hat{\Psi}^{(k)}, I],$$

for $x \in (d_{k-1}, d_k)$ and $k = 1, \dots, r$, where

$$\nu_j^{(k)} = \tilde{\pi}_+(d_{k-1}+)C_+N_j^{(k)} + \tilde{\pi}_-(d_k-)C_-N_{j+2}^{(k)},$$

for $j = 1, 2$ and

$$\begin{bmatrix} N_1^{(k)} & N_2^{(k)} \\ N_3^{(k)} & N_4^{(k)} \end{bmatrix} = \begin{bmatrix} I & e^{K^{(k)}b_k}\Psi^{(k)} \\ e^{\hat{K}^{(k)}b_k}\hat{\Psi}^{(k)} & I \end{bmatrix}^{-1},$$

and $b_k = d_k - d_{k-1}$. The densities $\pi_0(x)$ for the states belonging to the set $\{(x, i), x \in (d_{k-1}, d_k), i \in \mathcal{S}_0\}$ can be retrieved from $\pi_+(x)$ and $\pi_-(x)$ as follows

$$\pi_0(x) = (\pi_+(x)F_{+0} + \pi_-(x)F_{-0})(-F_{00})^{-1}.$$

Notice, as $\pi(x)$ is expressed in terms of a matrix exponential, the normalization constant and the probabilities $\Pi(x)$ can be expressed in closed form (see [20], [9]).

Remark: Let $\tilde{\xi}^{(k)} = (\tilde{\xi}_+^{(k)}, \tilde{\xi}_-^{(k)})$ be the unique stochastic invariant vector of $\tilde{F}^{(k)}$, for $k = 1, \dots, r+1$. The process $\{\tilde{X}(t), \tilde{Z}(t), t \geq 0\}$ is positive recurrent if $\tilde{\xi}_+^{(r+1)} \mathbb{1} < \tilde{\xi}_-^{(r+1)} \mathbb{1}$, while the process $\{X(t), Z(t), t \geq 0\}$ is positive recurrent if and only if $\{\tilde{X}(t), \tilde{Z}(t), t \geq 0\}$ is. Further, the matrices $N_i^{(k)}$, for $i = 1, \dots, 4$, are only properly defined if $\tilde{\xi}_+^{(k)} \mathbb{1} \neq \tilde{\xi}_-^{(k)} \mathbb{1}$.